INAUGURAL-DISSERTATION

zur

Erlangung der Doktorwürde

der

Gesamtfakultät für Mathematik, Ingenieur- und Naturwissenschaften

der

Ruprecht-Karls-Universität Heidelberg

vorgelegt von

Annika Reinke, M. Sc.

aus Leer (Ostfriesland)

Tag der mündlichen Prüfung : _____

Eliminating flaws in biomedical image analysis validation

Betreuerin: Prof. Dr. Lena Maier-Hein
Zweitbetreuerin: Prof. Dr. Annette Kopp-Schneider

Abstract

The field of automatic biomedical image analysis substantially benefits from the rise of Artificial Intelligence (AI). Proper validation of those AI algorithms is, however, frequently neglected in favor of a strong focus on the development and exploration of new models. This research practice can, however, be risky since it may propagate poorly validated algorithms that could cause adverse outcomes for patients. Thus, a thorough and high-quality validation is crucial for any algorithm to potentially be used in clinical practice. This particularly holds true for biomedical image analysis competitions, so-called challenges, which have emerged as the state-of-the-art technique for comparative assessment of AI algorithms and determining which is the most effective in solving a certain research question. Challenges have strong implications. While challenge winners typically receive large monetary awards and are highly cited, the algorithm also stands a better chance of being translated into clinical practice. Given the tremendous importance of challenges, it is surprising that hardly any attention has so far been given to quality control.

The objective of the work presented in this thesis was to analyze common practice in challenges, to systematically reveal flaws in both challenges and general image analysis validation, to propose solutions to eliminate those issues and to improve general validation practice. Contributions related to the analysis of flaws and strategies for improvement are presented for four areas: challenge design, validation metrics, rankings, as well as reporting and result analysis.

First, we demonstrate that challenges are highly heterogeneous yet not standardized, making it difficult to assess their overall quality. We further show that the research community is concerned about critical quality issues of challenges. The community eagerly asked for more quality control and best practice recommendations. Moreover, we evidence how effortlessly both challenge participants and organizers could, in theory, manipulate challenges by taking advantage of potential security holes in the challenge design. To compensate for this issue, we introduce a structured challenge submission system to collect comprehensive information about the challenge design, which can then be critically reviewed by independent referees.

We further demonstrate that validation metrics, the key measures in the assessment of AI algorithms, come with critical limitations that are often not taken into account during validation. In fact, researchers typically favor the use of common metrics without being aware of the numerous pitfalls pertaining to their use. An exhaustive list of metric-related pitfalls in the context of image-level classification, semantic segmentation, instance segmentation, and object detection tasks is provided in this thesis. To promote the selection of validation metrics based on their suitability to the underlying research problem rather than popularity, we propose a problem-driven metric recommendation framework that empowers researchers to make educated decisions while being made aware of the pitfalls to avoid.

Since challenge rankings are an integral part of competitions, we place particular emphasis on analyzing the stability and robustness of rankings against changes in the ranking computation method. We demonstrate that rankings are typically unstable, meaning that an algorithm could win a challenge simply due to the nature of a ranking calculation scheme and not due to actually being the best fit for solving a particular research task. To enable uncertainty-based ranking analysis, we present an open-source toolkit that includes several analysis and advanced

visualization techniques for challenges and general benchmarking experiments.

Finally, the transparency of validation studies is one of the core elements of high-quality research and should thus be carefully considered. However, our analysis of the transparency and reproducibility of both challenge design and participating algorithms shows that this is often not the case, substantially decreasing the interpretability of challenge results. To facilitate and enhance challenge transparency, we present a guideline for challenge reporting. In addition, we introduce the concept of challenge registration, i.e. publishing the complete challenge design before execution. This concept is already successfully applied in clinical trials and increases the transparency and reliability of a challenge, as it makes substantial changes in the design traceable. Finally, we show that challenge results can be used for a dedicated strength-weakness analysis of participating algorithms, from which future algorithm development could heavily benefit in addressing unsolved issues.

In summary, this thesis uncovers several critical flaws in biomedical image analysis challenges and algorithm validation. In response, it also introduces several measures that have already proven their practice-changing impact and substantially increased the overall quality of challenges, especially for the well-known Medical Image Computing and Computer Assisted Interventions (MICCAI) and IEEE International Symposium on Biomedical Imaging (ISBI) conferences. The suggested advancements in challenge design promise to give rise to competitions with a higher level of reliability, interpretability, and trust. The overall findings and suggested improvements are not specific to challenges alone, but also generalize to the entire field of algorithm validation. The presented thesis thus paves the way for high-quality and thorough validation of AI algorithms, which is crucial to avoiding translating inefficient or clinically useless algorithms into clinical practice.

Zusammenfassung

Automatisierte biomedizinische Bildanalyseverfahren können erheblich von Künstlicher Intelligenz (KI) profitieren. Die Validierung solcher Algorithmen wird oft von Wissenschaftler:innen unterschätzt, die sich primär mit der Entwicklung und Erforschung neuer Modelle und Algorithmen beschäftigen. Diese Forschungspraxis kann durchaus riskant sein, wenn sie schlecht validierte Algorithmen propagiert, die für Patienten nachteilige Folgen haben könnten. Daher ist eine gründliche und qualitativ hochwertige Validierung von entscheidender Bedeutung für jeden Algorithmus, der potenziell in der klinischen Praxis eingesetzt werden soll. Dies gilt insbesondere für Wettbewerbe in der biomedizinischen Bildanalyse, sogenannte Challenges, welche als Standard-Methodik zur vergleichenden Analyse von KI-Algorithmen und zur Ermittlung des effektivsten Algorithmus für die Lösung einer bestimmten Forschungsfrage fungieren. Challenges haben starke Implikationen. Die Gewinner einer Challenge erhalten typischerweise hohe Preisgelder und werden häufig zitiert. Zudem ist die Wahrscheinlichkeit höher, dass der entsprechende Algorithmus in die klinische Praxis überführt wird. Angesichts der Relevanz von Challenges ist es überraschend, dass ihrer Qualitätskontrolle bisher kaum Aufmerksamkeit geschenkt wurde.

Ziel dieser Arbeit war es, die gängige Praxis von Challenges zu analysieren, systematisch Schwachstellen in Challenges und der allgemeinen Validierung von biomedizinischen Bildanalysealgorithmen aufzudecken und Lösungen zur Beseitigung dieser Probleme sowie zur Verbesserung der generellen Validierungspraxis vorzuschlagen. In dieser Arbeit werden Beiträge zur Analyse von Fehlern und Verbesserungsstrategien für vier Bereiche vorgestellt: Design von Challenges, Validierungsmetriken, Ranglisten sowie Berichterstattung und Ergebnisanalyse.

Wir demonstrieren zunächst die Heterogenität und fehlende Standardisierung von Challenges, welche eine Bewertung ihrer Gesamtqualität erschwert. Wir zeigen außerdem, dass die Forschungsgemeinschaft über kritische Qualitätsprobleme von Challenges besorgt ist. Die Forschungsgemeinschaft verlangt mehrheitlich mehr Qualitätskontrolle und Empfehlungen zu guter wissenschaftlicher Praxis. Darüber hinaus zeigen wir experimentell, dass es sowohl für Challenge-Teilnehmer:innen als auch -Organisator:innen theoretisch möglich wäre, Challenges durch die Ausnutzung potentieller Sicherheitslücken im Design zu manipulieren. Um dieses Problem zu kompensieren, führen wir ein strukturiertes Onlinesystem zur Einreichung von Challenges ein, um umfassende Informationen über den Aufbau einer Challenge zu sammeln, welche darüber hinaus von unabhängigen Gutachtern kritisch geprüft werden können.

Ferner weisen wir nach, dass Validierungsmetriken, die wichtigsten Maßstäbe für die Bewertung von KI-Algorithmen, mit kritischen Einschränkungen verbunden sind, die bei der Validierung oft nicht berücksichtigt werden. Tatsächlich bevorzugen Forscher:innen in der Regel gängige Metriken, ohne sich der zahlreichen Probleme bewusst zu sein, die mit ihrer Verwendung verbunden sein können. Diese Arbeit beinhaltet einen umfassenden Überblick über Fallstricke in Bezug auf Metriken im Kontext von Klassifizierungsproblemen, der semantischen und Instanzsegmentierung sowie der Objekterkennung. Um zu vermeiden, dass Validierungsmetriken nur aufgrund ihrer Popularität ausgewählt werden, präsentieren wir ein problemorientiertes Empfehlungssystem für Metriken, welches es Forscher:innen ermöglicht, fundierte Entscheidungen zu treffen, während sie gleichzeitig auf zu vermeidende Fallstricke aufmerksam gemacht werden.

Da Challenge-Ranglisten ein integraler Bestandteil von Wettbewerben sind, legen wir besonderen Wert auf die Analyse der Stabilität und Robustheit von Ranglisten gegenüber Änderungen in deren Berechnungsmethode. Wir zeigen, dass Ranglisten in der Regel instabil sind, was bedeutet, dass ein Algorithmus eine Challenge nur aufgrund der Beschaffenheit der Berechnungsmethode der Rangliste gewinnen könnte, und nicht aufgrund seiner Eignung für eine bestimmte Forschungsfrage. Um eine auf Unsicherheit basierende Analyse von Ranglisten zu ermöglichen, stellen wir ein Open-Source-Toolkit vor, das verschiedene Analyse- und fortgeschrittene Visualisierungstechniken für Challenges und allgemeine Benchmarking-Experimente enthält.

Schließlich befassen wir uns mit der Transparenz von Validierungsstudien, welche eines der Kernelemente qualitativ hochwertiger Forschung darstellt und kritisch geprüft werden sollte. Unsere Analyse der Transparenz und Reproduzierbarkeit sowohl des Challenge-Designs als auch der teilnehmenden Algorithmen zeigt jedoch, dass dies häufig nicht der Fall ist, wodurch die Interpretierbarkeit von Challenge-Ergebnissen erheblich einschränkt wird. Um die Transparenz von Challenges zu verbessern, stellen wir eine Leitlinie zur Beschreibung und Dokumentation von Challenges vor. Darüber hinaus präsentieren wir das Konzept der Challenge-Registrierung, bei dem das vollständige Challenge-Design bereits vor deren Durchführung veröffentlicht wird. Dieses Konzept wird bereits erfolgreich in klinischen Studien angewandt und erhöht die Transparenz und Zuverlässigkeit einer Challenge, da es gravierende Änderungen des Designs rückverfolgbar macht. Schließlich demonstrieren wir, dass die Ergebnisse von Challenges für eine dedizierte Stärken-Schwächen-Analyse der teilnehmenden Algorithmen verwendet werden können. Von einer solchen Analyse kann künftige Algorithmenentwicklung stark profitieren, um bislang ungelöste Probleme zu adressieren.

Zusammenfassend deckt diese Arbeit mehrere kritische Mängel auf dem Gebiet von Challenges und der Validierung biomedizinischer Bildanalyseverfahren auf. Die präsentierten Mängel werden um entsprechende Lösungsansätze ergänzt, welche bereits in der Praxis umgesetzt werden und die Gesamtqualität von Challenges erheblich verbessert haben, vor allem im Rahmen der bekannten Konferenzen Medical Image Computing and Computer Assisted Interventions (MICCAI) und IEEE International Symposium on Biomedical Imaging (ISBI). Die vorgeschlagenen Verbesserungen versprechen Challenges mit einem höheren Maß an Zuverlässigkeit, Interpretierbarkeit und Vertrauen. Die allgemeinen Erkenntnisse und Verbesserungsvorschläge sind nicht spezifisch für Challenges, sondern lassen sich auch auf das gesamte Gebiet der Algorithmenvalidierung übertragen. Die vorliegende Arbeit bereitet damit den Weg für die hochwertige und gründliche Validierung von KI-Algorithmen, was entscheidend dazu beiträgt, die Übertragung ineffizienter oder klinisch nutzloser Algorithmen in die klinische Praxis zu verhindern.

Acknowledgements

The past years working in the Division of Intelligent Medical Systems (IMSY) have been a wonderful and exciting time for me. I have encountered a variety of intriguing research challenges and have managed to grow, advance, and improve my knowledge and skills. During this time, I was notably taught how teamwork and engagement substantially boost motivation and scientific advancement. I want to express my gratitude to everyone who helped make my Ph.D. experience wonderful and who gave me excellent advice.

This thesis would not have been possible without the guidance and advice from my Ph.D. supervisor, Prof. Dr. Lena Maier-Hein. I am extremely grateful for her supervision and the possibility to work on a Ph.D. thesis, although I started as a research scientist. She listened to all my questions, was always happy to offer assistance, and took the time to discuss or come up with ideas for a new research topic. I would like to thank her for the opportunity to grow, to develop my strengths, and for the offer to lead a sub-group of the department. In addition, I am very grateful for the possibility of being engaged in so many international initiatives, making it possible to meet and work with even more excellent researchers.

Furthermore, I am very thankful for the support of my co-supervisor Prof. Dr. Annette Kopp-Schneider, who was always happy to help out with her statistical advice, even though our research was out of her comfort zone in the beginning, and ended up being one of our most important collaborators. Thus, I am even more grateful for her feedback on my projects. She has taught me the importance of a thorough statistical analysis, which is unfortunately not always done in current research practice. Also, thank you for being part of my thesis advisory committee. Of course, I would also like to thank the remaining members of the committee: Prof. Dr. Klaus Maier-Hein and Dr. Hannes Kenngott. It was an awesome experience to receive so much positive feedback from you on my projects and such great ideas and suggestions to improve them even more.

Many thanks to my sub-group for the Validation of Intelligent Systems (VIS): Matthias Eisenmann, Emre Kavur, Tim Rädsch and Evangelia Christodoulou. You are an amazing group of people, always happy to help out and of course, be in for a nice chat. A special thanks to Matthias, who has been with me during my whole time in the division, who supervised me in the beginning, and who is the most thorough person I know. Thanks, Matze, for always being "solid as a rock".

Thanks a lot, Minu Tizabi, for being an incredible help in writing texts. You are a genius when it comes to great phrasings and making the best out of a paper! Without your help, most of our publications would have not been the same.

I would like to thank Tim Adler and Leonardo Ayala Menjivar for our thesis writing club. It was so much fun to exchange experiences, help each other out or just talk about our writing time. Thanks should also go to my office, the Phantastisch produktiven Positivitätskanonen. It was so much fun working with you, chatting, and exchanging ideas on our individual projects. I will never forget our swear jar and our Christmas market visits.

My whole Ph.D. time would not have been the same without the help of the best office I could imagine. Thanks, Michaela Gelz, Stefanie Strzysch, Theresa Klocke, and Janina Dunning. It is

awesome how you manage everything, that you always directly offer advice and help and that I know that I can come to you with whatever question I have. Also thanks a lot to Dr. Alexander Seitel and Dr. Keno März for being a constant within the department and for all your help.

A big thank you goes to all members of the SYMIC group(s). You are a fantastic group of people with great team spirit and are always eager to help with any question. Many thanks to Sinan Onogur, Dr. Tobias Roß, Dr. Paul Jäger, and Patrick Scholz. You are excellent team players and working with you was a pleasure. Thanks to Melanie Schellenberg and Ina Kompan for making our retreats awesome. I would like to thank all of my collaborators, for enabling such great research and for the invitations to present our work on several occasions. Without all of you, my thesis would not have been possible.

Many thanks to my family, who have supported by during my whole life and for being so proud of me. Also, I would like to thank my best friend, Famé Bültge, for the possibility to call her at any time, whenever I want to talk about anything, for having so much fun and for proof-reading this thesis, although she is not familiar with the topic at all.

Lastly, I am extremely grateful to Jannik Piper. You have supported me during the whole time, provided advice on topics that you don't know, have managed to always motivate me, and have accepted my bad moods in times of failure. You are the best person I know. I love you.

Contents

Abstract **v**

Acknowledgements **ix**

Contents **xi**

1 Introduction **1**
1.1 Motivation 1
1.2 Objectives 4
1.3 Outline 5

2 Fundamentals **7**
2.1 Fundamentals of image analysis 8
2.2 Fundamentals of statistics 15
2.3 Fundamentals of machine learning 19
2.4 Common validation metrics 26
2.5 Related work 62

3 Descriptive analysis of Biomedical Image Analysis Validation **66**
3.1 Common practice of challenges 67

4 Revealing flaws in common practice **76**
4.1 Revealing flaws related to challenge design 77
4.2 Revealing flaws related to metrics 84
4.3 Revealing flaws related to rankings 133
4.4 Revealing flaws related to reporting and analyses 140
4.5 Conclusion related to flaws of biomedical challenges 153

5 Improving common practice **154**
5.1 Improving common practice of challenge design 155
5.2 Improving common practice of metrics 161
5.3 Improving common practice of rankings 209
5.4 Improving common practice of reporting and analyses 242

6 Discussion **262**

7 Summary of Contributions and Conclusion **267**
7.1 Summary of Contributions 267
7.2 Own publications 271
7.3 Conclusion 274

Bibliography **275**

List of Acronyms **295**

List of Figures **300**

List of Tables **304**

A Appendix **307**
A.1 Contributions 307
A.2 Ranking schemes used in this thesis 310
A.3 Profile of common validation metrics 312
A.4 Metric recommendations for common biomedical use cases 350
A.5 Overview of the descriptions from the RobustMIS teams 355
A.6 RobustMIS challenge design document 360
A.7 MSD challenge design document 365
A.8 BIAS reporting guideline 370
A.9 Search terms related to mixed model analysis 375

1 Introduction

1.1 Motivation

The relevance of Machine Learning (ML)-based analysis of biomedical images is continuously growing [Litjens et al., 2017]. Nowadays, automatic image analysis is used to solve a variety of problems, including tumor classification and detection (e.g. [Khan et al., 2020; Hu et al., 2018]), organ, lesion, or cell segmentation (e.g. [Antonelli/Reinke et al., 2022; Menze et al., 2014; Ulman et al., 2017]), or surgical skill assessment (e.g. [Wang and Majewicz Fey, 2018]). Given the high number of algorithms for solving biomedical image analysis tasks that are published every year [Litjens et al., 2017], it becomes challenging to keep track of which methodological strategy is most effective for a certain problem.

Since the conditions under which the research is conducted – for example the used data sets or the available hardware – differ between publications, methods cannot easily be compared against each other in a fair manner. In medicine, clinical trials are the state-of-the-art approach for assessing and comparing the effects of new medications or treatments. Benchmarking in the field of data science is accordingly conducted through international competitions, which are commonly known as *challenges* [Maier-Hein et al., 2018; Reinke, 2021].

Challenges allow for a fair comparison of the performance of participating algorithms by enforcing the same conditions for every competing method (for example same data sets). Challenge participants are usually provided with a training data set (and optionally a validation set), which can be used to train and tune the algorithms to the specific application. Their performance is then validated using a test data set, for which the reference annotations are hidden from the participants. Challenges typically generate a ranking or leaderboard based on the results of this test set, which informs the research community and the participants about the winning method and the respective performances.

Given the scarcity of large-scale medical data sets [Ker et al., 2017], arising for example from patient privacy concerns, challenges frequently make their training data sets public. As shown in Varoquaux and Cheplygina [2022], this practice hugely impacts and drives the research itself. In their specific example, the authors show that there was a large increase of lung cancer studies in the area of Artificial Intelligence (AI) publications compared to general medical oncology publications after the organization of the Kaggle lung cancer challenge [Kaggle, 2016] in 2016.

Challenges have become a standard technique for comparing algorithms, and they often receive tremendous attention in terms of citations and monetary incentives [Maier-Hein et al., 2018]. Meanwhile, many journals request authors to demonstrate the performance of their approaches on challenge data to improve comparability. This means that the acceptance of a publication, and consequently a scientist's career, may be dependent on challenge results and challenge data. Furthermore, challenge winners enjoy a higher likelihood of their methods being transferred into clinical practice [Maier-Hein et al., 2018].

Due to their immense scientific impact, challenge organizers are faced with a lot of responsibility concerning the organization, design, and data quality of a challenge. Figure 1.1 summarizes the advantages as well as typical main steps in organizing a challenge. As shown in this workflow, a substantial amount of work must be completed before the challenge itself is conducted. Organizers need to define the goal of the challenge, generate the data sets including reference annotations accordingly, and work out the details of the challenge design, such as its mission and assessment method. Another large part of the organization lies in the evaluation of the challenge results, by calculating rankings and providing insights based on statistical analyses of the algorithm performances.

Given their relevance in the scientific community, challenges should be subject to stringent quality control, equivalent to quality control measures in clinical trials. However, as will be demonstrated throughout this thesis, numerous fundamental weaknesses in challenge design, reporting, and implementation are prevalent. Due to the lack of quality control in challenges, a totally inefficient, clinically useless algorithm could, in theory, win a biomedical imaging challenge. This could notably explain the lack of translation of AI algorithms into clinical practice [Panch et al., 2019; Lennerz et al., 2022].

> *'However, "the inconvenient truth" is that at present the algorithms that feature prominently in research literature are in fact not, for the most part, executable at the frontlines of clinical practice.'*
>
> — [Panch et al., 2019]

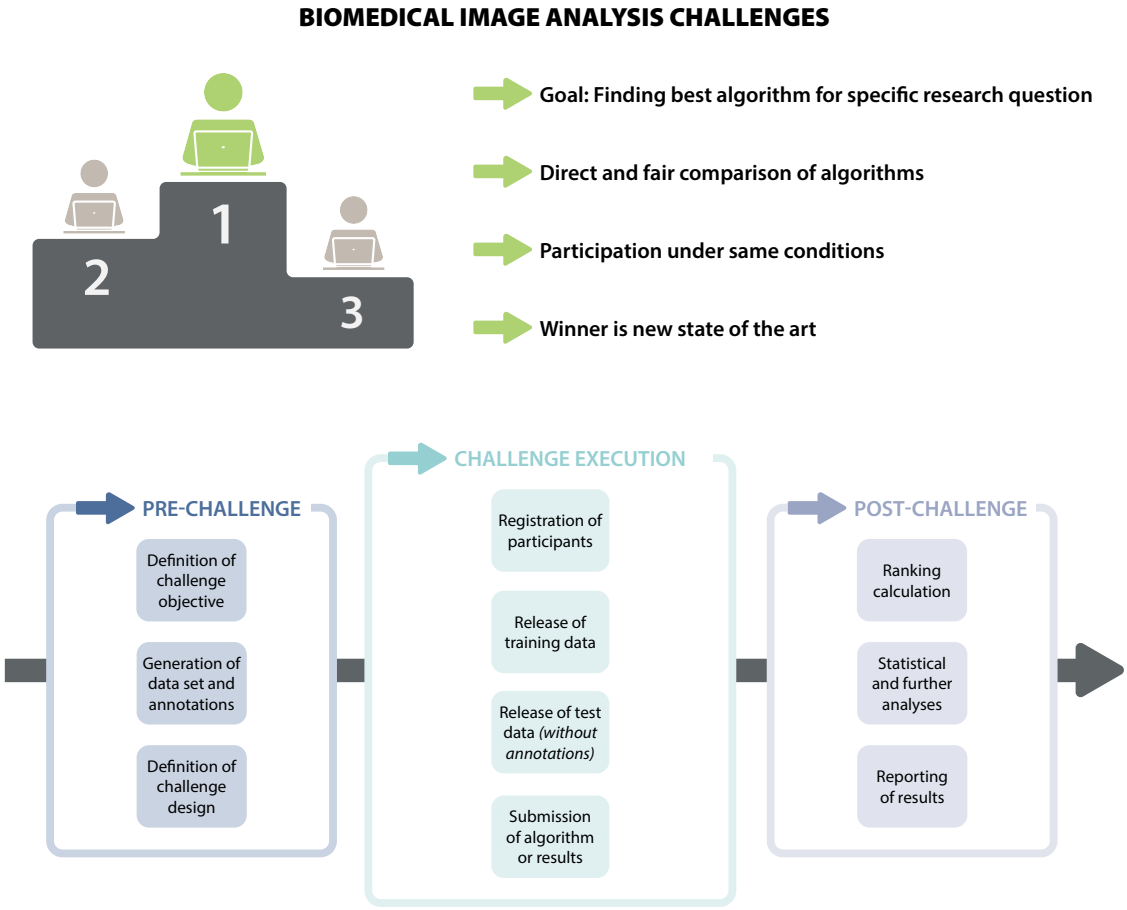


Figure 1.1: Top: Goals and advantages of organizing challenges. Bottom: Common steps in organizing a biomedical image analysis challenge.

Disclosure

The work presented in my thesis has been conducted within several interdisciplinary consortia. Therefore, other researchers were involved in most of the research. For that reason, I will use the "we" formulation throughout the thesis. In every section of the contribution Chapters 3 - 5, a disclosure states the core team members for the presented work. Furthermore, my specific contributions in the broader context are described in Chapter A.1.

1.2 Objectives

The objectives of the research presented in this thesis are to examine common practices in biomedical challenge design, metrics, rankings, and reporting, and subsequently provide solutions to the identified weaknesses as well as ensure that upcoming challenges will undergo a more rigorous quality control process, comparable to that of clinical trials. For this purpose, we first examine the status quo of challenges by reviewing more than 500 biomedical image analysis competitions to identify common practices. Moreover, we present the results of a community survey dedicated to gathering the general opinion of the community on challenge design [Maier-Hein et al., 2018]. In the ensuing sections of the thesis, we consider the following four hypotheses:

Hypothesis 1 (H1): Current biomedical challenge design is heavily flawed.

We hypothesize that past quality control mechanisms of challenges were not sufficient to ensure high-quality challenge design and organization. More specifically, we analyze weaknesses in challenge design to investigate whether challenge results are susceptible to manipulation [Reinke et al., 2018a]. To eliminate those weaknesses, we demonstrate the advantages of a structured challenge submission system including a dedicated peer review process for challenges [Reinke et al., 2018b].

Hypothesis 2 (H2): Common image analysis metrics do not reflect the biomedical domain interest.

We hypothesize that different choices of validation metrics may entirely change the winner(s) of a challenge. Moreover, we hypothesize that researchers are often not aware of the limitations of metrics. For this reason, we first show that there is no common access point to information on metric-related pitfalls, which is why this thesis includes the first comprehensive overview of that matter [Reinke et al., 2021a,b]. We conclude that validation metrics should not be chosen according to their popularity and introduce a problem-aware metric recommendation framework to ensure metric selection that reflects biomedical needs [Maier-Hein et al., 2022].

Hypothesis 3 (H3): Challenge rankings are highly unstable.

Given the importance of challenge rankings in determining a challenge winner, rankings should be computed carefully. We demonstrate that rankings are generally very sensitive to the chosen challenge design [Maier-Hein et al., 2018]. To address this problem, we highlight how the interpretability of challenge rankings can be enhanced by advanced visualization techniques [Wiesenfarth et al., 2021; Antonelli/Reinke et al., 2022; Roß/Reinke et al., 2020].

Hypothesis 4 (H4): Challenge results are not reproducible.

Lastly, we hypothesize that challenges are not transparent and their results can typically not be reproduced. We show that both challenge details provided by the organizers and methodological details from the participants are often not sufficient to reproduce challenge results [Maier-Hein et al., 2018]. To address the lack of transparent reporting, we adapt common practice in the context of clinical trials to challenges. First, we introduce a challenge reporting guideline; second, we apply the concept of challenge registration [Maier-Hein et al., 2020]. Finally, we show how we

can gain more insights from challenge results by introducing a mixed model-based framework for a strength-weakness analysis of participating algorithms [Roß et al., 2021].

Generalizability of results

Although the focus of this thesis are biomedical image analysis challenges, **the presented flaws and solutions can easily be generalized to a broad range of general validation problems** as indicated in the respective chapters. Thus, rather than solely maximizing the impact of competitions, the presented work aims to enhance the quality of AI validation in general.

1.3 Outline

In the following Chapter 2, we introduce fundamental concepts related to image analysis, statistics, ML, and validation metrics. The fundamentals chapter ends with an overview of related work.

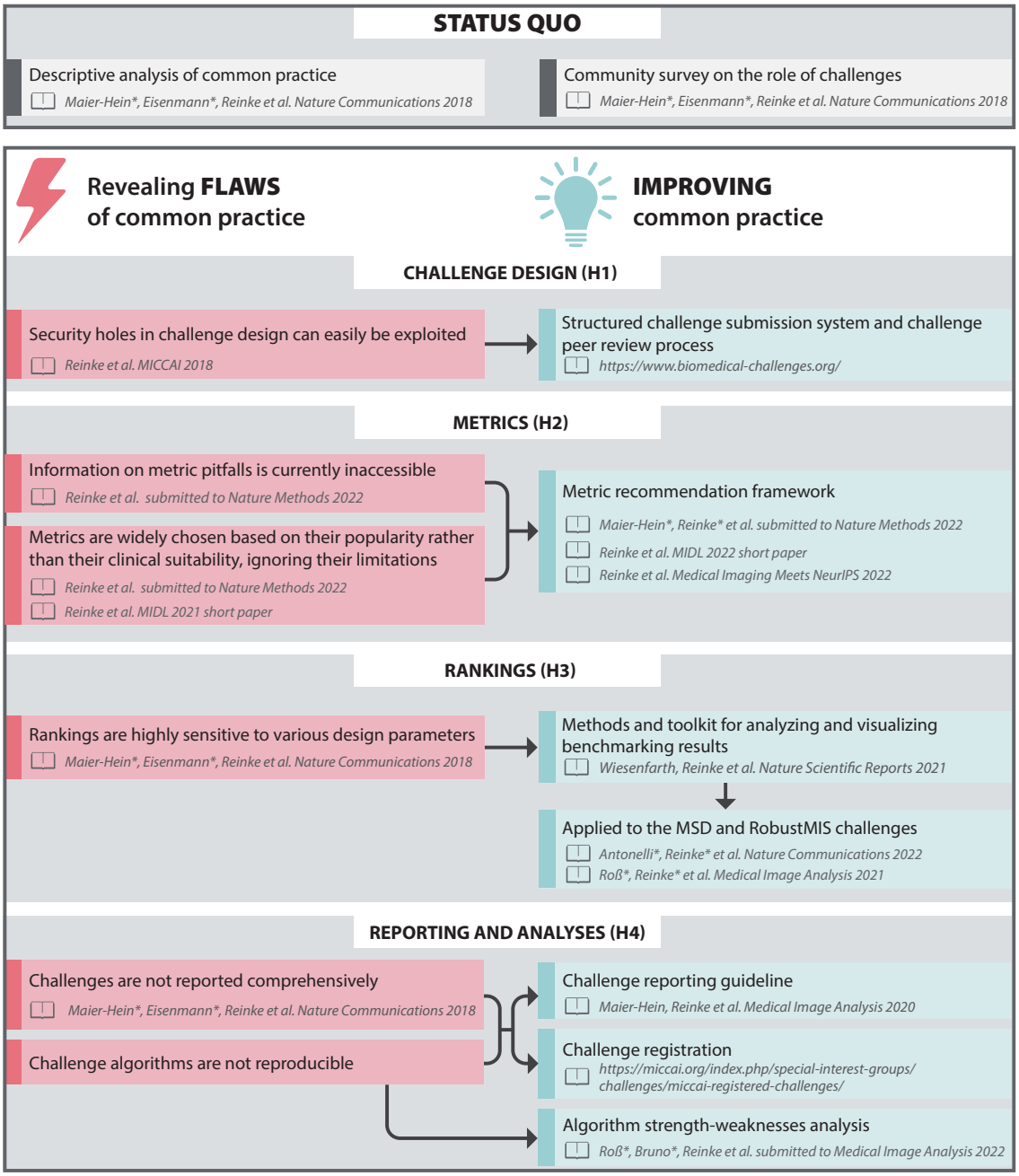
The contributions of this thesis are presented in Chapters 3 - 5. Every contribution section includes an introduction, a description of the used methods, results, and a brief discussion and conclusion. Contributions were partitioned into three parts, as summarized in Figure 1.2: Chapter 3 examines the role of challenges and current practices. This part concludes with the presentation of a community survey on challenges.

Chapter 4, including Sections 4.1 - 4.4, focuses on revealing flaws of common practice in challenges and is guided by Hypotheses 1 to 4. Flaws are examined related to the topics of challenge design (Section 4.1), validation metrics (Section 4.2), challenge rankings (Section 4.3), and the reporting and analyses of challenge results (Section 4.4). A conclusion of this part is given in Section 4.5.

For all of the presented issues, Chapter 5, which contains the Sections 5.1 - 5.4, provides strategies for improvement and mirrors the topics of the previous part, namely presenting solutions for issues in challenge design (Section 5.1), validation metrics (Section 5.2), challenge rankings (Section 5.3), and the reporting and analyses of challenge results (Section 5.4).

Chapter 6 discusses the findings of Chapters 3 - 5 in a broad context, showing how they relate to each other and bringing them into the context of general validation of AI algorithms, not only from the biomedical image analysis perspective.

My specific contributions in the broader context are described in Chapter 7, which also contains a listing of my publications. Finally, the Appendix contains complementary information to the presented results.



* shared first authors

Figure 1.2: Overview of contributions. The contributions of the thesis are presented in three parts. While Chapter 3 gives a general overview of challenges, including a community survey, Chapters 4 and 5 cover the topics of challenge design, metrics, rankings, and reporting and analyses. Chapter 4 focuses on revealing flaws related to the four topics while Chapter 5 presents solutions. The book icon refers to publications and H1 - H4 to Hypotheses 1 - 4, as presented in Section 1.2.

2 | Fundamentals

In this chapter, we present an overview of the concepts used throughout this thesis. The first section describes fundamentals in image analysis (Section 2.1), including the presentation of common problem categories and validation strategies, while the second section focuses on fundamentals in statistics (Section 2.2). Subsequently, we describe Machine Learning (ML) concepts in Section 2.3, followed by reviewing common validation metrics in Section 2.4. We conclude this chapter with a review of related work in Section 2.5.

2.1 Fundamentals of image analysis

The principles of image analysis and benchmarking are covered in this section. First, we describe common problem categories of image analysis algorithms in Section 2.1.1. The following Section 2.1.2 then presents general concepts related to the validation and benchmarking of image analysis algorithms. This includes the description of biomedical image analysis challenges as well as ranking methods.

It should be noted that we focus on the term *validation* rather than evaluation in this thesis. Validation involves determining whether the model actually achieved the goal for which it was designed. On the other hand, the objective of evaluation is to determine whether the model is effective and valuable for the intended purpose and is accepted by the end users [Jannin et al., 2006].

2.1.1 Problem categories

In this thesis, we focus on problem categories that are related to classification tasks. We refer to classification on image level as image-level classification, classification on object level as object detection and instance segmentation, and classification on pixel level as semantic and instance segmentation. Note that instance segmentation may operate on both, object and pixel level. A comparison of the four problem categories is provided in Figure 2.1.

Image-level classification

In image-level classification problems, class labels $y \in \{1, \ldots, C\}$ of C classes are assigned to an image. In the case of multiple labels per image, the labels $y \in \{0, 1\}^C$ indicate the presence or absence of every class. *Binary classification* refers to the case with only one class of interest. For instance, the decision of whether a patient shows a disease versus a healthy patient falls into this category. In the case of multiple classes of interest, we refer to *categorical* or *multi-class classification*. A common example is disease type classification, such as disease classification in dermoscopic images [Codella et al., 2019]. Nowadays, neural networks usually produce predicted class scores $\hat{y} \in [0, 1]^C$ as an output for every class instead of a fixed class label (see Section 2.3). Introducing a threshold on the predicted class scores divides the predictions into positive (greater than or equal to the threshold) and negative (less than the threshold) classes [Davis and Goadrich, 2006]. We provide further details on the thresholding in Section 2.4.

Semantic segmentation

The process of partitioning an image into multiple content-related regions or segments is known as semantic segmentation. To achieve this, each pixel P is assigned one or more labels $y_P \in \{1, \ldots, C\}$ or $y_P \in \{0, 1\}^C$ for C classes and pixels of the same class receive the same label [Guo et al., 2018]. Similar to image-level classification tasks (although less common in semantic segmentation), some neural networks output class scores $\hat{y}_P \in [0, 1]^C$, which are assigned to a class label based on a threshold [Asgari Taghanaki et al., 2021]. However, in semantic segmentation, many networks directly output a binary label $\hat{y}_P \in \{0, 1\}^C$ for every pixel for every

class. It is important to note that semantic segmentation problems do not distinguish between different object instances. Instances that belong to the same class receive the same labels. If instances should be differentiated, one should consider an instance segmentation task instead (see below). Common examples of semantic segmentation are liver [Heimann et al., 2009] or brain tumor segmentation [Menze et al., 2014].

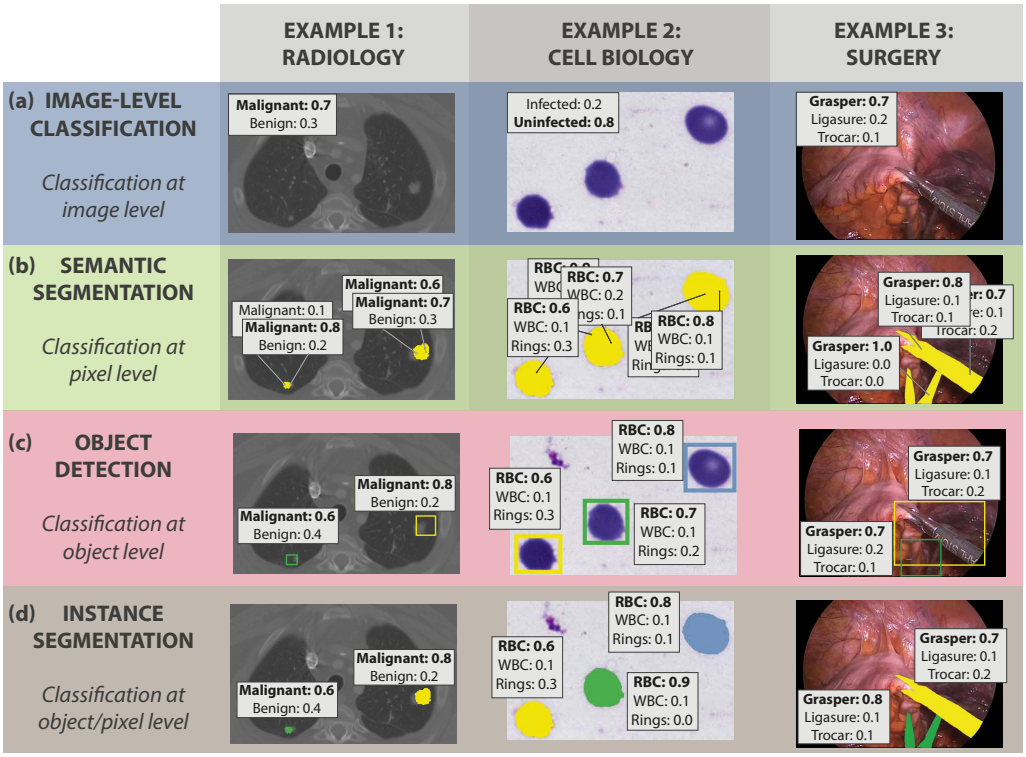


Figure 2.1: Problem categories covered in this thesis that are illustrated for three different application domains: radiology (left), cell biology (middle), and surgery (right). The fact that they can be interpreted as classification tasks is a common denominator among the underlying research problems. They assign a class label to the image or (multiple) components of it: A class label is assigned to the whole image in image-level classification tasks (a), to each individual pixel in semantic segmentation tasks (b), to identified objects in object detection tasks (c), and to identified objects made up of multiple pixels in instance segmentation tasks (d). Gray boxes show the predicted class scores on image, pixel, or object level. The class with the highest probability is highlighted in boldface. Abbreviations: Red Blood Cell (RBC), White Blood Cell (WBC). Figure adapted from Reinke et al. [2021a] and Maier-Hein et al. [2022].

Object detection

The detection and (rough) localization of objects in an image are referred to as object detection [Lin et al., 2014]. Object detection can therefore be interpreted as object-level classification. Each object O is assigned with a class label $y_O \in \{1, \ldots, C\}$ (or $y_O \in \{0, 1\}^C$ for multiple labels per object), along with localization information l_O, which may be bounding box coordinates, center points, or similar. For each object, algorithms typically output predicted class scores $\hat{y}_O \in [0, 1]^C$. In the context of object detection, they are often referred to as confidence scores.

Objects can be represented in different ways. The most common representation are bounding boxes, typically provided as coordinates of the boxes, that are drawn around the object of interest. Alternatively, a rough approximation or the center of an object can be used. In the case of pixel-level annotations, the full pixel masks can serve as the representation of the object. Object detection algorithms involve two additional steps, first, deciding on whether a reference object was correctly detected, and second, which predicted objects are matched with which reference objects. Section 2.4.3 explains both processes in more detail. Common examples of object detection are lung nodule detection [Setio et al., 2017] or polyp detection in colonoscopy videos [Sánchez-Montes et al., 2019].

Instance segmentation

While semantic segmentation does not differentiate between objects of the same class, this is done in instance segmentation [Gupta et al., 2019; Reinke et al., 2021a]. Each object or instance is detected and localized with pixel-level accuracy, in addition to drawing accurate object boundaries, making instance segmentation a combination of object detection and semantic segmentation. Instance segmentation problems are therefore typically validated similarly to object detection tasks, with additional semantic segmentation performance metrics applied to each instance (see Section 2.4.4 for details). Often, instance segmentation outputs are generated from semantic segmentation outputs by applying connected component analysis [Rosenfeld and Pfaltz, 1966; Reinke et al., 2021a], that provides the segmentations for every instance. Common examples of instance segmentation are cell nuclei instance segmentation [Ulman et al., 2017] or spine vertebrae instance segmentation [Sekuboyina et al., 2021].

Other problem categories

In contrast to image-level classification, *regression* problems require a continuous output rather than a categorical class label. Predictive regression of the cardiac output is a common example [Critchley et al., 2010]. *Prediction* refers to predicting the probability of a disease or outcome and is often interpreted as a classification or regression task. Popular examples include patient outcome prediction of progression-free survival [Andrearczyk et al., 2021] or survival prediction [Bakas et al., 2018]. *Image retrieval* is the process of retrieving images from an image database via a search query. A prominent example in medical image retrieval is finding similar images from a database based on different features such as the imaging modality or anatomical region [Clough et al., 2004]. The problem of automated object *tracking* can be expressed as detecting (and optionally segmenting) all object occurrences in a time series. A common example is cell tracking [Matula et al., 2015]. In a *registration* task, different images are transformed into a single coordinate system. Registration of thoracic images is a common example [Murphy et al., 2011].

2.1.2 Validation and benchmarking of image analysis algorithms

in the following paragraphs, we explain how biomedical image analysis challenges are typically constructed. We further cover methods for evaluating algorithm performance in the form of rankings.

Biomedical image analysis challenges

A **biomedical image analysis challenge** or **benchmarking competition** refers to a competition organized in the scope of a certain research problem in the field of biomedical image analysis. They typically aim to gain insight into a research problem or to find the best algorithm for a given scientific question or problem [Mendrik and Aylward, 2019]. Challenges are the standard method to compare algorithms under similar requirements, such as using the same data for performance validation [Maier-Hein et al., 2018]. Challenges are often organized at conferences and attract participants who submit their algorithms and methods to the challenge. They may be organized as one-time or repeated events and may feature an open call, i.e. accepting submissions even after a dedicated conference event [Kozubek, 2016]. The main steps for organizing, executing, and analyzing challenges are provided in Figure 1.1.

A challenge may encompass several **tasks** or **sub-challenges** to solve a sub-problem of the challenge [Maier-Hein et al., 2018]. For example, the overall goal of a challenge may be segmentation with several segmentation tasks of different organs or structures of interest (e.g. [Antonelli/Reinke et al., 2022; Goksel et al., 2015]). Another challenge may feature different problem categories, such as one task for the detection of cancer metastases and a second task for classifying them (e.g. [Bejnordi et al., 2017]). Typically, the assessment method (see below) or validation is performed individually for every task. In addition, some challenges compute an aggregated ranking across tasks.

A challenge is typically split into a **training** and a **test phase** [Maier-Hein et al., 2018; Kozubek, 2016]. First, challenge participants are provided with a training data set, commonly supplemented with a reference annotation, which can be used to train their algorithms. Depending on the challenge's design, either the challenge organizers give a validation data set or it is up to the competitors to separate the training data set into training and validation sets. Occasionally, participants may be allowed to complement the provided training data with publicly available or private data sets, subject to the challenge design. Some challenges feature a pre-evaluation based on results on a validation data set (if any).

To avoid potential misuse by participants, most challenges do not include reference annotations in the test data sets. The test phase commonly serves the purpose of the final assessment of participating methods and, if any, to calculate rankings. The test phase may be additionally divided into **off-site** and **on-site** phases. The off-site phase defines the standard challenge setting, in which participants submit their results on the test set or methods online (details on the submission process are given below). An optional on-site phase includes an in-person workshop, in which participants work on an additional test set.

The challenge data set(s) are typically composed of several **cases**. A training or test case is defined as the data for which the challenge participants produce an output for the challenge [Maier-Hein et al., 2018]. For example, a case may refer to one or multiple images or may be

complemented by additional meta information. In contrast to test cases, training cases usually also include a reference annotation.

The **submission** of results may be organized in various ways. For example, the algorithm outputs can be directly sent to the organizers or uploaded to a challenge platform or cloud. On the other hand, organizers may opt for the methods to be directly sent to them, for instance in the form of docker containers [Docker, 2022]. This setup comes with the advantage that the test data can be kept secret, which may be beneficial in terms of data privacy or in avoiding potential cheating or overfitting of methods [Goodfellow et al., 2016]. Participants may have the possibility to upload their results or methods multiple times and get feedback on their performances either individually or in the form of a public leaderboard.

The **assessment method** of a challenge includes the chosen performance metrics, ranking schemes, and statistical analysis methods, if any. A challenge should follow an assessment aim, corresponding to the properties of algorithms that should be assessed. The metrics should be chosen to reflect this aim. If challenge organizers choose to calculate a ranking (see below for details), it determines the winner or winning groups of a challenge or task and is commonly based on one or multiple metrics being assessed. The challenge results may be complemented by statistical analyses, for example investigating whether the challenge winner's performance is significantly superior to the other participants.

Ranking methods

From a mathematical perspective, a ranking is a **partially ordered set**, as defined in order theory [Davey and Priestley, 2002]:

Definition 2.1.1 (Ordered set as defined by [Davey and Priestley, 2002])**.** Let P be a set. An **order** (or **partial order**) on P is a binary relation $\leq$ on P such that, for all x, y and $z \in P$,

(i) $x \leq x$ (*reflexivity*),

(ii) $x \leq y$ and $y \leq x$ implies $x = y$ (*antisymmetry*),

(iii) $x \leq y$ and $y \leq z$ implies $x \leq z$ (*transitivity*).

A ranking may contain ranked objects sharing the same rank, so-called **ties**. In this case, different strategies may be applied to calculate a ranking, as presented in the following:

Standard competition ranking The standard competition ranking is the most commonly applied ranking method in cases of tied ranks [Madani et al., 2014]. Equally ranked objects receive the *same rank, followed by a gap*, which is equal to the number of the tied objects for the current rank. This ranking method is also referred to as the "1-2-2-4" rule [Mishra, 2009]. In this specific example, two objects share the same rank ("2"). Rank 2 appears twice, followed by a gap of two ranks, continuing with rank 4.

Modified competition ranking The modified competition ranking is very similar to the standard competition ranking except for *adding the gap before the set of tied ranks*, not afterward [Madani et al., 2014]. This ranking method is also referred to as the "1-3-3-4" rule [Mishra, 2009]. In this specific example, two objects share the same rank ("3"). After the first rank, a gap of two is included, followed by twice the rank of 3 for the equally ranked objects.

Dense ranking The dense ranking assigns the *same rank to the tied objects without adding a gap* before or afterward [Madani et al., 2014]. This ranking method is also referred to as the "1-2-2-3" rule [Mishra, 2009]. In this specific example, two objects share the same rank ("2"). After the shared ranks, no gap is included, so the ties are directly followed by rank 3.

Ordinal (strict order) ranking The ordinal or strictly ordered ranking assigns *different ranks to every ranked object* even if objects are observed equally according to the ranking criterion [Madani et al., 2014]. The ranks for the tied objects are assigned either randomly or based on another ranking criterion (for example the runtime of an algorithm or any other secondary validation metric). This ranking method is also referred to as the "1-2-3-4" rule [Mishra, 2009]. In this specific example, objects may share the same rank based on the ranking criterion but are assigned different ranks.

Fractional ranking The fractional ranking assigns the *average of ranks* that tied objects would receive with an ordinal ranking. This ranking method is also referred to as the "1-2.5-2.5-4" rule [Mishra, 2009]. In this specific example, two objects share the same rank ("2.5"). The rank 2.5 appears twice, other than that the ranking is continued as an ordinal ranking.

In a challenge (task), participating teams are typically ranked by the performances of their submitted methods to identify the winning algorithm(s). Determining a winner and deciding on the ranking scheme depends on the underlying problem category. In binary image-level classification problems, for example, the performance metrics are typically calculated over the complete data set [Russakovsky et al., 2015]. For instance, the number of True Positives (TPs), True Negatives (TNs), False Positives (FPs), and False Negatives (FNs) are determined over the whole data set and the corresponding metric scores (like Sensitivity or Accuracy) are calculated, resulting in one score per metric per data set. In this case, no further aggregation needs to be applied and participating methods can directly be ranked according to their metric scores.

However, one may choose to calculate metric values per case to allow for a case- or patient-sensitive validation, since a per-data set validation might oversee results for a specific case or patient. In this situation, multiple metric values are present for every participating team, which have to be aggregated to generate a concrete ordering of participants. In the following, three possible ranking schemes are presented. Please note that several other ranking methods exist. In some challenges, even heuristics or specific weighting schemes are applied.

Metric-based ranking scheme In a *metric-based ranking* or *aggregate-then-rank method*, the metric values per participating team are aggregated first (for example using the mean, median, or a specific percentile). Afterwards, the ranking is computed based on the aggregated result per team [Wiesenfarth et al., 2021; Maier-Hein et al., 2018]. Alternatively to the presented algorithm, one may choose to calculate one ranking per metric without aggregation. We present this ranking method more formally in Algorithm 1 in Appendix A.2.

Case-based ranking scheme In a *case-based ranking* or *rank-then-aggregate method*, a ranking is computed for every test case over all participating teams. Those ranks are then aggregated (for example using the mean, median, or a specific percentile) for the final ranking [Wiesenfarth et al., 2021; Maier-Hein et al., 2018]. Alternatively to the presented algorithm, one may choose

to calculate one ranking per metric without aggregation. We present this ranking method more formally in Algorithm 2 in Appendix A.2.

Test-based ranking scheme In a *test-based method* or *significance ranking*, differences in the performances between teams are determined by pairwise comparisons using statistical significance testing. The ranking may then be based on the number of significant test results [Wiesenfarth et al., 2021; Maier-Hein et al., 2018]. Alternatively, relations obtained from the testing may be used for the ranking, as suggested by Demšar [2006]. Alternatively to the presented algorithm, one may choose to calculate one ranking per metric without aggregation. We present this ranking method more formally in Algorithm 3 in Appendix A.2 with the example of using a Wilcoxon signed rank test and the number of statistical test results as the ranking measure.

2.2 Fundamentals of statistics

This section serves as an overview of the statistical concepts used in this thesis. The first two parts 2.2.1 and 2.2.2 present techniques that can be leveraged for ranking comparisons and for measuring ranking variability. We conclude by explaining the concepts of a mixed model analysis in Section 2.2.3.

2.2.1 Ranking comparison

In the course of the thesis, we compare rankings at various points. This is done to identify the influence of various ranking schemes or to analyze ranking uncertainty. In the following, we present two measures that can be used for ranking comparisons: Kendall's tau and Spearman's footrule.

Kendall's tau rank correlation coefficient

Kendall's τ measures the correlation between two rankings by determining how often one rank order deviates from the other [Kendall, 1938]. More formally, consider a set of observations $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ from the random variables X and Y. The values of (x_i) and $(y_i), i = 1, \ldots, n$ describe two rankings of same lengths n with unique values of (x_i) and (y_i) in each ordering [Szmidt and Kacprzyk, 2011].

Definition 2.2.1 (Concordant and discordant pairs as defined by Szmidt and Kacprzyk [2011] and Nelsen [2001])**.** A pair of (x_i, y_i) and $(x_j, y_j), i < j$ is **concordant** if either

(i) $x_i > x_j$ and $y_i > y_j$ or

(ii) $x_i < x_j$ and $y_i < y_j$,

meaning that the order of sorting is the same in both rankings. Otherwise, the pair is called **discordant**.

When calculating Kendall's τ, we compute the S, namely the difference between the number of concordant pairs C and the number of discordant pairs D, and divide it by the total amount of possible pairs of observations $\binom{n}{2} = \frac{n(n-1)}{2}$ [Kendall, 1938]:

$$\tau = \frac{C - D}{\binom{n}{2}} = \frac{C - D}{\frac{n(n-1)}{2}} = \frac{2(C - D)}{n(n-1)} = \frac{2S}{n(n-1)} \text{ [Kendall, 1938]} \tag{2.1}$$

By definition, Kendall's τ is bounded between $[-1, 1]$. A value of 1 corresponds to a perfect agreement between rankings (similar rankings; all pairs are concordant) and -1 to a complete disagreement (reverse rankings; all pairs are discordant). Independent rankings yield a value of 0 (same number of concordant and discordant pairs) [Szmidt and Kacprzyk, 2011].

Spearman's footrule

Spearman's footrule denotes the absolute distance between two ranking vectors and can thus be used to compare the distance between two rankings. It is defined as

$$D(x, y) = \sum_{i=1}^{n} |x_i - y_i| \text{ [Spearman, 1987]} \tag{2.2}$$

where x_i and y_i refer to the ith rank of two rankings x and y. However, this distance does not consider the correlation between rankings. Thus, we mostly calculate Kendall's τ for ranking comparisons throughout this thesis.

2.2.2 Ranking variability

Throughout this thesis, we examine the stability and robustness of rankings. Bootstrapping refers to the process of resampling with replacement [Efron and Tibshirani, 1994]. Such a technique can be used to estimate errors, confidence intervals, or variances from an approximate distribution [Stine, 1989]. Bootstrapping can also be used to measure the uncertainty and variability of rankings. In this context, we simulate small variations to the original data set that was used to compute the ranking by sampling with replacement, as shown in Figure 2.2. The resulting rankings may differ from the original ranking due to those perturbations. The ranking uncertainty can then be assessed, for example by measuring Kendall's τ (see Section 2.2.1) between the original and each bootstrap ranking. Calculating several bootstrap rankings (we typically assume 1,000 bootstraps) can be used to identify the uncertainty of a ranking [Wiesenfarth et al., 2021].

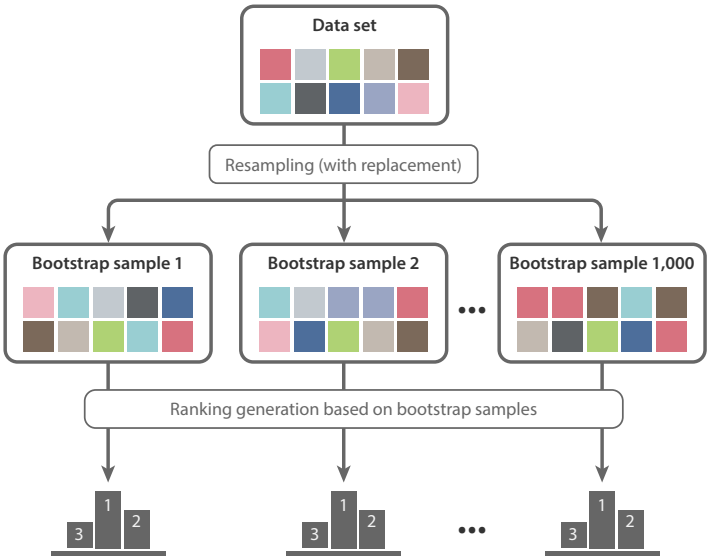


Figure 2.2: Illustration of bootstrapping in the case of ranking variability. Different cases from the data set are shown in different colors. To generate one bootstrap sample (ranking), samples from the original data set are randomly drawn with replacement until the original data set size is achieved. This means, a bootstrap sample may contain one case several times or not once, simulating small perturbations from the original data set. For every bootstrap sample, a ranking is computed based on the same ranking scheme.

2.2.3 Mixed model analysis

Biomedical image data often features a hierarchical data structure. Data from the same patient, device, or hospital is typically correlated and cannot be assumed to be independent. However, this property is required for classical linear regression models [West et al., 2006]. In such a case, we can leverage *Linear Mixed Models (LMMs)* [West et al., 2006]. LMMs have been designed as a generalization of linear models used for the regression of continuous variables in the case of non-independent data and are typically defined as:

$$Y = \mathbf{1}\alpha + X\beta + Zu + \epsilon \tag{2.3}$$

Y denotes the *outcome variable*, also known as the *dependent variable*, and refers to the outcome that should be regressed. It is an $N \times 1$ vector, with N being the number of data points. The outcome variable could, for example, refer to the performance of an algorithm for N images.

LMMs incorporate fixed and random effects. The *fixed effects* estimate the effect or influence of independent explanatory variables that are not random (i.e. fixed). X, an $N \times p$ matrix, denotes the *design matrix for the fixed effects*. The rows of X correspond to the data points (e.g. images), whereas the columns represent p explanatory variables. For instance, if we aim to measure how specific image artifacts affect algorithm performance, the fixed effects would be the p image artifacts (see for example [Roß et al., 2021]). The design matrix is weighted by the regression coefficients β, which are of $p \times 1$ shape. In the mentioned binary example of the presence of a specific image characteristic a, it would yield a change in the expected algorithm performance by β_a in comparison to the case in which a is not present in an image (see for example [Roß et al., 2021]). The scalar *intercept* α constitutes the average outcome in the case that all explanatory variables are not present and is multiplied by an $N \times 1$ vector of ones.

Random effects, on the other hand, refer to the non-independent variables that are correlated. Thus, they seek to explain the data hierarchy. In the example above, the image and patient would be modeled as a random effect. For q random effects, Z is the corresponding $N \times q$ *design matrix*. Every random effect, e.g. the image and patient, corresponds to one column of Z. The $q \times 1$ vector u represents the respective coefficients of regression and quantifies the how the random effects impact the outcome variable Y. In LMMs, we assume that u follows a normal distribution.

The error term, called *residuals*, is incorporated by the $N \times 1$ vector ϵ and refers to the variation that is not explained by the random and fixed effects. For LMMs, the residuals are assumed to follow a normal distribution, i.e. the outcome variables should be normally distributed given the explanatory variables. For the outcome variable, a transformation can be applied to make sure that it approximately follows a normal distribution. For example, to achieve a value range of $(-\infty, \infty)$, a *logit* transformation could be used for values that are bounded by $[0, 1]$. After the model fitting, the distribution of the residuals needs to be checked for normality. This could be done by analyzing the Quantile-Quantile plot (Q-Q plot) of residuals [Thode, 2002].

If this normality assumption is violated, we cannot utilize LMMs. Generalized Linear Mixed Models (GLMMs) [McCulloch and Searle, 2004] can be used as an alternative, as they generalize from outcome variables of a normal distribution to various other distributions, such as Bernoulli or Poisson. This is accomplished by including a *link function* [McCulloch and Searle, 2004; Roß et al., 2021]:

$$g(\mathbf{E}[Y]) = g(\pi) = \mathbf{1}\alpha + X\beta + Zu + \epsilon \tag{2.4}$$

A logit link function is a canonical choice in a binary environment based on a Bernoulli distribution, i.e. $g = \frac{\pi}{1-\pi}$.

2.3 Fundamentals of machine learning

There is no standardized definition of *Artificial Intelligence (AI)*. As stated by Luger [2005], definitions of AI "suffer from the fact that intelligence itself is not very well defined or understood". Poole and Mackworth [2010] describe AI as "the field that studies the synthesis and analysis of computational agents that act intelligently". Another opinion can be found in Russell and Norvig [2020], for instance, who discuss the differences between human or rational thinking and human or rational acting when defining AI. Moreover, Fan et al. [2020] and others discuss the connections between the human brain and AI. A broad discussion on various approaches to defining AI is out of the scope of this thesis and can be followed by reading the provided references. Generally, we can note that AI deals with machines that somehow act in an intelligent way. This thesis focuses on two sub-fields of AI: Machine Learning (ML) and Deep Learning (DL). The sub-field of AI called *ML* concentrates on learning from data and makes use of principles from both statistics and optimization [Mitchell and Mitchell, 1997; Sun et al., 2019; Goodfellow et al., 2016]. *DL*, on the other hand, is a sub-group of ML algorithms, which make use of Neural Networks (NNs) with multiple layers [Goodfellow et al., 2016].

ML algorithms can be subdivided into *supervised*, *unsupervised*, and *reinforcement learning* [Buduma et al., 2022; Goodfellow et al., 2016]. Supervised problems link a given input x to a label or target y. They aim to predict the labels of an input image. Unsupervised methods operate without labels and typically cluster the data based on similarities of intrinsic properties by recognizing patterns. Reinforcement learning, on the other hand, refers to the interaction between an agent and an environment in order to learn independently based on a reward system [Goodfellow et al., 2016]. In this thesis, we focus on supervised methods, which are often subdivided into regression and classification problems. We focus on the classification aspect.

2.3.1 Fundamentals of neural networks

The concept of NNs was derived from the functionality of neurons in the human brain [Rosenblatt, 1958]. The input neurons are modeled as a weighted sum, which is given to a non-linear activation function, mimicking the action potential of neurons [McCulloch and Pitts, 1943]:

$$\hat{y} = g\left(\sum_{i=1}^{n} x_i \cdot w_i + b \cdot w_0\right) \text{[Goodfellow et al., 2016]} \tag{2.5}$$

The output of the model is given by $\hat{y}$, predicting the target y. The n input variables are given by $x_i, i = 1, \ldots, n$ and w_i refer to the weights of the model that are learned during model training. The bias of the model is given by b and g denotes its non-linear activation function. NNs often produce pseudo probabilities between 0 and 1, which represent the likelihood of belonging to a particular class, rather than directly predicting a specific label or class. We refer to those pseudo probabilities as *predicted class scores*. They are also called confidence scores or continuous class scores. It should be noted that the predicted class scores should not automatically be interpreted as probabilities. More details on this aspect are provided in Section 2.4.5.

NNs that are composed of several layers of neurons are called deep NNs, thus referring to DL. DL algorithms are typically very powerful in modeling complex non-linear situations and may comprise intricate operations such as pooling, normalizations, or convolutional layers. Convolutional layers are especially useful in image analysis since they are able to capture the spatial information of an image in contrast to encoding an image as a vector, as it would by done in a standard neural network, for example. A network composed of convolutional layers is called a Convolutional Neural Network (CNN), and is used as a standard technique in biomedical image analysis (e.g. [Li et al., 2014; Shin et al., 2016]). In a convolutional layer, a so-called kernel or filter, containing weights, is used to capture specific information about the image in a sliding-window approach. Importantly, the kernel is always applied to a small region of the image only, before moving on to the next region. Following Goodfellow et al. [2016], a convolution operation for a two-dimensional image I and a kernel K is defined as:

$$S(i,j) = (K * I)(i,j) = \sum_m \sum_n I(i-m, j-n)K(m,n) \tag{2.6}$$

The weights of the kernel are learned by the network during training. CNNs are usually composed of three components. First, the network is typically composed of multiple convolutions, whose outputs are secondly given to a non-linear activation function. Finally, a pooling layer is applied, which aims for making the representation invariant to translations of the image [Goodfellow et al., 2016]. A pooling layer acts as a summary or aggregation function, for example by computing the maximum (max pooling) or the average (average pooling) of a specific region in the image. In this way, the dimension of the input is reduced.

2.3.2 Training and inference of neural networks

Data split and data augmentation

ML models learn to predict the labels of a data set during the process of training. Thus, the given data set should be split into a training set, comprising roughly 80% of the data, a validation, and a test set, comprising roughly 20% of the data [Goodfellow et al., 2016]. The training data is then solely used for training and optimizing the model, while the performance is reviewed on the validation set. This step is important to avoid the model overfitting on the training data. This indicates that the model cannot generalize to unseen data when it is strongly optimized on the training set. Finally, the test set should remain untouched until the final validation is performed.

For small data sets, one may choose to not define a specific validation set, but rather use other techniques, such as k-fold cross-validation [Stone, 1974]. In cross-validation, the data is randomly split into k subsets. For each subset, one of the subsets is considered as the hold-out test set, while the remaining data is treated as the training data. This process is repeated such that each of the subsets is considered as the test set once and the scores for each of the k folds are aggregated [Gareth et al., 2013]. Additionally, in the case of limited training data, the amount of data points can be synthetically increased by using *data augmentation*. . In this process, new data is generated with a certain probability such that the input images are slightly transformed. Common data augmentations involve the rotation of images, changing the color, or adding blur [Shorten and Khoshgoftaar, 2019].

Weight initialization and pre-training

The network weights must first be initialized before the training can begin. This could be done either randomly or with pre-defined weights. In the Xavier normalization, for example, the weights are initialized as random numbers from a uniform distribution in the range of $\pm\frac{1}{\sqrt{n}}$ [Glorot and Bengio, 2010]. Another common practice lies in the *pre-training* of networks, for which the weights are taken from a model that was already trained on a different task. Nowadays, pre-trained networks are available in common DL libraries and are often based on the famous ImageNet [Deng et al., 2009] or Microsoft-Common Objects in COntext (COCO) [Lin et al., 2014] data sets.

Loss functions

The goal of the training step lies in minimizing a *loss function* in order to optimize the network weights. The score of the loss function reflects the performance of the model in predicting the reference labels. In the following, we present some popular loss functions that will be used in this thesis.

The *Cross Entropy (CE) loss* (also known as *logarithmic loss*) [Cybenko et al., 1998] is a popular loss function in classification problems, which is defined as

$$\mathcal{L}_{CE}(y, \hat{y}) = -\sum_{i=1}^{N} y_i \cdot \log(\hat{y}) \tag{2.7}$$

for N classes. Here, y and $\hat{y}$ denote the reference and predicted vector, in which every component refers to a specific class. The loss is calculated and summed up for the different classes i. It is known as *Binary Cross Entropy (BCE) loss* for binary classification tasks and simplifies to

$$\mathcal{L}_{BCE}(y, \hat{y}) = -\big(y \cdot \log(\hat{y}) + (1-y) \cdot \log(1-\hat{y})\big), \tag{2.8}$$

with y and $\hat{y}$ reducing to scalars.

The *focal loss* [Lin et al., 2017b] has been suggested as an improvement of the CE loss, especially for binary object detection problems. The advantages include better handling of class imbalances via the parameter α_t and the ability to differentiate easy and hard cases by the focusing parameter γ:

$$\mathcal{L}_{FL}(\hat{y}_t) = -\alpha_t \cdot (1-\hat{y}_t)^{\gamma} \cdot \log(\hat{y}_t) \tag{2.9}$$

with

$$\hat{y}_t = \begin{cases} \hat{y} & y = 1 \\ 1-\hat{y} & \text{otherwise} \end{cases} \tag{2.10}$$

Note that the focal loss is identical to the BCE loss for $\gamma = 0$ and $\alpha_t = 1$.

The *Smooth L_1 loss*, also known as *Huber loss* [Huber, 1992; Girshick, 2015] can be interpreted as an L_1 function with a smooth transition for values near zero to avoid exploding gradients and is defined as

$$\mathcal{L}_{smooth-L1}(y, \hat{y}) = \begin{cases} 0.5 \cdot (y - \hat{y})^2 & |y| < 1 \\ |y| - 0.5 & \text{otherwise} \end{cases} \tag{2.11}$$

It is often used in object detection tasks for the prediction of bounding boxes.

The *Dice Similarity Coefficient (DSC) loss* [Milletari et al., 2016], used for segmentation tasks, follows the definition of the DSC validation metric (see Section 2.4) and measures the overlap between a prediction and the reference on pixel level. In this case, it is iterated over pixels i. The loss is defined as:

$$\mathcal{L}_{DSC}(y, \hat{y}) = \frac{2 \cdot \sum_{i=1}^{N} y_i \cdot \hat{y}_i}{\sum_{i=1}^{N} y_i^2 + \sum_{i=1}^{N} \hat{y}_i^2}. \tag{2.12}$$

Finally, the Jaccard loss [Iglovikov and Shvets, 2018] for segmentation tasks similarly follows the definition of the Intersection over Union (IoU) metric (also known as Jaccard Index) (see Section 2.4), calculating the overlap of the prediction and reference. It is given by

$$\mathcal{L}_{Jacc}(y, \hat{y}) = \frac{1}{N} \sum_{i=1}^{N} \frac{y_i \cdot \hat{y}_i}{y_i + \hat{y}_i - y_i \cdot \hat{y}_i}. \tag{2.13}$$

Both the DSC and Jaccard losses are typically calculated per class.

Optimization

Given the typically high amount of data, the data set is usually further split into several *(mini) batches*. The number of batches is one of the hyperparameters of a NN, the so-called *batch size*. In every training step, the trainable weights of the network are updated via the gradient of the loss. This step is achieved by the concept of *backpropagation* [Goodfellow et al., 2016]. In an optimization step, the loss function is minimized, typically done by gradient descent. The optimizer of a network is based on the *learning rate*, another hyperparameter, which defines the step length of the optimization and should be carefully chosen. A learning rate that is too high may miss the minimum, while a low learning rate slows down the training process. The commonly used *Stochastic Gradient Descent (SGD)* optimizer [Kiefer and Wolfowitz, 1952] relies on the basic concept of gradient descent, but rather than calculating the gradient descent for the whole data set, it is computed over randomly selected batches. Another standard optimizer is the *Adaptive Moment Estimation (Adam)* optimizer [Kingma and Ba, 2014], which is similarly computed over batches. It further uses adaptive learning rates instead of a single learning rate (as for the SGD). The training step is repeated for several *epochs*[1], until the maximum number of epochs is reached or a pre-defined stop criterion is reached.

[1] One epoch may be finished once either all data or a specific number of batches was presented to the network.

Model selection and inference

Finally, after the model training is completed, the final model is selected based on specific criteria, for example, a specific score of the loss function of the validation set or based on application-related metrics. The model is then tested against the unseen test data set. This step is typically referred to as *inference*. The final performance of the model on the hidden test set is given by the validation metrics.

2.3.3 Model architectures used in this thesis

Residual Neural Network (ResNet)

Very deep NNs often show a degradation in training performance since they suffer from gradients that converge to zero very quickly [He et al., 2016]. This concept is known as "vanishing gradients". Vanishing gradients cause the learning of the network to stop because the weights are not updated anymore given the gradients turned to zero. Thus, He et al. [2016] have introduced *residual blocks* in their *Residual Neural Network (ResNet)* architecture. Those blocks help to overcome the problem of vanishing gradients by using so-called *skip connections* or *identity connections*, as shown in Figure 2.3. Those connections skip some layers of the network by connecting them via the identity function. With this modification, the gradients are enabled to directly flow through the skip connections in the backpropagation step. In this way, very deep architectures can be created, such as the ResNet-50 or ResNet-101, with 50 or 101 layers.

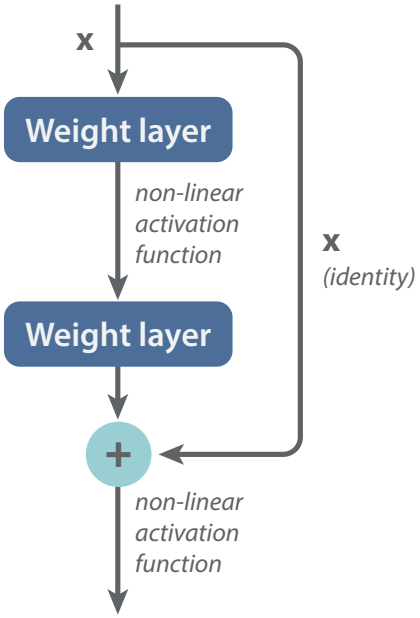


Figure 2.3: Residual block as proposed by He et al. [2016].

U-Net

Ronneberger et al. [2015] suggested the *U-Net* architecture in the context of semantic segmentation of biomedical images as a fully convolutional NN. It adheres to the principles of an Autoencoder by using an encoder-decoder structure, which, together with the below-mentioned skip connections, naturally introduce the "U" shape (see Figure 2.4), the basis for the architecture's name. The downsampling path of the network (encoder; left part of Figure 2.4) serves the purpose of capturing the context of the input image by compressing the spatial information (bottleneck), but increasing the number of feature channels. This is achieved by step-wise convolutions, followed by non-linear activation functions and a pooling operation, reducing the spatial resolution of the image. The encoder results in an increasing semantic explanatory ability of the network at the expense of a reduced spatial resolution of the image. Encoders are typically used for the classification part of a model.

The upsampling path (decoder; right part of Figure 2.4) is symmetric to the encoder and serves the purpose of creating a segmentation map at increasing image resolutions from the bottleneck to ensure a correct localization of objects in the original image. It utilizes step-wise upsampling operations instead of pooling. To combine the local and global information, the upsampling also involves the concatenation of the encoder and decoder layers of the same sizes (skip connections) to overcome the problem of information loss in the bottleneck. As the U-Net was designed for semantic segmentation problems, it naturally cannot distinguish between different instances or objects. If used for instance segmentation problems, the network output is usually post-processed with a *connected component analysis*, which separates different objects into multiple instances [Chen et al., 2006].

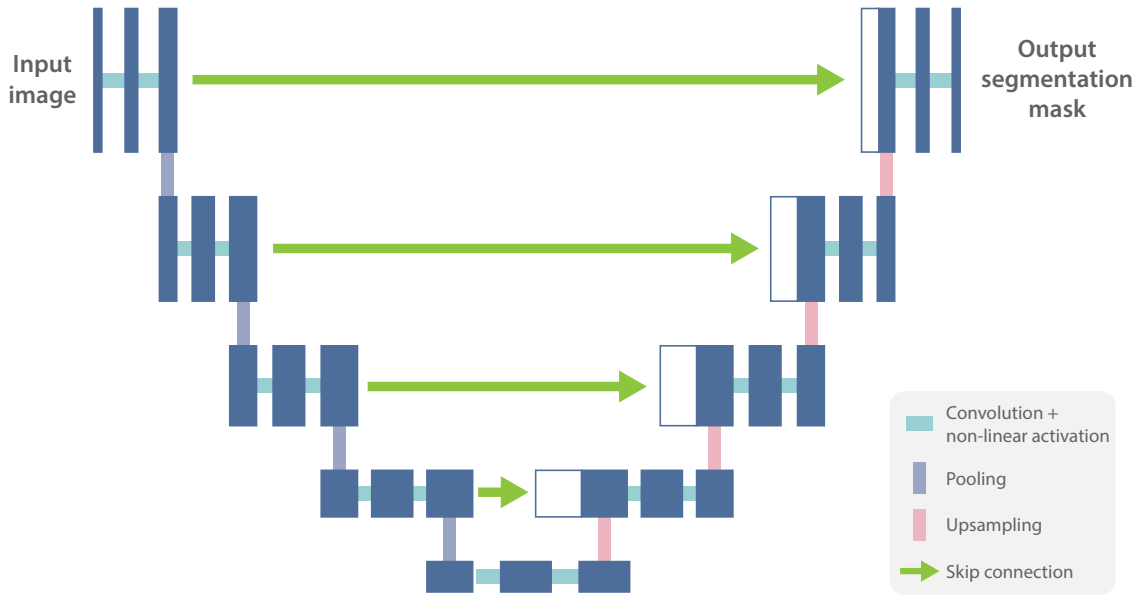


Figure 2.4: U-Net architecture as proposed by Ronneberger et al. [2015].

Mask R-CNN

In contrast to the U-Net, the *Mask R-CNN* [He et al., 2017] was directly designed for solving instance segmentation problems. The architecture is composed of several small NNs handling different steps of the pipeline. First, the *backbone* is used to convert the input image to a latent representation, aiming to extract relevant features. It is typically a pre-trained classification or detection network, such as the ResNet, a VGG [Simonyan and Zisserman, 2014] or a Feature Pyramid Network (FPN) [Lin et al., 2017a]. Based on the feature representation, the so-called *Region Proposal Network (RPN)* predicts several regions in the image, which may be candidates for the objects that should be segmented. This network is typically a small binary classifier. In the following steps, bounding box coordinates for the different regions and their classes are predicted for the different regions and the objects within the boxes are segmented. All of those steps are done by individual small CNNs. The architecture is trained via a multi-task loss:

$$\mathcal{L}_{maskrcnn}(y, \hat{y}) = \mathcal{L}_{cls}(y, \hat{y}) + \mathcal{L}_{box}(y, \hat{y}) + \mathcal{L}_{mask}(y, \hat{y}) \tag{2.14}$$

The classification loss $\mathcal{L}_{cls}$ is typically defined as a CE loss (see Equation 2.7) and assesses the performance of the object classifier. The bounding box regression loss $\mathcal{L}_{box}(y, \hat{y})$ rates the accuracy of the bounding boxes and was introduced as a smooth L_1 loss (see Equation 2.11). Finally, $\mathcal{L}_{mask}(y, \hat{y})$ measures the segmentation quality and was defined as the average BCE loss (see Equation 2.8) [He et al., 2017].

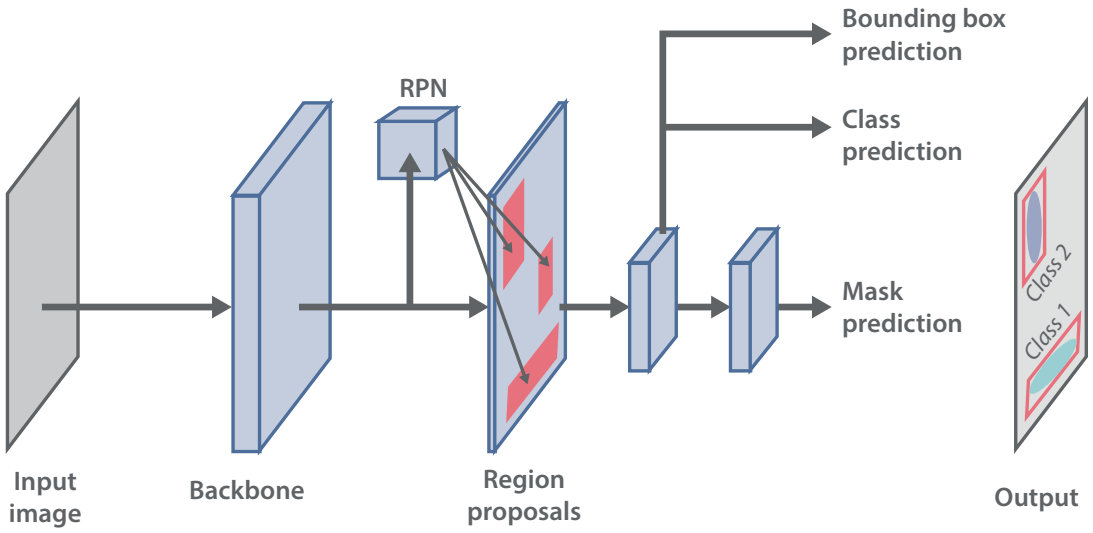


Figure 2.5: Mask R-CNN architecture as proposed by He et al. [2017]. The architecture is composed of several sub-networks: the backbone, the Region Proposal Network (RPN) and several small Convolutional Neural Networks (CNNs), which predict the bounding box, class, and segmentation mask for every proposed region.

2.4 Common validation metrics

Disclosure

This section presents common validation metrics used in Sections 4.2 and 5.2 of this thesis. We follow the fundamentals presented in the work of Reinke et al. [2021a] for a comprehensive overview.

In this section, we present common validation metrics used to assess the performance of algorithms. We focus on validation metrics that assess the classification abilities of a prediction. This can be done at various levels, based on the underlying problem category. Particularly, we consider the following levels: Per image (image-level classification), per object (object detection), per pixel (semantic segmentation) as well as per object and per pixel (instance segmentation).

In this thesis, our focus are supervised Machine Learning (ML) algorithms that link a given input x to a label or target y [Goodfellow et al., 2016] (see Section 2.3). The algorithms aim to predict the labels of a reference annotation, which typically involves manual annotation efforts. In this section, we consider reference-based validation metrics that analyze how well the algorithm replicated the reference by comparing the prediction to the reference and providing a score.

In the following, we introduce several common reference-based validation metrics related to the four problem categories. Some of them overlap between categories and can be used for different problem formulations, as indicated below. Appendix A.3 provides a profile of the most-relevant metrics presented in this section including relevant information such as the formula, definition, information on cardinalities, prevalence dependency, and metric family. The profile further contains relevant limitations and recommendations based on the results of Sections 4.2 and 5.2.

Most of the presented metrics rely on the creation of the *confusion matrix*, also known as the *error matrix* [Stehman, 1997; Tharwat, 2020], by comparing an algorithm prediction to a reference annotation as illustrated in Figure 2.6 for the binary case. In the binary case, the confusion matrix divides a prediction into the following cardinalities:

True Positives (TPs) i.e. positively predicted samples that are actually positive,

False Negatives (FNs) i.e. negatively predicted samples that are actually positive,

False Positives (FPs) i.e. positively predicted samples that are actually negative and

True Negatives (TNs) i.e. negatively predicted samples that are actually negative.

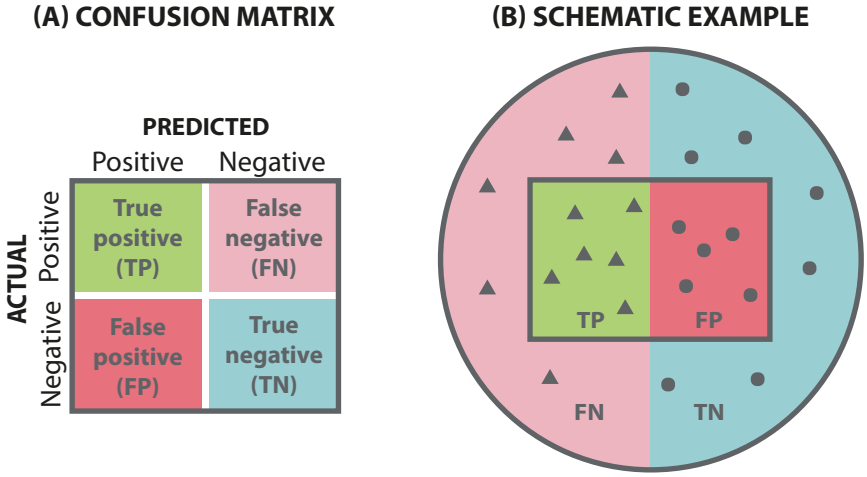


Figure 2.6: Visualization of a binary confusion matrix. (A) The confusion matrix shows the cardinalities True Positive (TP) (green), False Negative (FN) (pink), False Positive (FP) (red), and True Negative (TN) (blue). (B) Schematic example of a binary classification into positive (triangle) and negative (circle) classes.

For non-binary problems with C classes, $C > 2$, the confusion matrix is extended to a $C \times C$ matrix, from which the entries are also referred to as cardinalities. To introduce the concepts of the presented metrics, we restrict ourselves to the binary case for illustration purposes. However, most metrics can be generalized to the multi-class case, as described below.

Neural networks nowadays typically output predicted class scores instead of a fixed class [Vaicenavicius et al., 2019] (see Section 2.3). In order to calculate the entries of the confusion matrix, a threshold or cutoff value τ needs to be set. Predictions greater than or equal to τ are considered positive, while predictions below the cutoff value are interpreted as negative. In comparison to the actual positive and negative sample distributions, this results in the confusion matrix for this specific threshold. For a different threshold, the confusion matrix may change, since the number of positive and negative predicted samples may differ. The concept is illustrated in Figure 2.7 for $\tau = 0.5$. It should be noted that setting a cutoff value is more complicated for $C > 2$ (multi-class), for which one might need to define decision regions as a partition of $[0, 1]^{C-1}$. Alternatively, a global decision threshold needs to be defined across all classes.

For image-level classification, semantic segmentation, object detection, and instance segmentation, validation metrics can roughly be classified into **counting metrics**, **multi-threshold metrics**, **distance metrics**, and **calibration metrics**, as described in the following paragraphs.

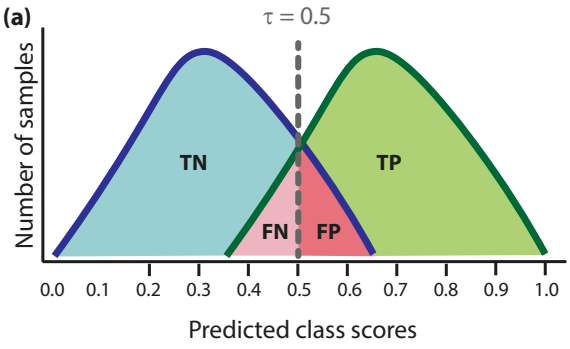
Counting metrics Counting metrics rely on a fixed classification threshold, based on which the confusion matrix is calculated. Counting metrics then compute different ratios of its cardinalities. We further distinguish between per-class counting metrics and multi-class counting metrics. *Per-class counting metrics* require the definition of a concrete positive (or foreground) and negative (or background) class. *Multi-class counting metrics* are defined for multiple classes while being invariant to the order of classes. In the case of segmentation problems, counting metrics are often referred to as ***overlap-based metrics***, as those metrics focus on the predicted

pixels that are overlapping with the reference segmentation pixels (TP).

Multi-threshold metrics In contrast to counting metrics, multi-threshold metrics operate on multiple thresholds. Usually, predictions are ranked by the predicted class scores and the confusion matrix is dynamically calculated based on the current threshold. Multi-threshold metrics are typically the (approximated) area under a curve drawn from the multiple thresholds.

Distance-based metrics In the case of semantic and instance segmentation problems, the distance from the object boundaries are often of specific interest, measured by distance-based metrics, often referred to as ***boundary-based metrics***.

Calibration metrics Calibration metrics focus on the accuracy of the predicted class scores and how close they are to the actual probabilities for a certain class.



Actual classe	Predicted class score	Predicted class	Cardinality
1 (positive)	0.8	1 (positive)	TP
1 (positive)	0.6	1 (positive)	TP
0 (negative)	0.7	1 (positive)	FP
0 (negative)	0.3	0 (negative)	TN
1 (positive)	0.4	0 (negative)	FN
...	...	...	...

Figure 2.7: Illustration of thresholding of predicted class scores. (a) The green curve shows the distribution of the positive, the dark blue curve the distribution of the negative class. An exemplary threshold τ of 0.5 is set to be able to compute the cardinalities of the confusion matrix, i.e. True Positive (TP) (light green), False Positive (FP) (dark red), False Negative (FN) (light red) and True Negative (TN) (light blue). Samples with class scores above the threshold are predicted as positive, while values below the threshold are predicted as negative. (b) Example for one prediction. The table shows the actual class, the predicted class scores, their class based on $\tau = 0.5$, and the resulting cardinalities. Figure adapted from [Reinke et al., 2021a].

2.4.1 Image-level classification metrics

In classification tasks, it is important to be aware of the **prevalence** ϕ. In the context of epidemiology, prevalence refers to the proportion of a population that is affected by a particular disease [Rothman, 2012]. More generally, prevalence is defined as the proportion of actual positive samples in a data set:

$$\phi = \frac{\mathrm{TP} + \mathrm{FN}}{N} \in [0, 1], N = \mathrm{TP} + \mathrm{FN} + \mathrm{FP} + \mathrm{TN} \tag{2.15}$$

Several metrics are dependent on the underlying prevalence of the data set. In this case, metric values can not directly be compared across data sets with different ϕ. Throughout this section, we indicate which metrics depend on the prevalence. Figure 2.8 illustrates the behavior of the presented image-level classification counting metrics as functions of the prevalence for given Sensitivity and Specificity values.

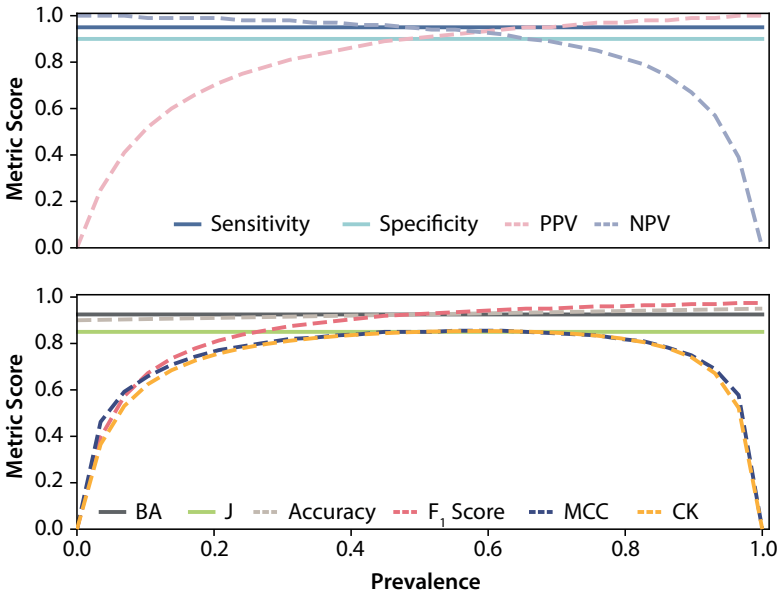


Figure 2.8: Image-level counting metrics as functions of the prevalence for a fixed Sensitivity of 0.95 and a fixed Specificity of 0.90. Used abbreviations: Balanced Accuracy (BA), Cohen's Kappa (CK), Youden's Index (J), Matthews Correlation Coefficient (MCC), Negative Predictive Value (NPV), Positive Predictive Value (PPV). Figure adapted from [Reinke et al., 2021a].

Per-class counting metrics

An illustration of all presented per-class counting metrics is given in Figure 2.9.
The **Sensitivity** (also known as **Recall**, **True Positive Rate (TPR)** or **Hit rate**) [Tharwat, 2020] measures how many of the actual positive samples were predicted as such. Sensitivity is independent of the prevalence and calculated as follows:

$$\text{Sensitivity} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \in [0, 1] \tag{2.16}$$

The **Specificity** (also known as **True Negative Rate (TNR)** or **Selectivity**) [Tharwat, 2020] measures how many of the actual negative samples were predicted as such. Similar to Sensitivity, Specificity is independent of the prevalence and calculated as follows:

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \in [0, 1] \tag{2.17}$$

Sensitivity and Specificity are usually calculated together to assess the full performance of a prediction to classify images into positive and negative samples.

The **PPV** (also known as **Precision**) metric [Tharwat, 2020] represents the likelihood that a prediction marked as positive matches an actual positive sample. For $\phi = 0.5$, PPV can be computed as

$$\text{PPV} = \frac{\text{TP}}{\text{TP} + \text{FP}} \in [0, 1] \tag{2.18}$$

However, PPV crucially depends on the prevalence ϕ, as shown in Figure 2.8. Thus, if the prevalence is different from 0.5, the following formula should be used for prevalence correction:

$$\text{PPV}_{\text{corrected}} = \frac{\text{Sensitivity} \cdot \phi}{\text{Sensitivity} \cdot \phi + (1 - \text{Specificity}) \cdot (1 - \phi)} \in [0, 1] \tag{2.19}$$

Please note that typical sample implementations (e.g. [Maier, 2013; MONAI-developers, 2019; Scikitlearn, 2007; PyTorchLightning, 2020]) often only refer to Equation 2.18, ignoring the necessary prevalence-correction given in Equation 2.19 for prevalences unequal to 0.5. It should be noted that we focus on the term PPV rather than Precision in this thesis because different scientific communities use the word "precision" in different ways. For instance, it frequently relates to the degree of output confidence in the medical community.

NPV [Tharwat, 2020] is the negative equivalent of PPV and represents likelihood that a prediction marked as negative matches an actual negative sample. For a prevalence of 50%, NPV can be computed as

$$\text{NPV} = \frac{\text{TN}}{\text{TN} + \text{FN}} \in [0, 1] \tag{2.20}$$

Similar to PPV and as shown in Figure 2.8, NPV depends on ϕ:

$$\text{NPV}_{\text{corrected}} = \frac{\text{Specificity} \cdot (1 - \phi)}{\text{Specificity} \cdot (1 - \phi) + (1 - \text{Sensitivity}) \cdot \phi} \in [0, 1] \tag{2.21}$$

As described for PPV, common sample implementations of NPV ([Maier, 2013; MONAI-developers, 2019]) refer to Equation 2.20 rather than using the corrected formula, displayed in Equation 2.21.

The $\mathbf{F}_\beta$ **Score** [Chinchor, 1992] weights PPV, i.e. the FP samples, and Sensitivity, i.e. the FN samples, by the parameter β:

$$\text{F}_\beta = (1 + \beta^2) \frac{\text{PPV} \cdot \text{Sensitivity}}{\beta^2 \cdot \text{PPV} + \text{Sensitivity}} = \frac{(1 + \beta^2) \cdot \text{TP}}{(1 + \beta^2) \cdot \text{TP} + \beta^2 \cdot \text{FN} + \text{FP}} \in [0, 1], \beta \in \mathbb{R}_+ \tag{2.22}$$

A higher value of β (e.g. $\beta = 2$) gives Sensitivity or FN samples a higher weight over PPV or FP samples. Similarly, a value lower than 1 (e.g. $\beta = 0.5$) weights PPV or FP over Sensitivity or

FN samples. It is quite common to set β to 1, which results in the harmonic mean of PPV and Sensitivity:

$$\mathsf{F}_1 = \frac{2 \cdot \mathsf{PPV} \cdot \mathsf{Sensitivity}}{\mathsf{PPV} + \mathsf{Sensitivity}} = \frac{2 \cdot \mathsf{TP}}{2 \cdot \mathsf{TP} + \mathsf{FN} + \mathsf{FP}} \in [0, 1] \tag{2.23}$$

The F_β Score depends on PPV, implying prevalence dependency (see Figure 2.8).

The **Positive Likelihood Ratio (LR+)** [Attia, 2003] indicates the factor by which a positive prediction occurs more frequently among positive samples than among negative samples. LR+ is the odds ratio of Sensitivity and Specificity and thus invariant to the prevalence [Šimundić, 2009]:

$$\mathsf{LR+} = \frac{\mathsf{Sensitivity}}{1 - \mathsf{Specificity}} \in [0, \infty) \tag{2.24}$$

Finally, the **Net Benefit (NB)** [Vickers and Elkin, 2006] has been introduced as "a simple type of decision analysis, with benefits and harms put on the same scale so that they can be compared directly" [Vickers et al., 2016]. More specifically, the benefit is given by the fraction of TPs and the harm by the fraction of FPs, which is multiplied by an exchange rate, defined via the decision thresholds p_t:

$$\mathsf{NB} = \frac{\mathsf{TP}}{\mathsf{TP} + \mathsf{FN} + \mathsf{FP} + \mathsf{TN}} - \frac{\mathsf{FP}}{\mathsf{TP} + \mathsf{FN} + \mathsf{FP} + \mathsf{TN}} \cdot \frac{p_t}{1 - p_t} \in [-1, 1] \tag{2.25}$$

NB can also be used to tune the decision threshold of a model and is typically calculated for a range of clinically-relevant thresholds.

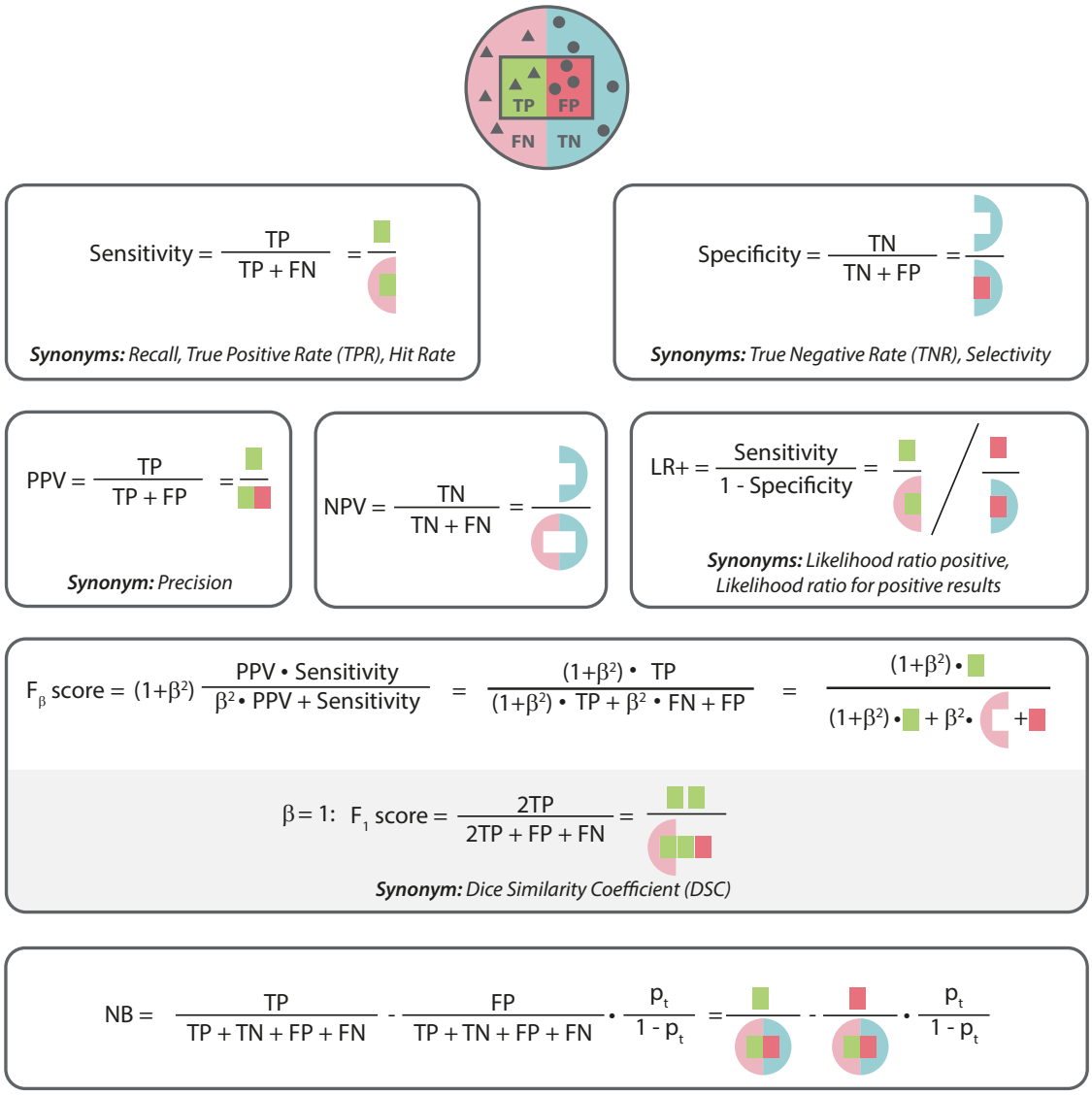


Figure 2.9: Overview of the most frequent per-class counting classification metrics that are based on the confusion matrix's cardinalities True Positive (TP), False Negative (FN), False Positive (FP) and True Negative (TN), as illustrated in Figure 2.6. Used abbreviations: Positive Likelihood Ratio (LR+), Negative Predictive Value (NPV), Positive Predictive Value (PPV). Figure adapted from [Reinke et al., 2021a].

Multi-class counting metrics (binary)

An illustration of all presented multi-class counting metrics is given in Figure 2.10.

Accuracy [Tharwat, 2020] is a well-known metric in image-level classification and measures the ratio of samples that were correctly predicted over all predictions made. The metric can be calculated as:

$$\begin{aligned} \text{Accuracy} &= \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FN} + \text{FP} + \text{TN}} \\ &= \text{Sensitivity} \cdot \phi + \text{Specificity} \cdot (1 - \phi) \in [0, 1] \end{aligned} \tag{2.26}$$

As can be seen from Equation 2.26, Accuracy is dependent on ϕ. In contrast to other multi-class counting metrics, Accuracy does not weight different classes equally.

The **Balanced Accuracy (BA)** [Tharwat, 2020] is often used as an alternative metric given its prevalence independence. It measures the arithmetic mean of Sensitivity and Specificity:

$$\text{BA} = \frac{\text{Sensitivity} + \text{Specificity}}{2} \in [0, 1] \tag{2.27}$$

In the case of $\phi = 0.5$, Accuracy is equivalent to the BA.

A related metric is the **Youden's Index (J)** [Youden, 1950] (also known as **Bookmaker Informedness (BM)**), measuring the sum of Sensitivity and Specificity:

$$\text{J} = \text{Sensitivity} + \text{Specificity} - 1 \in [-1, 1] \tag{2.28}$$

Both BA and J are independent of the prevalence (see Figure 2.8). J can directly be derived from the BA score:

$$\text{BA} = \frac{\text{J} + 1}{2} \tag{2.29}$$

$$\text{J} = 2 \cdot \text{BA} - 1 \tag{2.30}$$

Matthews Correlation Coefficient (MCC) [Matthews, 1975; Chicco et al., 2021] (also known as **Phi Coefficient**) calculates how the prediction is correlated with the actual outcome. The metric is calculated as follows:

$$\begin{aligned} \text{MCC} =& \frac{\text{TP} \cdot \text{TN} - \text{FP} \cdot \text{FN}}{\sqrt{(\text{TP} + \text{FP}) \cdot (\text{TP} + \text{FN}) \cdot (\text{TN} + \text{FP}) \cdot (\text{TN} + \text{FN})}} \in [-1, 1] \\ =& \sqrt{\text{PPV} \cdot \text{Sensitivity} \cdot \text{Specificity} \cdot \text{NPV}} \\ & - \sqrt{(1 - \text{PPV}) \cdot (1 - \text{Sensitivity}) \cdot (1 - \text{Specificity}) \cdot (1 - \text{NPV})} \end{aligned} \tag{2.31}$$

Given the dependency on PPV and NPV, MCC is similarly dependent on ϕ, as shown in Figure 2.8.

Cohen's Kappa (CK) [Cohen, 1960] was introduced to measure the agreement between two raters. In the ML community, it is often used for a comparison between the reference and the prediction. It incorporates the agreement given by change with the term p_e in Equation 2.32:

$$\begin{aligned} \mathsf{CK} &= \frac{p_0 - p_e}{1 - p_e} = \frac{2 \cdot (\mathsf{TP} \cdot \mathsf{TN} - \mathsf{FP} \cdot \mathsf{FN})}{(\mathsf{TP} + \mathsf{FP}) \cdot (\mathsf{TN} + \mathsf{FP}) + (\mathsf{TP} + \mathsf{FN}) \cdot (\mathsf{TN} + \mathsf{FN})} \in [-1, 1] \\ p_0 &= \mathsf{Accuracy} \\ p_e &= \frac{(\mathsf{TP} + \mathsf{FP}) \cdot (\mathsf{TP} + \mathsf{FN})}{\mathsf{TP} + \mathsf{FN} + \mathsf{FP} + \mathsf{TN}} + \frac{(\mathsf{TN} + \mathsf{FP}) \cdot (\mathsf{TN} + \mathsf{FN})}{\mathsf{TP} + \mathsf{FN} + \mathsf{FP} + \mathsf{TN}} \end{aligned} \tag{2.32}$$

CK can also be rewritten in terms of Sensitivity, Specificity, and prevalence ϕ [Feuerman and Miller, 2008]:

$$\mathsf{CK} = \frac{2 \cdot \phi \cdot (1 - \phi) \cdot (\mathsf{Sensitivity} + \mathsf{Specificity} - 1)}{\phi^2 + (1 - \phi)^2 + (1 - 2 \cdot \phi) \cdot (\phi \cdot \mathsf{Sensitivity} - (1 - \phi) \cdot \mathsf{Specificity})} \tag{2.33}$$

For a prevalence of 0.5, CK is equal to J.

An extension of CK is its weighted version **Weighted Cohen's Kappa (WCK)** [Cohen, 1960]. In contrast to the unweighted version, it takes into account the degree of disagreement between reference and prediction.

$$\begin{aligned} \mathsf{WCK} &= \frac{p_0^w - p_e^w}{1 - p_e^w} \in (-\infty, \infty) \\ p_0^w &= \frac{w_{\mathsf{TP}} \cdot \mathsf{TP} + w_{\mathsf{FN}} \cdot \mathsf{FN} + w_{\mathsf{FP}} \cdot \mathsf{FP} + w_{\mathsf{TN}} \cdot \mathsf{TN}}{} \\ p_e^w &= w_{\mathsf{TP}} \cdot \frac{(\mathsf{TP} + \mathsf{FP}) \cdot (\mathsf{TP} + \mathsf{FN})}{\mathsf{TP} + \mathsf{FN} + \mathsf{FP} + \mathsf{TN}} + w_{\mathsf{TN}} \cdot \frac{(\mathsf{TN} + \mathsf{FP}) \cdot (\mathsf{TN} + \mathsf{FN})}{\mathsf{TP} + \mathsf{FN} + \mathsf{FP} + \mathsf{TN}} \\ &\quad + w_{\mathsf{FN}} \cdot \frac{(\mathsf{FN} + \mathsf{FP}) \cdot (\mathsf{FN} + \mathsf{TN})}{\mathsf{TP} + \mathsf{FN} + \mathsf{FP} + \mathsf{TN}} + w_{\mathsf{FP}} \cdot \frac{(\mathsf{FP} + \mathsf{TP}) \cdot (\mathsf{FP} + \mathsf{TN})}{\mathsf{TP} + \mathsf{FN} + \mathsf{FP} + \mathsf{TN}} \end{aligned} \tag{2.34}$$

Usually, w_{TP} and w_{TN} are chosen to be equal to zero, while w_{FP} and $w_{\mathsf{FN}} > 0$, yielding WCK bounded between -1 and 1, similar to CK.

Unlike all other metrics presented so far, MCC, CK, and WCK do not give us an easy-to-interpret value. For all three of them, a value of 1 indicates a perfect prediction, -1 means total disagreement and a value of 0 refers to a prediction that is not better than a naive classifier, which would always predict only one class (random performance). However, there is no standard for interpreting the values in between.

Finally, the **Expected Cost (EC)** [Bishop and Nasrabadi, 2006; Hastie et al., 2009; Ferrer, 2022] is a metric describing a weighted sum of error rates, in which each error rate can be weighted individually given its severity. It is defined as

$$\begin{aligned} \mathsf{EC} &= \mathsf{C}_{\mathsf{miss}} \cdot \mathsf{P}_{\mathsf{miss}} \cdot \mathsf{P}_{\mathsf{tar}} + \mathsf{C}_{\mathsf{FA}} \cdot \mathsf{P}_{\mathsf{FA}} \cdot (1 - \mathsf{P}_{\mathsf{tar}}) \in (0, \infty) \\ \mathsf{P}_{\mathsf{miss}} &= \frac{\mathsf{FN}}{\mathsf{TP} + \mathsf{FN}} \\ \mathsf{P}_{\mathsf{FA}} &= \frac{\mathsf{FP}}{\mathsf{FP} + \mathsf{TN}}, \end{aligned} \tag{2.35}$$

where P_{miss} describes the miss rate or FN rate and P_{FA} the false alarm or FP rate. P_{tar} is the prior probability of the positives, equivalent to the prevalence. Finally, C_{miss} and C_{FA} refer to the estimated costs of both error rates, which are typically positive real numbers. The EC can be reformulated as

$$\begin{aligned} \theta &= \frac{P_{tar}}{1 - P_{tar}} \cdot \frac{C_{miss}}{C_{FA}} \\ P_{eff} &= \frac{\theta}{1 - \theta} \\ EC &= P_{eff} \cdot P_{miss} + (1 - P_{eff}) \cdot P_{FA}, \end{aligned} \tag{2.36}$$

such that the deduced parameter P_{eff} becomes a hyperparameter, which can be used to individually weight the error rates. In EC, P_{tar} can be adjusted to reflect the intended behavior, such as using the prevalence ϕ, setting $\phi = 0.5$, or inserting any anticipated future prevalence. Furthermore, EC can be used to assess the discrimination and calibration ability of a model in a single parameter. By calculating the difference of an EC based on an analytically ideal threshold value τ_1 and an empirically optimized threshold τ_2, the calibration performance of a prediction can be measured.

Based on the choice of the costs and priors, the values of EC may change, making a direct comparison difficult. For this reason, EC can be normalized by dividing it using the value that EC would take for a random classifier, i.e. a classifier that always takes the same decision d. For a classifier always predicting the positive class, FNs and TNs are equal to zero, meaning that EC reduces to

$$d = \text{positive} : EC = C_{FA} \cdot (1 - P_{tar}) \text{ given that } P_{miss} = 0 \ (FN = 0), P_{FA} = 1 \ (TN = 0) \tag{2.37}$$

Similarly, for always predicting the negative class, TPs and FPs are equal to zero, such that EC reduces to

$$d = \text{negative} : EC = C_{miss} \cdot P_{tar} \text{ given that } P_{miss} = 1 \ (TP = 0), P_{FA} = 0 \ (FP = 0) \tag{2.38}$$

The normalized variant of EC divides EC by the value it takes for such a naive system [Ferrer, 2022]:

$$EC_{norm} = \frac{C_{miss} \cdot P_{miss} \cdot P_{tar} + C_{FA} \cdot P_{FA} \cdot (1 - P_{tar})}{\min(C_{miss} \cdot P_{tar}, C_{FA} \cdot (1 - P_{tar}))} \tag{2.39}$$

Lastly, according to Ferrer [2022] and based on the underlying conditions (choice of costs and priors), EC can be reformulated such that:

$$\begin{aligned} EC &= 1 - \text{Accuracy} \\ EC &= 1 - BA \\ EC &= (1 - F_\beta) \cdot \left(\beta^2 \cdot P_{tar} \cdot \frac{TP + FP}{N} \right) \\ EC &= 1 - (LR+ - 1) \cdot P_{FA} \\ EC &= 1 - \frac{MCC}{\sqrt{\frac{(FP+TN)\cdot(TP+FN)}{(FN+TN)\cdot(TP+FP)}}} \end{aligned} \tag{2.40}$$

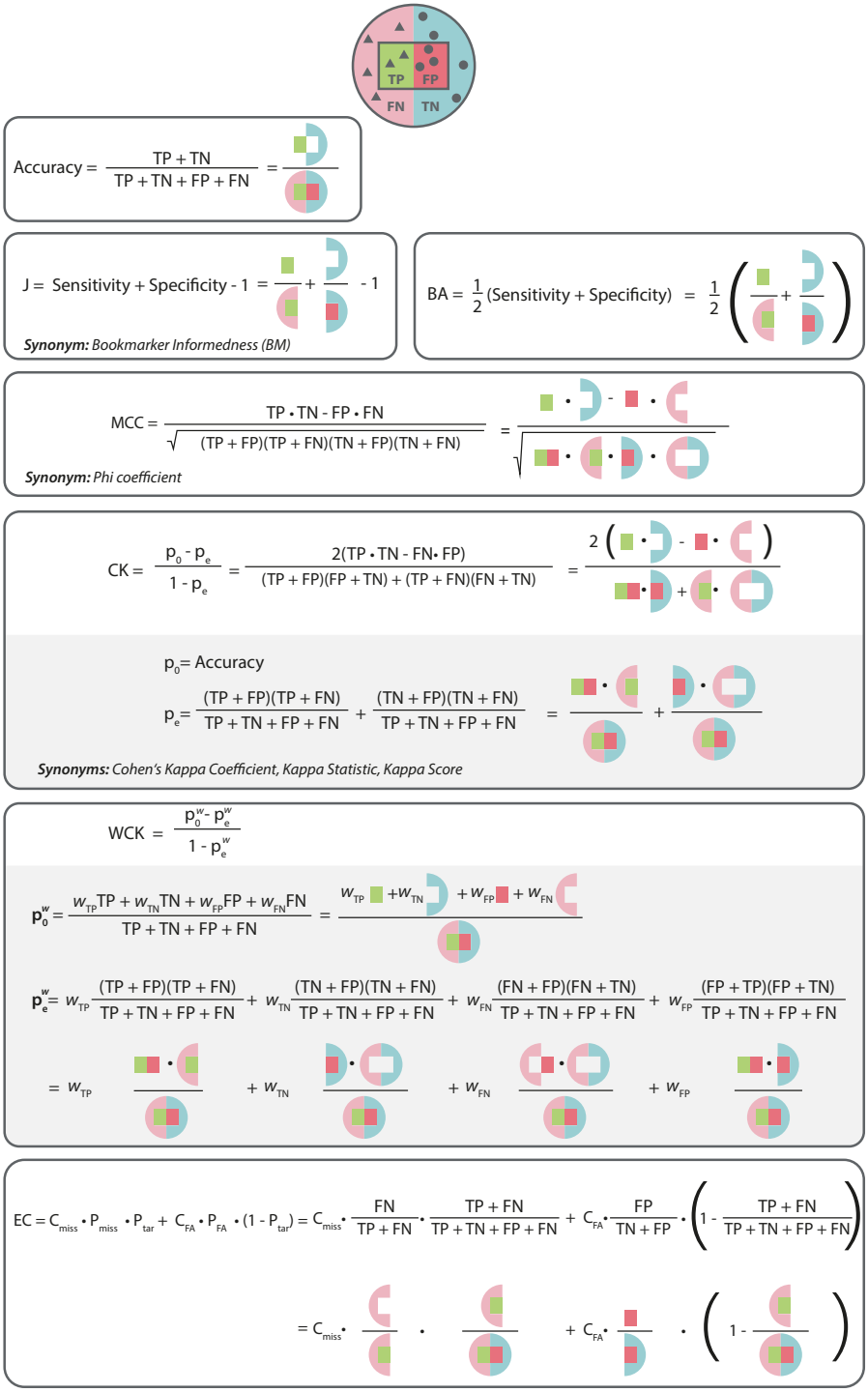


Figure 2.10: Overview of the most frequent multi-class counting classification metrics that are based on the confusion matrix's cardinalities True Positive (TP), False Negative (FN), False Positive (FP) and True Negative (TN), as illustrated in Figure 2.6. Used abbreviations: Balanced Accuracy (BA), Cohen's Kappa (CK), Expected Cost (EC), Youden's Index (J), Matthews Correlation Coefficient (MCC), Weighted Cohen's Kappa (WCK). Figure adapted from [Reinke et al., 2021a].

Multi-class counting metrics (multi-class)

For the case of more than two classes C, two strategies can be applied. On the one hand, the setting can be evaluated by considering one class at a time and merging all other samples into the negative class ("one-versus-rest"), finally averaging over all perspectives. This allows for using the metric definitions of the binary case. On the other hand, the binary confusion matrix can be extended to a $C \times C$ matrix, as shown in Figure 2.11.

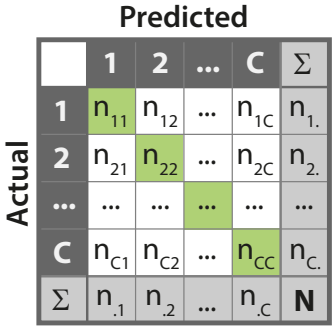


Figure 2.11: Visualization of a confusion matrix with C classes and N samples. The last row and column show the sum of all respective row and column entries.

The generalized version of the Accuracy can then be written as:

$$\text{Accuracy} = \frac{\sum_{i=1}^{C} n_{ii}}{N} \tag{2.41}$$

For the multi-class case of BA and J, the metrics can be interpreted as either the arithmetic mean of Sensitivities per class (BA) or the sum of Sensitivities per class minus $(C-1)$ (J).

The generalized MCC uses the following intermediate variables:

$c = \sum_{k=1}^{C} n_{kk}$: Number of correctly predicted samples

$s = \sum_{i=1}^{C} \sum_{j=1}^{C} n_{ij}$: Total number of samples

$p_k = \sum_{i=1}^{C} n_{ki}$: Number of times that class k was predicted

$t_k = \sum_{i=1}^{C} n_{ik}$: Number of times that class k actually occurred

and result in the following formula [Gorodkin, 2004]:

$$\text{MCC} = \frac{c \cdot s - \sum_{k=1}^{C} p_k \cdot t_k}{s^2 - \sum_{k=1}^{C} p_k \cdot t_k} \tag{2.42}$$

The formulations of CK [Brennan and Prediger, 1981] and WCK [Schuster, 2004] are generalized to the following terms:

$$\text{CK} = \frac{p_0 - p_e}{1 - p_e}, p_0 = \frac{\sum_{i=1}^{C} n_{ii}}{N}, p_e = \frac{\sum_{i=1}^{C} n_{i\cdot} \cdot n_{\cdot i}}{N^2} \tag{2.43}$$

$$\text{WCK} = 1 - \frac{\sum_{i=1}^{C}\sum_{j=1}^{C} w_{ij} \cdot n_{ij}}{\sum_{i=1}^{C}\sum_{j=1}^{C} w_{ij} \cdot \frac{n_{i\cdot} \cdot n_{\cdot j}}{N^2}} \tag{2.44}$$

The multi-class case of the EC can be interpreted as the weighted sum of the resulting error rates for all classes [Ferrer, 2022]:

$$\text{EC} = \sum_{i=1}^{C}\sum_{j=1}^{C} P_i \cdot w_{ij} \cdot \frac{n_{ij}}{n_{i\cdot}} \tag{2.45}$$

Here, w_{ij} refers to the chosen weights or costs and P_i to the priors (prevalences) of class i. The normalized variant of EC is given by [Ferrer, 2022]:

$$\text{EC}_{\text{norm}} = \frac{\sum_{i=1}^{C}\sum_{j=1}^{C} P_i \cdot w_{ij} \cdot \frac{n_{ij}}{n_{i\cdot}}}{\min_d \sum_{i=1}^{C} w_{id} P_i} \tag{2.46}$$

Multi-threshold metrics

Per-class counting metrics rely on setting one cutoff value as a threshold on the predicted class scores to calculate the confusion matrix and subsequent metrics. However, results may change substantially when changing the threshold. To compensate for this issue, multi-threshold metrics can be applied.

The **Area under the Receiver Operating Characteristic Curve (AUROC)** [Hanley and McNeil, 1982] (also referred to as **Area under the Curve (AUC)**) of a prediction indicates how well it differentiates between the positive and negative classes. It is measured as the area under the Receiver Operating Characteristic (ROC) curve. The ROC curve plots 1 - Specificity against the Sensitivity (see Figure 2.12). Samples are ordered descendingly by their predicted class scores for the ROC curve computation, with each score regarded as a potential threshold. Each point on the curve then corresponds to one threshold, for which Sensitivity and 1 - Specificity are calculated. The points are connected via a linear interpolation [Davis and Goadrich, 2006]. Given its reliance on Sensitivity and Specificity only, AUROC is independent of the prevalence.

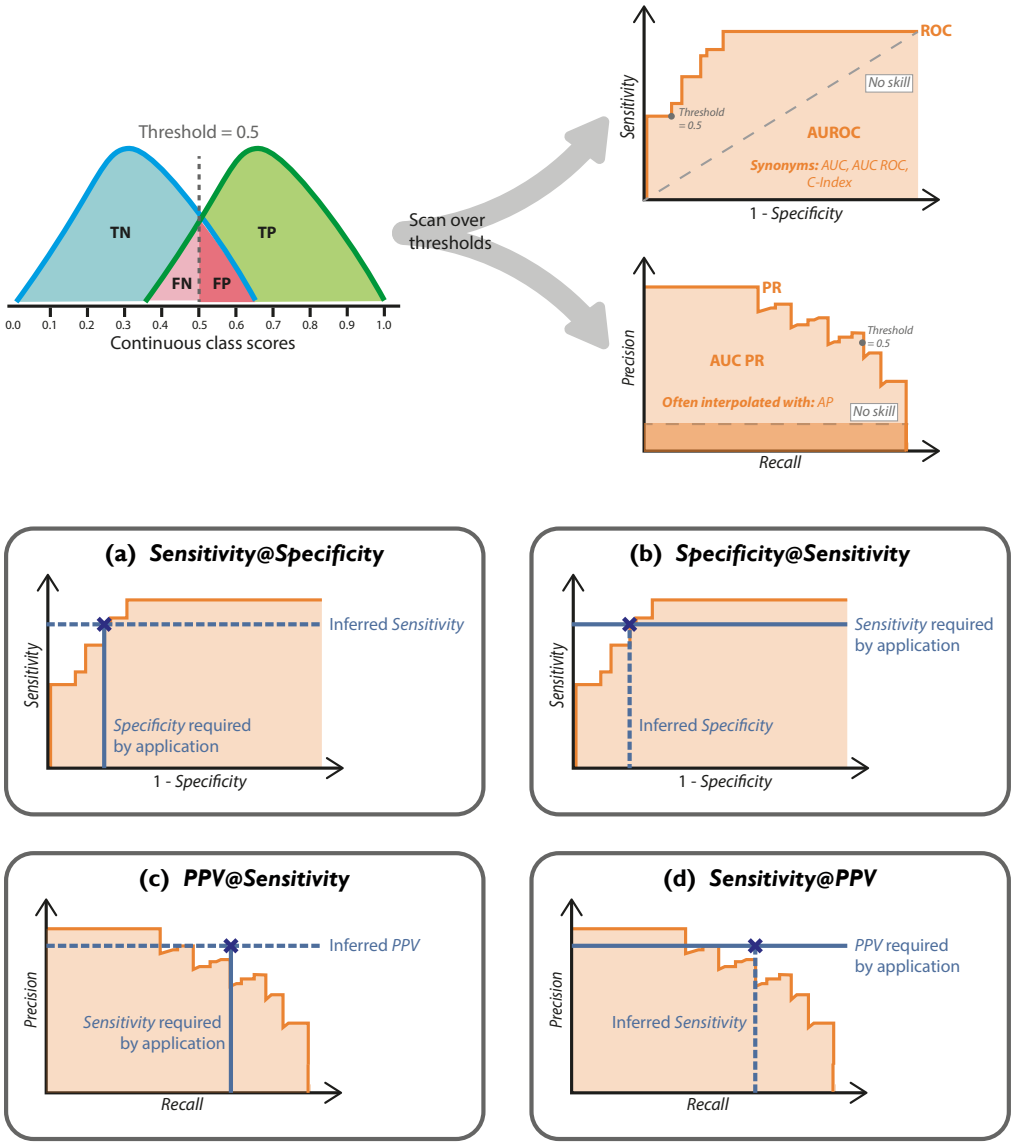


Figure 2.12: Multi-threshold metrics Area under the Receiver Operating Characteristic Curve (AUROC) and Average Precision (AP) (top) and per-class counting metrics with application-driven thresholds (bottom). Multi-threshold-based metrics integrate across a range of thresholds to generate a dynamic confusion matrix rather than being based on a single threshold. Based on the cardinalities for each threshold, i.e. the true (T)/false (F) positives (P)/negatives (N) values, Sensitivity (Recall), and 1 - Specificity or Positive Predictive Value (PPV) are calculated and plotted against each other (top right). Repeated for several thresholds, this results in the Receiver Operating Characteristic (ROC) or Precision-Recall (PR) curves. The area under the ROC/PR curve is referred to as AUROC/AUC PR. The Average Precision (AP) metric is frequently used to interpolate AUC PR. A classifier with no skill level (random guessing) is represented by dashed gray lines. In the case of an application-driven threshold (for example, required Sensitivity of 0.9), the metrics Sensitivity@Specificity, Specificity@Sensitivity, PPV@Sensitivity and Sensitivity@PPV can be computed using the ROC/PR curves. Figure adapted from [Reinke et al., 2021a].

The **Average Precision (AP)** [Lin et al., 2014; Everingham et al., 2015] refers to an interpolation of the area under the Precision-Recall (PR) curve. For the PR curve, we plot the Sensitivity against PPV based on multiple thresholds. Please note that we use the term Precision-Recall curve instead of PPV-Sensitivity curve, as this term is more prominent in the research community. Similar to the ROC curve considerations, samples are ranked by the predicted class scores, and Sensitivity and PPV are calculated for every threshold to form a point of the PR curve. A zigzag shape is common for the PR curve, as illustrated in Figure 2.12. Because FP substitutes FN in the denominator of the PPV, PPV does not necessarily vary linearly as the level of Sensitivity changes [Davis and Goadrich, 2006]. As discussed in [Davis and Goadrich, 2006], this means that a linear interpolation would be unduly optimistic, necessitating a more complicated interpolation. AP denotes a conservative curve interpolation simplification and is defined as

$$\text{AP} = \sum_i (R_i - R_{i-1}) P_i, \tag{2.47}$$

with R_i and P_i denoting the Recall (Sensitivity) and Precision (PPV) at the ith threshold, with R_0 being equal to zero.

Metric1@Metric2

Per definition, AUROC and AP measure the complete area under their respective curves. It is, however, possible to further utilize their curves in order to derive per-class counting metrics for a pre-defined target value at a single working point of the curves. For example, an application might require a Sensitivity of 0.98. We could use the AUROC to find the corresponding Specificity value linked to a Sensitivity of 0.98, denoted as Specificity@Sensitivity. We refer to such an approach as *Metric1@Metric2* for a specific target value in Section 5.2. Similarly, AP can be used to find the respective PPV (see the bottom of Figure 2.12). However, it is not necessary to draw a curve for such a setup. When optimizing an algorithm and threshold to yield a pre-defined target value (e.g. Sensitivity of 0.98), other metrics can be directly calculated for the threshold yielding the respective target score.

2.4.2 Semantic segmentation metrics

Overlap-based metrics

In semantic segmentation, counting metrics are the most frequently used type of metrics. In this context, TPs refer to the overlapping pixels between the reference and predicted segmentation, FPs to all pixels that were incorrectly predicted as positive pixels, and FNs to all reference pixels that were not predicted as positive. Thus, counting metrics are often referred to as overlap-based metrics in the context of segmentation.

The **Dice Similarity Coefficient (DSC)** [Dice, 1945] is a frequently used metric in segmentation problems [Maier-Hein et al., 2018]. It is equal to the pixel-level **F_1 Score** and is defined as

$$\begin{aligned} \text{DSC}(A,B) &= \frac{2 \cdot \text{PPV} \cdot \text{Sensitivity}}{\text{PPV} + \text{Sensitivity}} = \frac{2 \cdot \text{TP}}{2 \cdot \text{TP} + \text{FN} + \text{FP}} \\ &= \frac{2|A \cap B|}{|A| + |B|} \in [0,1], \end{aligned} \tag{2.48}$$

with A referring to the reference structure and B to the predicted structure, $|A|$ and $|B|$ denoting the cardinality of both structures, i.e. their total number of pixels, and $|A \cap B|$ the number of pixels of the intersection between structures A and B.

The **Intersection over Union (IoU)** (also known as **Jaccard Index**) [Jaccard, 1912] is closely related to the DSC and defined as

$$\begin{aligned} \mathsf{IoU}(A,B) &= \frac{\mathsf{PPV} \cdot \mathsf{Sensitivity}}{\mathsf{PPV} + \mathsf{Sensitivity} - \mathsf{PPV} \cdot \mathsf{Sensitivity}} = \frac{\mathsf{TP}}{\mathsf{TP} + \mathsf{FP} + \mathsf{FN}} \\ &= \frac{|A \cap B|}{|A| + |B| - |A \cap B|} = \frac{|A \cap B|}{|A \cup B|} \in [0,1], \end{aligned} \tag{2.49}$$

with $|A \cup B|$ being the number of pixels of the union of structures A and B. DSC is more common in the biomedical community, while the general computer vision community prefers the IoU. Irrespective of that use, both metrics basically measure the same properties and can directly be translated into each other:

$$\begin{aligned} \mathsf{DSC} &= \frac{2 \cdot \mathsf{IoU}}{1 + \mathsf{IoU}} \\ \mathsf{IoU} &= \frac{\mathsf{DSC}}{2 - \mathsf{DSC}} \end{aligned} \tag{2.50}$$

Similar to image-level classification, the **F_β Score** is a generalization of the DSC, which can also be applied on pixel level, with $A \setminus B$ being the difference of structure A and B:

$$\begin{aligned} \mathsf{F}_\beta(A,B) &= \frac{(1+\beta^2) \cdot \mathsf{PPV} \cdot \mathsf{Sensitivity}}{\beta^2 \cdot \mathsf{PPV} + \mathsf{Sensitivity}} = \frac{(1+\beta^2) \cdot \mathsf{TP}}{(1+\beta^2) \cdot \mathsf{TP} + \beta^2 \cdot \mathsf{FN} + \mathsf{FP}} \\ &= \frac{(1+\beta^2) \cdot |A \cap B|}{(1+\beta^2) \cdot |A \cap B| + \beta^2 \cdot |A \setminus B| + |B \setminus A|} \in [0,1], \end{aligned} \tag{2.51}$$

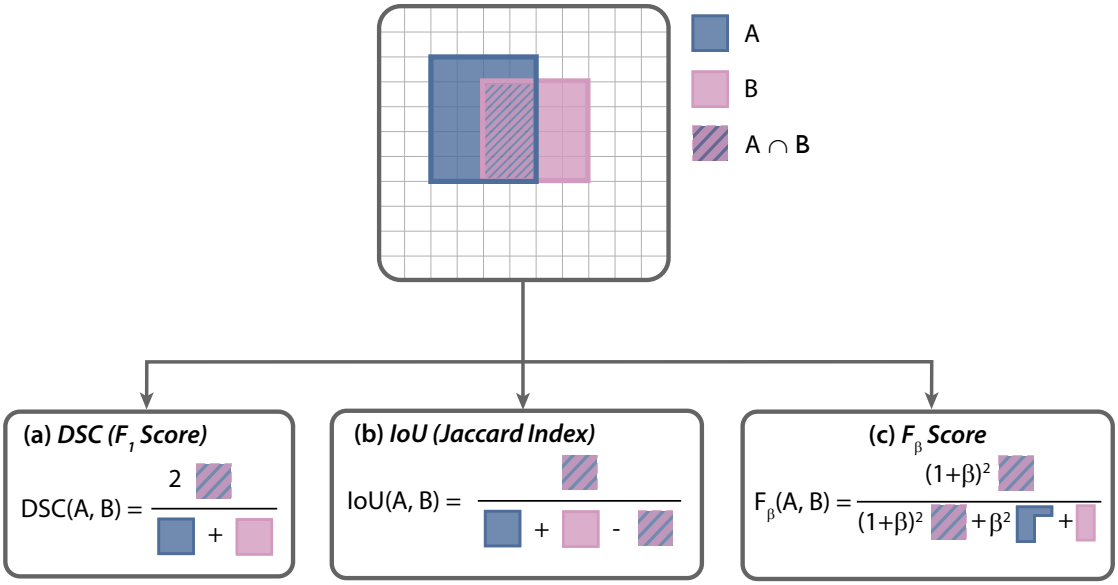


Figure 2.13: Most frequently used overlap-based metrics: (a) Dice Similarity Coefficient (DSC), (b) Intersection over Union (IoU) and (c) F_β Score for two structures A and B. Figure adapted from [Reinke et al., 2021a]. Figure adapted from [Reinke et al., 2021a].

In the case of tubular structures, such as airway trees or vessels, in which we are particularly interested in their center or centerline, connectivity, or network topology, the **Centerline Dice Similarity Coefficient (clDice)** [Shit et al., 2021] offers an alternative to DSC or IoU. For the computation of clDice, it is necessary to calculate the skeletons from the binary segmentation masks. Please refer to [Shit et al., 2021] for a detailed description of how to derive the skeletons, as this is out of the scope of this thesis. The clDice is similarly defined as the standard DSC, namely the harmonic mean of the PPV and Sensitivity. However, clDice uses the definition of the *Topology Precision* and *Topology Sensitivity* instead, as illustrated in Figure 2.14. The most important difference between the standard definition is the use of skeletons in contrast to whole masks. The Topology Precision measures the FP pixels of the predicted skeleton, while Topology Sensitivity focuses on the FN pixels.

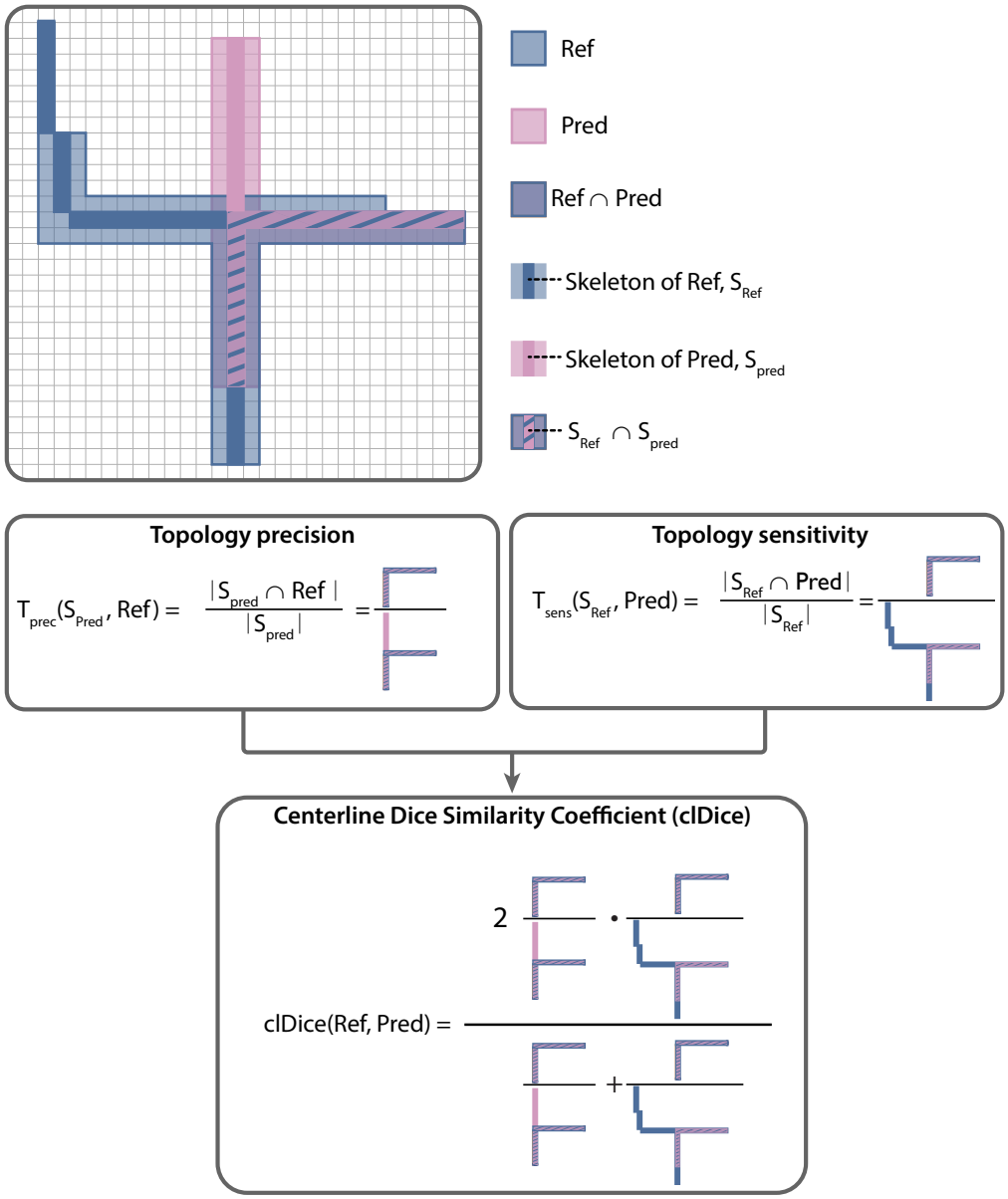


Figure 2.14: Illustration of the Centerline Dice Similarity Coefficient (clDice), measuring the connectivity of structures. The metric is designed for tubular structures and relies on the skeletons of the structures S_{Ref} and S_{Pred} for structures Ref and Pred. The clDice is the harmonic mean of the Topology Precision and Topology Sensitivity. Figure adapted from [Reinke et al., 2021a].

Boundary-based metrics

Boundary-based metrics measure the distances between the reference boundary and the predicted boundary (therefore also referred to as distance-based metrics) and are focusing on the accuracy of a predicted structure boundary.

Many boundary-based metrics build upon calculation the minimal distance from one boundary to the other. For two boundaries A and B and a boundary pixel $a \in A$, the shortest distance is defined as

$$d(a, B) = \min_{b \in B} d(a, b), \tag{2.52}$$

with d being the euclidean distance between a and b.

The **Hausdorff Distance (HD)** [Huttenlocher et al., 1993] (also known as **Maximum Symmetric Surface Distance**, **Hausdorff Metric** or **Pompeiu–Hausdorff Distance**) is the maximum of all shortest distances (as defined in Equation 2.52):

$$\mathrm{HD}(A, B) = \max\Big\{\max_{a \in A} d(a, B), \max_{b \in B} d(A, b)\Big\} \in [0, \infty) \tag{2.53}$$

HD is unbounded to the top and the lower its value, the better the predicted boundary.

Since calculating the maximum of all shortest distances heavily penalizes spatial outliers, an alternative metric is the **Hausdorff Distance 95 Percentile (HD95)** [Huttenlocher et al., 1993]. Instead of calculating the total maximum of all shortest distances, HD95 measures the maximum of the 95 percentiles x_{95} of distances for each boundary:

$$\begin{aligned} d_{95}(A, B) &= \underset{a \in A}{x_{95}}\Big\{\min_{b \in B} d(a, b)\Big\} \\ \mathrm{HD95}(A, B) &= \max\Big\{d_{95}(A, B), d_{95}(B, A)\Big\} \in [0, \infty) \end{aligned} \tag{2.54}$$

Please note that the 95 percentile can be replaced by any other percentile.

The **Average Symmetric Surface Distance (ASSD)** [Yeghiazaryan and Voiculescu, 2015] symmetrically measures the average over all shortest distances. Based on Equation 2.52, it is defined as:

$$\mathrm{ASSD}(A, B) = \frac{1}{|A| + |B|} \cdot \left(\sum_{a \in A} d(a, B) + \sum_{b \in B} d(A, b)\right) \in [0, \infty) \tag{2.55}$$

The **Mean Absolute Surface Distance (MASD)** [Beneš and Zitová, 2015] also calculates the average of shortest paths, but independently for every boundary to the other. The averages per boundary are then averaged:

$$\mathrm{MASD}(A, B) = \frac{1}{2} \cdot \left(\frac{\sum_{a \in A} d(a, B)}{|A|} + \frac{\sum_{b \in B} d(A, b)}{|B|}\right) \in [0, \infty) \tag{2.56}$$

Figure 2.15 illustrates the differences between the four metrics.

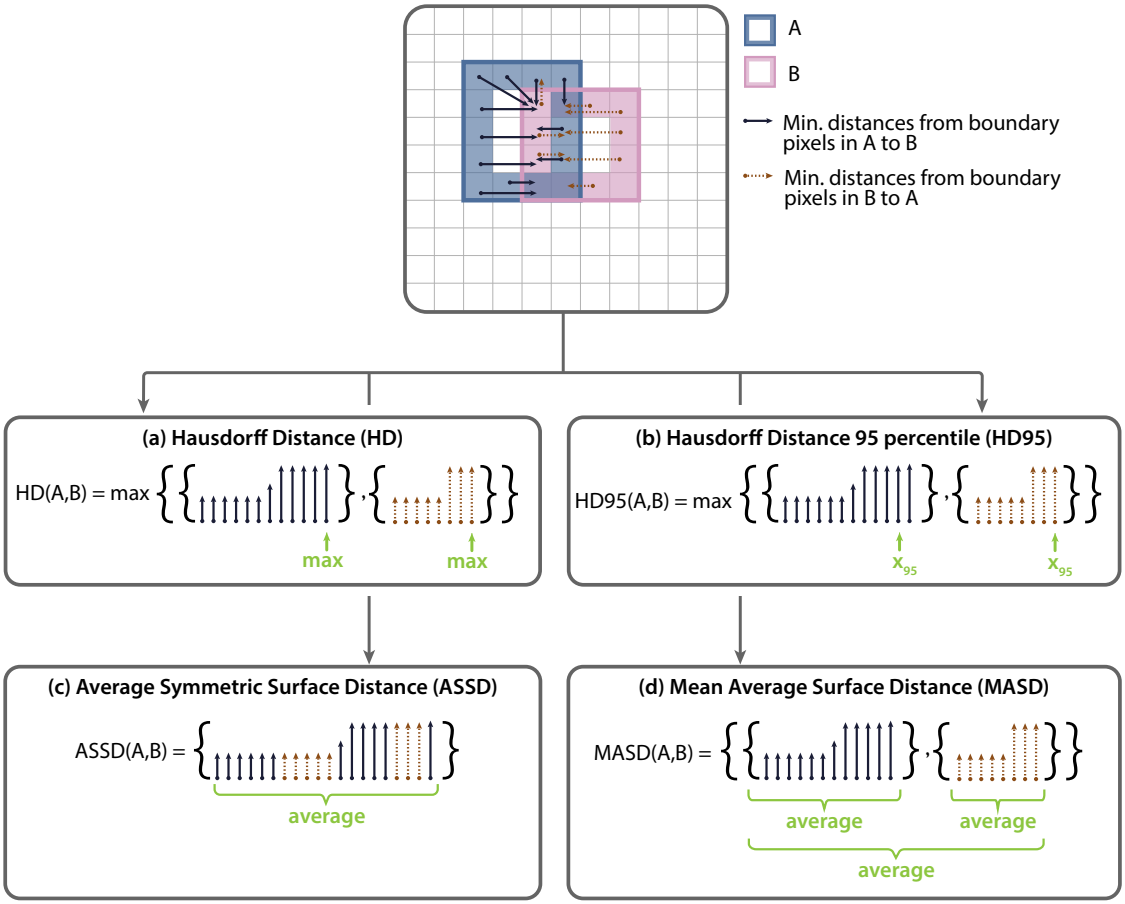


Figure 2.15: Boundary-based metrics that aggregate shortest distances of boundaries A and B: (a) Hausdorff Distance (HD), (b) Hausdorff Distance 95 Percentile (HD95), (c) Average Symmetric Surface Distance (ASSD), Mean Absolute Surface Distance (MASD). Figure adapted from [Reinke et al., 2021a].

HD, HD95, ASSD, and MASD are not able to compensate for labeling imprecisions or inter-rater variability. However, we often face inter-rater variability in biomedical image analysis problems [Joskowicz et al., 2019]. In situations in which the annotators themselves are not certain, we might not want to penalize our predictions for errors within the inter-rater variability. For this purpose, the **Normalized Surface Distance (NSD)** [Nikolov et al., 2021] (also known as **Surface Dice** or **Normalized Surface Dice**) has been designed. Annotation imprecisions or noise in the reference are captured by the hyperparameter τ. It measures the maximum tolerated distance from the reference boundary and controls the rigor with which a boundary pixel is defined. Only boundary parts within this range are considered as TP boundary pixels. Given two boundaries S_A and S_B of structures A and B, we consider the border regions, i.e. pixels that are within the maximum tolerated distance of the boundaries, as $\mathcal{B}_A^{(\tau)}$ and $\mathcal{B}_B^{(\tau)}$ (see Figure 2.16a). The NSD is then calculated as:

$$\text{NSD}(A,B)^{(\tau)} = \frac{|S_A \cap \mathcal{B}_B^{(\tau)}| + |S_B \cap \mathcal{B}_A^{(\tau)}|}{|S_A| + |S_B|} \in [0,1] \tag{2.57}$$

NSD can be considered as the DSC on boundary pixels and can be interpreted as the ratio of correctly predicted boundary pixels [Nikolov et al., 2021]. Similarly to its original, the metric is bounded between 0 and 1, with 1 indicating a perfect prediction.

The **Boundary Intersection over Union (Boundary IoU)** [Cheng et al., 2021] directly calculates the overlap between predicted and reference boundaries without imposing a tolerance. It computes the IoU between two boundaries up to a pre-defined width, determined by the distance parameter d. There is a direct relationship between the width and boundary errors, i.e. the smaller the distance, the higher the sensitivity to boundary errors. In contrast to NSD, Boundary IoU is not concerned about noise or uncertainties, but rather assesses errors at the boundaries or wider areas around the boundary lines (see Figure 2.16b). Let A_d and B_d be the boundary areas up to the width d. Boundary IoU is then defined as:

$$\text{Boundary IoU}(A,B) = \frac{|A_d \cap B_d|}{|A_d| + |B_d| - |A_d \cap B_d|} = \frac{|A_d \cap B_d|}{|A_d \cup B_d|} \in [0,1] \tag{2.58}$$

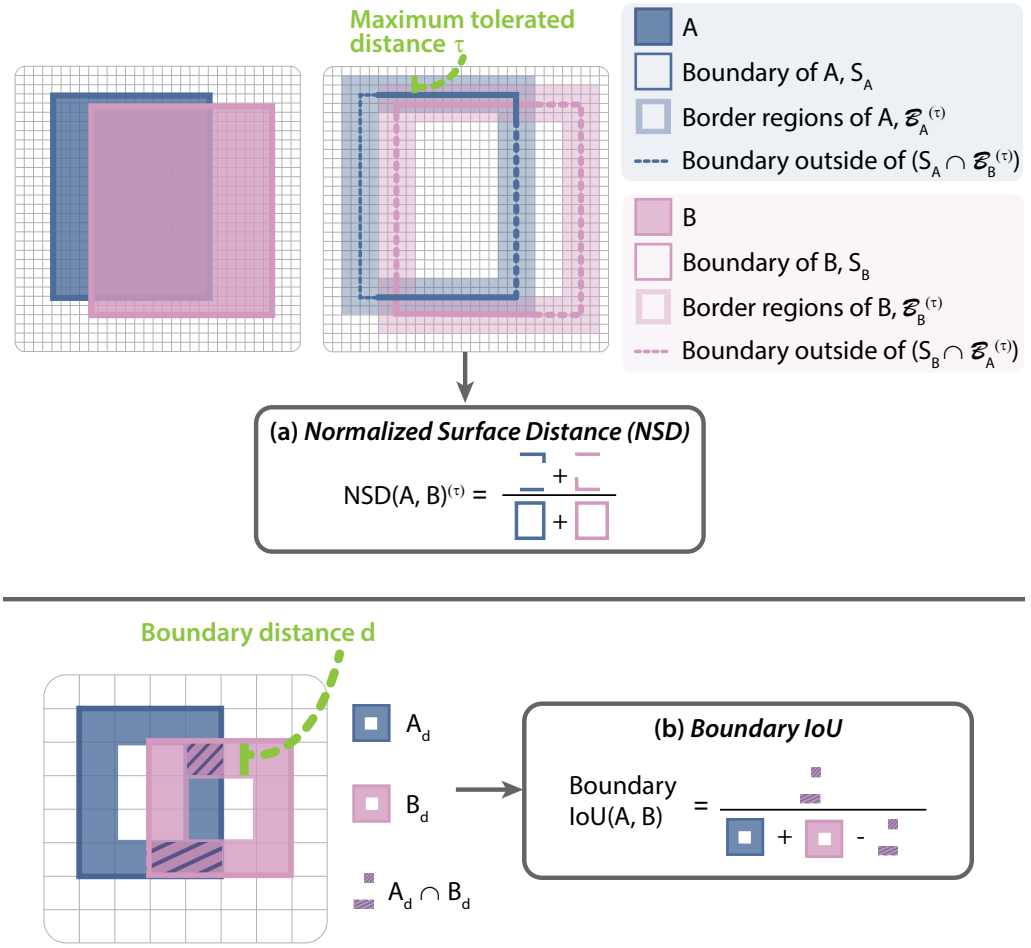


Figure 2.16: Boundary-based metrics based on hyperparameters. (a) Normalized Surface Distance (NSD) calculates the overlap between two boundaries, while the parameter τ represents the maximum tolerated difference between the reference and predicted boundary. (b) Boundary Intersection over Union (Boundary IoU) calculates the overlap between two boundaries up to a certain width d. Figure adapted from [Reinke et al., 2021a].

2.4.3 Object detection metrics

The first step for validating an object detection problem lies in the determination of whether a prediction correctly detected a reference object, which can be measured by calculating the localization criterion. This step is achieved by the localization criterion. In addition, we need to decide how to deal with assignment ambiguities. This is important for cases, where multiple predictions can be assigned to the same reference or multiple references which may correspond to the same prediction. After both steps were applied, performance metrics can be calculated.

Localization criteria

The localization or hit criterion determines whether a prediction detected/hit the reference object (TP) or whether it missed any reference (FP). Remaining reference objects, for which no prediction could be assigned are defined asFN. Localization criteria can be defined at various levels based on the available reference annotation. An overview of the presented localization criteria can be found in Figure 2.17.

The simplest way to define a hit is via the **Point inside Mask** or **Point inside Box/Approximation** criterion. In this case, a prediction is defined as TP once there is a point inside the reference annotation. The latter may be expressed as a mask, bounding box, or approximation (such as an convex hull) of the structure.

If we are interested only in the position of an object, a localization criterion operating on the object centers is useful. The **Center Distance** (or **distance-based hit criterion**) [Gurcan et al., 2010; Piccinini et al., 2012] measures the distance between the reference and prediction center and considers the prediction as TP if the distance is below a pre-defined localization threshold. The **center-cover criterion** classifies the prediction as TP if the center of the reference structure is contained within it. Contrarily, the **center-hit criterion** defines a TP prediction if the predicted object's center is located inside the reference.

Overlap-based localization criteria are more fine-granular and measure the overlap between reference and prediction. This is most often done by using the IoU (see Section 2.4.2) at various scales depending on the required granularity of localization. Once the IoU is above a certain threshold, the prediction is considered as TP. In the case of bounding boxes or approximations of the desired structure, the **Box/Approximation Intersection over Union (Box/Approx IoU)** is calculated, which equals the IoU over a bounding box or approximation. For pixel-level localization, the **Mask Intersection over Union (Mask IoU)** calculates the IoU over two segmentation masks. If the concrete structure boundary is of particular importance, the **Boundary IoU** (see Section 2.4.2) can be used as an alternative overlap-based localization criterion. The **Intersection over Reference (IoR)** [Maška et al., 2014] depicts an alternative to the IoU criteria. It is similar to the Sensitivity on pixel level and therefore only considers FN pixels:

$$\mathsf{IoR}(A, B) = \frac{\mathsf{TP}}{\mathsf{TP} + \mathsf{FN}} = \frac{|A \cap B|}{|A|} \in [0, 1] \tag{2.59}$$

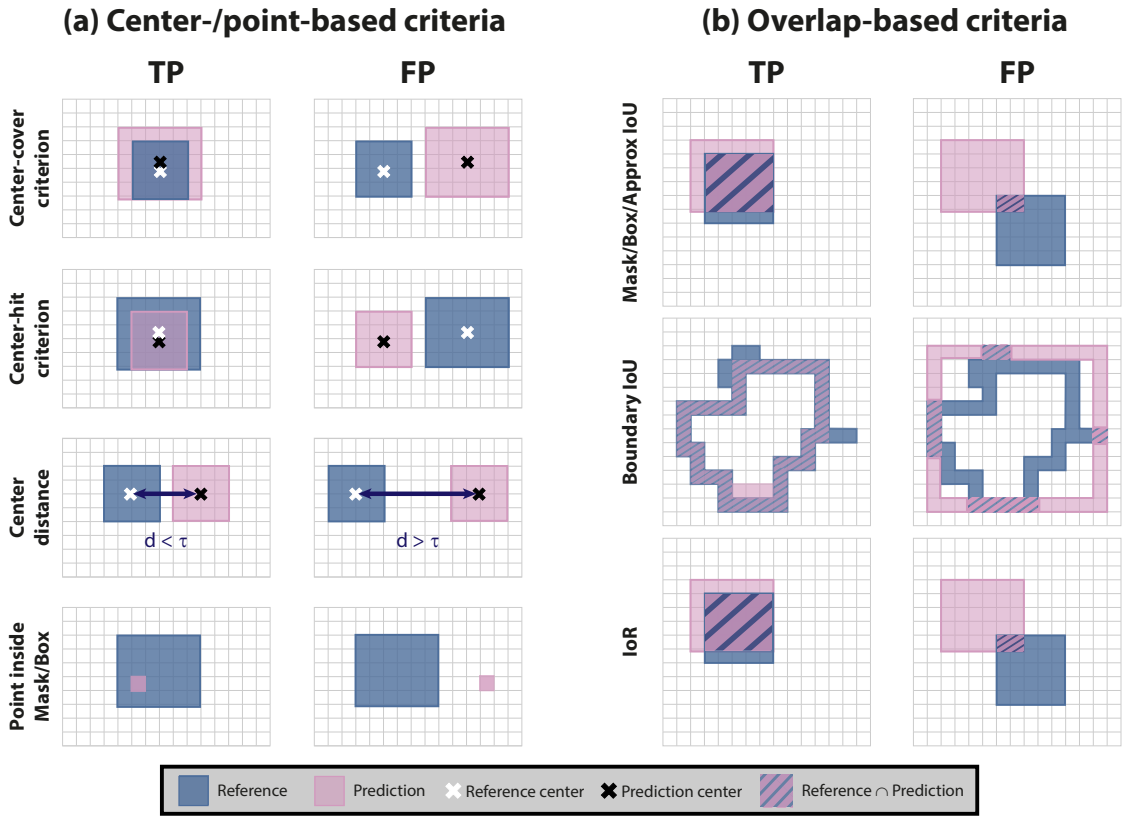


Figure 2.17: Overview of (a) center-/point-based and (b) overlap-based localization criteria. For each criterion, an example of a True Positive (TP) (detected/hit) and False Positive (FP) (missed) is presented. Center-/point-based include the Center-cover Criterion, Center-hit Criterion, Centroid Distance, and Point inside Mask/Box/Approx(imation), while overlap-based criteria include the Mask/Box/Approx(imation) Intersection over Union (IoU), Boundary IoU, and Intersection over Reference (IoR).

The localization criterion heavily influences the final metric scores and must be carefully chosen. For instance, it should be noted that a localization at a lower resolution, such as a bounding box or center point, leads to an information loss, as illustrated in Figure 2.18. Similarly, in the case of overlap-based localization criteria, the chosen cutoff value may result in substantial differences in the final metric scores and should be selected carefully. For example, if precise boundaries are of particular interest, higher thresholds should be used. A low cutoff value, on the other hand, is preferable for noisy reference standards.

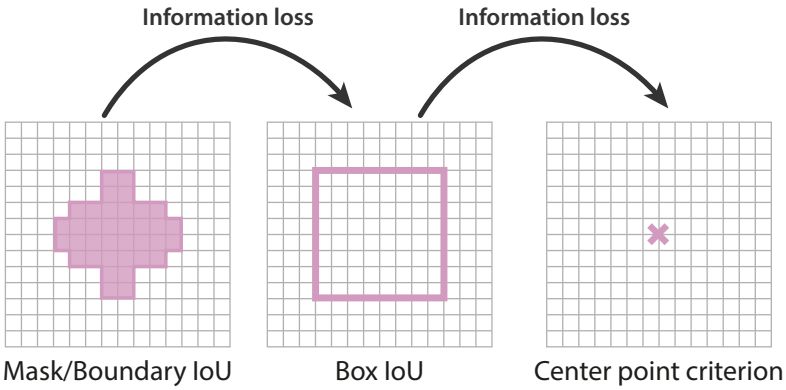


Figure 2.18: Localization criteria may discard spatial information and should be well motivated by the underlying task. Figure adapted from [Reinke et al., 2021a].

Assignment strategies

During object detection validation, assignment ambiguities may occur. During the localization step, it may happen that multiple predictions could be assigned to the same reference object or multiple reference objects may correspond to only one prediction. In such situations, the assignment strategy decides how such ambiguities are handled. An example is shown in Figure 2.19.

A common strategy is the **Greedy by Score Matching** [Everingham et al., 2015], in which all predicted objects are ordered given their predicted class scores. Predictions are iteratively allocated to the reference object for which they offer the highest localization criterion, starting with the prediction with the highest predicted class score. Because it was already assigned, the selected reference object is removed from the process. In the case that no predicted class scores are available, the strategy can be replaced with the **Greedy by "localization criterion"**[2] **Matching**, for which the prediction with the highest overlap-based localization criterion for a reference object is matched.

In order to find the best match between predicted and reference objects, the **Optimal Hungarian Matching** [Kuhn, 1955] strategy is more sophisticated and minimizes a cost function that is typically dependent on the localization criterion.

In the biomedical domain, more simplified approaches are often desired. One option is the **Matching via "localization criterion"**[2], in which only non-overlapping predictions are allowed, thereby preventing matching conflicts. In the case of many touching reference objects and one prediction overlapping with multiple reference objects, the **Matching via IoR > 0.5** [Matula et al., 2015] strategy should be preferred.

[2] "Localization criterion" refers to the selected localization criterion.

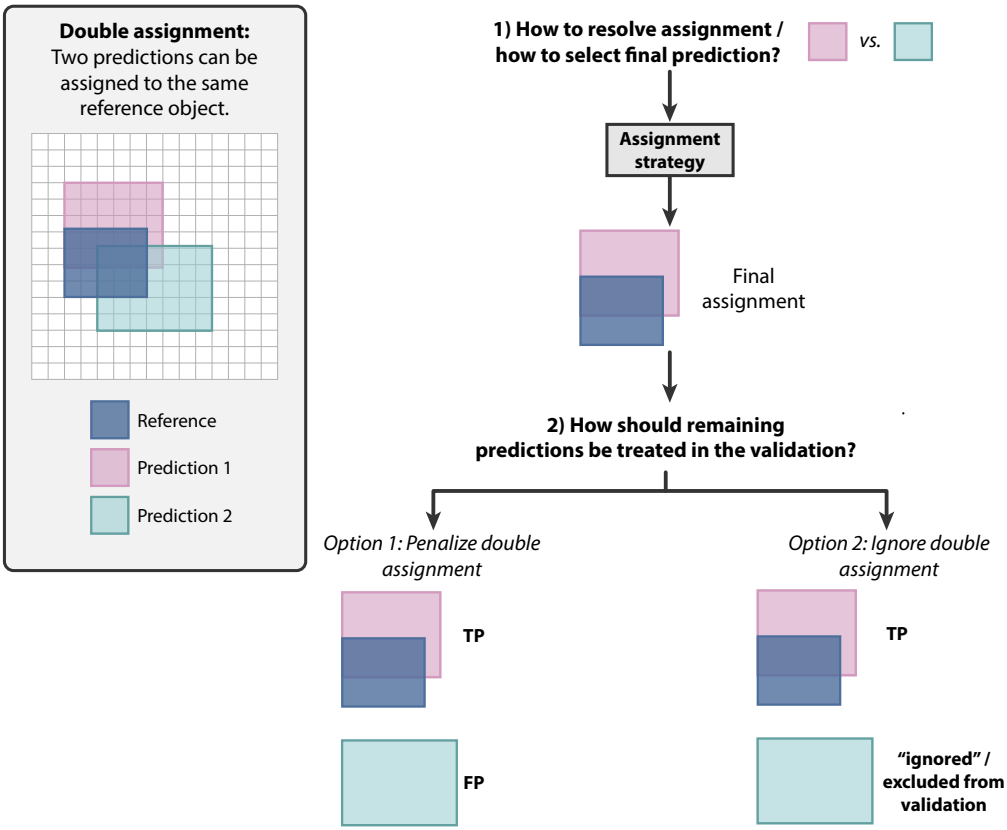


Figure 2.19: Example of an assignment strategy in the case of two predictions that have been assigned to the same reference object. This step entails deciding how to proceed with the remaining prediction. For instance, remaining predictions can be penalized with a False Positive (FP) assignment or may be ignored. Further abbreviations: True Positive (TP). Figure from [Reinke et al., 2021a].

Multi-threshold metrics

Similarly to image-level classification, **AP** is a popular metric for object detection problems as the PR curve can be calculated for objects rather than images. Contrary, the ROC curve is usually drawn on image level, hindering a per-object validation. As described above, the cutoff value chosen for the localization criterion impacts the resulting metric scores. For this reason, it has become common practice to average the metric values over multiple localization criterion cutoff values. For example, the default for the IoU localization criterion is AP@0.50:0.05:0.95 [Lin et al., 2014]. This means that the AP is calculated for the localization thresholds, IoU$\leq$0.50, IoU$\leq$0.55, . . ., IoU$\leq$0.95. The resulting scores are averaged over all thresholds. The **mean Average Precision (mAP)** [Lin et al., 2014] is a related commonly known metric, which averages the class-wise AP over classes.

In the biomedical context, the **Free-Response Receiver Operating Characteristic (FROC) Score** [Van Ginneken et al., 2010; Bandos et al., 2009] is often preferred given its simpler interpretation. As shown in Figure 2.20, it operates on object level. For every threshold, the average number

of False Positives per Image (FPPI) and the average Sensitivity per image are calculated and added to the curve as one point. The procedure is repeated over multiple thresholds. The FROC Score then measures the area under the FROC curve and indicates how well the probabilities of the positive class are separated from those of the negative class while considering object-level information. In contrast to other multi-threshold metrics, the FROC Score is unbounded to the top, given the FPPI on the abscissa. The concrete values on the abscissa are not standardized, meaning that FROC Scores cannot easily be compared across data sets.

Per-class counting metrics

In contrast to image-level classification, cardinalities in object detection problems are computed per object rather than per image. Based on this, TN are typically not defined for object detection problems. Thus, not all counting metrics presented in Section 2.4.1 can be used in object detection problems. We can similarly compute **Sensitivity**, **PPV**, and **F_β Score** for detection tasks. However, Specificity and all metrics based on it (LR+, Accuracy, BA and J) in addition to EC, MCC and (W)CK can *not* be measured. Furthermore, we can use the **FPPI** as a per-class counting metric. As with image-level classification metrics, we can calculate the metrics based on an application-specific threshold as a single working point of the PR or FROC curves. For instance, for a required Sensitivity of 0.9, we can calculate the FPPI@Sensitivity for the threshold yielding this Sensitivity value.

Counting cardinalities at different scales

Validation is generally performed on the entire data set in image-level classification problems. In semantic segmentation, the metric values are usually computed image-wise, while they are aggregated to produce a single score over the data set. Because of the small number of objects per image rather than the thousands of pixels in semantic segmentation, the latter approach is not easily applicable to object detection. In most cases, object detection problems are validated on the entire data set. The difference between the validation scales is illustrated in Figure 2.21. If one chooses the per-image (or per-patient) validation for object detection problems, counting metrics can easily be applied (see Figure 2.21). For multi-threshold metrics, the Sensitivity and PPV or FPPI are computed for every threshold for every image (or patient) and aggregated over the data set. The aggregated metric pairs can then be used to form the respective curves, from which the metric scores can be derived (e.g. AP and FROC Score).

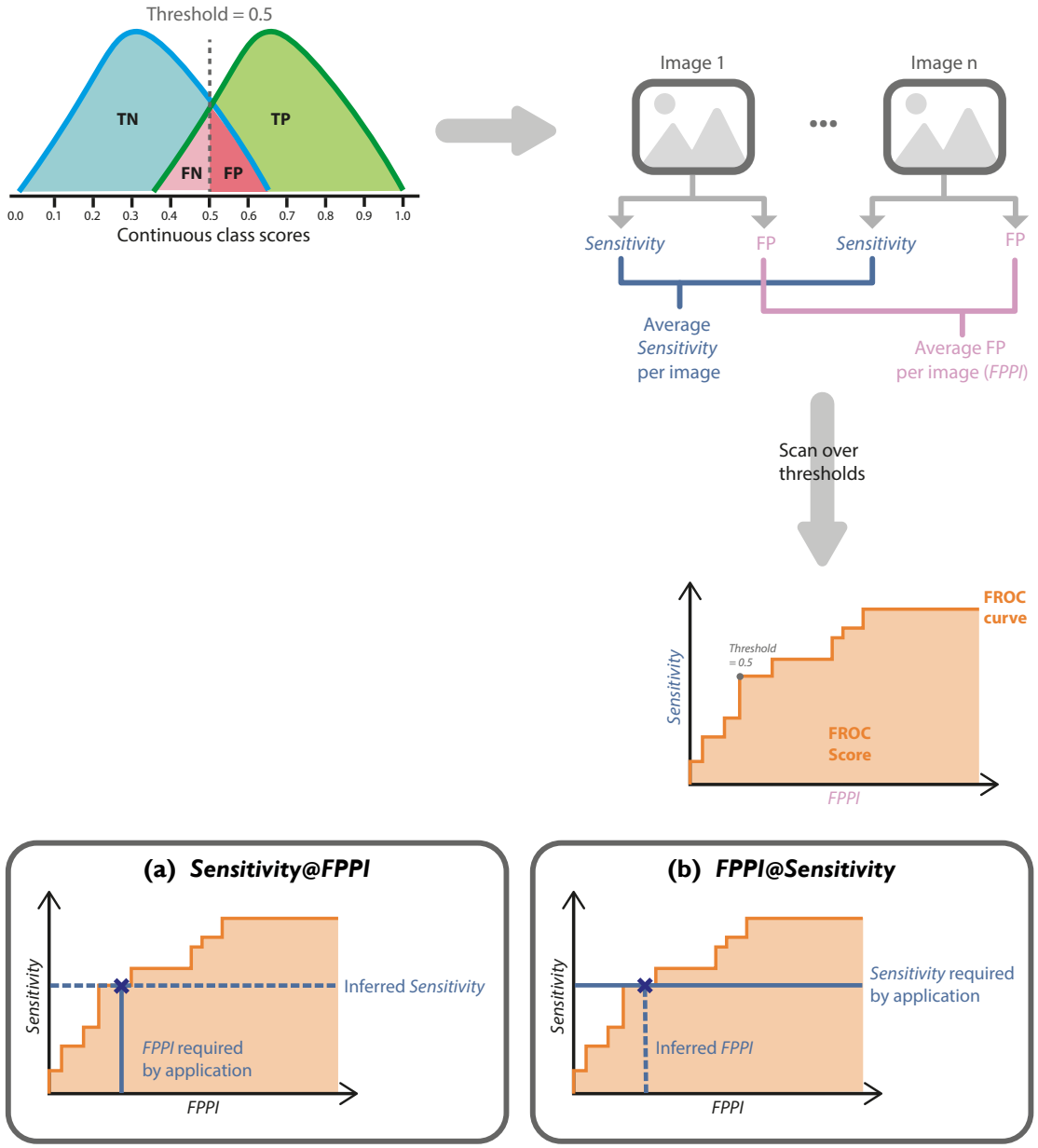


Figure 2.20: Principle of the Free-Response Receiver Operating Characteristic (FROC) Score. The FROC curve integrates across a range of thresholds to generate a dynamic confusion matrix rather than being based on a single threshold. Based on the cardinalities for each threshold, i.e. the true (T)/false (F) positives (P)/negatives (N) values, Sensitivity and False Positives per Image (FPPI) are calculated and plotted against each other (middle). FROC operates on object level and the FROC Score measures the area under the FROC curve. In the case of an application-driven threshold (for example, required Sensitivity of 0.9), the metrics Sensitivity@FPPI and FPPI@Sensitivity can be computed using the FROC curve. Figure from [Reinke et al., 2021a].

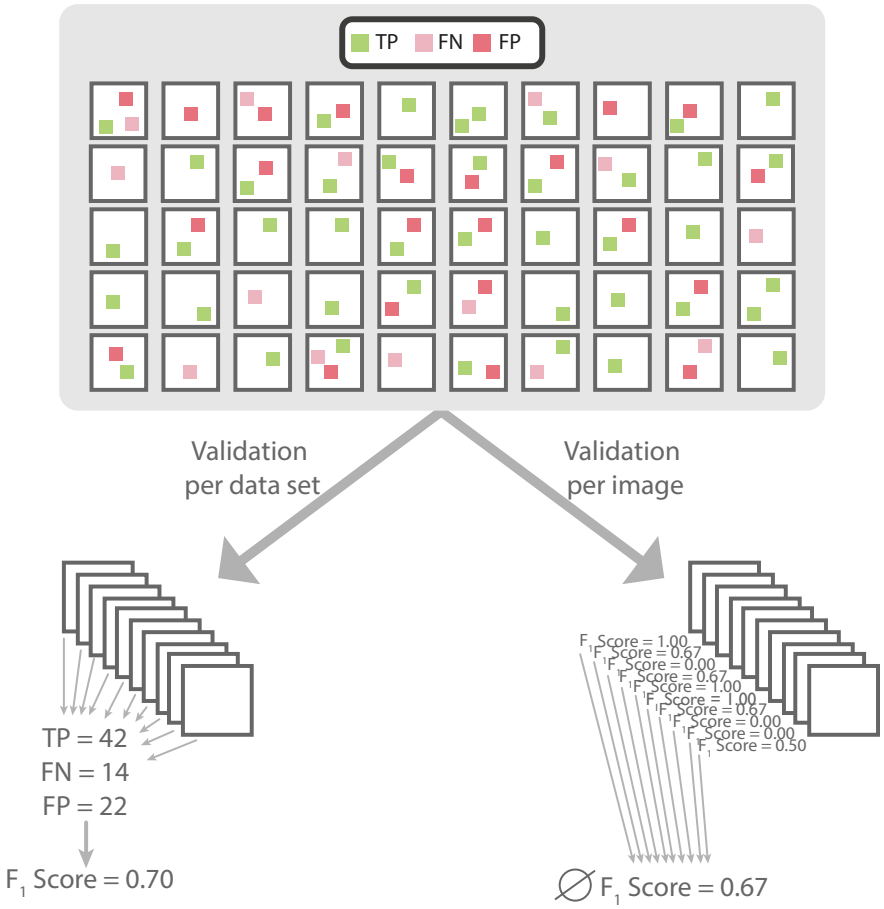


Figure 2.21: Validation at object level can be done on a per-data-set (left) or per-image (right) basis. When validated per data set, cardinalities are counted over all objects of the entire data set, from which the metric scores can be derived (here: F_1 Score). Contrary, when validated per image, the metric values are directly computed for every image (here: F_1 Score) and aggregated afterwards. $\varnothing$ refers to the average F_1 Score. Figure adapted from [Reinke et al., 2021a].

2.4.4 Instance segmentation metrics

In instance segmentation problems, object detection metrics are typically combined with semantic segmentation metrics that are applied per instance. Thus, the same localization criteria (**Mask IoU (> 0)**, **Box/Approx IoU**, **Centroid Distance**, **Point inside Mask/Box/Approximation**, **IoR** and **Boundary IoU**) and assignment strategies (**Greedy by Score Matching**, **Greedy by "localization criterion" Matching**, **Optimal Hungarian Matching**, **Matching via "localization criterion" > 0.5**) as presented in Section 2.4.3 can be applied to instance segmentation problems to find out which instances were detected and how to deal with ambiguities.

Per-class counting metrics (**Sensitivity**, **PPV**, **F_β Score**, and **FPPI**) can be applied to assess the detection performance of the instance segmentation algorithm, while overlap-based metrics (**DSC**, **IoU**, **F_β Score**, and **clDice**) and boundary-based metrics (**HD**, **HD95**, **ASSD**, **MASD**, **NSD**, and **Boundary IoU**) assess the pixel-wise segmentation performance per instance.

Panoptic Quality (PQ) [Kirillov et al., 2019] was designed to assess detection and segmentation performance in one score. The detection quality is measured by the F_1 Score, taking into account TP, FP, and FN, while the segmentation performance is measured by averaging IoU scores of all TP instances (see Figure 2.22 for illustration):

$$\begin{aligned} \mathrm{PQ} &= \frac{\sum_{(\mathrm{Ref},\mathrm{Pred})\in\mathrm{TP}} \mathrm{IoU}(\mathrm{Ref},\mathrm{Pred})}{|\mathrm{TP}| + 0.5\cdot|\mathrm{FP}| + 0.5\cdot|\mathrm{FN}|} \\ &= \underbrace{\frac{\sum_{(\mathrm{Ref},\mathrm{Pred})\in\mathrm{TP}} \mathrm{IoU}(\mathrm{Ref},\mathrm{Pred})}{|\mathrm{TP}|}}_{\text{Segmentation quality}} \cdot \underbrace{\frac{|\mathrm{TP}|}{|\mathrm{TP}| + 0.5\cdot|\mathrm{FP}| + 0.5\cdot|\mathrm{FN}|}}_{\text{Detection quality}} \in [0,1] \end{aligned} \tag{2.60}$$

PQ was initially used to validate panoptic segmentation problems, which combine semantic and instance segmentation [Kirillov et al., 2019]. The metric is calculated for each class, including the background class, and can be used to solve instance segmentation problems directly by validating only foreground instances.

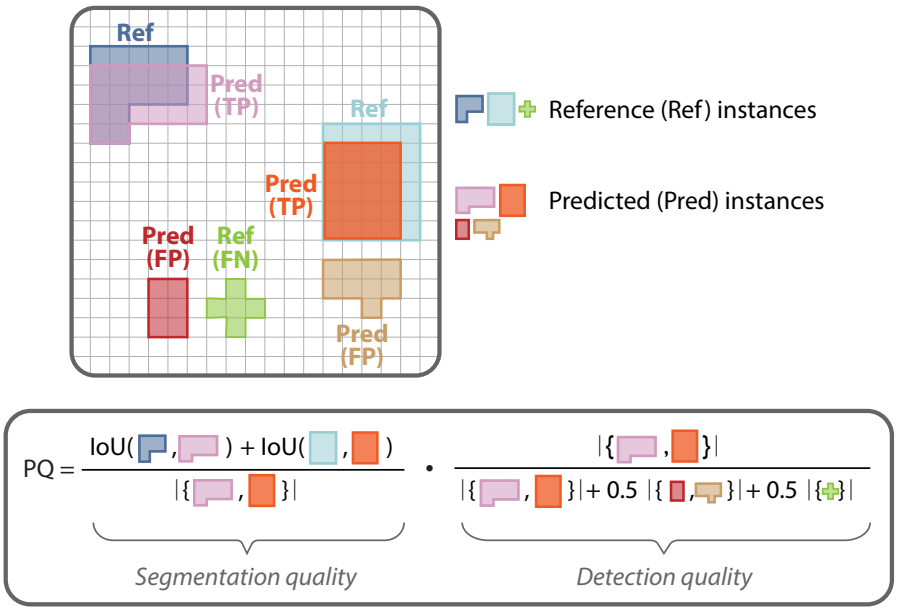


Figure 2.22: Principle of the Panoptic Quality (PQ). The metric measures the segmentation and detection quality simultaneously. Figure adapted from [Reinke et al., 2021a].

2.4.5 Calibration metrics

A model's **discrimination capability** is defined as its ability to distinguish between samples containing and excluding a specific class. If the model discriminates perfectly, for example as indicated by an AUROC score of 1.0, then all predicted class scores for the samples that belong to the class of interest are higher than for the other samples. In contrast, predicted class scores from a **calibrated model** match the empirical success rate. For instance, if a model generates a predicted class score of 0.7 for a particular class, this indicates that it empirically belongs to this class in 70% of cases [Cook, 2007]. Indeed, a model discriminating perfectly does not automatically imply that it is also well-calibrated.

For this reason, calibration metrics measure the calibration of a predictor. When assessing the calibration performance of a method, the concrete validation question should be clearly defined. In many cases, a *comparative calibration assessment* is desired, i.e. ranking different models on how well they perform the calibration. This type of calibration can be further subdivided into three main use cases, as illustrated in Figure 2.23 (U1-U3). One may be interested a comparison of different re-calibration methods for one fixed classifier (U1). In such a case, one would not be interested in interpreting the calibration errors, but rather in the comparison of the re-calibrations. Alternatively, the calibration performance should be compared across multiple classifiers with optional re-calibration (U2). Finally, one may wish to assess the overall performance of different models, both in terms of discrimination and calibration (U3). Comparative calibration assessment could be either complemented or replaced by an *interpretable estimate of the calibration error* (U4), i.e. an understandable and shareable indicator of the calibration performance. In such a use case, the interest lies in quantifying and communicating the reliability of the predicted class scores for a fixed model.

It should be noted that different publications refer to varying definitions of calibration based on the output vectors of a model containing predicted class scores. An example illustrating the difference between them for the computation of the calibration error (see Definition 2.4.1) is provided in Figure 2.24. Based on Vaicenavicius et al. [2019] and Posocco and Bonnefoy [2021], we refer two random variables $X \in \mathcal{X}$ as the input and $Y \in \mathcal{Y}$ as the label, and distinguish between:

Top-label or max-confidence calibration: In top-label calibration, only the maximum predicted class score (top prediction) is considered. All other scores are neglected. A model g is calibrated if the following condition holds: $P[Y = \operatorname{argmax} g(X) \mid \max g(X)] = \max g(X)$.

Marginal or class-wise calibration: In the class-wise calibration, the marginal predictions of an algorithm should be calibrated. The predicted class scores are compared to the reference class-wise, neglecting all other scores that do not belong to the current class. A model g is calibrated if the following condition holds: $P[Y = k \mid g_k(X)] = g_k(X)$ for all classes k.

Canonical calibration: The canonical calibration definition inspects the full vectors of predicted class scores and is therefore the strictest definition. A model g is calibrated if the following condition holds: $P[Y \in \cdot \mid g(X)] = g(X)$.

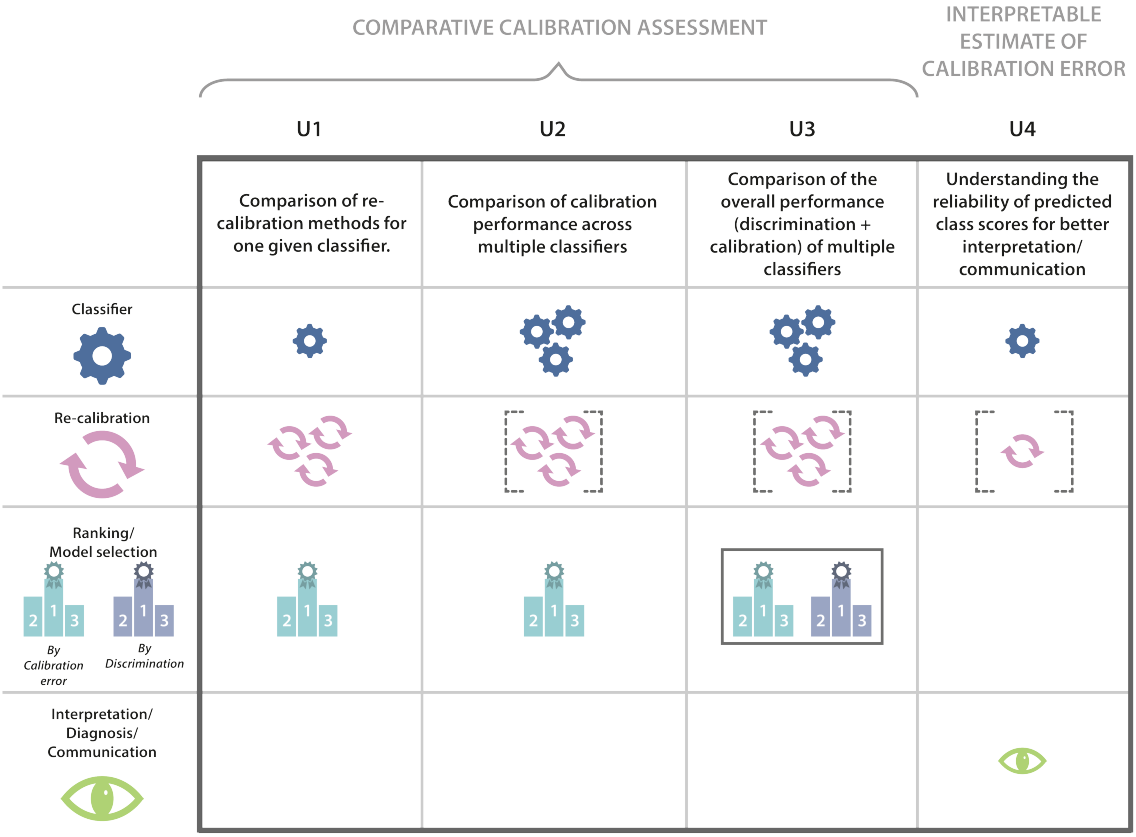


Figure 2.23: Use cases of calibration assessment. Depending on the use case U, either a *comparative assessment of calibration* (U1-U3) or an *interpretable estimate of the calibration error* (U4) may be desired. For the four use cases U1-U4, the type of classifier (fixed classifier versus multiple classifiers), re-calibration (once or multiple; may be optional as indicated by dotted lines), potential ranking and/or model selection strategies (by calibration error or by discrimination), and potential interpretation, diagnosis, or communication is listed.

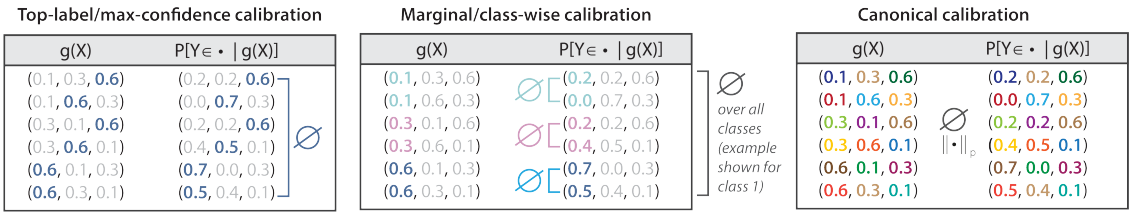


Figure 2.24: Illustration of different definitions of calibration for the exemplary computation of the calibration error. ∅ denotes the average or expectation of values.

Calibration errors [Guo et al., 2017] measure the difference between the predicted class scores of a model g and the actual outcome for one of the three definitions of calibration. According to Kumar et al. [2019], an ℓ_p calibration error for a binary situation is defined as follows:

Definition 2.4.1 (Calibration error as defined by Kumar et al. [2019])**.** Let $\mathcal{X}$ be the input space and $\mathcal{Y} = \{0, 1\}$ be the label space. The ℓ_p calibration error CE_p of a model $g : \mathcal{X} \to [0, 1]$ is given by

$$\mathsf{CE}_p = \left(\mathbb{E}\Big(|g(X) - \mathbb{E}\big(Y \mid g(X)\big)|^p \Big) \right)^{1/p} \tag{2.61}$$

Since the conditional expectation for the calibration error is unknown and must be estimated using the validation data, determining the true calibration error is usually an intractable problem. In order to empirically compute the conditional expectation for the calibration error, density estimation is typically applied. In many cases, this is done by subdividing the interval of predicted class scores [0, 1] into bins and quantify deviations from the described matching [Naeini et al., 2015; Popordanoska et al., 2022]. They often rely on so-called *reliability diagrams*, which "are a visual representation of model calibration" [Guo et al., 2017].

The ℓ_1 calibration error is commonly known as the **Expected Calibration Error (ECE)** [Naeini et al., 2015]. In the course of the thesis, we use ECE for top-label calibration. More concretely, for different bins of confidence $\mathsf{conf}(B_m)$, ECE considers the expected accuracy $\mathsf{acc}(B_m)$ of the calibration. It is defined as:

$$\begin{aligned} \mathsf{ECE} &= \sum_{m=1}^{M} \frac{|B_m|}{n} |\mathsf{acc}(B_m) - \mathsf{conf}(B_m)| \in [0, 1] \\ \mathsf{acc}(B_m) &= \frac{1}{|B_m|} \sum_{i \in B_m} \mathbf{1}(\hat{y}_i = y_i) \\ \mathsf{conf}(B_m) &= \frac{1}{|B_m|} \sum_{i \in B_m} \hat{p}_i, \end{aligned} \tag{2.62}$$

where B_m denotes the indices of all n samples, which are put into one bin, $\hat{y}_i$ is the predicted and y_i the true class label, $\mathbf{1}$ is referred to as the indicator function, which is one if $\hat{y}_i = y_i$ and zero otherwise, and $\hat{p}_i$ refers to the predicted class score. The score can be interpreted as the weighted average of the difference between the average predicted class scores in a bin and the number of correct predictions in this bin [Nixon et al., 2019].

Instead of measuring an average of the difference between accuracy and confidence, the **Maximum Calibration Error (MCE)** [Guo et al., 2017] measures the maximum of this difference:

$$\mathsf{MCE} = \max_{m \in \{1, \ldots, M\}} |\mathsf{acc}(B_m) - \mathsf{conf}(B_m)| \in [0, 1] \tag{2.63}$$

Small values of the ECE and MCE correspond to a better-calibrated model.

The definition of calibration is not clearly defined in a non-binary environment with K classes [Kumar et al., 2019]. The **Top-label Calibration Error (TCE)** [Kumar et al., 2019] is the calibration error when only considering the top predicted class scores:

$$\text{TCE} = \left(\mathbb{E}\left(P\Big[Y = \underset{j \in |K|}{\operatorname{argmax}}\, g(X)_j \mid \max_{j \in |K|} g(X)_j \Big] - \max_{j \in |K|} g(X)_j \right)^2 \right)^{1/2} \tag{2.64}$$

The **Marginal** or **Class-wise Calibration Error (CWCE)** [Kumar et al., 2019], on the other hand, focuses on the class-wise calibration performance and is defined as

$$\text{CWCE} = \left(\sum_{k=1}^{K} w_k \mathbb{E}\Big(\big(g(X)_k - P[Y = k | g(X)_k] \big)^2 \Big) \right)^{1/2} \tag{2.65}$$

with w_k being a class-specific weight. With $w_k = 1/k$, all classes are given the same weight.

In contrast, the **Kernel Calibration Error (KCE)** [Widmann et al., 2019; Gruber and Buettner, 2022] includes the whole predicted class score vector in its calculation and is based on a matrix-valued kernel k, instead of binning, from a reproducing kernel Hilbert space $\mathcal{H}$. Based on [Micchelli and Pontil, 2005; Caponnetto et al., 2008], reproducing kernel Hilbert spaces and matrix-valued kernels are defined as follows:

Definition 2.4.2 (Reproducing Hilbert space as defined by [Caponnetto et al., 2008])**.** Let $\mathcal{Y}$ be a real Hilbert space with inner product $(\cdot, \cdot)$, $\mathcal{X}$ be a set, and $\mathcal{H}$ a linear space of functions on $\mathcal{X}$ and values from $\mathcal{Y}$. $\mathcal{H}$ is a reproducing kernel Hilbert space when for any $y \in \mathcal{Y}$ and $x \in \mathcal{X}$ the linear functional that maps $g \in \mathcal{H}$ to $(y, g(x))$ is continuous.

Definition 2.4.3 (Matrix-valued kernel as defined by [Widmann et al., 2019; Micchelli and Pontil, 2005])**.** A function $k : \Delta^m \times \Delta^m \to \mathbb{R}^{m \times m}$ is called a *matrix-valued kernel* if $k(s, t) = k(t, s)^T$ for all $s, t \in \Delta^m$ and it is positive semi-definite, that means for all $n \in \mathbb{N}, t_1, \ldots, t_n \in \Delta^m$, and $u_1, \ldots, u_n \in \mathbb{R}^m$:

$$\sum_{i,j=1}^{n} u_i^T k(t_i, t_j) u_j \leq 0. \tag{2.66}$$

Based on these definitions, the KCE can be defined with the matrix-value based kernel k from a reproducing kernel Hilbert space $\mathcal{H}$:

$$\text{KCE} = \Big\| \mathbb{E}\Big(\big(g(X) - P[Y \mid g(X)] \big) k(g(X), \cdot) \Big) \Big\|_{\mathcal{H}} \tag{2.67}$$

The proposed discretization of the conditional expectation of the calibration error by binned estimates of the predicted class scores comes with the problem of not being differentiable, thus, the calibration error cannot be easily integrated into the training of a Neural Network (NN) [Popordanoska et al., 2022]. Therefore, Popordanoska et al. [2022] proposed a different strategy, namely the **Expected Calibration Error Kernel Density Estimate (ECE$^{\text{KDE}}$)**, which is based on kernel density estimates instead of binning. With the kernel density estimate, ECE can be reformulated as

$$\begin{aligned} \text{ECE}^{\text{KDE}} &= \frac{1}{n} \sum_{j=1}^{n} \Big| \widehat{\mathbb{E}(y \mid g(X))}|_{g(x_j)} - g(x_j) \Big| \\ \mathbb{E}(y \mid g(X)) &\approx \frac{\sum_{j=1}^{n} k(g(X), g(x_j)) y_j}{\sum_{j=1}^{n} k(g(X), g(x_j))} =: \widehat{\mathbb{E}(y \mid g(X))} \end{aligned} \tag{2.68}$$

with k being the kernel of a kernel density estimate and $\mathbb{E}(y \mid \widehat{g(X)})|_{g(x_j)}$ referring to $\mathbb{E}(y \mid \widehat{g(X)})|$ being evaluated at $g(X) = g(x_j)$.

Proper Scoring Rules (PSR) [Gneiting and Raftery, 2007] validate how closely the predicted class scores match the actual probability for each class given the reference. An ideal score of a PSR reflects the accuracy of the predictions, and PSR measures both calibration and discrimination ability. A prominent example is the **Brier Score (BS)** [Brier et al., 1950], defined as the mean squared error of a predicted class score p_t for one of n events t and the actual outcome y_t. The actual outcome is typically defined as 1 if the event holds true, i.e. if a specific class is correct, and 0 otherwise. Small values of the BS correspond to a better-calibrated model.

$$\text{BS} = \frac{1}{n}\sum_{t=1}^{n}(p_t - y_t)^2 \in [0, 1] \tag{2.69}$$

For use case U1 (see Figure 2.23), BS can be decomposed into discrimination and calibration terms showing that the calibration term exactly resembles the canonical calibration error (see Definition 2.4.1) [Gneiting and Raftery, 2007].

A normalization variant of BS is the **Brier Skill Score (BSS)**, which quantifies how far off a forecast is from an ideal prediction [Varoquaux and Colliot, 2022]. A value of zero for this metric indicates that the predicted class scores are only as insightful as the Prevalence ϕ. It is defined as:

$$\text{BSS} = 1 - \frac{\text{BS}(p, y)}{\text{BS}(\phi, y)} \in (-\infty, 1] \tag{2.70}$$

The **Root Brier Score (RBS)** [Gruber and Buettner, 2022] denotes the square root of the BS:

$$\text{RBS} = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(p_t - y_t)^2} \in [0, 1] \tag{2.71}$$

According to Gruber and Buettner [2022], RBS calculates an asymptotically tight and unbiased upper bound of the canonical calibration error (see Definition 2.4.1).

Another example is the **Negative Log Likelihood (NLL)** [Cybenko et al., 1998], which calculates the negative logarithms of the predicted class scores p_t multiplied by the actual outcome y_t:

$$\text{NLL} = -\frac{1}{n}\sum_{t=1}^{n} y_t \cdot \log(p_t) \in (-\infty, 0] \tag{2.72}$$

NLL validates a weighted version of the canonical calibration condition, since the logarithm introduces a great emphasis on penalizing tail-probabilities [Quinonero-Candela et al., 2005]. For use case U3 (see Figure 2.23), both BS and NLL imply that an optimal score corresponds to perfect prediction and perfect discrimination [Gneiting and Raftery, 2007].

While calibration is typically considered for image-level classification problems, it can also be applied to object detection and instance segmentation problems with available predicted class scores. For these problems, the background class is typically neglected from the calibration measurement. Only the NLL cannot be computed for the two problem categories, due to the its

consideration of the predicted class scores for the correct reference class. For FP predictions that could not be matched to a specific reference object, the correct reference class would be the background, which is not used for calibration assessment [Maier-Hein et al., 2022].

2.5 Related work

In this section, we present an overview of related work concerning the topics of this thesis. Concretely, we review literature on the four topics of challenge design, validation metrics, rankings, and reporting. While there is a large amount of work on network architectures and Artificial Intelligence (AI) models themselves, the available literature on the validation of those models is limited. Nonetheless, we present the most important current work related to biomedical challenges and the validation of biomedical image analysis algorithms.

2.5.1 Challenge design

Only limited research is available on the design of biomedical image analysis competitions. Kozubek [2016] provides a review on challenges and benchmark design in bioimage analysis, abstracting from medical use cases and focusing on biological applications. The author describes the design of challenges similarly to our description in Section 2.1.2. The focus, however, is on design choices related to bioimaging.

Mendrik and Aylward [2019] categorize challenges into *insight* and *development challenges*. Gaining a comprehensive understanding of a biomedical research problem is the aim of insight challenges. They provide insights into how the variables of interest affect the performance of the participants. Their primary goal is to understand certain parameters, such as the impact of a certain metric to the greatest possible extent rather than to generalize to a larger population. Given that they typically operate on only a small subset of the general population, the rankings of insight challenges only provide a rough estimate of algorithm performance. Deployment challenges, on the other hand, generalize to a broader population, possibly based on the insights gained from an insight challenge, aiming to identify whether the submitted methods solve the primary research question of the challenge. The work highlights the fact that the challenge design and objective should match and mentions important problems such as 'leaderboard climbing', in which challenge participants only incrementally increase their metric values rather than actually working towards solving a particular problem.

2.5.2 Metrics

Taha and Hanbury [2015] present an overview of 20 validation metrics for segmentation problems, including a definition and explanation of each presented measure. In addition, all metrics were implemented in a toolkit in C++. The metrics were categorized into

- Spatial overlap-based metrics (e.g. Dice Similarity Coefficient (DSC)),
- Volume-based metrics (e.g. Volumetric Similarity (VS) [Cárdenes et al., 2009]),
- Pair counting based metrics (e.g. Rand Index (RI) [Reddy et al., 2013]),
- Information theoretic-based metrics (e.g. Mutual Information (MI) [Zou et al., 2004]),
- Probabilistic metrics (e.g. Cohen's Kappa (CK)) and

- Spatial distance-based metrics (e.g. Hausdorff Distance (HD)).

The authors further present some insights into metric correlation and limitations.

Nai et al. [2021] provides another review of segmentation metrics, for the special case of prostate segmentation. They compared nine previously proposed measures with 24 standard metrics to check which metrics work best for clinical validation of segmentation problems. The performance assessment was conducted with three deep learning methods and two additional human raters who further assessed the visual agreement of segmentations and whether the metric values agreed with their inspection. The authors found that the Interclass Correlation (ICC) [Gerig et al., 2001] achieved the highest correlation with the visual agreement of the human raters and that previously proposed metrics did not outperform older validation measures. However, this review is only based on a single data set for segmentation problems and it is unclear whether the results transfer to a different study.

A similar overview of image-level classification metrics is provided by Hicks et al. [2022] with a specific focus on gastroenterology use cases. They describe and define several metrics and briefly mention some limitations along with publishing an open-source toolkit for calculating several counting metrics from problem-specific confusion matrices.

Several publications mention the limitations of metrics. For example, Kofler et al. [2021] show that the DSC only moderately correlates with results assessed by human raters. Yeghiazaryan and Voiculescu [2015] highlight the shape unawareness of overlap-based metrics. Moreover, some flaws of image-level classification metrics are highlighted in [Chicco and Jurman, 2020; Chicco et al., 2021]. More frequently, blog posts have been utilized to discuss limitations of metrics in more depth (e.g. [Kooi, 2021a; Widmann, 2020; Brownlee, 2020]).

2.5.3 Rankings

Rankings in general have been clearly defined in the literature, e.g. in [Davey and Priestley, 2002], as described in Section 2.1.2. Some attention has been drawn to the question of how to visualize ranking variability. Demšar [2006] presents basic diagrams for the visualization of post-hoc tests across multiple data sets. Changes in rankings were visualized by parallel coordinates diagrams by Gratzl et al. [2013]. More advanced visualization of benchmark results is presented in Eugster et al. [2008, 2012]. The authors illustrate benchmark results using boxplots, podium plots, and cluster dendrograms, for example.

In the case of multiple tasks or metrics, aggregating tasks is often desired. Fishbaugh et al. [2017] presents a data-driven approach to combining aggregate rankings for different metrics in biomedical image analysis challenges. Metrics that are robust to small changes in the data will be given higher weights compared to metrics whose rankings change more drastically. In addition, metrics that yield similar rankings will receive higher weights. Although this approach seems reasonable, it has not yet been implemented in practice. Lin [2010]; Hornik and Meyer [2007] present a more convenient way to achieve a consensus ranking, for example by finding the ranking that minimizes distances between individual rankings.

2.5.4 Reporting

Several reporting guidelines, checklists, and statements for standardization exist outside the field of biomedical image analysis, such as

- 'Consolidated Standards of Reporting Trials (CONSORT)' for reporting randomized controlled trials [Begg et al., 1996; Schulz et al., 2010],
- 'Standards for Reporting of Diagnostic Accuracy Studies (STARD)' for reporting diagnostic or prognostic studies [Bossuyt and Reitsma, 2003; Bossuyt et al., 2003, 2015; Cohen et al., 2016],
- 'Strengthening the Reporting of Observational studies in Epidemiology (STROBE)' [Vandenbroucke et al., 2007] for reporting observational studies, or
- 'Case Reporting Guidelines (CARE)' [Gagnier et al., 2013] for case reports.

All of those guidelines have been registered with the 'Enhancing the QUAlity and Transparency Of health Research (EQUATOR)' network [Altman et al., 2008], an umbrella organization for high quality and transparency Of health research. Guidelines in the context of Machine Learning (ML) and AI are much less common. The CONSORT statement, for instance, released an extension for clinical trial reports for AI research [Liu et al., 2020]. Very recently, the 'Developmental and Exploratory Clinical Investigations of DEcision support systems driven by Artificial Intelligence (DECIDE-AI)' reporting guideline was published for early-stage clinical validation of support systems using AI [Vasey et al., 2022]. The 'Checklist for Artificial Intelligence in Medical Imaging (CLAIM)' [Mongan et al., 2020] was specifically developed for reporting the results of medical image analysis algorithms. It includes AI-specific items, such as detailed descriptions of the model architecture, inputs and outputs, or potentially used ensembling techniques. It further gives advice on how to report the model performance. However, none of those guidelines cover challenge-specific parameters.

Jannin et al. [2006] propose a model for reporting the results of validation of image processing methods for segmentation and registration problems. Their model relies on an ontological approach and was instantiated for several use cases. The authors highlight the importance of the definition of a *validation objective*, which includes both the clinical context and the clinical goal of the validation research. However, the checklist provided alongside the model by the authors does not cover challenge-specific parameters such as a challenge's life cycle, rankings, or submission methods. It further mainly focuses on segmentation and registration problems.

2.5.5 Conclusion

The literature concretely focusing on biomedical challenges is sparse. Mendrik and Aylward [2019] and Kozubek [2016] briefly mention some limitations of challenges, but without analyzing them in a structured fashion. In this thesis, we build upon their definitions and expand their analysis on limitations.

Several publications mention metric limitations. However, previous work usually did not cover multiple problem categories and only provided limited insight into potential issues regarding metrics. Nonetheless, we collected metric pitfalls from previous work that we include in Section 4.2. In addition, we used the work of Taha and Hanbury [2015], Nai et al. [2021] and Hicks et al. [2022] as a basic list for the metric selection and families in Section 5.2.

Although there is some literature on the visualization of rankings, we are not aware of challenges or benchmarking experiments in the biomedical image analysis community that were using the presented plots and diagrams before our intervention and proposed work on ranking uncertainty, presented in Section 5.3. This work builds upon the proposed visualization of Eugster et al. [2008, 2012].

In comparison to other research fields such as epidemiology and general diagnostic testing, biomedical image analysis lacks a variety of reporting guidelines. Jannin et al. [2006] provides a highly relevant basic variant for reporting validation studies that is, however, restricted to segmentation and registration approaches. In addition, this and other checklists for AI research, such as CLAIM, do not cover challenge-specific parameters that are important to report. Still, the model proposed by [Jannin et al., 2006] served as the basis for our considerations for a challenge design-specific parameter list presented in Chapter 3 and Sections 4.4 and 5.4.

Summary:

The literature on the design, validation, and reporting of challenges is sparse. To our knowledge, there is neither a comprehensive work on collecting limitations of challenges or general validation of biomedical image analysis algorithms nor comprehensive work on solving potential issues. We used related work from other communities, such that related to clinical trials, as the basis for our work, adapting it for use in biomedical image analysis applications, such as challenges.

3 Descriptive analysis of Biomedical Image Analysis Validation

3.1 Common practice of challenges

Disclosure

This work has been supervised by Lena Maier-Hein. It has been conducted together with several co-authors, especially Matthias Eisenmann and Annette Kopp-Schneider. The work has been published in *Nature Communications* [Maier-Hein et al., 2018]. Please refer to Chapter A.1 for full disclosure.

3.1.1 Introduction

The Machine Learning (ML) research community has historically relied primarily on authors' personal data sets for the validation and assessment of new methodologies. This meant that a direct and fair comparison of algorithms was not possible, as other researchers typically did not have access to the exact same data sets or hardware setup. Comparisons between data sets are challenging: On the one hand, the conditions are not comparable. Data sets may differ regarding the difficulty of their processing, may be different in volume, or may be subject to domain shifts. On the other hand, many of the common performance metrics rely on the prevalence of samples in the data (see Section 2.4) and behave differently for different data sets.

Comparing various methods for brain registration was the focus of the first attempt at a fair comparison of algorithms in the biomedical community, carried out in 1997 [West et al., 1997]. In this comparative study, methods were benchmarked on the same data set. An interesting modification to previous research practice was that the teams taking part were not given access to the reference annotations of the test data. Since 2004, there have been yearly competitions in the field of medical image retrieval in the context of the Image Cross Language Evaluation Forum (ImageCLEF). The largest conference for computer-assisted interventions and biomedical image processing (Medical Image Computing and Computer Assisted Interventions (MICCAI)) adopted this novel idea and presented its first two segmentation challenges in 2007 [Heimann et al., 2009; Van Ginneken et al., 2007]. Since then, the number of challenges has increased yearly, with challenges becoming the state-of-the-art method for assessing the performance of ML algorithms comparatively.

In this section, we review more than 500 competitions from the biomedical image analysis domain to assess the impact of challenges as well as common practices. Finally, we present the findings of an international survey that reflects the viewpoint of the general scientific community.

Research questions investigated in this chapter

We contend that prior quality assurance mechanisms for challenges were insufficient to guarantee high-quality challenge design and organization. We focus on the following three research questions:

1. What is the role of challenges?
2. What are common practices in designing challenges?
3. What is the general opinion of the community on challenge design?

3.1.2 Methods

With the following experiment, we aimed to analyze common practices of challenges. For this purpose, we captured several challenges from different sources, as described below. For those challenges, we conducted a descriptive analysis of their designs and results. This step was achieved by creating a list of challenge parameters that was instantiated for all challenges. The challenge parameter list is presented in Section 4.4 in more depth. Finally, we examined the opinion of the research community with an international survey.

Challenges inclusion criteria

To assess the role of challenges in the research community, we collected all competitions from the biomedical image analysis area that were conducted in the period from 2004–2016. Challenges organized at a later point in time were omitted given that their reports were not available at the time of the study. To collect all data, we first gathered challenges organized at biomedical image analysis conferences.

In the second step, we sourced challenges from three popular challenge platforms that were not already included:

1. The Grand Challenge[1] website is a well-known platform for hosting and organizing biomedical challenges. Many challenges organized at the above-mentioned conferences were hosted on Grand Challenge (83% of the acquired challenges).

2. The Dialogue on Reverse Engineering Assessment and Methods (DREAM)[2] is a platform for crowdsourced challenges, especially focusing on biological algorithms, but also hosting several challenges from the image analysis domain which were included in our list.

3. Kaggle[3] is a large-scale organization for data scientists, hosting thousands of competitions for different applications and use cases. We scanned the platform for biomedical image analysis challenges.

[1] grand-challenge.org
[2] dreamchallenges.org
[3] kaggle.com/competitions

Descriptive analysis

The research questions presented at the beginning of this section address the role of challenges in the community and common practice. With these research questions, we were aiming to identify a comprehensive overview of organized challenges, including the number of challenges, tasks and data, fields of application, problem categories, and similar.

We implemented a meta-model in `Java 11`, using `Eclipse Modeling Tools`, in which we modeled important challenge design parameters (see Section 4.4 for details). A total of four engineers, including myself, and one medical student served as observers for challenge instantiation, supported by four engineering Master students. At least two individual observers instantiated the parameters for each challenge, such that all competitions in the period from 2004 to 2016 were instantiated by two independent observers. In this process, information from a dedicated publication of the challenge results, if available, was preferred over the challenge website content, since the latter was outdated in some cases. Using the implemented meta-model, we then automatically compared the instantiation of parameters for every challenge done by the independent observers and checked for ambiguities. In such a case, we consulted an independent third observer for the final decision. In cases of adaptation of the parameter list, we repeated the process and updated the challenge instantiation. We based our descriptive analysis addressing the role and common practice of challenges on this ontological data.

International community survey

To identify the broader community view on biomedical image analysis challenges, we prepared a survey, which was taken by all co-authors of the study presented by Maier-Hein et al. [2018]. Furthermore, the chairs of the MICCAI 2015–2017 challenges distributed the survey among challenge organizers. In addition, it was shared via mailing lists of international recognized conferences or societies, such as ImageWorld and the MICCAI society. Finally, the URL for the survey was available on the Grand Challenge website. The survey was composed of 34 questions covering the following question categories:

Background with respect to challenge participation: We asked survey respondents if they had participated in any challenge so far. Those who did were asked about the number of challenges they took part in and their main motivation for participating. In addition, we asked if they ever registered for a challenge but did not submit results and, if so, for what reason. Lastly, we asked them to indicate any major issues they faced during their participation in challenges, as well as whether they could easily interpret their challenge ranks.

Background with respect to challenge organization: We asked survey respondents if they organized any challenges. Those who did were asked about the number of organized challenges and the temporal effort put into challenge organization. Moreover, we asked them to indicate whether they ever participated in their own challenge(s) and whether they struggled with the creation of the data set, reference annotation, or the choice of validation metrics. Finally, we advised them to list potential issues related to challenge design and organization.

General view on challenges: We asked survey respondents about their general views on challenges. Specifically, we asked whether pre-evaluation results should be provided to

participants, whether organizers should be permitted to participate in their own challenges, and whether they think that rankings adequately reflect algorithm performance. Furthermore, we asked them to indicate if they think that the design of biomedical image analysis challenges should be improved in general.

Open issues and recommendations: We asked survey respondents about open issues in challenges related to data, reference annotations, validation, and documentation. We finally asked for concrete recommendations and whether they think challenges should undergo more quality control.

3.1.3 Results

Descriptive analysis of common practices in challenge design

Based on the above-mentioned search criteria, we captured a total of 150 challenges organized in the period from 2004 to 2016, including 549 tasks. We present most of the numbers below on task rather than challenge level. This is due to the fact that a challenge is often organized in terms of multiple tasks with different focuses, problem categories, or data sets. A challenge-level validation is therefore often not possible. An overview of the results is given in Figure 3.1. A comparison to more recent challenges conducted within the past five years is provided in Section 5.4 and Figure 5.1.

A median of seven challenges (Interquartile Range (IQR): (3, 18)) and 13 tasks (IQR: (4, 18)) were organized in the period from 2004 to 2016 (cf. Figure 3.1(a)). 57% of challenges were published in journals or conference proceedings. Half of the challenges were organized along the MICCAI conference, followed by 34% of challenges organized at ISBI. More than half of the challenges (57%) were organized as repeated events with a fixed conference submission deadline, followed by challenges organized as one-time events (26%). Challenges opening for new submissions after the conference deadline were less common, with 10% being repeated and 7% organized as one-time events. For most of the challenges, challenge participants were supposed to submit their results via a cloud (40%), followed by directly sending their results to the organizers (26%). Semantic or instance segmentation were by far the most commonly used problem categories for challenge tasks (70%), followed by image-level classification (10%) and object detection tasks (7%) (see Figure 3.1(b)).

Challenges were very heterogeneous in their goals and applications. Magnetic Resonance Imaging (MRI) (40%) and Computed Tomography (CT) (26%) were the most commonly applied imaging techniques, but less common modalities such as Endoscopic Imaging (2%) or Ultrasound (US) (2%) could also be found (see Figure 3.1(d)). Most challenges' application fields were cross-topic (29%), followed by tasks designed for diagnosis (28%) and assistance purposes (13%) (see Figure 3.1(e)).

Challenge data was usually acquired from a median of one center. Only 5% of challenges provided the concrete ethical approval ID for the data, with 6% of challenges claiming that no ethical approval was needed, for instance due to complete anonymization of data or simulations. The number of training and test cases in the data sets varied substantially across challenge tasks, as shown in Figure 3.1(c). A training set contained of a median of 15 cases, an IQR of (7, 30) cases and a maximum number of 32,468 cases. The number of test cases was usually higher, with a median of 20 cases, an IQR of (21, 33) cases, and a maximum number of 30,804 cases.

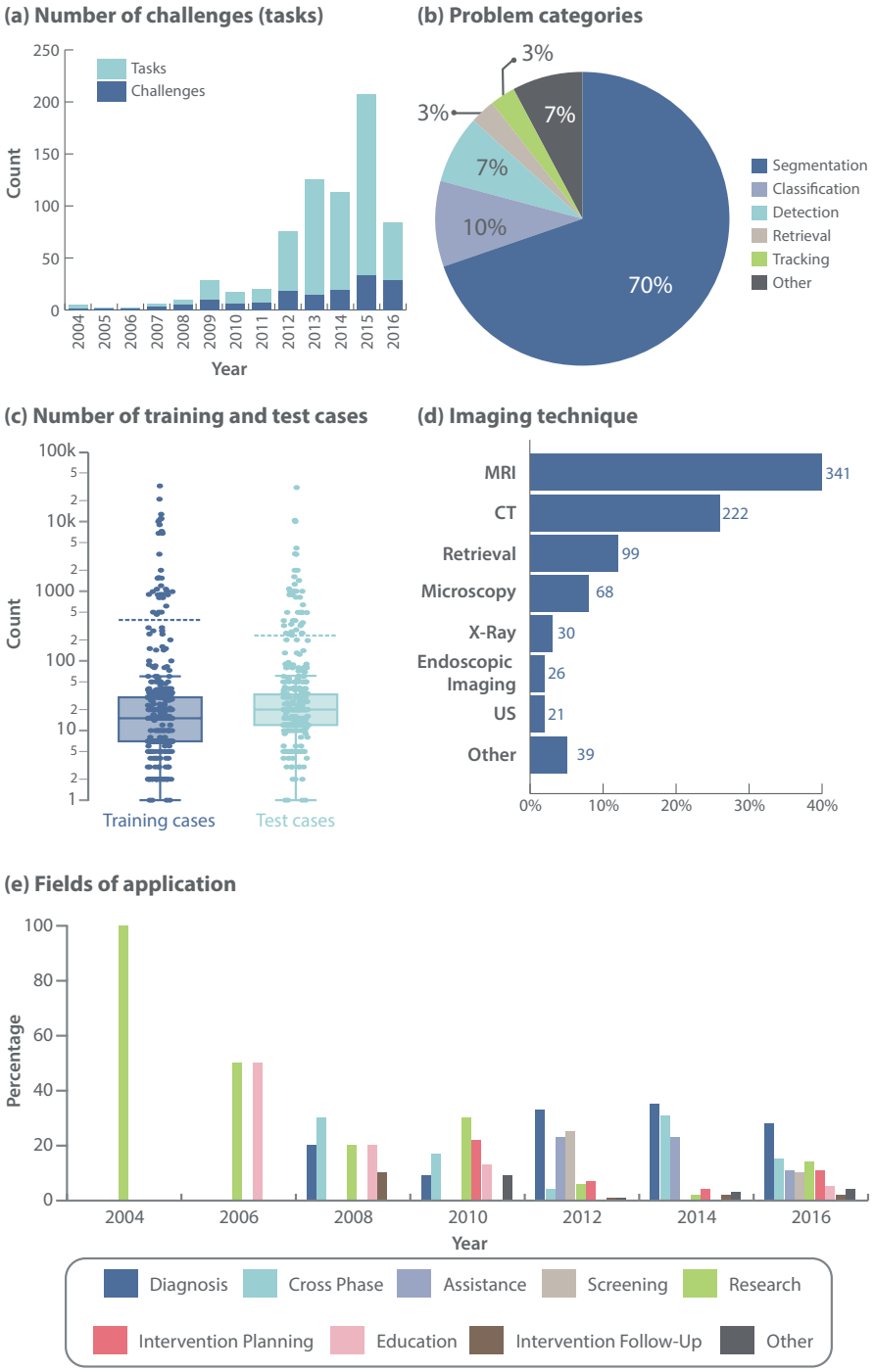


Figure 3.1: Overview and statistics of biomedical image analysis challenges. **(a)** Number of challenges and tasks per year. **(b)** Percentage of problem categories assessed in challenge tasks. **(c)** Dots- and boxplots of the number of training and test cases used in challenge data sets. **(d)** Percentage of used imaging techniques. **(e)** Percentage of fields of application per year. For better readability, we only show results for every second year. Figure adapted from [Maier-Hein et al., 2018].

The overall median ratio of training and test cases was 0.75. The data was annotated by a median of three annotators (IQR: (3, 4); maximum: 9).

Similar to the data, the challenge validation was highly heterogeneous. A total of 97 different performance metrics were used, with the Dice Similarity Coefficient (DSC) being the most commonly used metric, employed by 64% of tasks (92% of the segmentation tasks), followed by the Hausdorff Distance (HD) or the Hausdorff Distance 95 Percentile (HD95), utilized in 34% of tasks (47% of segmentation tasks). 46% of the metrics were only used in a single task, and 58% of tasks did not provide a justification for why a specific metric was chosen. A challenge winner, determined by computing a ranking, was announced in only 39% of challenge tasks. This was not possible in 20% of tasks since only a single team participated in those challenge tasks. From the challenge tasks computing a ranking, 57% calculated the ranking only based on one single metric (45%). Ten different ranking schemes were identified, with a metric-based aggregation being the most common scheme (employed in 76% of tasks). The ranking analysis was done by statistical testing for only 6% of tasks. The t-test was the most commonly applied statistical test (36%), followed by a Wilcoxon signed rank test (17%).

General opinion of the community on challenge design

A total of 295 researchers from 23 countries worldwide participated in the international community survey (see Figure 3.2 for a summary of results). The majority of participants (87%) were from academia (30% professors, 24% Ph.D. students, 14% postdocs, 13% staff scientists, and 6% junior group leader or equivalent), 7% from industry, and 6% from other positions. Most of them (94%) had their primary background in engineering, maths, computer science, or physics, but researchers from medicine and biology (5%) also participated, with 1% of researchers having backgrounds in both areas.

63% of respondents participated in at least one challenge. The main motivation for challenge participation was to gain insights into the algorithm performance in competition with other researchers (47%). For some researchers, a publication required participation in the challenge (19%), and 17% participated to get data access. After submitting their results to the challenge, 18% of challenge participants struggled with interpreting their challenge ranks and 18% of all survey respondents (including non-challenge-participants) agreed that rankings did not represent algorithm performance well. 38% of challenge participants further stated that they registered for a challenge one or multiple times but did not submit their algorithms or results. This was mostly due to strict deadlines and the fact that participating in a challenge is very time-consuming (41%).

Roughly one third of survey respondents (31%) had already organized at least one challenge, of which 37% had participated in their own challenge(s). However, 36% of the survey participants thought that this should not be allowed or only under specific conditions (39%). 27% of challenge organizers indicated issues with the generation of the reference data, 23% with choosing the performance metrics, and 22% with finding the challenge data set and deciding on how to create the challenge ranking.

92% of the survey respondents agreed that the design of current biomedical challenges should be improved (44%/48% voted for slight/significant improvement), and 84% of respondents thought there should be greater quality control applied to challenges held as part of major conferences. Best practice recommendations would have been appreciated by 87% of respondents.

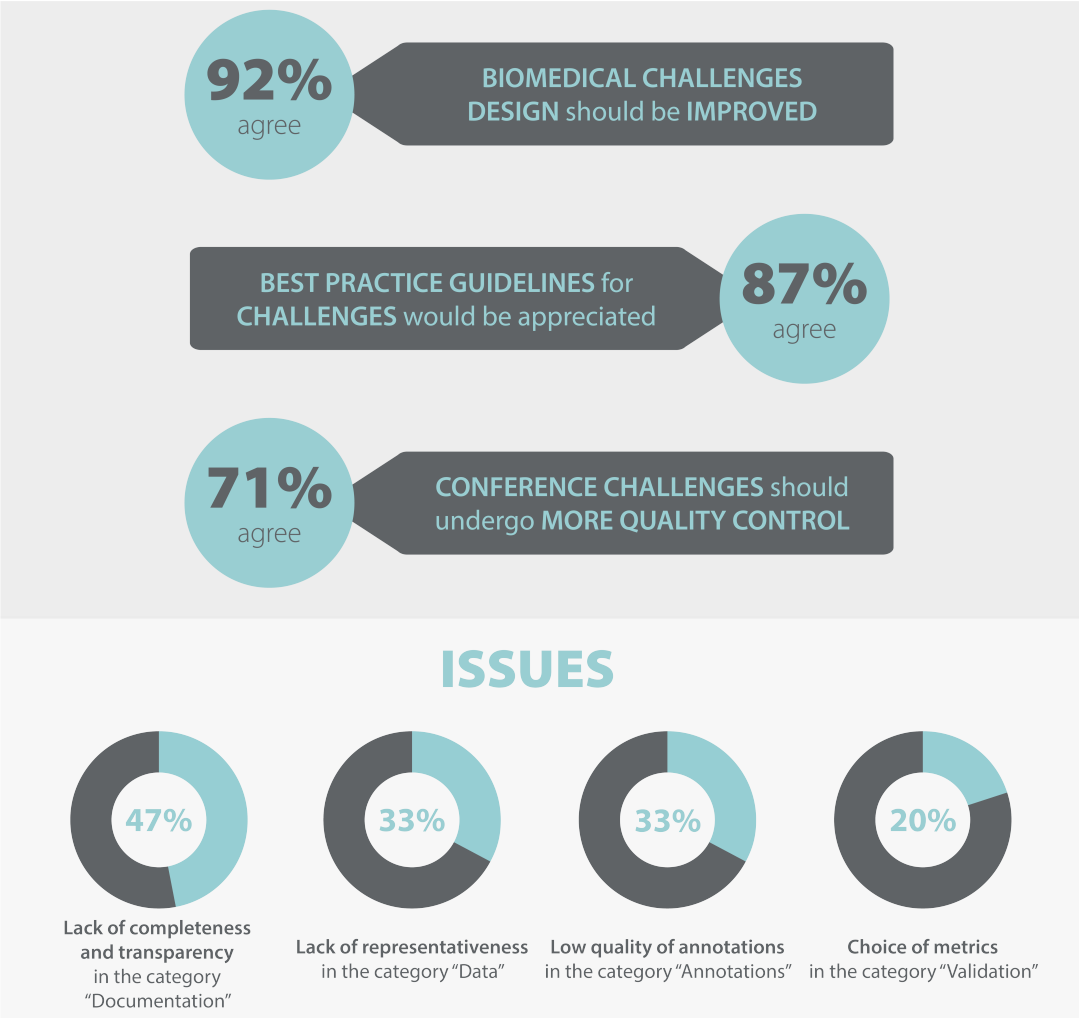


Figure 3.2: Overview of the results of the community survey with $n = 295$ respondents. The upper part (dark gray) summarizes the opinion of the survey respondents relating to design improvements, best practice guidelines, and quality control. The lower part (light gray) summarizes issues that were raised by the survey respondents related to challenge design for the categories documentation, data, annotations, and validation.

The second part of the survey aimed to identify the most pressing issues related to (1) the data, (2) the reference annotation, (3) the validation, and (4) the documentation of biomedical challenges, as summarized in the following:

Data Most concerns were raised with respect to *data representativeness*, realism, data balance in the training, validation, and test sets as well as selection bias in the data set generation (33%). Furthermore, it was mentioned that data was often only acquired by a single center and/or

device, limiting its representativeness (cf. results in *Descriptive statistics of past and recent challenges*). *Acquisition of data*, which is time-consuming and costly and often hindered by legal barriers, was found to be one of the main problems for challenge organizers (17%).

Reference annotation The *quality of reference annotations* was considered crucial for challenge data sets, but also one of the most important issues (33%). Survey respondents mentioned that annotators may often be biased or subjective when generating the annotations, therefore the results depend on the people working on the references. The *method on how to create the reference annotations* was found problematic as well by 16% of respondents. While annotations may be generated by multiple annotators to overcome individual differences, this is often not done due to time and cost issues.

Validation *Choosing the right metric* for a challenge was determined as the most problematic issue for challenge validation by 20% of the respondents. The facts that metrics often do not reflect the clinical or biological context, that metrics measuring the runtime or computational complexity of algorithms were used very rarely and that the aggregation of metrics was not straightforward were criticized. 19% of survey respondents further *missed standards* in the metric choice and implementation as well as the frameworks used for validation.

Documentation Nearly half of the survey respondents (47%) criticized the *lack of completeness and transparency* in challenge documentation, although survey respondents wished to document a challenge as exhaustively as possible. Similarly, the submitted algorithms are in many cases not easily reproducible, although they may become the new state-of-the-art methods. Finally, respondents mentioned issues regarding the *publication of challenge results* (13%). Publishing a challenge paper often takes time, with it possibly only appearing years after the actual submission deadline. Furthermore, the frequent and unexplained differences in challenge website content versus the publication were criticized.

3.1.4 Discussion

We showed that challenges have become an integral part of the validation of biomedical image analysis algorithms. Researchers are able to find a challenge for nearly all research problems and can easily test their algorithms against other methods. Challenges have been designed around a range of modalities and fields of application. While their focus lies on segmentation problems, there are multiple challenges for other common problems such as image-level classification and object detection problems. Indeed, it is harder to find a suitable challenge for less common problem categories such as registration or reconstruction.

However, the high heterogeneity of challenges may also be a problem since there are no common standards related to challenge design. For instance, the number of training and test cases substantially differs across challenges. While there are several challenges with only a single training or test case, others use more than 30,000 cases. Similarly, nearly 100 different performance metrics have been employed, of which half have only been used for a single challenge.

Due to the high importance of challenges and their wide applicability, the biomedical image analysis community agrees that challenges should undergo more quality control, that the design should be standardized and improved, and that best practice recommendations are

necessary. The community raised several concerns about the data and annotation quality. This issue is especially important given the fact that challenge data sets are often publicly available and used as new benchmarking data sets, compensating for the sparsity of available data in health research. Nonetheless, many researchers rate the data quality and representativeness as low although of immense importance. In addition, many researchers rate the choice of the right performance metric as the most critical issue related to the validation of challenge results. Although there numerous metrics are employed in challenges, they often do not reflect the biomedical needs for a specific problem or application. This opens up the question of whether standard metrics are actually suited for validating challenges. A performance metric suitable for one research problem may be absolutely inadequate for another. Problem-aware recommendations could be a feasible step towards solving this specific issue.

Finally, researchers complained about the lack of completeness and transparency in challenge documentation. This comprises the individual challenge websites, especially important for active challenge participants, as well as the following challenge reports. Missing information on the websites may be very frustrating for a challenge participant and may even discourage from participation. Lack of transparency is also critical for the challenge design and post-challenge analysis since the results may not be reproducible and interpretable – which is highly relevant for validation studies. Clinical trials, for example, have very high standards related to study design and reporting to ensure high-quality results.

3.1.5 Conclusion

Biomedical image analysis challenges have led to a tremendous increase in the comparability of algorithms. They are the equivalent of clinical trials in the assessment of image analysis algorithms. However, they are very heterogeneous in their design, raising the question of whether challenge results are actually meaningful. The research community shares the opinion that the design and quality of challenges need to be improved.

In the following Chapter 4, we examine several problems of challenges related to their design, rankings, metrics, and reporting, providing evidence for the issues raised by the research community. While uncovering problems and creating awareness among researchers is a good step, solving them is even better. This is why we provide solutions for the presented problems in Chapter 5 of this thesis. For this, we build upon the best practice recommendations compiled by the survey respondents.

4 | Revealing flaws in common practice of Biomedical Image Analysis Validation

4.1 Revealing flaws related to challenge design

Disclosure

This work has been supervised by Lena Maier-Hein. It has been conducted together with several co-authors, especially Matthias Eisenmann and Annette Kopp-Schneider. The work has been published at the *Medical Image Computing and Computer Assisted Interventions (MICCAI) conference* [Reinke et al., 2018a]. Please refer to Chapter A.1 for full disclosure.

4.1.1 Introduction

In the previous Chapter 3, we demonstrated that challenges have become an essential component of the current research practice. However, we also found that they are very heterogeneous in design and lack standardization, for example in terms of validation metrics or rankings. In fact, 92% of 295 researchers felt that the design of the biomedical challenges needed to be improved (see Chapter 3).

In this section, we review the consequences of low-quality challenge design. From the perspectives of both the challenge organizer and participant, we particularly examine whether challenge design could be exploited for rank manipulation. As cheating is unfortunately quite common among competitions in general [Chui et al., 2021; Dannenberg and Khachatryan, 2020] and practices like 'leaderboard climbing' are not uncommon in challenges [Mendrik and Aylward, 2019], we think that this bad practice may also occur in the research context. Young researchers and Ph.D. candidates in particular often experience extreme pressure to produce high-quality results. Thus, it may be reasonable to suppose that researchers regard rank manipulation as a valid alternative to achieve a higher rank in a challenge.

For this reason, we present the results of two theoretical experiments: First, from the perspective of a challenge organizer, altering a ranking that would see their competitor as the challenge winner may appear worthwhile. We hence simulated a decrease in the rank position of the main competitor by altering the ranking scheme – after the challenge participants submitted their results, and without informing them. Second, from the perspective of a challenge participant, it may appear beneficial to only submit results for those images on which the algorithm worked well. Other, subpar results might be disregarded and not submitted. Based on current challenge design, both situations are possible and may represent a potential security hole and a large risk in terms of cheating.

Hypothesis investigated in this chapter

H1: Current biomedical challenge design is heavily flawed.

In our community survey presented in the last section, many researchers mentioned potential flaws in challenge design. We thus investigate whether such weaknesses can be exploited in order to achieve higher ranks or to decrease the rank of an unpopular competitor.

4.1.2 Methods

The following experiments aim to analyze whether challenge rankings could be easily manipulated by both challenge organizers and participants. We first describe the challenge inclusion criteria and subsequently the concrete experimental setups.

Challenge inclusion criteria

As shown in Chapter 3, we identified segmentation as the most prominent problem category for challenges. To investigate the above-mentioned hypothesis, we aimed for an in-depth analysis of all 14 Medical Image Computing and Computer Assisted Interventions (MICCAI) 2015 segmentation challenges. We approached all challenge organizers and asked them to compute the confusion matrices (True Positives (TPs), True Negatives (TNs), False Positives (FPs), and False Negatives (FNs)) per image based on the segmentation masks. We asked for the results to be provided for all available annotators who annotated the references in the challenge data set and the resulting scores for every challenge participant based on every annotator. We found the Dice Similarity Coefficient (DSC) and Hausdorff Distance (HD) to be the most commonly used metrics across all challenges; therefore, we requested the organizers to calculate these scores as well in addition to the 95% percentile of the HD metric, the Hausdorff Distance 95 Percentile (HD95), as a metric variant. For one of the challenges, the required HD(95) values could not be computed, and we therefore excluded this challenge from the analysis. This resulted in 13 challenges with 124 tasks.

In the case of more than one annotator generating the reference annotation, we used the metric scores based on the merged annotation to calculate the rankings (if available). Two challenges provided results for multiple annotators but without a merged annotation. For one of those two challenges, the segmentation masks generated by the first annotator varied substantially from those of all other raters. Therefore, we used the metric scores based on the segmentation masks provided by the second annotator for the two challenges. Another challenge provided on- and off-site results. In this case, we used the results with most challenge participants. Lastly, we excluded algorithms outputting invalid metric scores for all test cases of a challenge task from the ranking calculation.

For ranking analyses using bootstrapping (ranking robustness analysis; see paragraph *reducing ranks via altering the ranking scheme*), we only considered challenge tasks with three or more participating algorithms (42 tasks excluded) and more than one test case (25 tasks excluded), yielding 56 challenge tasks.

Rank manipulation

In Chapter 3, we showed that challenges were exposed to several weaknesses. To investigate whether and how these may, in theory, be exploited, we conducted the following experiments:

Reducing ranks via altering the ranking scheme. As we show in Section 4.4, the ranking scheme was not published for 20% of the challenge tasks from 2004 to 2016. Indeed, 40% of the MICCAI 2018 challenges did not publish the complete ranking schemes before the challenge was conducted. In theory, this may easily be exploited by a challenge organizer wishing to hinder a specific participant (e.g. their main competitor) from winning by adjusting the ranking scheme until the desired ranking is achieved.

Given the results from Chapter 3, we found metric-based aggregation based on a single metric with an aggregation based on the mean as aggregation operator (used by 72% of all segmentation challenges) as the most commonly used ranking scheme. Furthermore, we identified the DSC to be the most commonly used metric, which was used in all MICCAI 2015 segmentation challenges. Therefore, we defined the **default ranking scheme** as a **metric-based aggregation using the DSC as a ranking metric and the mean as the aggregation operator**. For all 56 MICCAI 2015 segmentation tasks meeting the inclusion criteria, we computed rankings with several variations in the ranking scheme:

- *Metric variant:* We computed rankings for the DSC, HD, and HD95 metrics separately.
- *Aggregation method:* We computed rankings for a metric-based and a case-based aggregation scheme.
- *Aggregation operator:* We computed rankings using the mean and median for aggregation.

We applied these twelve different ranking schemes to all MICCAI 2015 segmentation tasks. Subsequently, we analyzed the winning and non-winning teams for the default ranking scheme by computing the maximal rank differences over all rank variations in bootstrapping simulations. Finally, we calculated Kendall's τ to measure the deviation from the default ranking scheme. We repeated the same experiment for the use case in which challenge organizers did not publish the full ranking scheme but announced the metric(s) used for evaluation. In this case, we considered the metrics as fixed and three default ranking schemes were calculated for the three metrics DSC, HD, and HD95.

Increasing ranks via exclusive submission of test cases. In Section 4.4, we demonstrate that 82% of challenge tasks did not describe their missing value handling, i.e. how they processed invalid submissions of participants. A strategy for missing values is needed for metric-based rankings. Simply ignoring a missing test case submission leads to different aggregated results compared to, for example, assigning the worst possible metric values to these cases.

Working with the 56 MICCAI 2015 segmentation tasks meeting the inclusion criteria, we simulated this problem by manually removing the test cases with DSC values smaller than 0.5 from the individual participants. With this strategy, we aimed to mimic a challenge participant only submitting the most plausible results with the highest metric scores. Although participants usually do not have access to the test case's reference annotations, it can be assumed that participants may clearly identify and separate those cases by visually inspecting their results before the

submission. We compared the default rankings based on a selective test case submission to the default rankings for all test cases.

4.1.3 Results

In the following, we present the results of the ranking analyses for 56 MICCAI 2015 segmentation tasks, for which we performed the rank manipulation experiments.

Reducing ranks via altering the ranking scheme

We computed twelve different ranking schemes for each of the 56 MICCAI 2015 segmentation tasks. Figure 4.1 illustrates the resulting rankings and changes for one example challenge task with 13 algorithms *A1* to *A13*. In this example, the winner of the default ranking scheme *A* is stable for the DSC for all variations of the aggregation method and operator. Kendall's τ values were between 0.79 and 0.90, indicating a high correlation to the default ranking scheme. However, changing the metric (HD in ranking schemes *E* - *H* and HD95 in ranking schemes *I* - *L*) results in a dramatic change in rankings 4 to 11, forcing the original winner down to rank 11, with Kendall's τ values in the range of 0.28 and 0.65.

In general, the winner of the default ranking scheme changed in 84% of tasks. Furthermore, a shared first rank was achieved in 11% of tasks that did not show a tie for the winner in the default ranking. On average, the default winner remained the same in only 57% of rank variations. For one task, the winner even dropped down to rank 11. 20% of the non-winning algorithms could have been ranked first in at least one ranking scheme.

The experiments were repeated for the assumption that the metric has been published before the challenge but the ranking scheme itself was unknown. In this case, the winner was stable in 63% (DSC), 46% (HD) and 41% (HD95) of tasks. In 5% (DSC), 9% (HD), and 8% (HD95) of all tasks, it was possible that the non-winning participants could have been the winner.

		A	B	C	D	E	F	G	H	I	J	K	L
	Metric	DSC	DSC	DSC	DSC	HD	HD	HD	HD	HD95	HD95	HD95	HD95
	Aggr. method	Metric-based	Metric-based	Case-based	Case-based	Metric-based	Metric-based	Case-based	Case-based	Metric-based	Metric-based	Case-based	Case-based
	Aggr. operator	Mean	Median	Mean	Median	Mean	Median	Mean	Median	Mean	Median	Mean	Median
Rank	1	A1	A1	A1	A1	A5	A6	A5	A6	A6	A1	A6	A1
	2	A2	A3	A2	A2	A6	A5	A6	A5	A5	A9	A1	A6
	3	A3	A4	A6	A3	A4	A4	A4	A4	A3	A6	A5	A4
	4	A4	A2	A3	A6	A11	A11	A2	A2	A1	A3	A4	A5
	5	A5	A5	A4	A5	A2	A3	A7	A3	A2	A7	A3	A3
	6	A6	A6	A5	A7	A3	A2	A11	A11	A4	A4	A2	A2
	7	A7	A7	A7	A4	A10	A9	A3	A7	A11	A11	A9	A9
	8	A8	A8	A8	A8	A7	A7	A9	A9	A7	A8	A7	A7
	9	A9	A9	A11	A9	A9	A10	A10	A10	A9	A5	A11	A11
	10	A10	A11	A12	A10	A8	A1	A1	A8	A10	A2	A10	A8
	11	A11	A12	A10	A11	A1	A8v	A8	A1	A8	A10	A8	A10
	12	A12	A10	A9	A12	A12	A12	A12	A12	A12	A12	A12	A12
	13	A13	A13	A13	A13	A13	A13	A13	A13	A13	A13	A13	A13
Kendall's tau			0.90	0.79	0.88	0.28	0.28	0.38	0.33	0.56	0.44	0.59	0.65

Figure 4.1: Effect of exchanging parameters of the ranking designs for one example MICCAI 2015 segmentation task. Ranking scheme A (first column, bold box) is considered as default for the algorithms *A1* to *A13*. Ranking schemes B - L are varied by the metric $\in$ {DSC, HD, HD95}, aggregation (aggr.) method $\in$ {Metric-based, Case-based} and the aggregation operator $\in$ {Mean, Median}. Algorithms *A1* (green), *A6* (blue), and *A11* (red) are highlighted to accentuate the changes in ranks across ranking schemes. Kendall's τ for every ranking and the default is provided in the last column. Figure adapted from [Reinke et al., 2018a].

Increasing ranks via exclusive submission of test cases

82% of challenge tasks did not explain how they dealt with missing values (see Section 4.4.3) and only 33% of MICCAI segmentation tasks explain how to penalize missing data. The choice of missing value handling, however, has a crucial impact on the resulting aggregated metric scores and subsequently on the resulting ranks (see Figures 4.14(b) and (c) in Section 4.2.3). Of the MICCAI 2015 segmentation tasks, 27% of participants did not submit results for every test case. The average proportion of missing values was 20% (maximum: 80%). By simulating a selective test case submission, 25% of all 419 non-winning algorithms of the MICCAI 2015 segmentation tasks could have been ranked first by only submitting the most plausible test case results. Furthermore, every single participant could have been ranked first by selective test case submission in 9% of the tasks. The threshold for implausible test cases in the experiments was a *DSC* value of 0.5.

Still, even by considering withholding only the worst 5% of test cases, 11% of the non-winning algorithms could have been the winner.

4.1.4 Discussion

The presented results show that rankings are indeed susceptible to manipulation due to weaknesses in the challenge design. One of the factors enabling this is the lack of reporting of the full ranking scheme calculation. According to our experiments, challenge rankings appear to be sensitive to the concrete design, in particular the aggregation scheme and operator as well as the metric choice. Combined with the reporting problem for ranking design, we showed that cheating is, in theory, very simple from a challenge organizer perspective. This may happen on purpose to either avoid or ensure that a specific team wins. It may also happen accidentally when organizers find issues with their original ranking scheme.

Another weakness in challenge design is the fact that organizers sometimes require the submission of concrete algorithm outputs for their validation. In two-thirds of the tasks in the period 2004 to 2016, participants were required to directly submit their results to a cloud or send the results to the organizers via email (see Chapter 3). In this case, participants could easily submit only a subset of their outputs and may receive a better rank for selective test case submission. In fact, during our assessment (see Chapter 3), we found challenges for which participants only submitted a fraction of the test cases. For example, the ranking tables of the Ischemic Stroke Lesion Segmentation (ISLES) challenge 2015 contained a column that lists the number of cases that were submitted by each team. For the task of sub-acute ischemic stroke lesion segmentation, none of the teams submitted results for the whole data set [Maier et al., 2015]. This concept was also chosen for the Brain Tumor Image Segmentation (BRATS) 2015 challenge [BRATS2015, 2015]. While most of the participants submitted results for all test cases, the third-ranked algorithm only submitted 2 out of 110 images. It is therefore reasonable to assume that this practice may be exploited by challenge participants.

A limitation of our experimental setup may be the fact that we used a threshold for the DSC metric to remove test cases. However, we think that a result yielding such a low metric score could easily have been found by visual inspection of the algorithm outputs. Indeed, the current experimental setup avoids a subjective rating of "bad" images, and instead uses an objective and scalable, easily reproducible approach. This experiment was only performed for the DSC metric as thresholding would not have been easily possible for the HD/HD95 metrics. Both metrics are unbounded to the top and a bad score highly depends on the image and task. Nevertheless, we think the experimental setup and results are sufficient to highlight our point: It is, in theory, possible to easily increase a rank through selective test case submission in the case of ignoring missing values. This issue can be aggravated by the fact that, where challenge participants are only required to upload their outputs, organizers can not even check whether participants may have manipulated their results manually.

Notably, by requesting the concrete confusion matrices from the challenge organizers, we were able to detect errors in their metric implementations that were not identified by either organizers or participants. One could argue that we should trust our research colleagues and that they organize and participate fairly. However, given the tremendous pressure to perform in various research environments, we believe that weaknesses in challenge design or security flaws will

eventually be exploited [Ioannidis, 2005; Chui et al., 2021], either on purpose or unwittingly. Indeed, in a survey of the organizers of MICCAI 2020 and 2021 challenges, 21% of them reported the occurrence of cheating.

4.1.5 Conclusion

In this section, we presented evidence that challenge rankings are sensitive to the type of ranking calculation. In addition, we found that many challenges do not accurately report the ranking schemes. As both are very critical issues, we investigate them in depth in Sections 4.3 and 4.4. Their combination may lead to severe weaknesses or even security flaws in the challenge design, which can – in theory – be easily exploited to tune challenge rankings. Although we assume the scientific integrity of challenge organizers and participants, we think that specific circumstances and performance pressure in the scientific community may lead researchers to use unfair practices. Weaknesses in challenge design may even encourage them to do so. This is a very damaging practice that may negatively impact healthcare research in a severe manner, if for example poorly performing algorithms become the new state of the art or are even translated into clinical practice only due to cheating.

4.2 Revealing flaws related to metrics

Disclosure

This work has been supervised by Lena Maier-Hein. It has been conducted together with several co-authors, especially with Minu D. Tizabi and Paul F. Jäger. The work has been is available in a *dynamic preprint* [Reinke et al., 2021a] and was published at the *Medical Imaging with Deep Learning conference (MIDL) conference* [Reinke et al., 2021b] (short paper). Please refer to Chapter A.1 for full disclosure.

4.2.1 Introduction

One of the most important challenge design concerns, according to the research community, is the selection of the validation metrics (see Chapter 3). However, reviewing current research shows that most of the scientific work in the general image analysis community focuses on the development of new algorithms rather than their validation [Lin et al., 2020]. Certainly, it is crucial for scientific advancement and the application of methodological research that Machine Learning (ML) algorithms are properly validated.

The efficacy of new medication or new treatments can be easily assessed via the patient outcome. In contrast, the validation of algorithms is more complex and is done via performance metrics, which form an integral part of the validation pipeline. They inform the researchers about how good or bad the algorithms performed on specific data points or the whole data set and serve as indicators for scientific advancement. They are therefore anticipated to reflect a validation objective that is specific to the research question. Unfortunately, previous work has shown that metrics often do not reflect the underlying research interest [Saha et al., 2021]. These findings are very critical since the metric scores serve as the foundation for determining whether an algorithm is clinically appropriate, and as a result, have significant consequences for patients.

Metrics that are not chosen properly therefore not reflect the actual clinical or biological needs that an algorithm must fulfill. Indeed, an entirely subpar algorithm may win a challenge if the metrics are not appropriate. It is therefore crucial that researchers choose adequate metrics and are aware of their specific properties, advantages, and limitations. However, it is not straightforward to be aware of all the restrictions associated with a particular validation metric. Weaknesses of metrics are discussed in several publications (e.g. [Gooding et al., 2018; Vaassen et al., 2020; Cheng et al., 2021]), but there is no common entry point that comprehensively lists and describes metric-related pitfalls.

In this section, we present an extensive review of metric limitations in the context of the most common problem categories in challenges, namely segmentation, image-level classification, and object detection. We found that a visual presentation helps to better understand the actual problem. To illustrate the broad issues of category-metric mismatches, inherent metric properties, and the post-processing of metric results, we demonstrate the metric pitfalls in an illustrative manner.

Hypothesis investigated in this chapter

H2: Common image analysis metrics do not reflect the biomedical domain interest.

Previous work has found that metrics often do not reflect clinical interest. Furthermore, it is very hard to find comprehensive information on metric limitations. As choosing the wrong metric may also result in inadequate challenge winners, we investigate (1) whether the limitations of metrics can be easily assessed and (2) what potential pitfalls are regarding metrics.

4.2.2 Methods

First, we describe our experiments that simulate literature research from the perspective of a researcher trying to find an overview of metric limitations. Second, we describe a multi-stage Delphi process used as the basis for a comprehensive overview of metric-related pitfalls.

Literature research

In a first attempt to review the literature for metric limitations, we performed a structured and exploratory literature search. For every included metric, we collected possible synonyms and acronyms. All of them were combined with a logical disjunction (OR). In addition, we used the following words, also combined with a logical OR, to describe the limitations: pitfall, limitation, caveat, drawback, shortcoming, weakness, flaw, disadvantage, and suffer. The metric names, limitation terms, and the term "metric" were combined with logical conjunction (AND). Google Scholar was used as a search engine.

In a follow-up experiment, we simulated the situation that the rough problem or limitation regarding a certain metric is known to the researcher, who wishes to find an explicit publication or online resource (such as blog posts) detailing the specific limitation. This search was based on the results of the Delphi process (see the following paragraph). Own publications or online resources were excluded from this search.

Delphi process

We collected pitfalls regarding metrics in the form of a multi-stage survey approach, inspired by the concept of a Delphi process [Brown, 1968]. A Delphi process aims to combine the expertise of several experts and reaches a consensus agreement through a series of questionnaires. The expert panel was composed of international researchers with expertise in the field of biomedical image analysis validation. First, we gathered consensus on a selection of properties that affect the choice of the metrics (conducted by 30 experts). In the second step, we asked the experts to describe pitfalls that may occur when this property is present (26 experts). During this survey round, we specifically reached out to the experts with biological and clinical expertise, asking for general problems in validation from their points of view (9/26 experts). The pitfalls collected through this analysis were complemented by literature research and feedback from the scientific community (provided by eleven additional researchers that were no members of the consortium). Based on the expert feedback and suggestions, examples for proposed pitfalls were created and

illustrations were designed. We asked the following questions for every pitfall illustration in the following round of the Delphi process:

- Please indicate whether you consider the illustration to represent a relevant pitfall and *(if so)* how practically relevant you consider this pitfall.

- Please provide comments or suggestions for improvement *(if any)*.

In addition, we asked for general feedback for every problem category and whether we missed important pitfalls. 26 experts replied to this survey. All metric scores used in the examples and illustrations have been verified by a second observer. In a final survey, we asked the consortium about consensus on the pitfall taxonomy and the presented pitfalls. 52 experts voted in this round and the taxonomy and pitfalls reached a consensus agreement.

4.2.3 Results

In the following, we first present the results of the literature research. It is followed by the illustration of the concrete pitfalls that were found during the analysis and the Delphi process. The pitfalls were categorized into the following categories: Four pitfalls related to problem-category mismatch, 31 pitfalls related to metric selection, and 27 pitfalls related to the metric application.

In the pitfall illustrations, we highlight the most relevant metric values in order to maintain visual clarity. A "good" metric value is represented by a green number, whereas a "poor" value is represented by a red number. Red crosses denote poor metric behavior, whereas green checkmarks indicate desired behavior. Finally, for simplicity, we typically refrain from providing precisely the predicted class scores for each sample.

Literature research

The literature research yielded an extremely high amount of publications found for most of the metrics. The highest number of publications was found for the Sensitivity metric, for which roughly 962,000 publications were returned by the search engine, followed by Accuracy with 895,000 hits. The lowest number of publications was given for the Centerline Dice Similarity Coefficient (clDice) with 49 results. From the pitfalls presented in the following paragraphs, 51% were identified to be presented in the literature or in online resources (24%). 13% were present in both, publications and online resources. Of all findings, only 30% presented the pitfalls in an easily understandable visual way.

Context-agnostic pitfalls taxonomy

We defined a context-agnostic taxonomy for metric pitfalls in order to present them in a comprehensive and structured manner. An overview of the proposed taxonomy is given in Figure 4.2. Based on our definition, we distinguish three major types of metric pitfalls:

Pitfall Type 1: Pitfalls related to incorrect problem categorization

Pitfall Type 2: Pitfalls related to poor metric selection

Pitfall Type 3: Pitfalls related to poor metric application

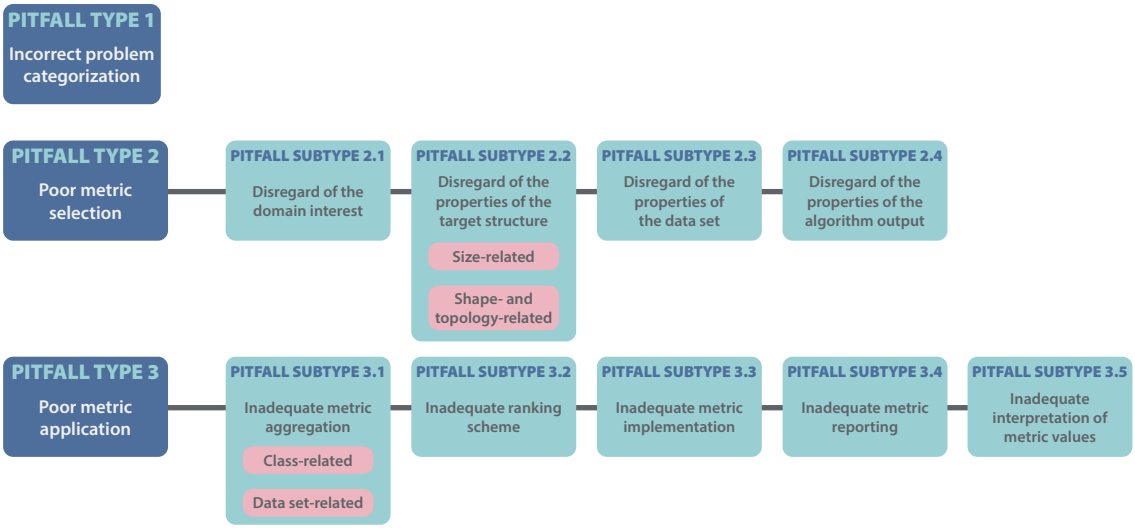


Figure 4.2: **Metric pitfalls taxonomy.** Pitfalls can be distinguished into three major types: Pitfall Type 1 – pitfalls related to incorrect problem categorization, pitfall Type 2 – pitfalls related to poor metric selection, pitfall Type 3 – pitfalls related to poor metric application. Types 2 and 3 can be further separated into subtypes.

Pitfall Type 1: Pitfalls related to incorrect problem categorization

Validation metrics typically assess a specific property for a specific problem category. While some metrics may be used at different classification scales, for example at pixel or object level, care should be taken when directly transferring a metric to a different problem category. We divide pitfalls of type 1 into those related to a *wrong choice of the problem category* and those related to a *lack of a missing problem category*. In the following, we describe exemplary pitfalls of both types that are presented in Figure 4.3:

Wrong choice of the problem category: Segmentation metrics used for object detection. In Section 2.4.3, the F_β Score was listed as a metric validating object detection problems at object level. However, the Dice Similarity Coefficient (DSC) (which is equivalent to the F_β Score, $\beta = 1$) or other pixel-level segmentation measures are often used to validate object detection problems at pixel level [Carass et al., 2020; Jäger, 2020]. Figure 4.3(a) provides an example of a reference with three objects. While *Prediction 2* successfully detects all objects, *Prediction 1* only hits the large object and misses the two tiny structures. The DSC, however, would rate *Prediction 1* as superior because this prediction has a larger overlap in terms of counted pixels. This would be a very critical pitfall for clinical applications. For instance, several missed polyps during colonoscopy may negatively impact a patient's survival. Because the pixel-level DSC generally has a strong bias against single objects, it is inappropriate for the detection of multiple structures [Yeghiazaryan and Voiculescu, 2018; Kirillov et al., 2019].

Wrong choice of the problem category: Image-level metrics used for object-level problems. The Receiver Operating Characteristic (ROC) curve and its respective multi-threshold metric, the Area under the Receiver Operating Characteristic Curve (AUROC), have been designed for image-level assessment

of problems. They rely on the computation of the Specificity, thus, on True Negatives (TNs) of the confusion matrix, which are typically undefined in object detection problems. Still, ROC and AUROC are sometimes employed for object detection problems. In these settings, typically, only one object in an image with the highest predicted class score is considered, neglecting all others. This practice may lead to severe pitfalls, as illustrated in Figure 4.3(b): The ROC curve was designed for image-level classification problems, in which no additional localization step is needed. Thus, it neglects the localization performance when applied to object detection problems. Thus, objects may be at the completely wrong location and still be interpreted as True Positive (TP) (top part of the figure). In addition, the ROC curve neglects the number of reference objects. A single detected object is automatically considered a TP for this image, although other objects may have been missed (middle part of the figure). Finally, as already mentioned, the curve is similarly unable to distinguish between a model that detects all objects in an image and a model that only detects one object and misses several others (bottom part of the figure). In clinical practice, such a misuse of a metric could yield very misleading results. Tumors could be localized at an entirely incorrect position or completely missed.

Wrong choice of the problem category: Semantic segmentation reference for instance segmentation objective. A prediction can only be as accurate as the provided reference. Special care therefore needs to be taken when preparing the reference annotation to make sure that it reflects the underlying research question. For example, in the case of many touching or overlapping structures, often present for cell or surgical instrument segmentation problems, a semantic segmentation reference may cause multiple cells or instruments to be merged into one object although a distinction would be desirable. Thus, an algorithm predicting a semantic segmentation reference may yield perfect metric scores although the desired output was not achieved, as shown in the example in Figure 4.3(c), which could lead to incorrect conclusions in clinical practice.

Lack of a missing problem category: Metrics not designed for specific property-related applications. Validation metrics should be chosen to reflect a specific biomedical domain interest. However, depending on the actual application, use cases may occur that are not assessed and reflected by standard technical metrics, such as the DSC. For instance, in [Bamira and Picard, 2018], the ejection fraction in a cardiac cycle may be the property of interest, i.e. how accurately the ratio between blood volumes match. Such an example is given in Figure 4.3(d). Despite the substantial differences in the volumes predicted, both predictions yield DSC scores with a similar average. Thus, typical segmentation measures do not capture the property of interest in such use cases.

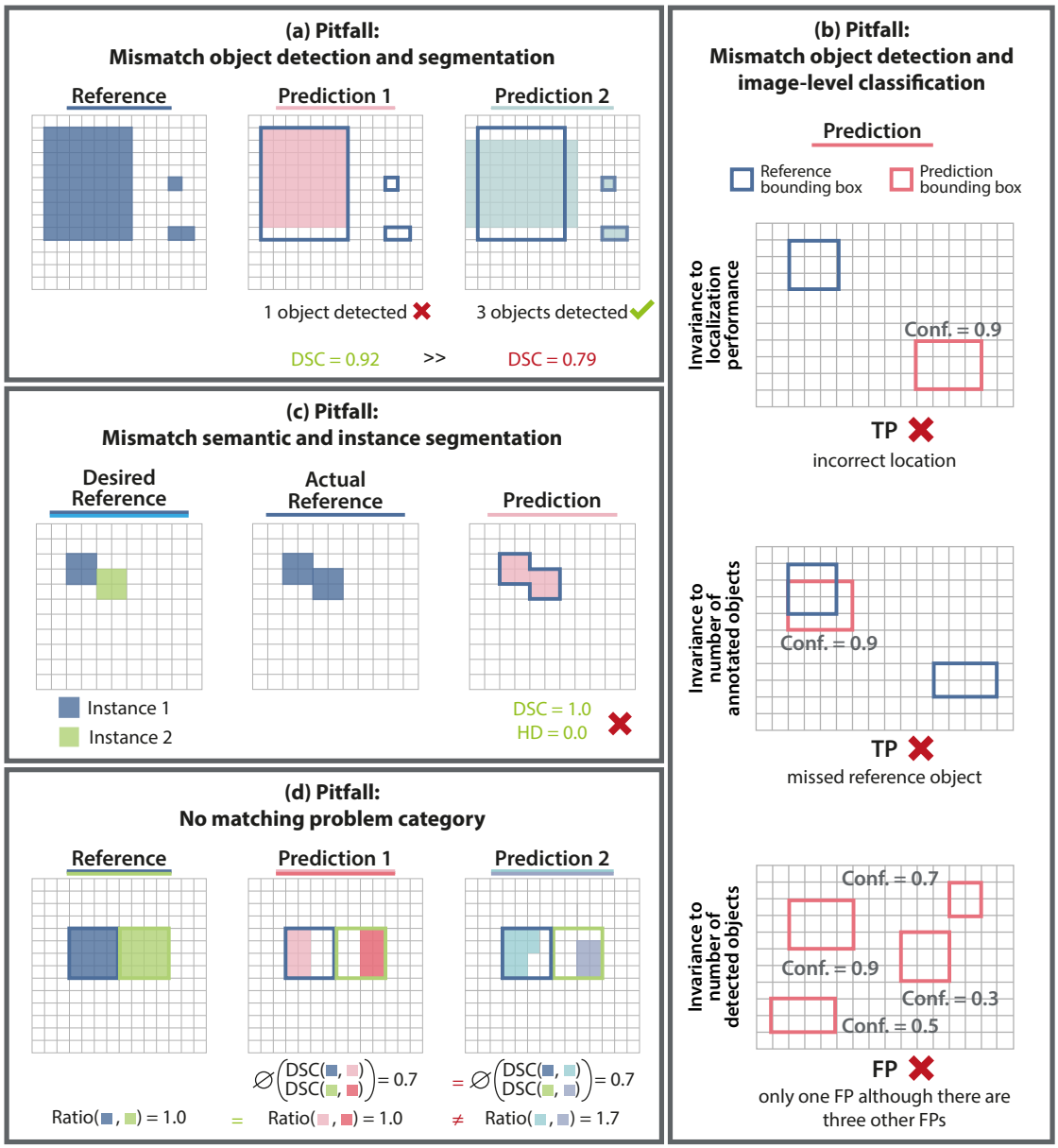


Figure 4.3: Pitfalls related to a mismatch between the problem category and the metric. **(a) Effect of a mismatch of semantic segmentation and object detection.** A prediction that only recognizes one of three structures (*Prediction 1*) yields a higher Dice Similarity Coefficient (DSC) than a prediction made by an algorithm that recognizes every structure (*Prediction 2*). **(b) Effect of a mismatch of image-level classification and object detection.** The Receiver Operating Characteristic (ROC) curve does not assess the localization performance of a prediction, which may lead to predictions at the wrong position being categorized as True Positive (TP). It is further invariant to the number of reference and detected objects in an image. Predicted class scores are indicated by confidence scores (conf.). **(c) Effect of a mismatch of semantic and instance segmentation.** If instance information is not available (although desired), connected components are often treated as one instance. Even if the task is not solved, the metrics may yield perfect results, here shown for the DSC and Hausdorff Distance (HD). **(d) Metrics may be poor proxies for computing properties of interest.** The property of interest is the accuracy of the ratio between two volumes. Although both predictions lead to a distinct ratio of structure volumes – the parameter of interest – they produce similar averaged DSC metric values, which do not reflect the actual interest. Figure adapted from [Reinke et al., 2021a].

Pitfall Type 2: Pitfalls related to poor metric selection

Pitfalls of type 2 relate to all limitations resulting from a poor selection of metrics, such as ignoring inherent properties of a specific metric. We further distinguish type 2 into four subtypes: Disregard of the domain interest (2.1), of properties of the target structure (2.2), data set (2.3), and algorithm output (2.4).

Pitfall Subtype 2.1: Disregard of the domain interest

Generally, metrics should reflect the domain interest of the practitioner. Since every metric was designed for a specific purpose, not every metric assesses the correct properties. The actual interest should not be confused with other inherent properties of a use case. This is why we distinguish subtype 2.1 particularly from the other subtypes such as disregarding properties of the data set. In the following, we present pitfalls that relate to a disregard of the domain interest.

Importance of structure boundaries Whenever the concrete boundary of a structure is the domain interest, the metrics should be chosen accordingly. However, common overlap-based metrics such as the DSC or volume-based metrics, such as the volumetric difference, do not consider object boundaries in their calculation. Example 1 in Figure 4.4(a) shows two predictions, of which one perfectly matches the outline, while the other misses it. However, both yield the same DSC value, since overlap-based metrics do not penalize incorrect predictions of boundaries. On the other hand, both predictions from Example 2 completely miss the object and are at the incorrect position, however, *Prediction 2* should be penalized much more strongly since the distance of the boundary is higher. While the volumetric difference is correct for both predictions, the predictions are not overlapping at all, yielding a DSC score of zero as well. Nonetheless, the DSC does not measure the extent of mislocalization and distance to the object boundaries.

Importance of structure center(line) Some applications, such as nerve segmentation [Mlynarski et al., 2020], require a specific focus on the center of a structure. The exact outline or overlap with the reference is less important in those cases. Overlap-based metrics are typically not good proxies for assessing the center point or line of a structure, as indicated in Figure 4.4(b). While both predictions have the same amount of overlapping pixels, yielding the same DSC score, the center point prediction is worse for *Prediction 2*. However, this is not reflected when using the DSC as the only metric. This pitfall is also relevant for object detection problems.

Importance of structure volume Radiologists are often solely interested in structure volumes rather than technical measures such as overlap-based metrics. Boundary-based metrics do not focus on the object volumes themselves, but rather on inspecting the object boundaries. The Boundary Intersection over Union (Boundary IoU), for example, only considers the pixels that are inside the defined distance from the boundary (see Section 2.4.2). Thus, it does not recognize whether a prediction shows a large hole inside the object (incorrect volume), as presented in Figure 4.4(c) and still yields a perfect score [Cheng et al., 2021] that may be misleading if the volume is of interest. A similar example is provided in Figure 4.4(d). Moreover, predictions with a spotted pattern of small holes are similarly not heavily penalized, although the volume and overlap are rather low [Taha and Hanbury, 2015]. This pitfall may be critical in clinical practice, especially if a particular interest lies in the object determining the exact volumes.

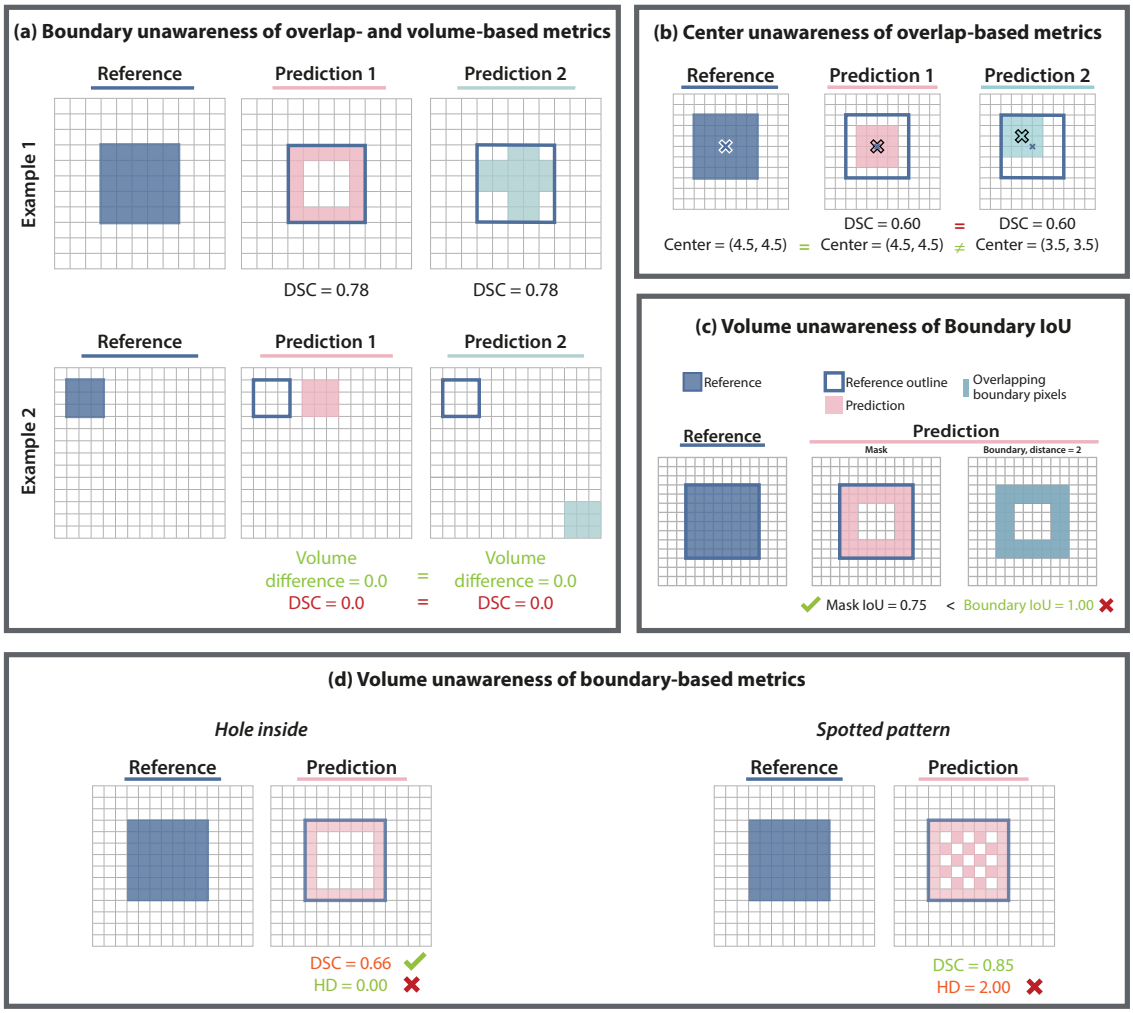


Figure 4.4: Pitfalls related to disregard of the domain interest: Importance of structure boundaries, center, and volume. **(a) Boundary unawareness of overlap- and volume-based metrics.** Common metrics that rely on the overlap or volume of structures do not consider the actual boundaries. In Example 1, *Prediction 1* perfectly hits the structure boundaries, while *Prediction 2* only hits parts of the boundaries. The Dice Similarity Coefficient (DSC) value, however, would be similar, not penalizing errors in the boundaries. In Example 2, neither the volumetric difference nor the DSC penalize the distance from the boundaries. **(b) Center unawareness of overlap-based metrics.** The most common overlap-based metrics are subpar substitutes for center point alignment. Although *Prediction 1* more closely approximates the object's center, *Prediction 2* nevertheless produces the same DSC value. **(c) Volume unawareness of Boundary Intersection over Union (Boundary IoU).** If the distance to the border contains all mask pixels (in this case, distance = 2), the Boundary IoU may yield a perfect score for a prediction with a large hole inside of the object. **(d) Volume unawareness of boundary-based metrics** Similarly, other boundary-based metrics (here: Hausdorff Distance (HD)) do not recognize holes within the objects. A spotted pattern of holes (left) is also not heavily penalized. Figure adapted from [Reinke et al., 2021a].

Unequal severity of class confusions In biomedical image analysis problems, classes and class confusions may be of unequal importance. For example, in colon polyp detection in the gastrointestinal tract, it is crucial for the survival of the patient to detect all polyps. In this use case, False Negatives (FNs), i.e. missing polyps, would be extremely critical. The Positive Predictive Value (PPV) yields misleading results for such a use case, since this metric does not account for FNs (see the top part of Figure 4.5(a)). On the other hand, in image retrieval tasks, the property of interest is not finding all images with certain characteristics, but the correct identification of all collected images. Thus, False Positives (FPs) should be particularly penalized. In this case, Sensitivity is not an adequate metric, since it is not concerned with FPs (see the bottom part of Figure 4.5(a)). This pitfall also translates to Specificity and Negative Predictive Value (NPV).

Over- vs. undersegmentation In radiotherapy, for example, it may be crucial to determine whether an algorithm systematically predicts larger (oversegmentation; more FPs) or smaller objects (undersegmentation; more FNs) compared to the reference. Overlap-based metrics such as the DSC treat both situations unequal [Yeghiazaryan and Voiculescu, 2018]. Figure 4.5(b) shows an example in which *Prediction 1* and *Prediction 2* both have a difference of a single layer of pixels compared to the reference, with *Prediction 1* presenting an undersegmentation and *Prediction 2* presenting an oversegmentation of the target structure. However, the overlap-based metrics yield substantially different scores for both predictions [Taha and Hanbury, 2015]. On the other hand, boundary-based metrics treat over- and undersegmentation similarly.

Ordinal classification (counting metrics) An unequal severity of class confusions is especially common in ordinal classification problems. In the case of ordinal classes, metrics should be carefully interpreted. Figure 4.5(c) presents two examples with a five-class severity scale, with zero referring to the lowest severity and four to the highest severity. This could represent the severity of a disease as determined in diabetic retinopathy grading, for example [Zachariah et al., 2015]. Both predictions correctly classify the disease severity for the first four patients. However, they misclassify the severity of *Patient 5*. While the predicted class from *Prediction 2* is close to the actual class 4, *Prediction 1* vastly underrates the disease severity for this patient. This issue is not recognized by the Accuracy and other multi-class metrics. Only the quadratic-weighted Cohen's Kappa (CK) detects the problem and heavily penalizes *Prediction 1*.

Ordinal classification (counting metrics) Similarly, calibration metrics that simultaneously assess the discrimination performance of a classifier should be used with caution in ordinal classification. Figure 4.5(d) provides an example of three levels, which may be interpreted as a very short survival time of zero to one year (pink circle), a longer survival time of two to three years (blue triangle), and a long survival time of more than three years (green square). Both predictions in this example fail to predict the right survival time. However, *Prediction 1* is closer to the actual survival time and should thus be penalized less than *Prediction 2*. However, the Brier Score (BS) does not recognize this difference, penalizing both predictions similarly. By only examining the score, this important information would be hidden from the practitioner and patient, although the expected survival time could make a tremendous difference.

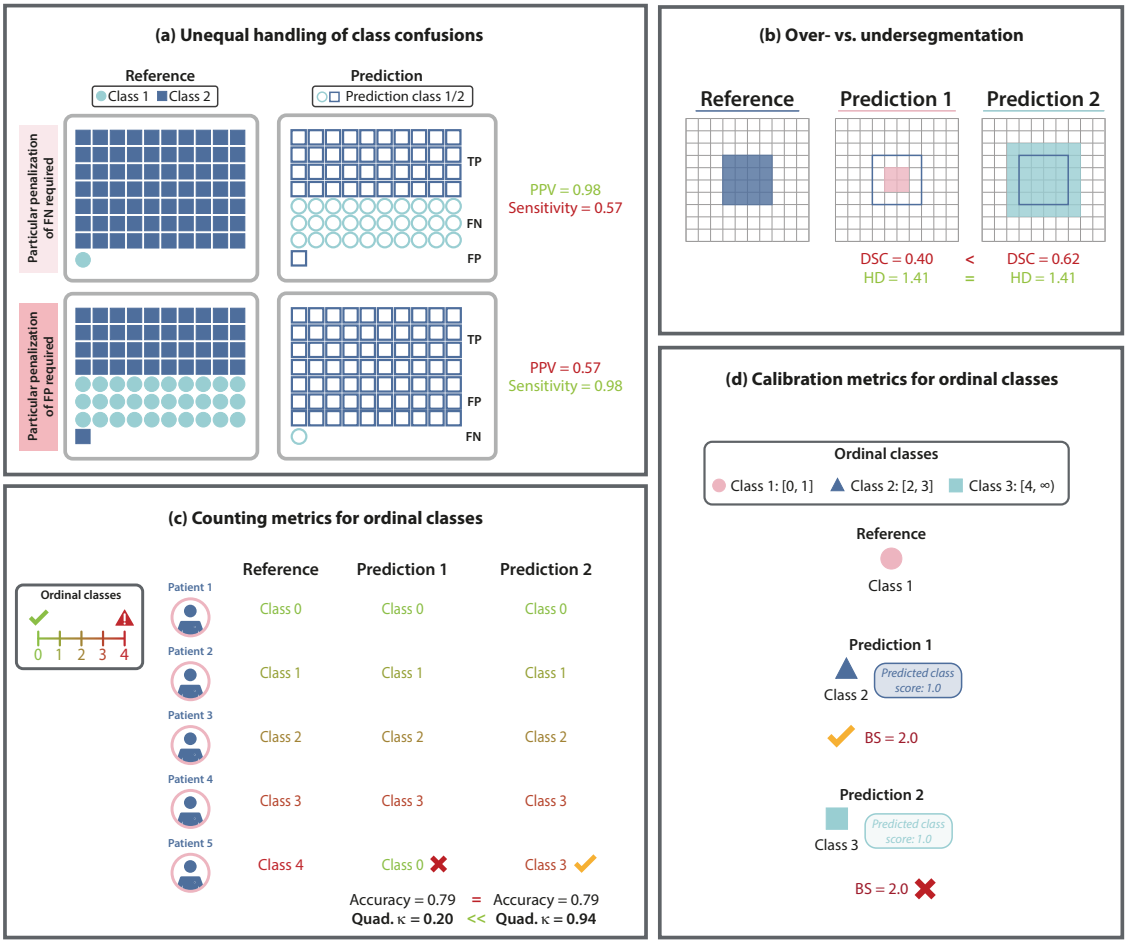


Figure 4.5: Pitfalls related to disregard of the domain interest: Unequal severity of class confusions. **(a) Effects of utilizing criteria that are inappropriate for False Negative (FN)/False Positive (FP) punishment.** The Positive Predictive Value (PPV) does not assess the number of FN and is therefore inappropriate to punish FN from a prediction. Similarly, Sensitivity does not penalize FP properly. **(b) Effect of over- and undersegmentation.** One layer of pixels separates the predictions of two algorithms compared to the reference. While the Hausdorff Distance (HD) treats the under- and oversegmentation the same, it yields a substantial difference in the Dice Similarity Coefficient (DSC). **(c) Effect of ordinal classes on counting metrics.** The Accuracy cannot distinguish between a prediction that is close to the actual scale (*Prediction 2*) and a prediction far away from the actual scale (*Prediction 1*). This difference is penalized by the quadratic-weighted CK (Quad. κ). **(d) Effect of ordinal classes on calibration metrics.** The Brier Score (BS) cannot distinguish between a prediction that is closer to the actual scale (*Prediction 1*) and a prediction far away from the actual scale (*Prediction 2*). Figure adapted from [Reinke et al., 2021a].

Importance of comparability across data sets - prevalence correction As previously discussed in Section 2.4.1, many metrics rely on the prevalence. Although PPV and NPV should be corrected for the respective prevalence, this is often not done in practice. In many cases, it is assumed that the prevalence of a study group is similar to the prevalence of the general population. However, due to the study design or clinician observations, case-control groups are frequently strongly biased, implying that this assumption is frequently untrue and resulting in higher study group prevalences than are observed in the general population. Such an example is provided in Figure 4.6(a), in which the assumed prevalence equals 50%, whereas the actual prevalence of the general population is only 0.5%. Without using the formulas for prevalence correction (see Equations 2.19 and 2.21), this practice can lead to misleading results or patient confusion. For example, in Covid-19 prediction models, the data sets are often created with a prevalence of roughly 50% (e.g. [Ko et al., 2020]). However, the general prevalence is usually much lower [Iacobucci, 2020].

Importance of comparability across data sets - prevalence dependency Whenever using prevalence-dependent metrics, one should be careful with comparisons across data sets. Figure 4.6(b) presents the results for two data sets with different prevalence values but similar Sensitivity and Specificity. Metrics that can be computed independently from the prevalence yield the same metric scores for both data sets. Prevalence-dependent metrics, however, produce different results and are thus not comparable across data sets. Appendix A.3 indicates which metrics depend on the prevalence.

Importance of comparability across data sets - rankings Metrics like the Balanced Accuracy (BA) may lead to different rankings of predictions compared to the Matthews Correlation Coefficient (MCC) for data sets with a prevalence dissimilar to 50% [Chicco et al., 2021]. Figure 4.6(c) shows two settings of prevalences (40% and 50%) as well as two data sets for each situation. In the case of a 50% prevalence, the rankings of all four metrics are the same, favoring the right prediction over the left one. For different prevalence values, MCC may lead to different rankings, preferring the left prediction. Thus, interpreting metric scores across data sets needs to be done with caution.

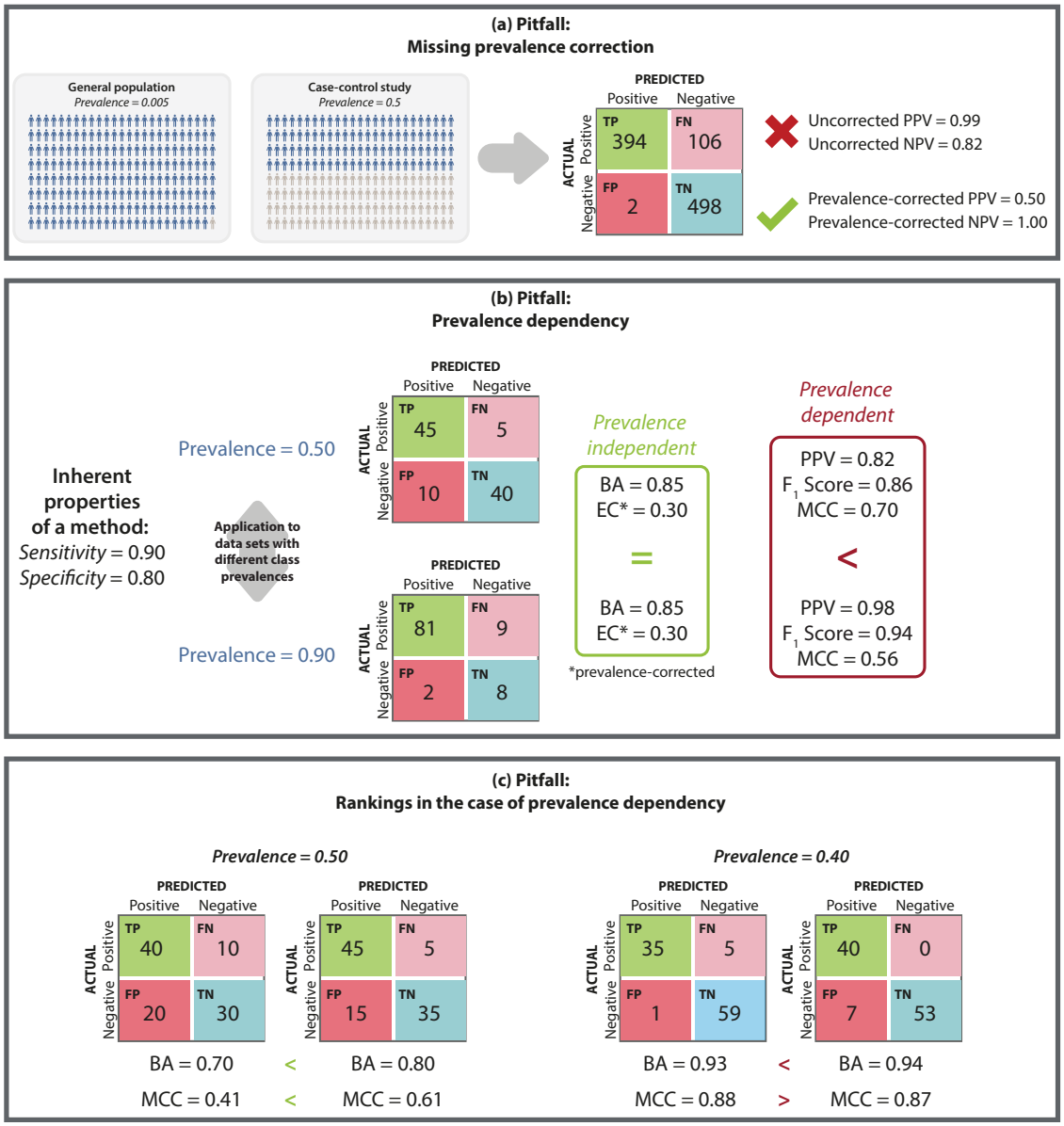


Figure 4.6: Pitfalls related to disregard of the domain interest: Importance of comparability across data sets. **(a) Effect of missing prevalence correction.** Prevalence correction based on the general population must be applied for case-control studies for the Positive Predictive Value (PPV) and Negative Predictive Value (NPV) metrics. Using the case-control prevalence incorrectly produces inaccurate metric scores as opposed to using the correct prevalence from the general population (which is often significantly lower compared to a study group). **(b) Effect of prevalence-dependent metrics.** Prevalence-dependent metrics (here: Positive Predictive Value (PPV), F_1 Score, Matthews Correlation Coefficient (MCC)) should not be used for data sets with different prevalence levels. Only prevalence-independent metrics (here: Balanced Accuracy (BA), Expected Cost (EC)) yield the same scores in this setting with similar Sensitivity and Specificity scores. **(c) Effect of rankings in the case of prevalence dependency.** The rankings produced by the Balanced Accuracy (BA) (favoring the right predictions) are different from those produced by Matthews Correlation Coefficient (MCC) (favoring the left predictions) for a prevalence different from 0.5. Only a prevalence of 0.5 maintains rankings (favoring the right prediction). Further abbreviations: True Positive (TP), False Negative (FN), False Positive (FP) and True Negative (TN). Figure adapted from [Reinke et al., 2021a].

Importance of confidence values Perfect discrimination does not guarantee that a model is also well-calibrated. Figure 4.7(a) shows an example that discriminates perfectly between the triangle and circle class (AUROC score of 1.0). The predicted class scores, however, are not representative of an empirical success rate, with scores of 0.51 for the circles and 0.52 for the triangles. This model is, unfortunately, not well-calibrated. In diagnostics, one should therefore be careful when interpreting the model outputs. Only the class scores of calibrated models can be interpreted as the probability of an event, such as a disease or the risk of post-surgery complications. This pitfall is also relevant for object detection and segmentation problems, which make use of predicted class scores.

Generally, the definition of calibration may influence the chosen calibration errors. As shown by Gruber and Buettner [2022] and Vaicenavicius et al. [2019], the Expected Calibration Error (ECE) can appear to indicate a perfect calibration for top-level and class-wise calibration, although the predicted class scores are not fully calibrated. This is illustrated in Figure 4.7(d), in which only the canonical definition of the ECE indicates imperfect calibration.

While providing predicted class scores per pixel is less common for segmentation problems, the use of fuzzy segmentation outputs is becoming more widespread [Nida et al., 2019; AlZu'bi et al., 2020]. Thus, a decision threshold is important to be defined for those use cases. Based on the threshold, the resulting segmentation masks vary and substantially influence the metric scores [Nair, 2018]. This is illustrated in Figure 4.7(b), in which results are shown for three different thresholds, which yield a difference in DSC scores of 0.44.

Importance of benefit-cost analysis Common validation metrics do not take into account the potential analysis of benefits of TP predictions and the harms of FP predictions. This is reflected by the Net Benefit (NB), which comes with an additional exchange rate parameter to incorporate individualized costs of FP harm. An example is illustrated in Figure 4.7(c), based on [Vickers et al., 2016]. Here, the specific exchange rate is defined by accepting about nine unnecessary biopsies for detecting one lesion (1/9). This is incorporated into the NB calculation as a weight of the fraction of FPs. When comparing two situations (directly applying the biopsy to all patients versus only having biopsies for patients with a specific marker), common metrics like Accuracy are in favor of the marker-based decision. Incorporating the exchange rate and benefit-harm analysis, however, favors the biopsy of all patients, indicated by a higher NB.

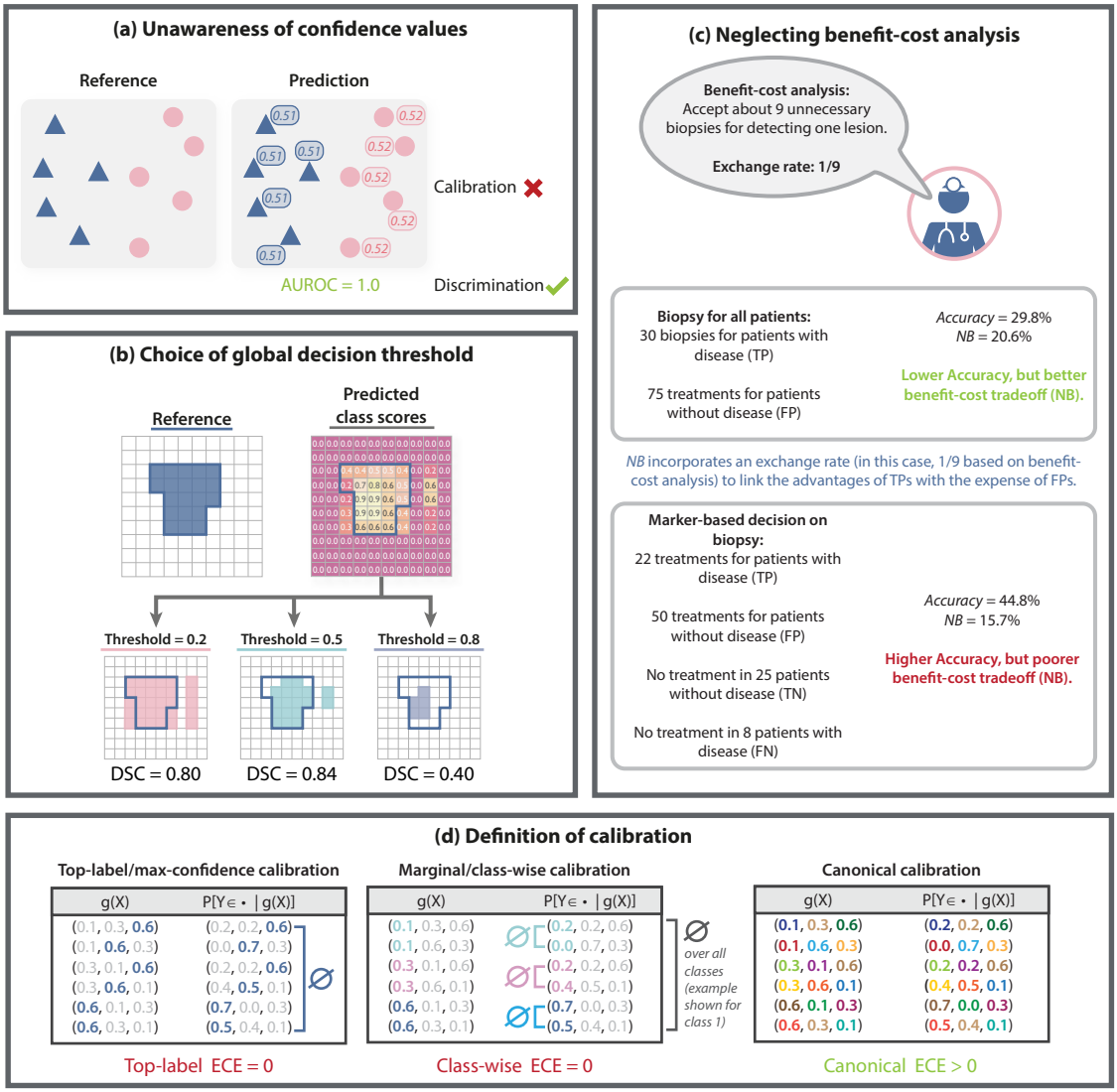


Figure 4.7: Pitfalls related to disregard of the domain interest: Importance of confidence values and benefit-cost analysis. **(a) Effect of perfect discrimination but poor calibration.** The example discriminates perfectly between the triangle and circle classes, yielding an Area under the Receiver Operating Characteristic Curve (AUROC) score of 1.0. However, there is very little association between the predicted class scores and the actual probability, therefore the predicted class scores are not well calibrated. **(b) Effect of choosing a global decision threshold**. The resulting segmentation masks for three decision thresholds (0.2, 0.5, and 0.8) are provided. The metric scores (here: Dice Similarity Coefficient (DSC)) are severely affected by the choice of thresholds. **(c) Effect of neglecting benefit-cost analysis.** Accuracy and similar metrics do not consider benefit-cost analyses, here resulting in favoring the marker-based biopsy. Incorporating the clinically defined exchange rate yields a higher Net Benefit (NB), favoring the biopsy of all patients in contrast to Accuracy. **(d) Definition of calibration.** Based on the used definition of calibration, the values of the Expected Calibration Error (ECE) change. For the top-label and class-wise calibration, the metric appears to indicate a perfect calibration, although the full probability vectors are not perfectly calibrated. This is only revealed by calculating the canonical calibration errors. Further abbreviations: True Positive (TP), False Negative (FN), False Positive (FP), True Negative (TN). Figure adapted from [Reinke et al., 2021a].

Pitfall Subtype 2.2: Disregard of the properties of the target structure

In image analysis, the appearance of the targeted structures may severely impact the metric values for improper metrics. We distinguish between size- and shape-related and topology-related pitfalls of this subtype. In the following, we present pitfalls that relate to a disregard of the properties of the target structure.

Small size of structures (size-related) In several biomedical use cases, we are faced with small structures of interest, such as stem cells [Maška et al., 2014] or brain lesions [Menze et al., 2014]. Metrics like the DSC and Intersection over Union (IoU) should be used carefully in these cases [Cheng et al., 2021], as illustrated in Figure 4.8(a). A difference of one or two pixels only slightly impacts the DSC (and IoU) value for a large structure, yielding a decrease of 0.01 and 0.03 in the presented example (top row). Yet, the same setting for small structures substantially varies the metric values, yielding a decrease of 0.20 and 0.50 in the presented example (bottom row). Those differences are even more relevant if one faces a high variability of structure sizes. In such a situation, one should report the results by stratification of sizes. It is often undesirable that few pixels influence the metrics that much since the correct boundaries of structures are unknown in most of the cases and reference annotations may be subject to high inter-observer variability [Joskowicz et al., 2019]. The same problems arise for F_β Score. Similarly, they affect the clDice when being applied to tiny tubular structures and are also relevant for localization criteria in object detection and instance segmentation problems. For boundary-based metrics, the effects are substantially less.

High variability of structure sizes (size-related) The Average Symmetric Surface Distance (ASSD) and Mean Absolute Surface Distance (MASD) both calculate the average of minimal distances between two boundaries (see Section 2.4.2). However, MASD greatly favors circumstances in which the predicted structure is much smaller than the reference object and is located close to the reference boundary, illustrated in Figure 4.8(b). MASD takes into account the average over minimal distances (1) from the prediction boundary to the reference boundary and (2) vice versa. In (1), the average would be approximately zero since the prediction boundary is closely related to the reference boundary. Thus, the MASD score is only determined by (2), i.e. the minimal distances from the reference to the prediction. As opposed to ASSD, this average distance is halved in the case of the MASD, which results in a substantial benefit compared to ASSD.

When overlap-based metrics are used as a localization criterion, the pitfalls regarding overlap-based metrics should be noted as well. For example, similarly to the DSC, the Mask/Box/Approx IoU treats large and small objects differently and is less sensitive to boundary errors of large objects, as shown in Figure 4.8(c), in which the Mask IoU is substantially less for the small structure. This effect is less problematic for the Boundary IoU. In particular, Boundary IoU penalizes errors in the boundaries more. However, it should be noted that the score heavily depends on the chosen distance parameter. If a very high distance is chosen, Boundary IoU score does converge towards the Mask IoU score.

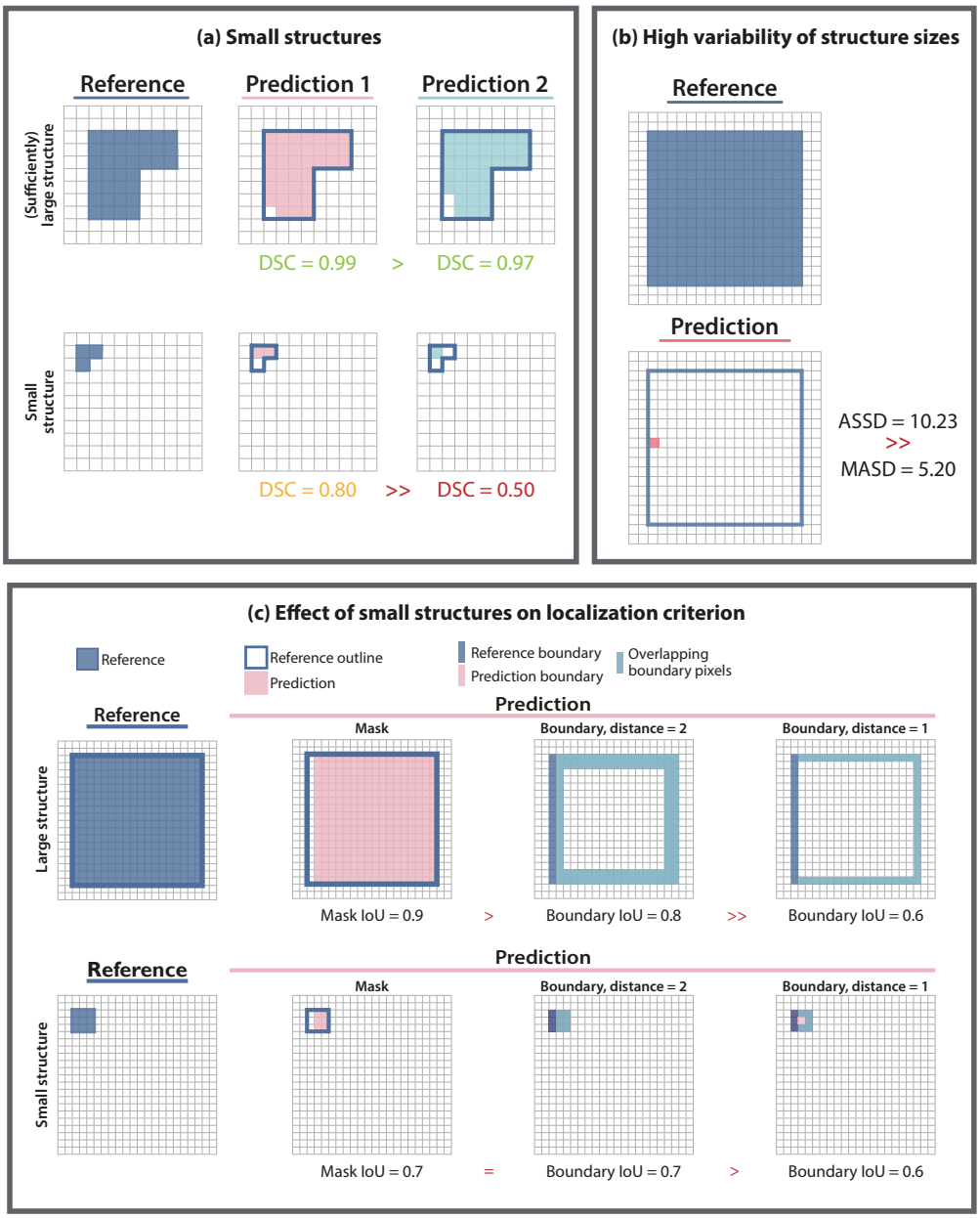


Figure 4.8: Pitfalls related to disregard of properties of the target structure: Small structure sizes and high variability of structure sizes. **(a) Effect of small structures.** Only one pixel distinguishes *Prediction 1* and *2* in the large and a small structures. This has a major impact on the corresponding Dice Similarity Coefficient (DSC) values for the small structure (bottom row). **(b) Effect of high variability of structure sizes.** The Mean Absolute Surface Distance (MASD) is substantially lower than the Average Symmetric Surface Distance (ASSD) if the *Prediction* is very small (in this case, one pixel) and close to the reference boundary. **(c) Effect of small structure on localization criterion.** Mask Intersection over Union (Mask IoU) is less susceptible to boundary errors of large objects compared to the Boundary IoU, which penalizes boundary errors and is more invariant to structure size. Figure adapted from [Reinke et al., 2021a].

Complex shapes (shape- and topology-related) Overlap-based metrics are not designed to accurately predict a segmentation boundary. As a result, they might not be able to distinguish between differences in the shapes and contours of objects, which would be crucial for applications like radiation, where locating and treating a tumor's entire surface is necessary to prevent recurrence [Burnet et al., 2004]. This issue is illustrated in Figure 4.9(a), which shows five different predictions with various shapes. The scores of the overlap-based metrics are the same for all predictions, neglecting the shapes of objects [Yeghiazaryan and Voiculescu, 2015]. This issue is better handled by boundary-based metrics.

Complex shapes often appear in biomedical images, such as images of the bronchial tube or hepatic vessels. In those cases, it is often desirable to segment the centerline rather than the full object. However, common metrics such as the DSC do not measure differences in shapes (see also Figure 4.9(a)). In Figure 4.9(b), we present an object inspired by bronchia. *Prediction 1* solely measures the root of the object, obviously not being a favorable approximation of the object. *Prediction 2* misses some pixels because it focuses solely on the structure's centerline, which would be favorable. The DSC, however, yields the same score for both predictions. The clDice was designed for such use cases and penalizes *Prediction 1* more for missing the centerline of the structure.

Occurrence of overlapping or touching structures (shape- and topology-related) Pixels may not only be assigned a single label. In brain tumor segmentation, for example, the tumor core lies inside the tumor, resulting in two labels for the tumor core pixels [Menze et al., 2014]. In such scenarios, prior knowledge indicates that an additional requirement should be added: The tumor core should lie inside the tumor. Those requirements, however, are not implemented in common overlap-based metrics. A similar example is illustrated in Figure 4.9(c), which shows two predictions. Despite not meeting the condition that *Label 2* must be inside of *Label 1*, *Prediction 2* still receives a higher DSC score than *Prediction 1*.

Occurrence of disconnected structures (shape- and topology-related) For object detection problems, the granularity of annotations should be chosen to reflect biomedical needs. However, often standard procedures such as the creation of bounding boxes are preferred. Bounding boxes may not be appropriate to approximate complex structures, such as the tubular object in the top example of Figure 4.9(d). Similarly, a box hides whether an object is disconnected, such as the one in the bottom example. For both examples, the Prediction is considered as a TP, although the object is not covered by the predicted bounding box. Moreover, the predicted bounding box of the right example only covers one part of the disconnected structure and is still interpreted as TP.

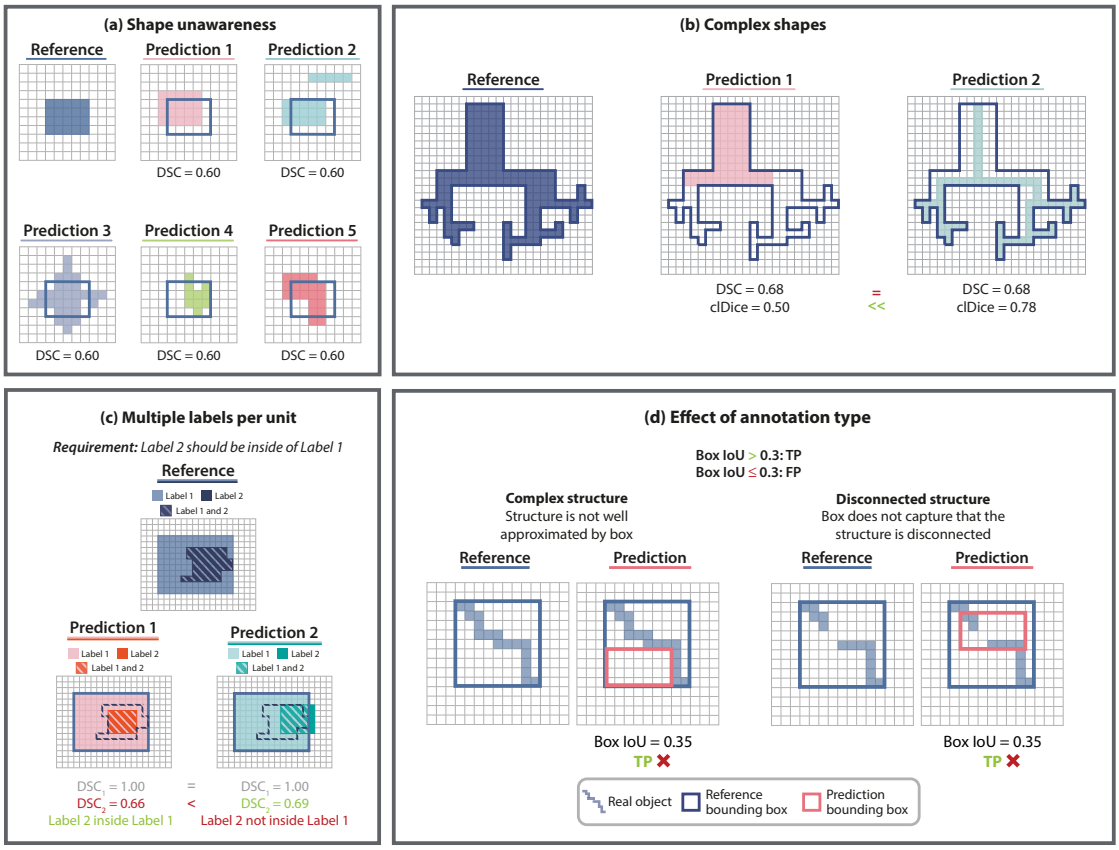


Figure 4.9: Pitfalls related to disregard of properties of the target structure: Complex shapes, overlapping and disconnected structures. **(a) Effect of shape unawareness.** The predictions of five algorithms (*Predictions 1* to *5*) have very different shapes but yield the same Dice Similarity Coefficient (DSC) scores. **(b) Effect of complex shapes.** The DSC does not distinguish between a prediction only segmenting the root of the structure (*Prediction 1*) and a prediction segmenting the centerlines (*Prediction 2*), which would be preferable. This is recognized by the Centerline Dice Similarity Coefficient (clDice). **(c) Effect of several labels per unit.** *Prediction 2* violates the requirement that *Label 2* be included within *Label 1*. Nonetheless, *Prediction 2* yields a greater DSC score than *Prediction 1*, which meets the condition. **(d) Effect of the annotation type**. Bounding boxes are not appropriate to approximate complex or disconnected objects. Although the *Prediction*'s bounding box is not hitting the object, it is still considered as True Positive (TP) as the localization criterion is fulfilled (here: Box Intersection over Union (Box IoU) > 0.3. Figure adapted from [Reinke et al., 2021a].

Pitfall Subtype 2.3: Disregard of the properties of the data set

The data sets used for performance assessment highly influence metrics. In the following, we present pitfalls that relate to a disregard of properties of the data set.

High class imbalance Class imbalance often occurs if one disease for example may be underrepresented in a data set. While Accuracy is one of the most commonly used metrics for image-level classification, one of its most critical limitations is that it is not designed to properly handle class imbalances. Figure 4.10(a) illustrates an example of an extremely imbalanced data set with only three occurrences of the positive and 97 samples of the negative class. *Prediction 1* provides a valid separation between both classes. However, *Prediction 2* only suggests a majority vote, always predicting a sample to be of the most frequently occurring class, namely class 2 (blue triangles). Nonetheless, the Accuracy grants both predictions the same near-perfect score of 0.97. This issue is spotted by metrics like the BA. Similarly, the MCC values indicate that *Prediction 2* is not better than random guessing [Chicco and Jurman, 2020]. Class imbalance is quite common in biomedical data sets. Thus, a high Accuracy score could be dangerous and misleading in clinical practice. For instance, in a data set mostly composed of healthy subjects, using Accuracy could quite easily lead to misdiagnosing those that are sick as healthy. This pitfall is also relevant for the AUROC metric, which can be overly optimistic in the case of class imbalances.

As shown in Figure 4.6(b), BA has desirable properties given its prevalence independence. Still, the metric might be deceptive in a very imbalanced situation, like in use case 2 from Chicco et al. [2021]. Such an example is illustrated in Figure 4.10(b). The BA score is very high, implying that the prediction is performing very well, although it predicts 499 FPs. This is because BA does not consider predictive values. This issue is revealed by the MCC (and CK). In clinical practice, focusing only on the BA score could thus be critical and lead to healthy subjects receiving harmful unnecessary treatments.

Small sample sizes Due to privacy and resource issues, sample sizes are often low in the biomedical domain. Severe problems may arise for the computation of some metrics in such situations, such as the AUROC. Six images, three positive and three negative samples, are shown in Figure 4.10(c) along with the predicted class scores per sample. The two predictions only differ in a single predicted class score. Still, even with this small change, the AUROC scores differ by 0.11. In the case of very small sample sizes, the respective 95% Confidence Interval (CI) [DeLong et al., 1988] is very large and does not allow for meaningful interpretation. This pitfall is also relevant for other multi-threshold metrics used in image-level classification, object detection, and instance segmentation problems. Moreover, the ECE depends on the sample size. Even for a perfectly calibrated model, the values may be unequal to zero for small sample sizes, as shown by Gruber and Buettner [2022] and illustrated in Figure 4.10(d).

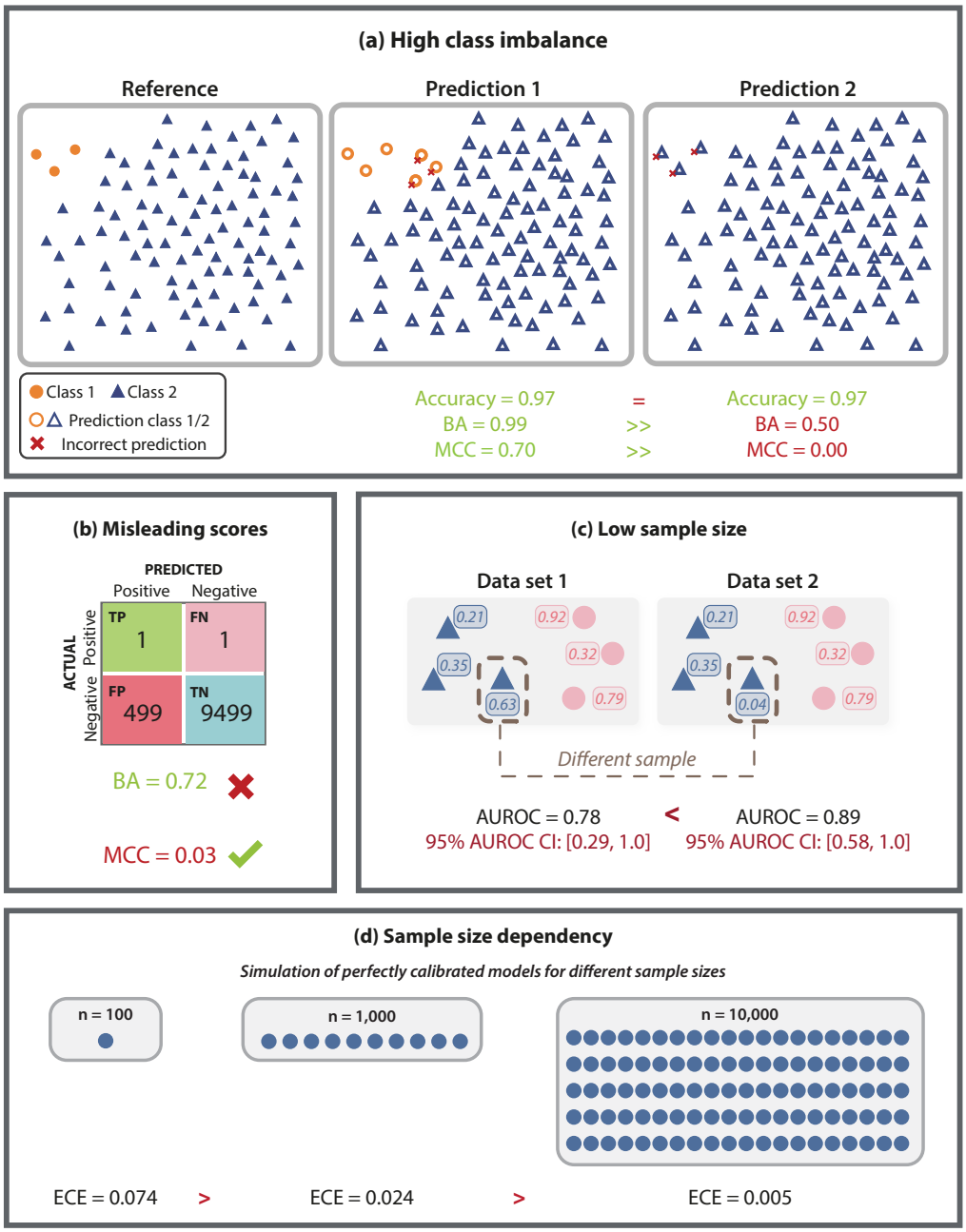


Figure 4.10: Pitfalls related to disregard of properties of the data set: Class imbalance, small sample sizes. **(a) Class imbalance.** Some metrics fail to properly reflect class imbalances (e.g. Accuracy). Even if the classifier performs very poorly for one of the classes (*Prediction 2*), the metric produces a high score for the underrepresented classes. **(b) Misleading Balanced Accuracy (BA) scores.** Even though the prediction is ineffective, as evidenced by the Matthews Correlation Coefficient (MCC) score, BA may still produce a nearly perfect score in imbalanced settings. **(c) Area under the Receiver Operating Characteristic Curve (AUROC) for small sample sizes.** When sample sizes are small, multi-threshold metrics are unstable. Although only one score differs between data sets 1 and 2, there is substantial variation in the AUROC scores between both sets. The 95% Confidence Interval (CI) demonstrates the wide range of potential value. **(d) Sample size dependency.** Even for a perfectly calibrated model, the Expected Calibration Error (ECE) may be unequal to zero (indicating imperfect calibration) for small sample sizes. Figure adapted from [Reinke et al., 2021a].

High inter-rater variability Errors in the annotations occur frequently for biomedical data, resulting in high inter-rater variability [Joskowicz et al., 2019]. An example of inter-rater variability is illustrated in Figure 4.11(a), which shows the annotations of two raters (which can also be interpreted as reference and prediction). The segmentation masks slightly differ in the outer pixels of the boundaries, yielding a DSC score of 0.68, although the main body of the structure is perfectly segmented. Using a sufficient threshold for the Normalized Surface Distance (NSD) metric resolves the issue with a perfect score of 1.0. This pitfall is also relevant to other metrics (e.g. the IoU, HD(95), ASSD or MASD) and for the localization criteria in object detection.

Outliers, noise, and artifacts in the reference annotation Metric scores may be substantially impacted by the existence of spatial outliers, for instance given by noise or reference annotation distortions. An example is shown in Figure 4.11(b), for which a single outlier pixel heavily affects the HD score. By the definition of using the 95% percentile instead of the maximum of shortest distances, outliers are handled better by the Hausdorff Distance 95 Percentile (HD95) compared to the regular HD.

Empty reference Biomedical image analysis data sets often contain data from healthy and sick patients, for example in the case of tumor monitoring. Thus, images of healthy patients do not contain a tumor and are empty. In those images, the number of TP predictions is naturally zero, as the target structure is not present. In the case of an empty reference, a non-empty prediction shows a FP with no TP. Vice versa, in the case of a non-empty reference, an empty prediction would yield a FN with no TP as well. If either the reference or prediction is empty, several boundary-based metrics are undefined, depending on the implementation. Similarly, if both of them are empty, this causes division by zero errors for metrics such as the DSC and NSD. These use cases are illustrated in Figure 4.11(c). In such cases, ad hoc rules need to be defined to handle those issues.

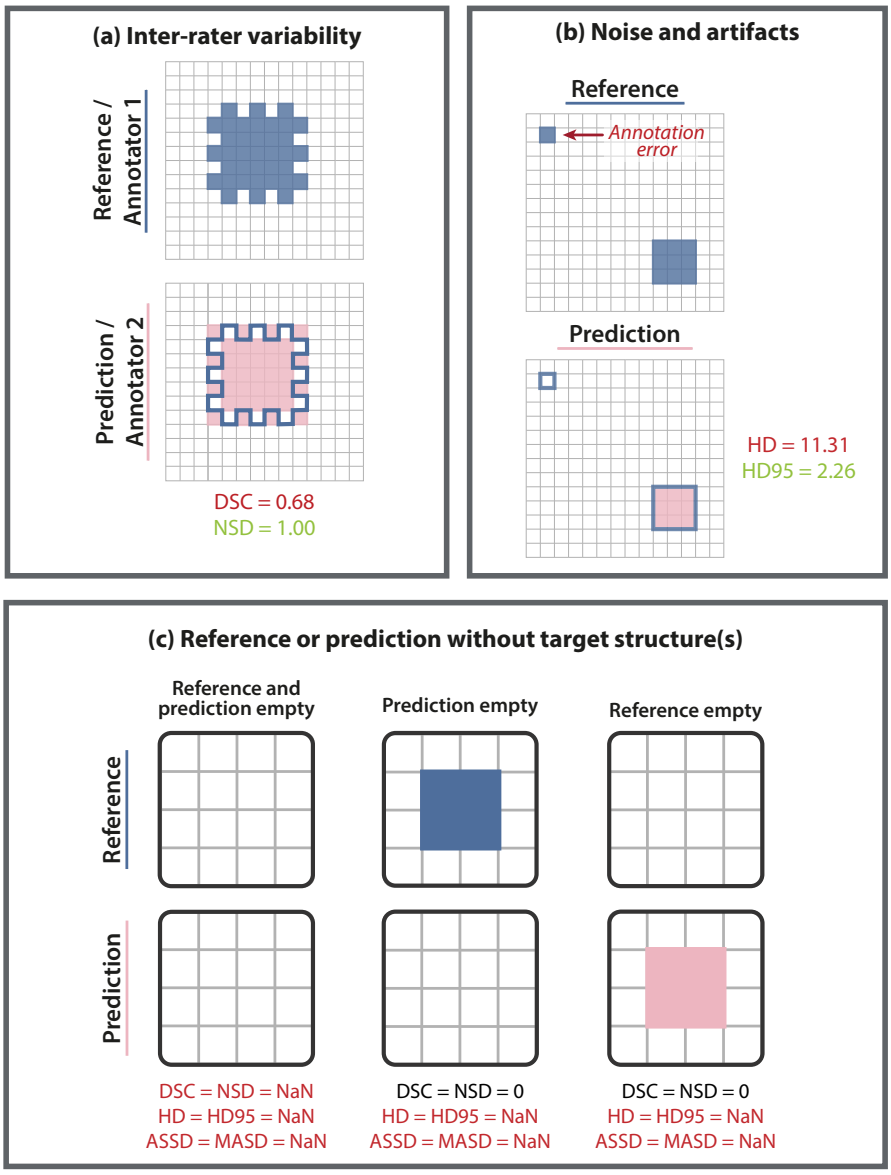


Figure 4.11: Pitfalls related to disregard of properties of the data set: High inter-rater variability, spatial outliers in the reference and empty reference or prediction. **(a) Effect of inter-rater variability.** Validating *Annotator 2*'s performance while utilizing *Annotator 1* to create the reference yields a poor Dice Similarity Coefficient (DSC) score, as inter-rater variability is not taken into account. The Normalized Surface Distance (NSD) captures this variability. **(b) Effect of annotation errors or noise.** Particularly in the case of the Hausdorff Distance (HD) when applied to small structures, a single incorrectly annotated pixel may result in a significant reduction in performance. In contrast, the Hausdorff Distance 95 Percentile (HD95) can handle spatial outliers. **(c) Effect of empty reference or prediction.** An empty reference and/or prediction yields division by zero errors, which in turn yield metric scores of `Not a Number (NaN)`. Further abbreviations: Average Symmetric Surface Distance (ASSD), Mean Absolute Surface Distance (MASD). Figure adapted from [Reinke et al., 2021a].

Pitfall Subtype 2.4: Disregard of the properties of the algorithm output

Depending on the output of algorithms, metric values may be affected. In the following, we present pitfalls that relate to a disregard of properties of the algorithm output.

Empty prediction Similar to the previous example, an empty prediction may cause division by zero errors, as shown in Figure 4.11(c).

Possibility of overlapping predictions Similar to how overlapping reference objects may cause pitfalls, overlapping predictions may be problematic. On the one hand, they may cause problems in the assignment strategy of object detection and instance segmentation tasks. On the other hand, they can lead to misleading results when phrasing the problem as semantic or instance segmentation. The two objects in Figure 4.12(a) end up being merged into a single object when being phrased as semantic segmentation, yielding ideal metric scores. However, the predictions are not perfect, which is shown by phrasing the problem as instance segmentation. Given the amount of overlap, they end up reflecting the merged shape perfectly. However, the prediction of *Instance 1* only shows a U-shape, not the actual rectangular shape, leading to a reduced DSC of 0.51. Also, the amount of overlap between the objects is much lower compared to the reference, which is not reflected in the metric scores.

Unavailability of predicted class scores If an algorithm outputs predicted class scores rather than a class label, multi-threshold metrics can be computed. If those scores are not available, other metrics should be used. However, some applications still calculate multi-threshold metrics (for example [Bai and Urtasun, 2017; Hirsch et al., 2020]). Often, it is assumed that all confidence scores are the same, meaning that the formula of the Average Precision (AP) would simplify to Sensitivity $\cdot$ PPV. Those assumptions are not standardized and may yield different results for different implementations, as shown in the example of Figure 4.12(b).

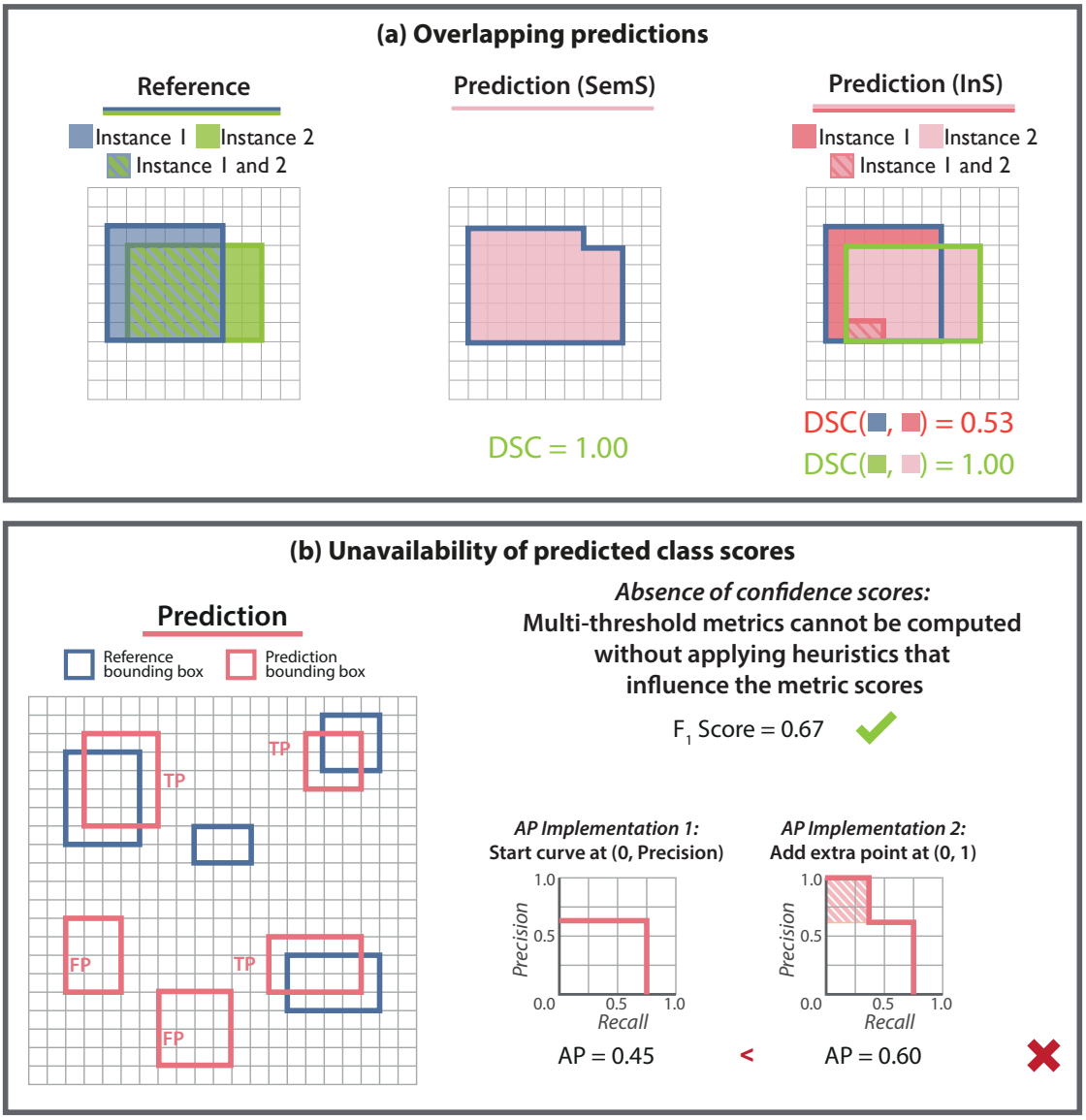


Figure 4.12: Pitfalls related to disregard of properties of the algorithm output: Possibility of overlapping predictions and unavailability of predicted class scores. **(a) Effect of overlapping predictions.** The overlapping instances of the prediction are combined into a single structure in Semantic Segmentation (SemS) problems, producing an ideal metric score. *Instance 1* is not at all well-approximated when the issue is framed as an Instance Segmentation (InS) problem. Common measures (here: Dice Similarity Coefficient (DSC)) do not highlight this problem. **(b) Effect of unavailability of predicted class scores.** Predicted class scores are necessary for being able to calculate multi-threshold metrics, such as the Average Precision (AP). Different strategies/heuristics for such situations (here: *implementations 1* and *2*) yield varying values. Figure adapted from [Reinke et al., 2021a].

Pitfall Type 3: Pitfalls related to poor metric application

Type 3 of pitfalls relate to all limitations resulting from a poor application of metrics. We further distinguish type 3 into five subtypes: Inadequate metric aggregation (3.1), ranking schemes (3.2), metric implementation (3.3), metric reporting (3.4), and interpretation of metric values.

Pitfall Subtype 3.1: Inadequate metric aggregation

In real situations, metric values are typically calculated for several images and are usually reported in an aggregated fashion. The type of aggregation, however, can have a huge impact on the resulting values. We distinguish between class-related and data set-related aggregation pitfalls, as presented in the following:

Hierarchical structure of classes (class-related) Many counting metrics are similarly defined for a setting with more than two classes (see Section 2.4.1). These classes are often set up in a hierarchical manner. For example, one class could refer to the absence of pathology, while other classes define different specifications of pathologies. In these cases, the performance assessment should be done per class and not for the binary case only. Figure 4.13(a) provides an illustration of how triangles and circles can be classified, with the circle class being further divided into two different classes (green and pink; for example representing different types of pathologies). The middle example's binary distinction between triangle and circle yields a high metric score (Accuracy of 0.88). The predictor, however, has trouble distinguishing between the circles when analyzing the three classes separately (pathologies), which results in a considerable decrease in their per-class Accuracy scores (0.63). This difference could be problematic in clinical practice if the classifier is assumed to perform well across all pathologies, missing that it may not be ideal to classify a specific pathology.

Non-binary situations (class-related) Different methods may be used to compute the metric values for multi-class problems. One could validate metrics per class and aggregate the values over the classes in a subsequent step. However, similar to class confusions, some classes may be more important than others. Such an example is provided in Figure 4.13(b), illustrating three classes with a pre-defined weight. Simple macro-averaging of classes (averaging scores over all classes) leads to misleading results compared to the weighted average. The pitfall is also relevant for object detection and segmentation problems and other metrics from all families that do not define a class aggregation by definition. For example, in the case of brain tumor segmentation, three classes are typically assumed, the contrast-enhancing tumor, the tumor core, and the whole tumor. From a clinical perspective, the contrast-enhancing tumor is more important than the other channels [Weller et al., 2014] and should be given a higher weighting.

A similar situation is displayed in Figure 4.13(c) for the segmentation of multiple classes. Simply averaging over all values and ignoring classes would not yield a good representation of the performance. Only a validation per class would reveal that the performance is very different over the different classes. A simple average would for example have hidden that the aggregated metric score for class 2 is only at 0.30.

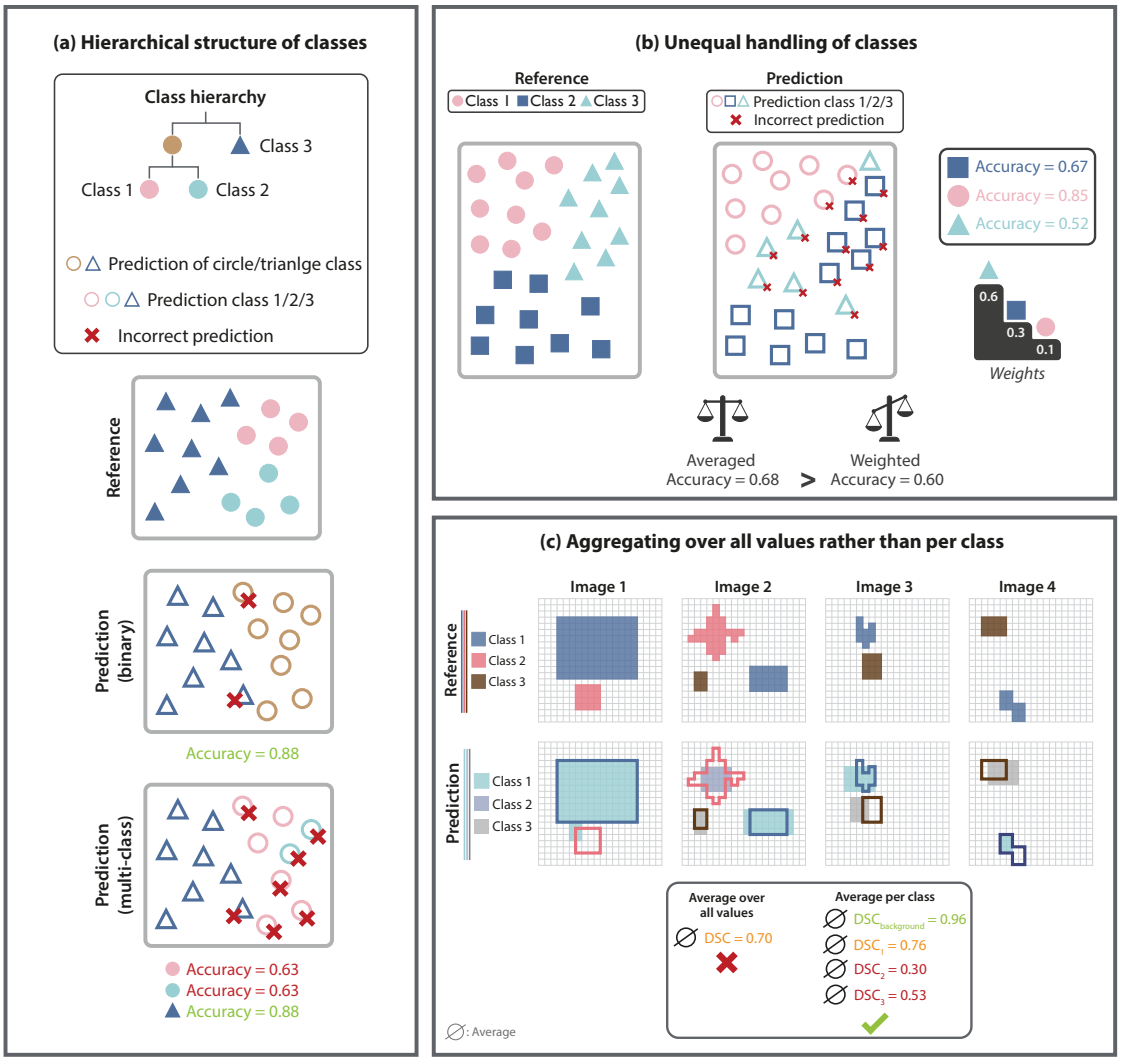


Figure 4.13: Pitfalls related to inadequate metric aggregation: Hierarchical structure of classes, unequal handling of classes, and per-class aggregation. **(a) Pitfall: Hierarchical structure of classes.** In categorical classification, classes may be arranged hierarchically, for instance, as a number of positive classes and a single negative class. The validation outcome is substantially impacted by whether the problem is stated as binary or multi-class. While a per-class validation produces a poor Accuracy score of 0.63 because the two circle classes cannot be separated well, binary classification (middle), which only distinguishes triangles from circles, produces a high Accuracy. A red cross denotes a prediction that was incorrect. **(b) Unequal handling of classes.** Unequal relevance of classes, as indicated by pre-defined weights of classes, is ignored by simple averaging (macro-averaging) of the Accuracy. A red cross denotes a prediction that was incorrect. **(c) Effect of aggregating over all scores.** Simply averaging over all metric values (here: Dice Similarity Coefficient (DSC)) rather than stratifying per class would not reveal the actual low performance for *Classes 2* and *3*. ∅ denotes the average DSC values. Figure adapted from [Reinke et al., 2021a].

Non-independence of test cases (data set-related) In biomedical image analysis, data is often structured hierarchically, such as data from different hospitals or patients, as shown in Figure 4.14(a). In this example, data was acquired from five patients with different amounts of images. A simple aggregation of metric scores (DSC_{all}) would not represent this hierarchy correctly, given the non-independence of data and the differing number of images per patient. As shown by computing the average metric scores per patient, the metric values are substantially different for the patients. *Patients 2* to *4* yield much lower metric scores compared to *Patients 1* and *5*. Especially given the high number of images for *Patient 1*, the aggregation would be overruled by a simple aggregation. Instead, aggregating the average metric scores per patient into a hierarchical average ($DSC_{hierarchical}$) would give a better representation of the actual data structure.

Metric aggregation and `NaN` handling (data set-related) In the previous paragraphs, we showed that `NaN` values may occur in several cases, for example, if one or both of reference and prediction is empty (see Figure 4.11(c)). In the case of challenges, participants may miss submitting results for some images, which also yields `NaN` values. Depending on the `NaN` type, a strategy should be defined on how to deal with those missing scores. A common strategy is to simply ignore them while aggregating, as in the examples presented in Figure 4.14(b) and (c). Another way to handle the values would be to set them to the worst possible value. For metrics like the DSC – and other metrics that are bounded between 0 and 1 and for which a value of 1 relates to a perfect score – the value would be set to 0. Depending on the chosen strategy, it should be noted that the aggregated value differs; in our example from Figure 4.14(b), the difference is quite substantial. Deciding on a missing value strategy is even more complex for metrics without an upper bound, such as the HD(95), the MASD, or ASSD. In these cases, there is no common worst possible value (see Figure 4.14(c)). One possibility would be to use the maximum distance of the images as the worst possible value. However, this distance would be different across data sets (or even within a data set) and may thus not be ideal for comparison or generalization. A case-based ranking scheme (see Section 2.1.2) would overcome the issue by assigning the last rank to every image with `NaN` values. Another strategy lies in normalizing the metric scores so that they are bounded (e.g. between 0 and 1) and using the highest score as the worst possible metric value. Similar to metrics with fixed boundaries, the missing value-handling strategy impacts the resulting aggregates.

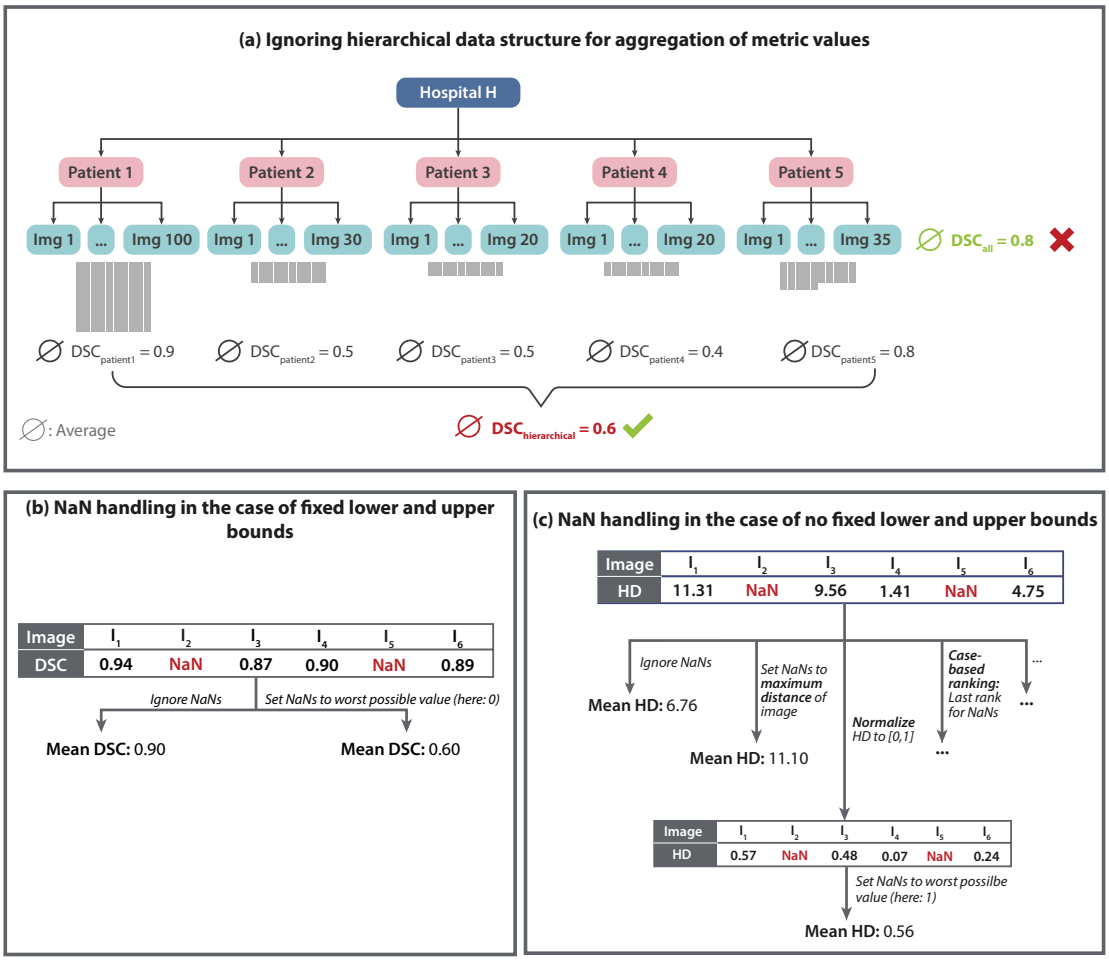


Figure 4.14: Pitfalls related to inadequate metric aggregation: Ignoring hierarchical data structure and missing value handling. **(a) Effect of ignoring the hierarchical structure of data.** In this example, images are taken from different patients with a varying number of images. A simple average of the metric scores (here: Dice Similarity Coefficient (DSC)) would yield a high value that does not represent the actual performance. In this case, the low performances for *Patients 2* to *4* would be overruled by the high scores of *Patient 1*, for which the highest amount of data points was collected. **(b) Effect of `Not a Number (NaN)` handling in the case of fixed bounds.** The strategy of missing value handling largely impacts the aggregated metric scores. For the DSC, ignoring `NaNs` yields a score of 0.90, setting the score to the worst possible value resulting in a mean score of 0.60. **(c) Effect of `NaN` handling in the case of no fixed bounds.** If no fixed upper (or lower) bounds are defined (as for the Hausdorff Distance (HD)), even more possibilities for the missing value handling may arise, with all of them yielding (substantially) different metric aggregates. Figure adapted from [Reinke et al., 2021a].

Lack of stratification and interdependencies between classes (data set-related) Data sets in biomedical image analysis are very often accompanied by different kinds of meta-data, such as specifications of acquisition protocols, image resolution (see Figure 4.22(a)), the presence of metal artifacts, or patient-related information such as age or gender. Figure 4.15(a) shows an example of ignoring the gender of patients, which was provided as meta-information along with the images. The overall Accuracy yields a value of 0.58. However, this value does not adequately represent the performance across the specific data set. It becomes clear that the performance is substantially worse for women than for images of men when the Accuracy is computed separately for both types of gender. This pitfall is also relevant for object detection and segmentation problems and generalizes to all metrics that do not take into account the stratification.

Additionally, if there are multiple classes present in a data set, one must carefully consider how classes are correlated. Classes may be interdependent by nature, for example the body fat percentage and Body Mass Index (BMI). On the other hand, interdependencies may occur from the data acquisition, for example by including images from the same patient. Figure 4.15(b) provides the example of a classifier aiming to classify the blue triangles. The prediction yields an Accuracy of 0.94, indicating a near-perfect prediction. When calculating the Accuracy for images with and without light blue squares, it becomes apparent that high Accuracy was only observed for images with visible light blue squares. Images without the square yielded a much lower performance. This pitfall is also relevant for object detection and segmentation problems and generalizes to all metrics that do not take into account the interdependencies.

Disregard of number of images (data set-related) The AP metric does not consider the total number of images. Figure 4.15(c) provides an example of two data sets. The second data set contains two additional empty images that were correctly predicted as empty. The AP score is similar for both situations, while data set *D1* receives a higher number of False Positives per Image (FPPI) and is thus penalized with a lower Free-Response Receiver Operating Characteristic (FROC) Score.

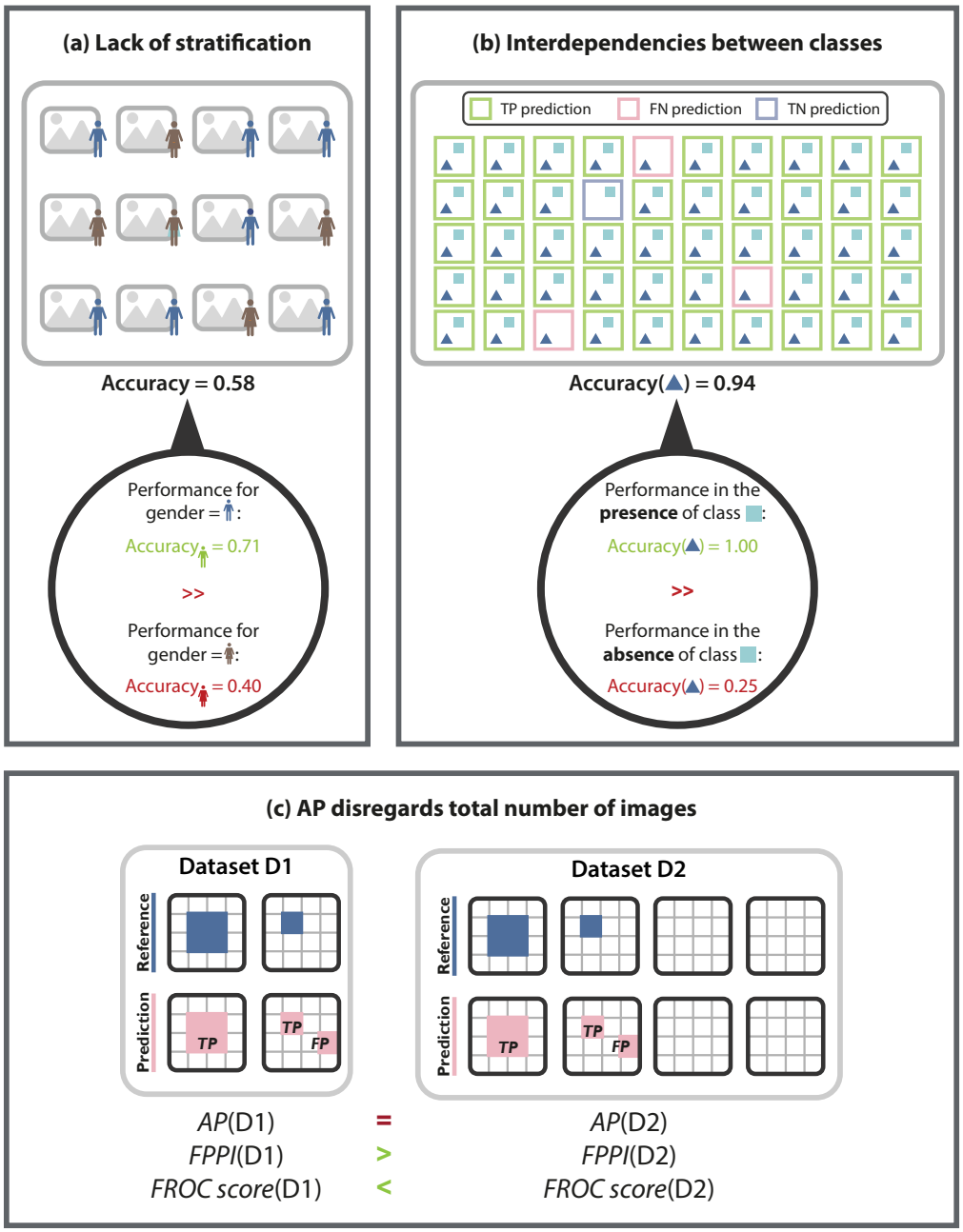


Figure 4.15: Pitfalls related to inadequate metric aggregation: Lack of stratification, interdependencies, and disregarding number of images. **(a) Effect of ignoring meta-information (here: gender).** The overall Accuracy does not indicate that the prediction yields much higher scores for men compared to women when the patient's gender is ignored from the available meta-information for each image. **(b) Effect of interdependencies between classes.** Given that the blue triangle regularly appears alongside the light blue square, a prediction may yield a nearly perfect Accuracy score of 0.94 for it. The algorithm only performs well when the pink class is present, apparent from computing the Accuracy in the presence and absence of the square class. **(c) Effect of the number of images.** Data sets *D1* and *D2* are similar, but *D2* contains additional empty images that were correctly predicted as such. While the Free-Response Receiver Operating Characteristic (FROC) Score correctly penalizes *D1*, the Average Precision (AP) ignores the total number of images, yielding the same score for both data sets. Figure adapted from [Reinke et al., 2021a].

Pitfall Subtype 3.2: Inadequate ranking schemes

In challenges or benchmarking experiments, algorithms are typically ranked to compare performances. As we discuss in the following sections, the chosen ranking scheme and metrics may substantially change.

Combination of related metrics in rankings Every metric reflects a certain property. When using multiple metrics, it should be made sure that the metrics highlight different properties and are not focused on the same effects. If so, a specific property might be overrated in the case of mathematically related metrics, such as the DSC and IoU (see Equation 2.50). Figure 4.16(a) provides an example of a ranking based on three metrics, the DSC, IoU, and HD. Since DSC and IoU focus on the overlap between structures and measuring the same properties, they lead to the same ranking scheme. The HD, reflecting the accuracy of the object boundaries, yields a different ranking. If all ranks would be aggregated into one single ranking, the property of overlap would receive a higher weight in this case. In addition, the DSC and IoU rankings won't add additional value to the other, thus, being redundant. In fact, when checking the challenge tasks that were analyzed in Sections 3.1 and 5.1 (with a total of 804 tasks in both Sections), we found that 15 tasks used both metrics, the DSC and IoU, simultaneously in their rankings. More severely, three tasks based their rankings on metrics that were identical because the organizers were not aware of metric synonyms: Two tasks used the F_1 Score and the DSC, and one task used the Sensitivity and Recall simultaneously.

Ranking variability As we show in the following Section 4.3, rankings are quite sensitive to the chosen calculation method. The common practice of only computing ranking tables (as we show in Section 5.3, 27% of challenges solely report ranking tables), hides important information on the distribution of metric values and differences between algorithms.

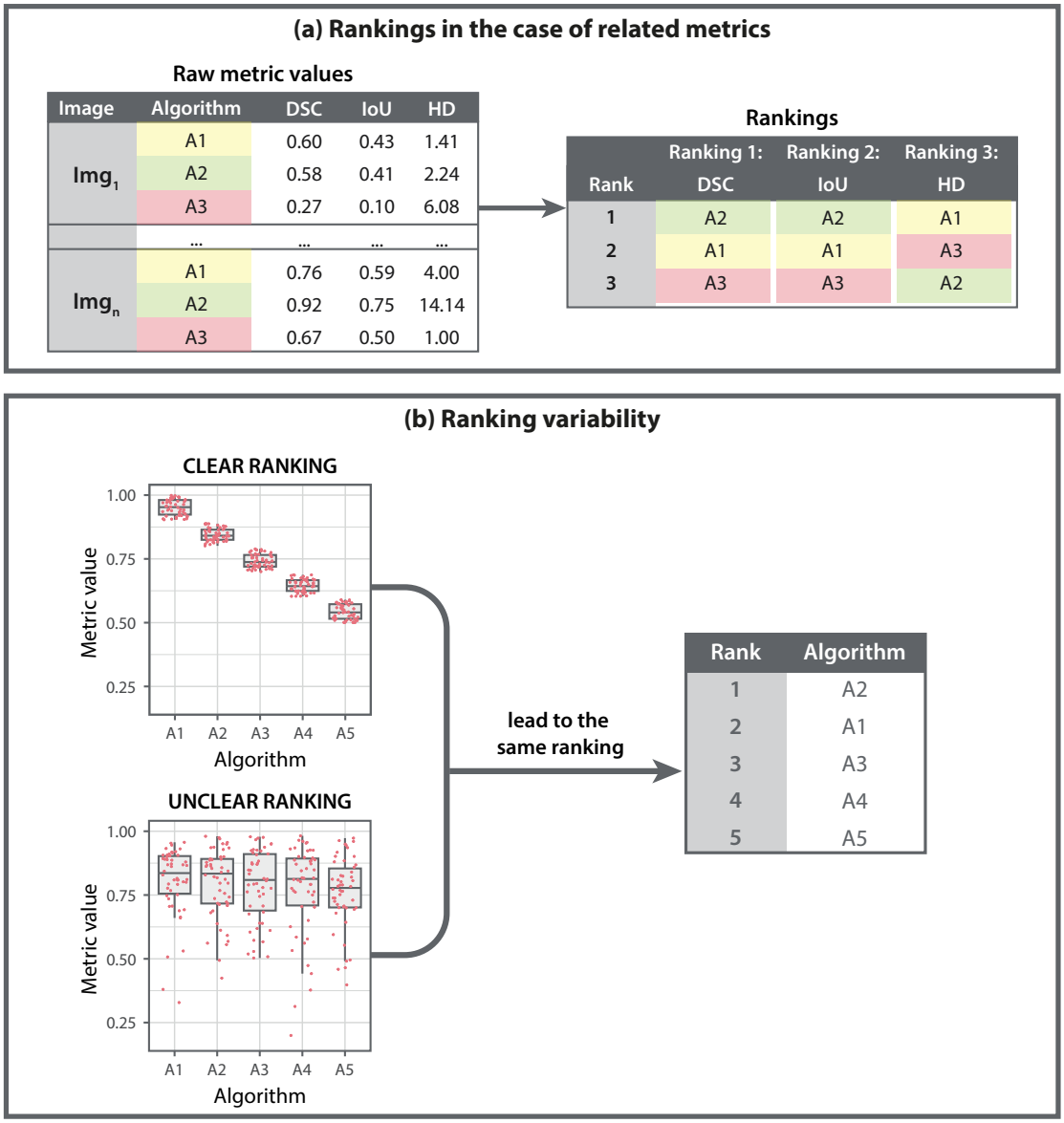


Figure 4.16: Pitfalls related to inadequate ranking scheme: Ignoring metric relationships and ranking variability. **(a) Effect of computing rankings for mathematically related metrics.** Related metrics, such as the Dice Similarity Coefficient (DSC) and Intersection over Union (IoU) usually yield the same rankings, while metrics focusing on other properties, such as the Hausdorff Distance (HD), may result in a completely different ranking. **(b) Effect of ranking variability.** Different benchmarking experiments (top: clear ranking, bottom: unclear ranking) may yield the same ranking tables. Only reporting tables may hide information on the distribution of metric scores. Figure adapted from [Reinke et al., 2021a].

Pitfall Subtype 3.3: Inadequate metric implementation

Theoretical knowledge of metrics is only one important step. The implementation of metrics is also not always straightforward and well-defined. In the following, we present several pitfalls related to metric implementation:

Non-standardized metric definition Although the FROC Score is often favored by clinicians given its easier interpretation, the abscissa of the FROC curve is not standardized. This results in different FROC Scores for different value ranges of the abscissa. Figure 4.17(a) shows the example of three different value ranges for the same prediction. The FROC Scores differ substantially from 0.69 to 0.78.

Using different implementations for one metric may result in different values. For example, popular repositories use varying approaches to handle the corner case of identical confidence scores in object detection or instance segmentation problems for the AP metric. While MS Common Objects in COntext (COCO) [Lin et al., 2014] applies one step for every prediction (also for duplicate scores), the CityScapes [Cordts et al., 2015] implementation treats all objects with duplicate scores in a single step. Also, they use different ways of interpolation. All these implementations yield different AP scores, as shown in Figure 4.17(b).

Sensitivity to hyperparameters As shown in Section 2.4, several metrics rely on defining additional hyperparameters. For example, the tolerated distance from a structure's boundary can be defined by the hyperparameter τ for the NSD, while the parameter β defines the treatment of FPs versus FNs in the F_β Score. In object detection or instance segmentation tasks, the chosen localization criterion (for example the (Box) IoU) includes the definition of a localization threshold. Only predictions whose localization criterion is higher than the defined threshold, are considered as TP instances. Yet, the threshold may highly influence the resulting metric scores. Figure 4.17(c) shows an example of a loose criterion, for which a TP is considered as soon as the IoU score is greater than zero. In this case, the prediction is far from perfect as it is vastly too large for the small reference instance and the overlap is only marginal with an IoU score of 0.05. Especially in the case of tubular structures, the threshold should be chosen carefully. Figure 4.17(d) illustrates an example with a diagonal, tubular shape. However, the bounding boxes quickly grow given the tubular nature of the object, resulting in *Prediction 1* being considered as TP and *Prediction 2* as FP, although both predictions look very similar.

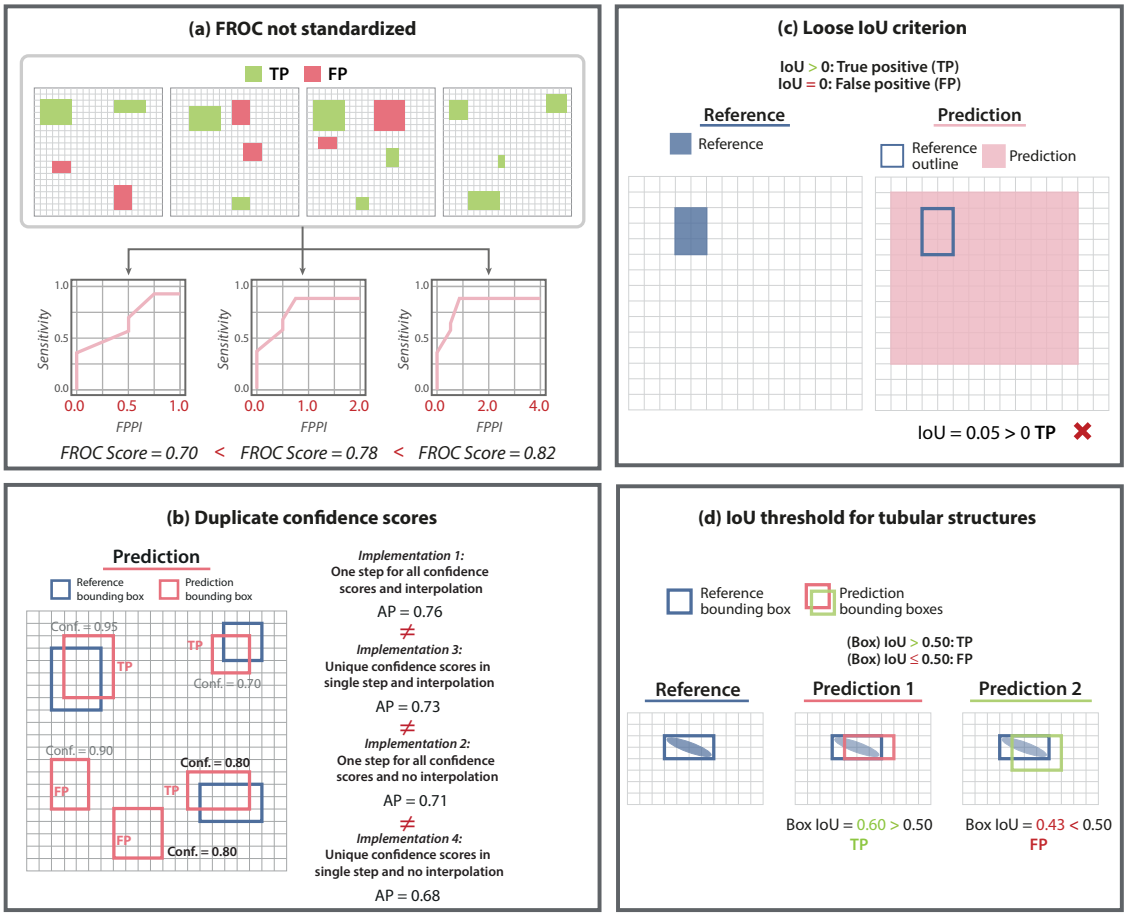


Figure 4.17: Pitfalls related to inadequate metric implementation: Non-standardized metric definition and sensitivity to hyperparameters. **(a) Effect of the abscissa for the FROC curve.** The Free-Response Receiver Operating Characteristic (FROC) curve is not standardized. Different choices of the abscissa yield different curves and substantially different FROC Scores. **(b) Effect of duplicate confidence scores.** In the case of duplicate confidence scores, implementations of the Average Precision (AP) vary and result in substantially different metric values. **(c) Effect of a loose threshold.** Large predictions may deceive the localizations when a True Positive (TP) is defined by an Intersection over Union (IoU) > 0. **(d) Effect of the threshold for tubular structures.** There may be a quadratic difference in the number of bounding box pixels, especially for diagonal, narrow objects, resulting in one prediction considered as TP and the other as False Positive (FP), although fairly similar. Figure adapted from [Reinke et al., 2021a].

Defining global decision thresholds In a scenario of multiple classes, a single cutoff value for a threshold must be determined for all classes. As the ideal threshold range for one class may be different from the ideal threshold range for another class, multi-threshold metrics like AUROC may be unduly optimistic. Three classes with three perfect discriminations are provided in Figure 4.18(a). The corresponding counting metrics (here: Accuracy) produce very poor results for classes 2 and 3 if a single threshold (0.7) is chosen based on the results of class 1. Other counting metrics would also be affected by this issue. Thus, if not chosen to represent all classes, this pitfall could yield a classifier that would only be able to, for example, predict one specific type of tumor, missing or wrongly classifying all others.

Discretization issues Calibration errors, such as the ECE and Maximum Calibration Error (MCE), often use a binning strategy to divide the probability interval of [0, 1] into bins. However, how to choose those bins is not clearly defined . The number could be chosen freely, similarly to the choice of whether it should be an equidistant split or not. The choice of the bins heavily affects the corresponding metric scores, as shown in Figure 4.18(b), which depicts three different numbers of equidistant bins for the same prediction. The calibration error scores vary substantially, especially the MCE is much higher for ten bins compared to three or five bins.

Definition of class labels In binary classification problems, classes are often defined as positive (e.g. cancer) and negative (e.g. no cancer). Per-class counting metrics heavily rely on the definition of those labels and yield completely different scores when exchanging them. Such an example is provided in Figure 4.18(c). When exchanging the positive and negative classes, the values of the presented per-class counting metrics substantially differ. Reversing the labels does not affect multi-class counting metrics. While defining the positive and negative class may be straightforward in some cases, there may be scenarios for which the interpretation can favor both ways. When classifying patients into sick and healthy, one could assume the sick class as positive (because it has the condition) as the goal would be to predict which patients suffer from a disease. On the other hand, one may be interested in identifying patients that are fully recovered, thus defining the healthy class as positive.

Assessing different properties simultaneously In instance segmentation problems, the Panoptic Quality (PQ) can be utilized to assess the detection and segmentation quality simultaneously in a single score. The metric value can, however, be ambiguous. Figure 4.18(d) provides an example with two predictions. While *Prediction 1* produces perfect segmentation results, the detection quality is low, as two additional FP objects are predicted. In contrast, *Prediction 2* is perfect in terms of object detection but rather poor in segmentation. However, the PQ is the same for both predictions.

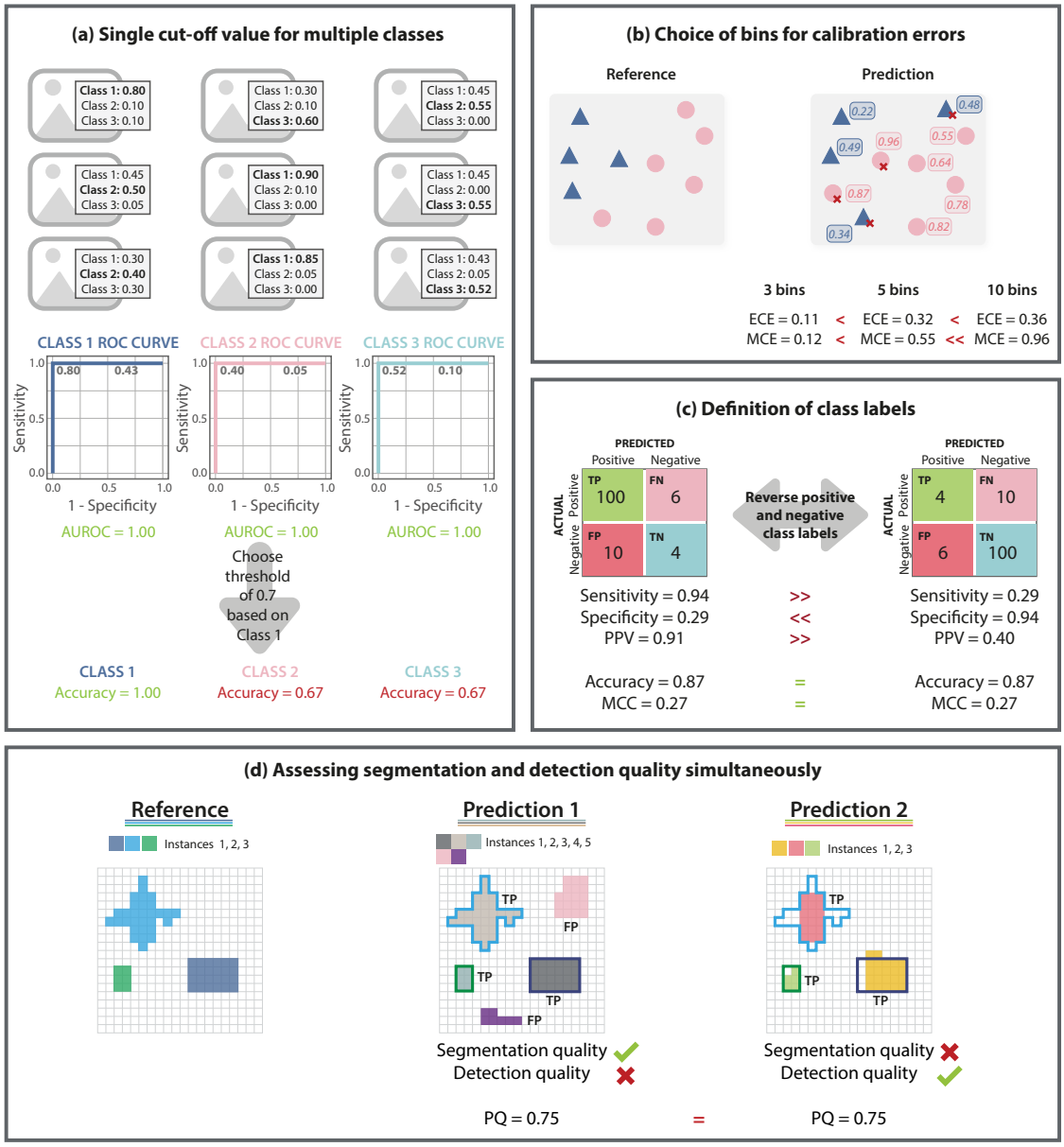


Figure 4.18: Pitfalls related to inadequate metric implementation: Cutoff value, discretization issues, definition of class labels, and assessing multiple properties. **(a) Effect of the determination of a global threshold for all classes based on a single class.** In this example, the class-specific Area under the Receiver Operating Characteristic Curve (AUROC) score is 1.0 for all three classes. The resulting counting metrics for classes 2 and 3 are subpar when selecting an overall threshold based on class 1 for all classes. **(b) Effect of the choice of bins.** The number of bins is not standardized for the Expected Calibration Error (ECE) and Maximum Calibration Error (MCE). However, the number of bins substantially affects the respective scores. **(c) Effect of the definition of class labels.** Per-class counting metrics (here: Sensitivity, Specificity, Positive Predictive Value (PPV)) depend on the definition of the positive class. Switching class labels yields substantially different metric scores. Multi-class metrics (here: Accuracy, Matthews Correlation Coefficient (MCC)) are not affected. **(d) Effect of assessing segmentation and detection quality simultaneously.** With many False Positive (FP) predictions, *Prediction 1* produces an excellent segmentation but shows low detection quality, whereas *Prediction 2* does the opposite (only predicting True Positive (TP) instances; no FP but low segmentation quality). Both, however, produce the same Panoptic Quality (PQ) score. Figure adapted from [Reinke et al., 2021a].

Metric-specific limitations Several pitfalls are related to the inherent mathematical properties of a metric. In the following, we will present two examples for multi-threshold metrics for the example of the AP and four examples for center-based localization criteria.

The AP is one of the most popular metrics for validating object detection problems and is based on the Precision-Recall (PR)-curve. For the creation of the curve, the predicted class scores (usually referred to as confidence scores in object detection) are ordered. Once such a ranking of confidence scores has been generated, the actual values of the confidence scores are no longer of importance.

Figure 4.19(a) shows an example of two predictions for the same two reference objects. The predictions are very similar to each other (both yielding two TPs and one FP), but they differ in one confidence score. However, they yield the exact same PR-curve and AP score because the confidence score that differs across predictions but do not change the ranking of confidence scores, thus not affecting the resulting AP score. Thereby, the predicted class scores are neglected within the ranking as long they are not changing the ordering.

In contrast, in other situations, very small changes in the confidence scores may substantially affect the AP if they are changing the ordering of predicted class scores. Such a situation is shown in Figure 4.19(b). In this example, the predictions again are very similar with small changes (0.04) in the confidence scores for two predicted objects. However, those differences cause a different ordering of class scores, thus affecting the PR-curve and AP scores.

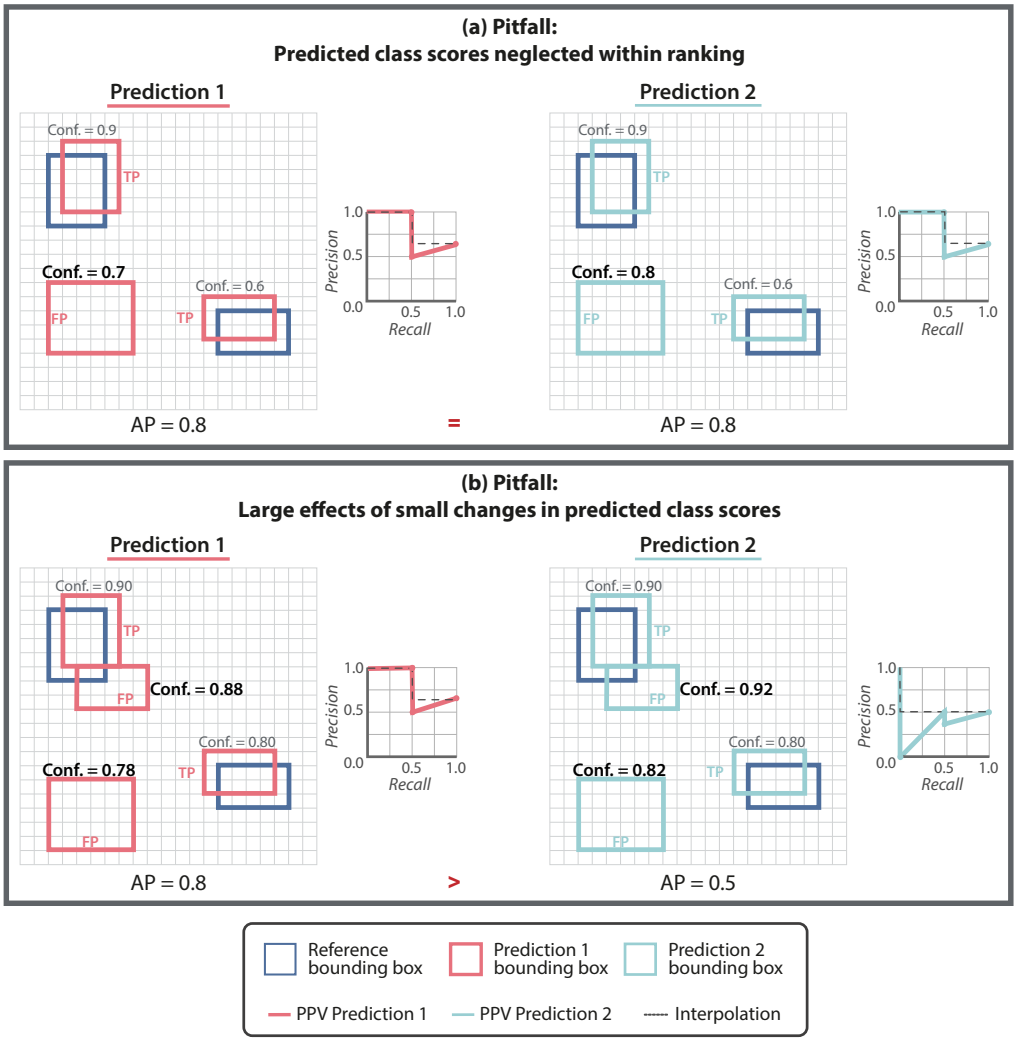


Figure 4.19: Pitfalls related to inadequate metric implementation: Pitfalls regarding multi-threshold metrics. **(a) Effect of predicted class scores that do not change the ranking.** *Predictions 1* and *2* show the same bounding boxes, but differ in one predicted class score. This is not recognized by the Average Precision (AP) if the score does not change the overall ranking of scores. **(b) Effect of small changes in the predicted class scores.** *Predictions 1* and *2* show the same bounding boxes and are only marginally different in their predicted class scores. However, the AP values differ substantially, since the small changes in the class scores change the ranking of predicted class scores. Figure adapted from [Reinke et al., 2021a].

Other mathematical implications arise from center- or point-based localization criteria. Since those criteria only rely on a single point (either a center point or a point inside a mask, box, or approximation), they are naturally susceptible to errors. Figure 4.20 presents four pitfalls of center- or point-based localization criteria.

The center-cover criterion, which assigns a TP if the reference center is inside the predicted object, and the point inside mask/box/approximation criterion can easily be deceived by very large predictions, as shown in Figure 4.20(a).

On the other hand, the Center Distance does not consider the overlap between objects. Thus, as shown in Figure 4.20(b), a prediction that does not overlap the reference could be considered as a TP object.

Similar to the center-cover criterion, the center-hit criterion assigns any prediction as TP for large reference objects since it assigns a TP as soon as the prediction center is inside the reference object (see Figure 4.20(c)).

For elongated, tubular objects, the Center Distance may assign a FP to a prediction that actually hits the object, as shown in Figure 4.20(d). This could be overcome by the point inside mask/box/approximation criterion.

In summary, center- or point-based localization criteria may be easily deceived, for example by always predicting an extremely large structure. Thus, this may yield detecting much more TPs than actually seem reasonable. For example, an algorithm predicting many very large structures could be translated into clinical practice although it is not able to accurately localize the actual tumor or similar objects.

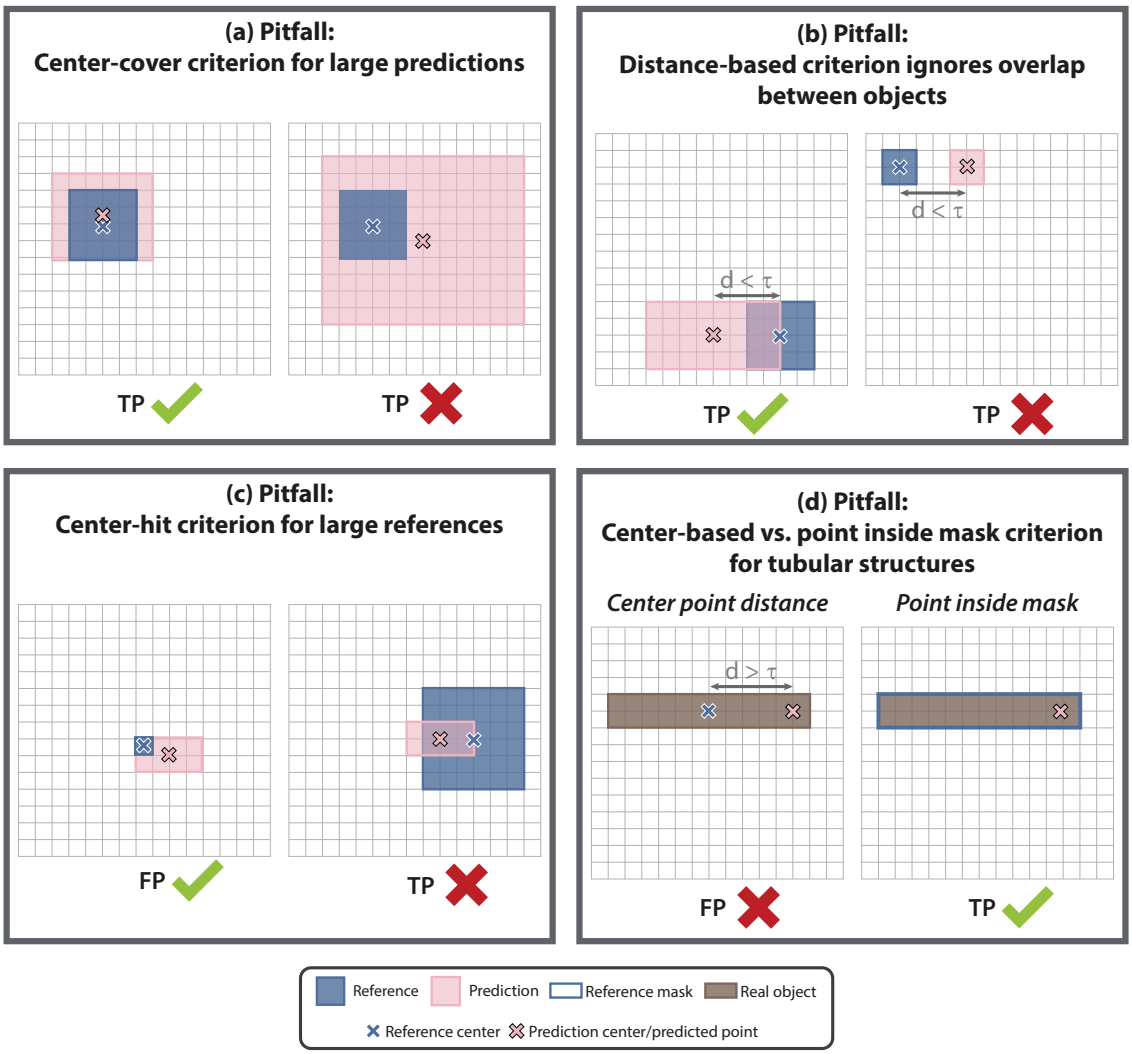


Figure 4.20: Pitfalls related to inadequate metric implementation: Center- or point-based localization criteria. **(a) Effect of large predictions for the center-cover criterion.** The center-cover criterion is more likely to assign large predictions True Positive (TP). **(b) Effect of no overlap for distance-based criteria.** The center distance-based criterion only considers the distance between two center points and ignores whether the prediction actually overlaps the reference object. **(c) Effect of large references for the center-hit criterion.** The center-hit criterion is more likely to assign an object TP if the reference is large. **(d) Effect of tubular structures on the distance-based criterion.** The center distance-based criterion may consider an object as False Positive (FP) for a tubular shape, although the point is inside the tube. Figure adapted from [Reinke et al., 2021a] and [Kooi, 2021b].

Pitfall Subtype 3.4: Inadequate metric reporting

Metric values should be carefully reported, keeping the above-mentioned limitations of metric application in mind. In the following, we present pitfalls specifically related to metric reporting:

Non-determinism of algorithms When reporting metric values, one should be aware of the fact that Artificial Intelligence (AI) algorithms are non-deterministic, i.e. several runs of the same algorithm may yield different results (see Figure 4.21(a)). This effect can be decreased by using fixed seeds. But even when using the exact same architecture, data splits, and random seeds, there is still variation in the metric scores, for example given by non-deterministic layers (e.g. dropout layers) or weight initialization [Pham et al., 2020].

Misleading visualization Visualizing the distribution of metric scores is a simple but powerful tool for analyzing the results of a challenge or even a single algorithm. This could be achieved via boxplots, as shown in the top left example in Figure 4.21(b). However, boxplots do not show the actual distribution of scores. The boxplot hides a cluster of images, for which only very low scores were achieved. This is better shown by either using violin plots (top right) or overlaying the plots with the individual metric scores as dots (bottom left). Yet, those two variants still hide important information and the meaningfulness of the plots can be further increased by color-coding the dots, as in the bottom right part of the figure. This kind of plot is especially useful for hierarchical data, which is not independent, such as videos or images from the same patient. Assuming that the metric scores were provided for images from four videos, only this last plot reveals that all low metric scores were achieved for *Video 4* (pink). Thus, the analysis of results could prompt investigation of why all images from this video were hard to process.

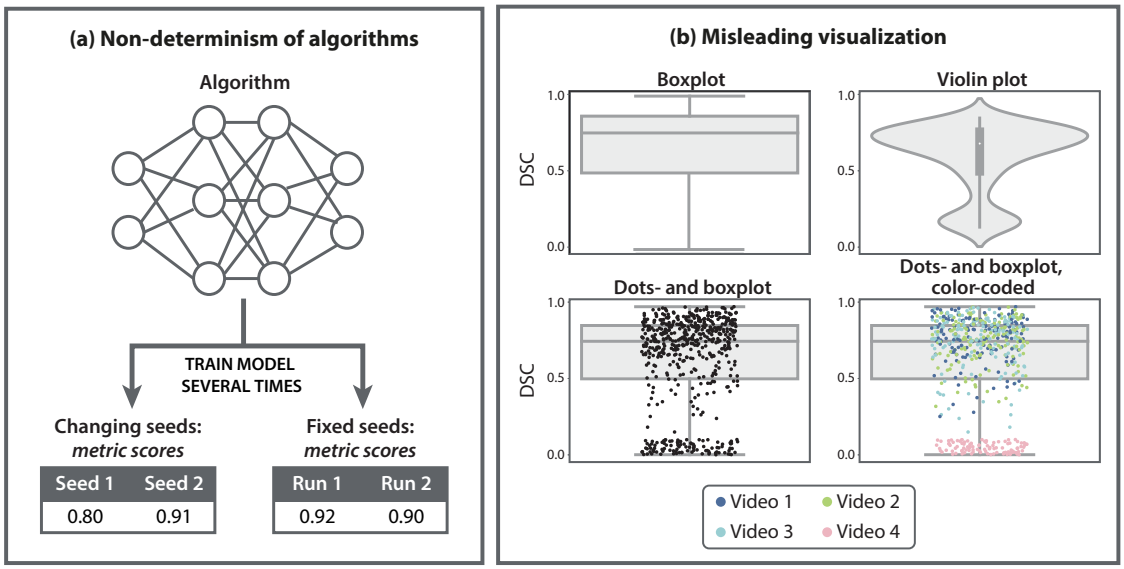


Figure 4.21: Pitfalls related to inadequate metric reporting. **(a) Effect of non-determinism of algorithms.** Artificial Intelligence (AI) algorithms are non-deterministic and yield different values in different runs, although network and data stay constant. Even when fixing random seeds, results still differ slightly. **(b) Effect of misleading visualization of raw metric scores.** Simple boxplots (top left) do not provide information on the distribution of the metric values (here: Dice Similarity Coefficient (DSC)), which would be given by violin plots (top right) or dots- and boxplots (bottom left). Yet, both plots would hide information in the case of hierarchical data and could be improved by color-coding the dots. Figure adapted from [Reinke et al., 2021a].

Pitfall Subtype 3.5: Inadequate interpretation of metric values

Whenever metric values are reported, they should be carefully interpreted, keeping the domain interest and current situation in mind. In the following, we present pitfalls related to metric value interpretation:

Image resolution and dimension The appearance and accuracy of the reference annotation and predictions highly depend on the image resolution, pixel size, and spacing. An example of different resolutions is shown in Figure 4.22(a). The resulting metric scores are different for low and high-resolution images, as the manifestation of the objects changes. This pitfall is also relevant for object detection problems.

Generally, it should be noted that the dimensionality of an image differently affects the metric values of overlap-based metrics. Figure 4.22(b) shows an example of two rectangles or bounding boxes in a 2D and 3D setting. In the 3D case, being off by one voxel in the z-dimension yields a much lower Box/Approximation Intersection over Union (Box/Approx IoU) score than in 2D.

Lack of upper or lower bounds CK examines how closely a prediction matches the distribution of the actual class. The resulting score can be accompanied by the calculation of the *maximum CK*, which represents a corner case where either of FP or FN are equal to zero [Umesh et al., 1989]. The maximum CK score decreases the more the predicted positive and negative sample numbers diverge from the actual positive and negative sample numbers, respectively [Widmann, 2020]. The distribution of positive and negative samples of the left prediction in Figure 4.22(c) closely follows the actual distribution of positive and negative samples (13 circle and 87 triangle predictions; 15 actual circles and 85 actual triangles), yielding a high maximum CK. With a larger difference to the actual number of circles and triangles, such as given by the right prediction, the maximum CK decreases, although the Accuracy and CK values increase.

Relevance of metric differences in rankings When ranking algorithms by aggregated metric scores, one should further note that extremely small differences in the scores may yield different ranks, but the differences may not be relevant from a biomedical perspective. An algorithm which surpasses another by a score difference of only 0.0001 (such as the algorithm in Figure 4.22(d)) may not actually be superior in practice, especially considering the non-determinism of algorithms (see above). Instead, assigning both algorithms a shared rank would be favorable.

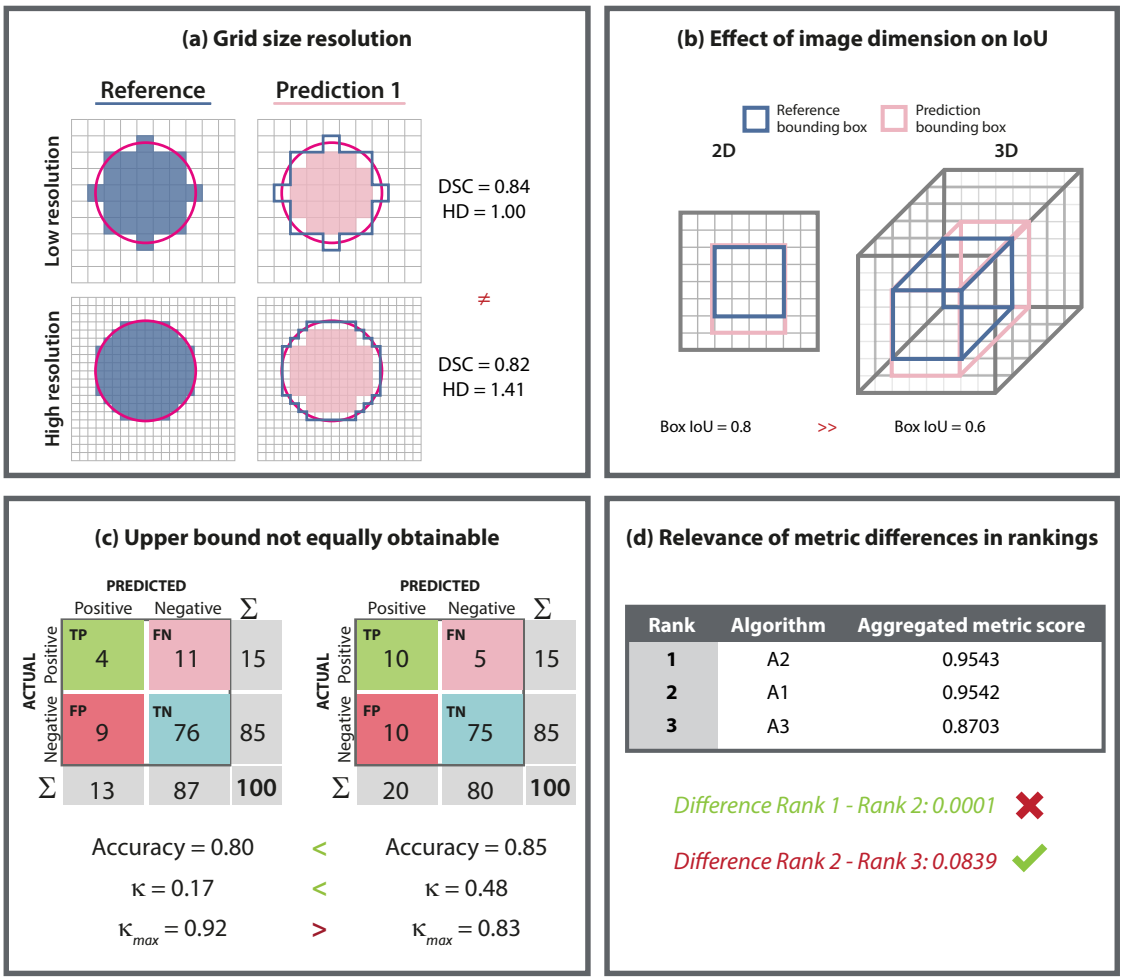


Figure 4.22: Pitfalls related to inadequate interpretation of metric values. **(a) Effect of different image resolutions.** The prediction and the reference annotation (blue shape (reference) vs. pink outline (desired circular shape)) are significantly influenced by variations in the image resolution. Due to the varied resolution, a prediction of the exact same shape yields different Dice Similarity Coefficient (DSC) and Hausdorff Distance (HD) scores. **(b) Effect of image dimension.** The additional z-dimension in 3D settings causes an increase in incorrect pixels that is cubical in size and thus impacts the Box Intersection over Union (Box IoU) more than in 2D settings. **(c) Upper bound not equally obtainable in Cohen's Kappa (CK) (κ).** In comparison to *Prediction 2* with a dissimilar distribution of positive and negative samples to the actual distribution, *Prediction 1* with a similar count of positive and negative samples to the actual distribution yields high maximum CK (κ_{max}) values. This is in contrast to the Accuracy and κ values. **(d) Effect of relevance of metric differences.** When calculating rankings, one should be aware that irrelevant metric differences (here: 0.0001) yield differences in the ranks. Figure adapted from [Reinke et al., 2021a].

Overview of pitfalls related to problem category

Here, we show which of the presented pitfalls apply to the metrics proposed in Section 2.4. We provide an overview of potential pitfalls and the metrics that are affected by the properties for image-level classification and calibration (Table 4.1), semantic and instance segmentation (Table 4.2), and object detection (Table 4.3).

Table 4.1: Overview of pitfalls related to **image-level classification** and **calibration** metrics. For each illustration, the metrics that are affected by the property are marked with a red cross. Abbreviations: Area under the Receiver Operating Characteristic Curve (AUROC), Balanced Accuracy (BA), Brier Score (BS), Cohen's Kappa (CK), Expected Cost (EC), Expected Calibration Error (ECE), Positive Likelihood Ratio (LR+), Matthews Correlation Coefficient (MCC), Net Benefit (NB), Negative Predictive Value (NPV), Positive Predictive Value (PPV) and Youden's Index (J).

Source of potential pitfall	Figure	Accuracy	AUROC	BA/J	BS	CK	EC	ECE	F_β Score	LR+	MCC	NB	NPV/ PPV	Sensitivity/ Specificity
Unequal importance of class confusions	Figs. 4.5(a), (c) and (d)	✗		✗	✗						✗	✗	✗	✗
Missing prevalence correction	Fig. 4.6(a)												✗	
Prevalence dependency	Fig. 4.6(b)	✗				✗	✗*		✗		✗		✗	
Rankings in the case of prevalence dependency	Fig. 4.6(c)					✗					✗			
Neglecting confidence values	Fig. 4.7(a)	✗	✗	✗		✗	✗		✗	✗	✗	✗	✗	✗
Neglecting of benefit-cost analysis	Fig. 4.7(c)	✗	✗	✗	✗	✗		✗	✗	✗	✗		✗	✗
High class imbalance	Figs. 4.10(a) and (b)	✗	✗				✗*						✗	
Misleading scores	Fig. 4.10(b)			✗										
Low sample sizes	Fig. 4.10(c)		✗											
Single cut-off value for multiple classes	Fig. 4.18(a)		✗											
Choice of bins for calibration errors	Fig. 4.18(b)							✗						
Definition of class labels	Fig. 4.18(c)								✗	✗		✗	✗	✗
Upper bound not equally obtainable	Fig. 4.22(c)					✗								

** Depending on the selected parameters, EC can be changed to become prevalence-independent.*

Table 4.2: Overview of pitfalls related to **segmentation** metrics. For each illustration, the metrics that are affected by the property are marked with a red cross. Abbreviations: Average Symmetric Surface Distance (ASSD), Centerline Dice Similarity Coefficient (clDice), Dice Similarity Coefficient (DSC), Hausdorff Distance (HD), Hausdorff Distance 95 Percentile (HD95), Intersection over Union (IoU), Mean Absolute Surface Distance (MASD), Normalized Surface Distance (NSD) and Panoptic Quality (PQ).

Source of potential pitfall	Figure	ASSD	clDice	DSC	F_β Score	HD	HD95	IoU	MASD	NSD/ Boundary IoU	PQ
Boundary unawareness	Fig. 4.4(a)		✗	✗	✗			✗			
Structure center unawareness	Fig. 4.4(b)		✗	✗	✗			✗			
Volume unawareness	Fig. 4.4(c) and (d)	✗				✗	✗		✗	✗	
Preference for over- *vs.* under-segmentation	Fig. 4.5(b)			✗				✗			
Neglecting confidence values	Fig. 4.7(a)	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗
Small size of structures relative to pixel size	Fig. 4.8(a)		✗	✗	✗			✗			
High variability of structure sizes	Figs. 4.8(b) and (c)		✗	✗	✗			✗	✗		
Complex shapes of structures	Figs. 4.9(a) and (b)			✗				✗			
Occurrence of overlapping structures	Fig. 4.9(c)		✗	✗	✗			✗			
High inter-rater variability	Fig. 4.11(a)	✗	✗	✗	✗	✗	✗	✗	✗		✗
Outliers, noise and artifacts in the reference annotation	Fig. 4.11(b)					✗					
Empty reference or prediction	Fig. 4.11(c)	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗
Possibility of overlapping predictions	Fig. 4.12(a)	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗
Sensitivity to hyperparameters	Fig. 4.17(c) and (d)				✗					✗	
Segmentation vs. detection quality	Fig. 4.18(d)										✗
Image dimension	Fig. 4.22(b)		✗	✗	✗			✗			

Table 4.3: Overview of pitfalls related to **object detection** metrics. For each illustration, the metrics that are affected by the property are marked with a red cross. Abbreviations: Average Precision (AP), Boundary Intersection over Union (Boundary IoU), Box/Approximation Intersection over Union (Box/Approx IoU), Center-/point-based localization criterion (Center/Point), Free-Response Receiver Operating Characteristic (FROC), Mask Intersection over Union (Mask IoU), and Positive Predictive Value (PPV).

Source of potential pitfall	Figure	AP	Boundary IoU	Box/Approx IoU	Center/ Point	FROC Score	Mask IoU	PPV	Sensitivity
Small size of structure sizes	Fig. 4.8(a)			✗			✗		
High variability of structure sizes	Fig. 4.8(c)			✗			✗		
Complex shapes of structures	Figs. 4.9(a) and (b)			✗			✗		
Occurrence of disconnected structures	Fig. 4.9(d)			✗					
Empty reference or prediction	Fig. 4.11(c)		✗	✗			✗	✗	✗
Unavailability of predicted class scores	Fig. 4.12(b)	✗				✗			
Disregard of number of images	Fig. 4.15(c)	✗							
Non-standardized metric definition	Figs. 4.17(a) and (b)	✗				✗			
Sensitivity to hyperparameters	Fig. 4.17(c) and (d)			✗			✗		
Predicted class scores neglected within ranking	Fig. 4.19(a)	✗							
Large effects of small changes in predicted class scores	Fig. 4.19(b)	✗							
Mathematical properties of center- or point-based localization criteria	Fig. 4.20				✗				
Image dimension	Fig. 4.22(b)			✗			✗		

4.2.4 Discussion

In this section, we presented a comprehensive compilation of metric limitations that are relevant for image-level classification, semantic segmentation, object detection, and instance segmentation problems. Researchers have rated metric selection as one of the most critical issues related to challenge design (see Chapter 3). Thus, we aimed to present a common access point for metric limitations, grouped by similar properties according to our presented taxonomy. While some of the pitfalls are mentioned in the literature or online resources, we did not find any publication that systematically presented them in a comprehensive way.

In addition, a search may not actually retrieve all relevant results, since there is no standardized format of how pitfalls are described. Terms like 'problem' or 'issue' are used in most research papers to describe the purpose of the presented methodology. Indeed, our search yielded hundreds of thousands of results, making it impossible to review all of them as a practitioner. Even the smallest number of publications (49 for the clDice) would be too much for a researcher to review in depth. Moreover, a metric acronym, such as the ASSD may be used for multiple metrics. For instance, the implementation of this metric in the `medpy` library [Maier, 2013] refers to our definition of the MASD (see Section 2.4), not to ASSD. Finally, several pitfalls were only briefly mentioned in the text or the appendix of publications, thus may be easily overlooked

by practitioners. Only a fraction of the pitfalls we found was presented in an easy-to-follow format, such as a figure. These findings support our hypothesis that it can be challenging for researchers retrieve information on metric limitations within reasonable time an efforts. In many circumstances, retrieval only appears viable if the specific problem and the exact wording in the respective publication are already known, which is naturally not the situation most researchers find themselves in.

We therefore decided to follow a different approach for our pitfall compilation and implemented a multi-stage Delphi process, leveraging the expert knowledge from a large international consortium. During our analysis, we found that 24% of challenges in the period of 2018 to 2022 (see Section 5.1 for details) justified their metric choice by stating that commonly used metrics were chosen. For segmentation tasks, the DSC is by far the most frequently used metric (see Chapter 3). However, in this section, we presented multiple pitfalls regarding this metric, such as a strong penalization of small objects, shape, center, and volume unawareness, or the inability to properly handle inter-rater variability. Similarly, in object detection problems, the AP is a common metric [Lin et al., 2014] which for example suffers from disregarding the total number of images. Moreover, common image-level classification metrics are largely influenced by their prevalence dependency (e.g. PPV and MCC) or are affected by class imbalances (e.g. Accuracy and AUROC). Finally, it should be noted that the metric implementation itself may also heavily impact the metric values. For example, boundary-based metrics rely on the extraction of the boundary from the reference and prediction. This step can be implemented in several ways, which affects the resulting scores. Similarly, the implementation of the AP metric related to corner cases differs for different repositories and is thus not standardized. Given these findings, it becomes clear that choosing a metric for its popularity may not properly reflect algorithm performance and produce meaningful challenge results on the properties of interest, which are determined by the underlying biomedical problem.

Of note, the conclusions from this chapter go beyond challenges. They apply to image analysis in general, irrespective of whether the focus is on biological, radiological, surgical, or even general computer vision problems. In fact, the source of a metric-related limitation or pitfall is independent of the target domain and often based on purely mathematical implications. However, in the biomedical domain, the consequences of choosing an improper metric not reflecting the actual interest may be more critical, since the well-being of patients is directly affected by translating an inappropriate algorithm into clinical practice. We thus argue that a collection of pitfalls is crucial to ensure safe validation of algorithms and to educate researchers to avoid common mistakes. In this thesis, we further provide an overview of common validation metrics in the form of a *metric profile* in Appendix A.3. This profile contains important information on metrics, such in the form of illustrations, explanations, statements on prevalence dependency as well as relevant limitations and recommendations.

4.2.5 Conclusion

Choosing the right metric for a specific problem is not straightforward and being aware of every single limitation of every metric is not possible in practice. With this collection of metric pitfalls, we hope to contribute to educating challenge organizers and researchers on which metrics should be avoided under certain circumstances. Our contribution is structured according to a pitfall taxonomy based on certain properties, so as to make the information easily retrievable. By providing several examples, we emphasized that the choice of a metric is very use case-specific and should be chosen accordingly. Selecting an inappropriate metric may yield misleading and incorrect conclusions for clinical applications. In the following, we show that the choice of metrics also heavily impacts challenge rankings (see Section 4.3) and finally present a problem-aware metric recommendations framework that builds upon the results of this section (see Section 5.2).

4.3 Revealing flaws related to rankings

Disclosure

This work has been supervised by Lena Maier-Hein. It has been conducted together with several co-authors, especially Matthias Eisenmann and Annette Kopp-Schneider. The work has been published in *Nature Communications* [Maier-Hein et al., 2018]. Please refer to Chapter A.1 for full disclosure.

4.3.1 Introduction

The goal of biomedical image analysis challenges is typically to identify the best algorithm for solving a specific biomedical problem. It is common practice to calculate one or multiple rankings to determine the winner of a challenge. Since there is only one score for each participant in a validation over the whole data set (see Section 2.4 and Figure 2.21), the winner can be determined directly from the metric scores. When the performances are validated per image, there is one metric score for each image for every participant in the challenge. To establish a ranking that orders the challenge participants according to their performance, the metric scores must be combined based on a predetermined aggregation scheme. Although the intuitive ranking scheme would be to average all metric scores per participant, there are several other possibilities to calculate a ranking (see Section 2.1.2). In this section, we focus on the *metric-based ranking scheme* (see Algorithm 1 for details), which is the most commonly used ranking method (see Chapter 3), and the *case-based ranking scheme* (see Algorithm 2 for details). Both ranking schemes rely on an aggregation of either metric values or ranks. Thus, an aggregation operator needs to be chosen, such as the mean or median.

A challenge winner should ideally remain the same no matter how the ranking scheme is changed and should be superior to the other algorithms regardless of the ranking method, aggregation operator, or other parameters used. If this is not the case, it is difficult for the organizers to decide which algorithm is truly the best at achieving the challenge's objective. This is crucial since challenge winners often receive much attention, will almost certainly replace an existing state-of-the-art method, acquire many citations, and may have their algorithms incorporated into clinical procedures.

However, in Section 4.1, we showed that challenge rankings can be manipulated easily to change the winner. In the present section, we further examine the sensitivity of challenge rankings to a range of design parameters. We specifically examine how frequently the winner changes when the ranking scheme, aggregation operator, metrics, and annotators are altered by using data from actual biomedical image analysis challenges. In addition, we explore how often non-winning algorithms could achieve the first rank for those perturbations.

Hypothesis investigated in this chapter

H3: Challenge rankings are highly unstable.

Considering the importance of challenge rankings in determining a challenge winner, caution should be applied when selecting the ranking schemes. We believe that rankings are highly influenced by the chosen calculation scheme. Thus, we investigate whether challenge rankings remain stable when the ranking scheme changes.

4.3.2 Methods

In this section, we build upon the results of Section 4.1 and analyze the robustness of rankings in more depth. For this purpose, we use the same challenge inclusion criteria as presented earlier. Namely, we considered all Medical Image Computing and Computer Assisted Interventions (MICCAI) 2015 segmentation challenges for a detailed analysis, yielding 13 challenges with 124 tasks. Similarly, for the ranking analyses using bootstrapping approaches (ranking robustness analysis; see below), we only considered challenge tasks with three or more participating algorithms (42 tasks excluded) and more than one test case (25 tasks excluded), yielding 56 challenge tasks. The experimental setup for the ranking robustness analysis is described in the following.

Ranking robustness analysis

For the ranking robustness analysis, we computed the twelve ranking schemes presented in Section 4.1 for the 56 MICCAI 2015 segmentation tasks meeting the inclusion criteria. For challenge tasks for which multiple annotators generated the reference segmentation, we additionally analyzed the effect of the annotator on the (default) ranking scheme. In order to calculate the differences between different ranking methods, we employed Kendall's τ rank correlation coefficient, presented in Section 2.2.1 (Equation 2.1). Kendall's τ is bounded between -1 (reverse ranking) and 1 (identical ranking), with 0 indicating no correlation between rankings.

To investigate whether specific ranking choices are more robust than others, we defined the robustness as a function of (1) the metric applied, (2) the aggregation method, and (3) the aggregation operator. For different ranking configurations, we simulated new rankings using a bootstrapping approach with 1,000 bootstrap samples. Thus, we investigated ranking variability against small perturbations within the test data set (cf. Section 2.2.2). We analyzed the robustness of every challenge task in terms of the percentage of bootstrap samples in which (a) the original winning method remained unchanged and (b) the original non-winning algorithms became the winner. For the same bootstrap sample, we computed different ranking schemes. In addition, we performed a Wilcoxon signed rank test with a significance level of $\alpha = 5\%$ for the proportion of challenge task winners staying robust. Tasks with more than one winner were omitted from this analysis.

Finally, we conducted a leave-one-out analysis in addition to the bootstrap simulations. To this end, the original ranking scheme was computed for all challenge tasks. For every task, the test data set was reduced by one case and the original rankings were re-computed. Similarly to the

simulations above, we considered the robustness as the percentage of bootstrap samples in which (a) the original winning method remained unchanged and (b) the original non-winning algorithms became the winner.

4.3.3 Results

In the following, we present the results of the ranking analyses for 56 MICCAI 2015 segmentation tasks, for which we inspected the robustness of rankings to several design choices. We use the Dice Similarity Coefficient (DSC), Hausdorff Distance (HD), and Hausdorff Distance 95 Percentile (HD95) for the ranking robustness analysis. Given the three metrics, we further consider the metric-based and case-based ranking schemes as well as mean and median as aggregation operators. This results in twelve different ranking schemes that were calculated for each of the 56 MICCAI 2015 segmentation tasks.

The comparison of the resulting rankings for different ranking design choices for all tasks is shown in Figure 4.23 in form of dots- and boxplots over Kendall's τ values. Large variability in the ranking can be seen for every change in the ranking scheme. Although a high number of tasks received Kendall's τ values of 1 and therefore showed identical rankings (49% of tasks for HD *vs.* HD95 in Figure 4.23(a); 34% of tasks for metric- *vs.* the case-based rankings in Figure 4.23(b); 36% of tasks for mean *vs.* median aggregation in Figure 4.23(c), and 60% of tasks for the different annotators in Figure 4.23(d)), the majority of tasks yielded a lower τ, indicating some (radical) changes in the rankings, as exemplarily shown in Figure 4.1 in Section 4.1. In one of the challenges, when switching between the HD and HD95 metrics, the last-ranked algorithm (rank 10) for one metric was the winner for the other.

As we show in Section 4.4, 62% of all challenge tasks did not indicate the exact number of annotators. Therefore, in addition to the general ranking design choices, we examined the influence of the annotator producing the reference annotations to which the participant's methods were compared. For challenge tasks with multiple annotators (n = 4), we compared the resulting ranking pairwise for each annotator to the others for all three metrics. It can be seen that the differences in rankings for the DSC were only marginal with a median Kendall's τ of 1 (Interquartile Range (IQR): (1, 1); minimum: 0.8). For the HD and HD95, however, the differences in the annotations had a larger effect on the rankings, leading to lower Kendall's τ values and a larger IQR, namely to a median of 1 (IQR: (0.33, 1); minimum: -0.02) for the HD and a median of 0.8 (IQR: (0.33, 1); minimum: 0.07) for the HD95. The winning algorithm in the metric-based ranking with the mean was changed in 15% (DSC), 46% (HD), and 62% (HD95) of tasks by changing the annotator.

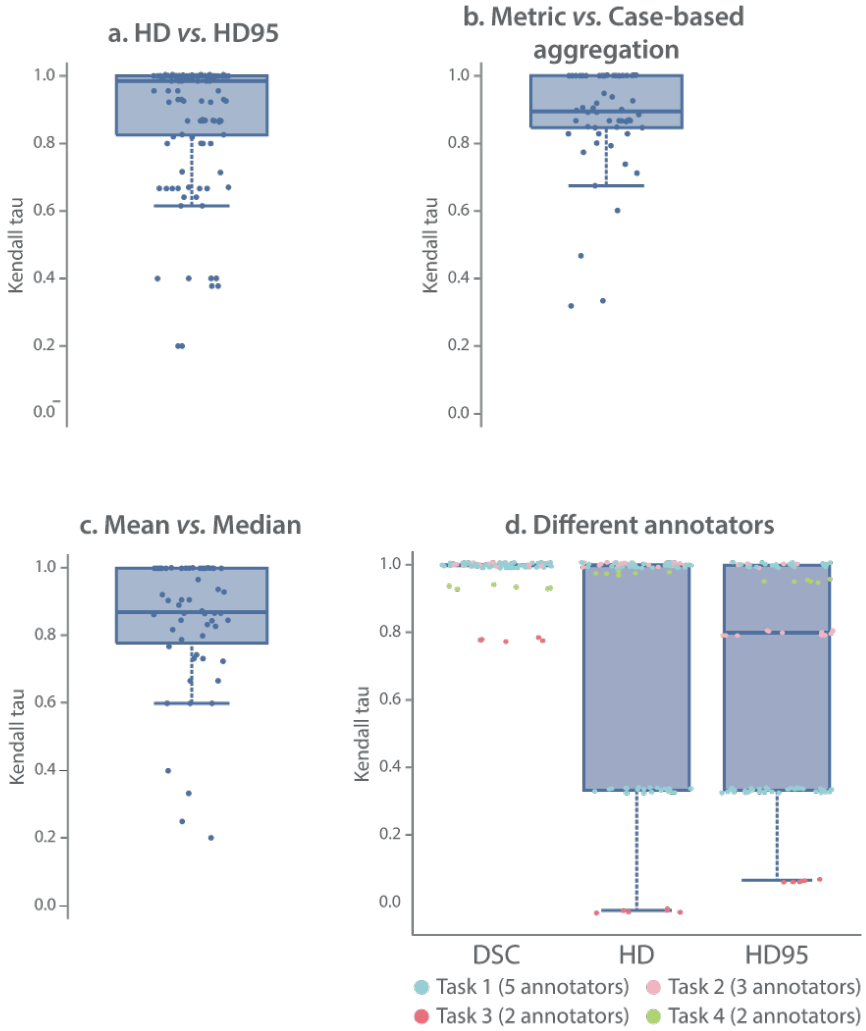


Figure 4.23: Ranking robustness for several ranking design changes for the MICCAI 2015 segmentation task (n = 56). Dots- and boxplots for which one data point represents Kendall's τ between two rankings for one task. **(a)** Kendall's τ values for the HD *vs.* the HD95 for a metric-based ranking with mean aggregation. **(b)** Kendall's τ values for the metric- *vs.* the cased-based ranking with mean aggregation for the HD. **(c)** Kendall's τ values for the mean *vs.* the median aggregation for a metric-based ranking for the HD. **(d)** Kendall's τ values for the comparison of rankings based on different annotators for the four tasks with multiple annotators. Metric-based rankings aggregated with the mean were compared for the DSC, HD, and HD95 metrics. Figure adapted from [Maier-Hein et al., 2018].

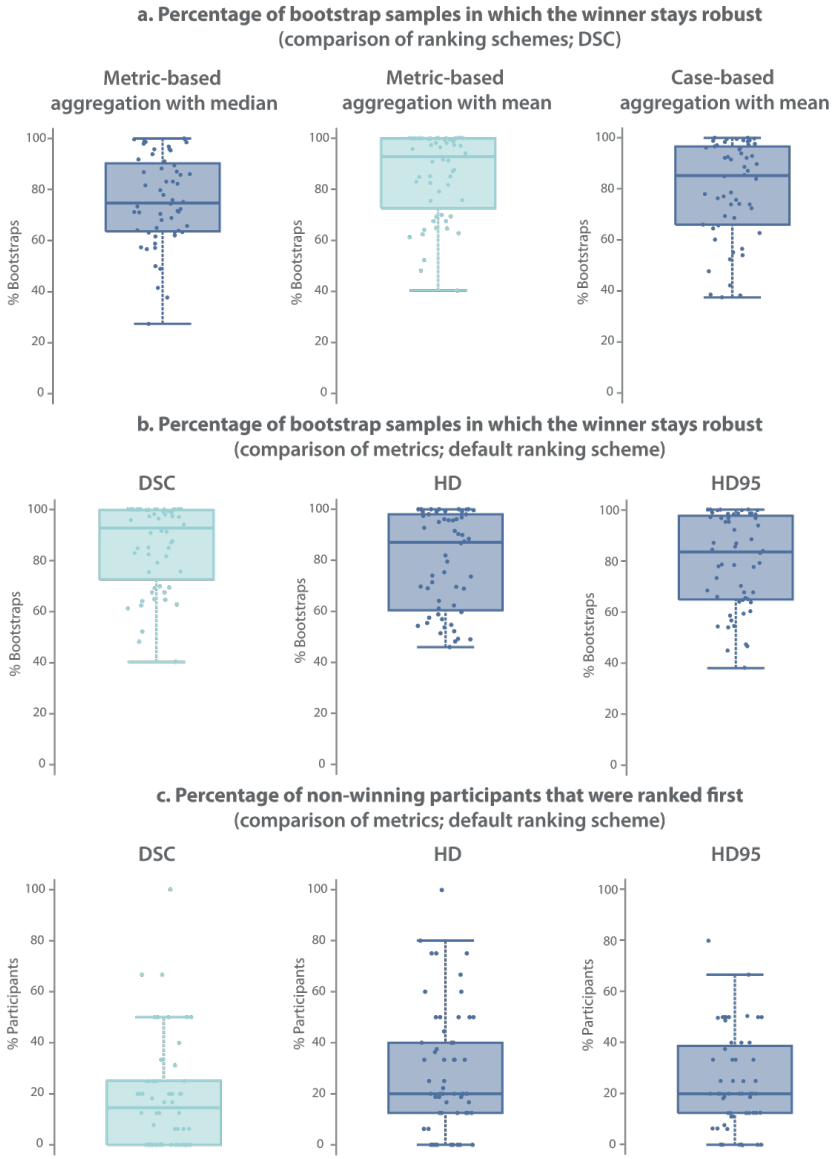


Figure 4.24: Ranking robustness shown for the MICCAI 2015 segmentation task (n = 56). Dots- and boxplots for which one data point represents the percentage value for one task. The most robust setting per row is marked in light blue. **(a)** Percentage of bootstrap samples for which the task's winner does not change. In this setting, three ranking schemes were calculated for the DSC metric: metric-based ranking aggregated with the median (left), metric-based ranking aggregated with the mean (middle), and case-based ranking aggregated with the mean (right). **(b)** Percentage of bootstrap samples for which the task's winner does not change. In this setting, three ranking metric-based ranking schemes aggregated with the mean were calculated for the DSC, HD and HD95 metrics. **(c)** Percentage of non-winning participants who were ranked first. In this setting, three ranking metric-based ranking schemes aggregated with the mean were calculated for the DSC, HD, and HD95 metrics. Figure adapted from [Maier-Hein et al., 2018].

We examined the ranking uncertainty by utilizing bootstrapping techniques and found that the rankings heavily depend on the test data set. Figure 4.24 shows the percentage of bootstrap samples in which the winner remained unchanged ((a) and (b)) and the percentage of non-winning algorithms that were ranked first in at least one bootstrap sample (c). The most robust configuration is highlighted in green. We compared different ranking configurations for the DSC metric with these simulations (Figure 4.24(a)), and found that the *metric-based aggregation with the mean* was much more robust compared to the other ranking schemes, resulting in a median of 93% (IQR: (74%, 100%); minimum: 40%) of robust winning methods. For the *metric-based aggregation with the median*, a median of only 75% (IQR: (64%, 90%); minimum: 28%) of winning methods was robust and a median of 85% (IQR: (67%, 97%); minimum: 38%) of winning methods was robust for *case-based aggregation with the mean*.

The same comparison was performed for the default ranking scheme but with different metrics (Figure 4.24(b)). In this case, the DSC again was more stable compared to the other metrics (same ranking scheme and median (IQR; minimum) as above). For the HD/HD95, a winning algorithm remained unchanged in a median of 85%/83% (IQR: (63%, 97%)/(65%, 97%); minimum: 46%/38%) of bootstrap samples.

Similarly, the percentage of non-winning algorithms that could have been ranked first was determined in the bootstrap simulations. Again, the DSC was more robust in this setting, with a median of only 15% (IQR: (0%, 25%); maximum: 100%) of non-winning algorithms becoming the winner. This number was higher for the HD/HD95 with a median of 20%/20% (IQR: (11%, 40%)/(13%, 38%); maximum: 100%/80%) of non-winning participants that were ranked first in at least one bootstrap sample.

Finally, in addition to the bootstrap simulations, we applied a leave-one-test-case-out approach to the ranking analysis. In this setting with removing one test case, 16% of non-winning algorithms were winners for the DSC. For one of the tasks employed, 67% of non-winning algorithms were ranked first.

4.3.4 Discussion

In this section, we provided evidence for our observations presented in Section 4.1, indicating that a variety of design parameters have a substantial impact on challenge rankings. Even small changes to the ranking scheme may dramatically change the rankings. In one challenge, for example, the last-ranked algorithm became the winner when a different metric was used. High variability was not only caused by the ranking scheme itself, but by also the annotator based on which the reference standard was generated, where changes hugely impacted rankings using distance-based metrics (HD(95)). This is probably due to the fact that even expert raters often disagree on an object's boundary [Joskowicz et al., 2019].

We further showed that a bootstrapping analysis reveals important and interesting insights into ranking robustness across challenges. For most of the investigated challenge tasks, the winner was not robust against small variations of the underlying data set or ranking scheme. We found the metric-based aggregation with the mean to be the most robust ranking scheme. However, this finding should be treated with caution. Case-based ranking schemes offer a straightforward missing value handling by assigning algorithms the last rank for missing submissions. This is

not as easy for metric-based aggregation schemes. For metric-based aggregation, a missing value-handling strategy needs to be determined before the start of the challenge and communicated accordingly. The chosen strategy can make a huge difference in the resulting aggregates (see Figures 4.14(b) and (c) and the experiments in Section 4.1). Surprisingly, rankings based on the mean were more robust than rankings based on the median, although the mean naturally includes outliers.

In comparison with the HD and HD95, the DSC yielded much more robust rankings. This finding should be viewed with caution. While the DSC is more robust for a ranking, it still comes with several limitations (see Section 4.2) and it should only be chosen for an appropriate problem, optimally in combination with other metrics, such as the HD(95). Generally, the metric choice for a ranking should depend on the underlying biomedical problem. For example, if object boundaries are of particular interest, a distance-based metric such as the HD(95) should still be chosen for the rankings. In this case, it is of particular importance to analyze the ranking variability.

Furthermore, the results of this section show that the general challenge design has a high influence on the resulting rankings. For challenges in which many participants yield quite narrow results, this opens up the question of whether a ranking generally makes sense or if we should announce a group of algorithms as winners. The sensitivity of rankings to its calculation scheme also hampers identifying why a specific algorithm was better than others. Finally, we may ask ourselves if challenge design may have a larger influence on determining a winner than the actual algorithms.

4.3.5 Conclusion

Challenge rankings heavily rely on the challenge design, especially the choice of metrics, ranking schemes, and annotators. Thus, rankings should be very carefully chosen and analyzed. Advanced analyzing techniques for rankings are be investigated in Chapter 5 of this thesis.

Since rankings are one of the most prominent building blocks of a challenge, it was an important step to raise the the research community's awareness of highly relevant related issues. Ranking sensitivity can be easily exploited, as we showed in Section 4.1. We should therefore not underrate the final choice of a ranking scheme and make sure that ranking uncertainty is analyzed in challenge reports. If not, an inadequate algorithm may end up being translated into clinical practice due to an improper assessment.

4.4 Revealing flaws related to reporting and analyses

Disclosure

This work has been supervised by Lena Maier-Hein. It has been conducted together with several co-authors, especially Matthias Eisenmann, Annette Kopp-Schneider, and Georg Grab. Parts of this work have been published in *Nature Communications* [Maier-Hein et al., 2018]. Please refer to Chapter A.1 for full disclosure.

4.4.1 Introduction

Biomedical image analysis challenges can be seen as the Artificial Intelligence (AI) algorithm equivalent to clinical trials which measure how well new medications or treatments perform compared to others. Clinical trials follow a strict quality control process with transparency and reproducibility of results being crucial for ensuring high-quality research [Baker, 2016; McNutt, 2014; Schulz et al., 2010]. Both aspects are important criteria for accepting clinical trials and their publications, which is in turn important for translating the medications or treatments into clinical practice.

> *'Trust and confidence are critical to the success of health care models. There are two main methods for achieving this: transparency (people can see how the model is built) and validation (how well the model reproduces reality).'*
>
> — [Eddy et al., 2012]

Eddy et al. [2012] contend that the most crucial factors in establishing trust and confidence in biomedical models are validation and transparency. In the course of this thesis, we showed that challenges are of tremendous importance for biomedical image analysis research but we also revealed several critical flaws related to cheating and the validation of challenges. In this section, we focus on the transparency aspect, especially regarding the provided general challenge details by the organizers and the submitted reports of the methods by the participants.

Since challenges are validation studies, challenge organizers should describe their design and organization transparently to the public. This means that they should describe all facets of the challenge, including more obvious parameters such as the challenge's main objective, the number of training and test cases, or the choice of the metric. However, there are also less obvious parameters that are important, for example, whether the provided training data could be complemented with other public or private data or who annotated the data. In the first case, it could make a huge difference in performance if one challenge participant would have access to additional several hundreds of images to train their algorithm. Other participants with fewer resources could heavily suffer in terms of performance. If not transparent, the challenge organizers could not decide if an algorithm won the challenge because of their model or pipeline or just because of the higher amount of training data or compute power. In the second example of who annotated the data, it could make a huge difference in the reliability of challenge results if expert

raters or professional annotation companies generated the reference annotation or untrained students or crowdworkers. Similarly, the number of annotators and the quality of provided labeling instructions may hugely impact the quality of the reference annotations [Joskowicz et al., 2019; Rädsch et al., 2022]. Specifically, challenge organizers should be transparent in the label-generation process. In this section, we review the challenges that were analyzed in Chapter 3 of this thesis and examine the degree of completeness and transparency of reporting challenge design parameters.

As described in previous sections, challenge winners receive large attention and often end up being the new state-of-the-art method for a specific problem. It is therefore reasonable to reuse the top-ranked challenge algorithms for similar problems and questions. Thus, the participant's method descriptions should be as detailed and transparent as possible to ensure a direct implementation. The best-performing algorithms are frequently described in challenge reports. Occasionally, these usually short explanations of the methods are supplemented by longer and more detailed descriptions in the appendix or on the challenge website. This would be the only source of knowledge for a reimplementation if there were no other method-specific publications or public source codes. As an example, we reimplemented the algorithms that participated in the Robust Medical Instrument Segmentation (RobustMIS) challenge 2019 based on the information provided by the challenge participants. We compare the original rankings with the rankings based on the reimplementation to explore whether their description was sufficient for a proper re-creation of the challenge results.

> **Hypothesis investigated in this chapter**
>
> *H4: Challenge results are not reproducible.*
>
> Transparency is crucial to ensure the reproducibility and interpretability of challenge results. However, based on previous findings, we hypothesize that challenge results and algorithms are not reproducible. We investigate whether (1) challenge organizers provide sufficient details and (2) whether challenge results are reproducible given the participants' submitted method descriptions.

4.4.2 Methods

Acquisition of challenge design details

In this experiment, we examined the question of whether challenge designs are reported with sufficient detail to ensure the transparency and reproducibility of the results. For this purpose, we reviewed a critical amount of challenges (150 challenges and 549 tasks) and analyzed how many important parameters were reported by the challenge organizers. The structured capturing of the challenges was based on a challenge parameter list, as described below. We instantiated the parameters for all collected challenges. Based on this instantiation, we could derive descriptive statistics such as the information for how many challenges and tasks each of the parameters were reported or the median number of reported parameters per challenge.

Challenge inclusion criteria Similar to our experiments from Chapter 3, we collected all biomedical image analysis challenges that were organized between 2004 and 2016, yielding a total of 150 challenges and 549 tasks.

Structured challenge parameter list For a structured review and comparison of challenges, we introduced a challenge parameter list, based on the work of Jannin et al. [2006] (see Section 2.5). We refined this initial list, which was originally designed for general validation studies, by adding challenge-specific parameters, which we noted as important while scanning through the 150 challenges. Subsequently, we provided the pre-final list to all co-authors of the study presented by Maier-Hein et al. [2018]. To achieve a consensus decision, we prepared a survey in which every single parameter was presented to the co-authors, asking for agreement with the parameter's name, an explanation as well as constructive feedback in case of disagreement. We required the co-authors to indicate the importance of each parameter (essential, important, optional) and to add exemplary instantiations. Lastly, they were allowed to propose additional parameters. Based on the survey feedback, we finalized the challenge parameter list and translated it into the `Eclipse Meta Model` introduced in Chapter 3.

Reproduction of challenge results given submitted reports

In a second experiment, we explored whether the challenge results are reproducible if we reimplement the participating algorithms based on their submitted method descriptions. This was done exemplarily for the RobustMIS challenge, which is briefly described in the following paragraph. The concrete experimental design is described subsequently.

Robust medical instrument segmentation (RobustMIS) challenge For the reimplementation, we chose a challenge that we organized ourselves, the RobustMIS challenge, such that we had access to the submitted method descriptions of the challenge participants. This challenge aimed to compare the robustness and generalizability of algorithms for medical instrument instance segmentation, which was assessed via two different ranking schemes (see below). We reimplemented the seven methods that participated in the challenge. A more detailed description of the RobustMIS challenge design can be found in Section 5.3 and Appendix A.6.

Only challenge participants who submitted a method description along with their algorithm were eligible for the final challenge rankings. Authors were provided with a method write-up template[1] to ensure that all relevant information was integrated into their descriptions. Participants were required to state their main motivation for the chosen approach, benefits over other common methods, and the (potential) novelty of the method. They were further asked give a detailed description of the submitted methods in a way that an independent researcher would be able to reproduce the results. We specifically asked them to describe all variables used in the model, to describe methods used for data filtering, processing, or normalization, and to fully specify the concrete algorithm architecture. The challenge participants were asked to submit a short paper based on the given template along with their results. However, they were not explicitly forced to provide all information, thus, some fields could have been missing or could have been only described very briefly.

[1] https://www.synapse.org/#!Synapse:syn18779624/wiki/592664

For the final publication [Roß/Reinke et al., 2020], we contacted the challenge participants and asked them to provide further details about their methods since we found many details to be missing. Specifically, we asked for clarification on whether temporal and/or additional training data was used, to describe all pre-processing steps, the network architecture, loss functions, hyperparameters, data augmentation techniques, optimizer, and training procedures.

In the original challenge, participants were validated based on the Multi-Instance Dice Similarity Coefficient (MI DSC) metric, defined as the Dice Similarity Coefficient (DSC) per instance. Two ranking schemes were employed: a) the accuracy ranking and b) the robustness ranking. The accuracy ranking refers to a test-based ranking scheme (see Section 2.1.2) based on a Wilcoxon signed rank test with a significance level of $\alpha = 0.05$. A metric-based ranking (see Section 2.1.2) served as the robustness ranking with the 5% percentile serving as the aggregation operator. This ranking aimed to reflect the worst-case performance.

Experimental design The seven submitted method descriptions of the participating algorithms along with additional information acquired from the challenge participants served as a basis for reimplementation. We followed the method descriptions as closely as possible. For example, in the case that concrete programming languages or libraries were provided, we also used those for our reimplementation. In the case of missing information or ambiguities in the method descriptions, we applied a three-stage procedure:

1. We inspected cited literature to find out if missing or ambiguous information was provided in those references.

2. If we could not find the information in the cited literature, we used secondary literature related to the task or network architecture.

3. Whenever 1. and 2. were not feasible, we reviewed publicly available implementations of similar approaches and selected the most popular approach that appeared to work reasonably well for the problem domain.

In the case of multiple possible interpretations, we implemented all of them and took the one with the best validation performance. All participating methods were reimplemented based on this approach and trained on a single Nvidia RTX 2080 Graphics Processing Unit (GPU), except for teams *Casial SRL* and *fisensee*, which required a high amount of GPU memory and were therefore trained on a single Nvidia V100 GPU. The two ranking schemes employed in the RobustMIS challenge were also computed for the reimplementation. To investigate ranking uncertainty, we calculated 1,000 bootstrap rankings for both ranking schemes and computed Kendall's τ between the original and the bootstrap ranking samples. The ranking computation was done with the `R` package *challengeR* [Wiesenfarth et al., 2021] (see Section 5.3). The participating methods based on their descriptions, missing information, and ambiguities we found during the reimplementation can be found in Appendix A.5. In cases of ambiguities, our assumptions made are explained. A summary of the most important design choices that could be retrieved from the method descriptions is provided in Table 4.4. It should be noted that even for the basic information in the table, not all fields could be filled.

Table 4.4: Summary of the most important design choices that could be retrieved from the method descriptions for the Robust Medical Instrument Segmentation (RobustMIS) challenge participants. Used abbreviations: Adaptive Moment Estimation (Adam), Binary Cross Entropy (BCE), Cross Entropy (CE), Dice Similarity Coefficient (DSC), Residual Neural Network (ResNet), Region Proposal Network (RPN), Stochastic Gradient Descent (SGD). A detailed description of all methods and assumptions drawn from the descriptions is provided in Appendix A.5. Table adapted from Roß/Reinke et al. [2020].

Team	Basic architecture	Usage of videos	Loss functions	Data augmentation	Optimizer
caresyntax	Mask R-CNN, backbone: ResNet-50	No	Smooth L_1 loss, CE loss, BCE loss	Applied in each epoch with 50% probability: horizontal random flip	SGD
CASIA SRL	U-Net with additional **unclear** components	No	CE loss $-\alpha \log(\text{Jaccard loss})$	Once before training: Random rotation, shifting, flipping	Adam
fisensee	U-Net with residual encoder	No	DSC and CE loss	Randomly applied for each batch: Rotation, elastic deformation, scaling, mirroring, Gaussian noise, brightness, contrast, gamma	SGD
SQUASH	Mask R-CNN, backbone: ResNet-50	Yes	ResNet-50: Focal loss, Mask R-CNN: Mask R-CNN loss and CE loss	Classification (35% of images): gaussian blur, sharpening, gamma contrast enhancement; additional 35% of images: mirroring; minority class: horizontal translation	SGD
Uniandes	Mask R-CNN, backbone: ResNet-101	Yes	Mask R-CNN losses	Applied for each batch: Random horizontal flips, propagating annotation backwards to previous video frames	SGD
VIE	Mask R-CNN, backbone: ResNet-50	Yes	RPN class loss, Mask R-CNN losses	Applied for each batch: image resizing, bounding boxes, label generation	**Unclear**
www	Mask R-CNN, backbone: ResNet-50	No	Smooth L_1 loss, focal loss, BCE loss	Applied for each batch: horizontal and vertical random flip, rotations of [0,10]°	Adam

4.4.3 Results

Acquisition of challenge design details

Nichols et al. [2017] identified transparent reporting as an integral part of a validation study. Therefore, we built a list of 53 important challenge design parameters, which cover the topics challenge organization (eight parameters), participation conditions (seven parameters), mission of the challenge (six parameters), study conditions (seven parameters), challenge data sets (fifteen parameters), assessment method (seven parameters), and challenge outcome (three parameters) [Maier-Hein et al., 2018]. The corresponding parameters can be found in Table 4.5. The following numbers are presented for the challenges organized in the period of 2004 to 2016.

From the 53 parameters, a median of 62% (Interquartile Range (IQR): (51%, 72%); minimum: 21%) was reported per challenge task, from which only three parameters (6%) were reported for all tasks, namely the challenge name, life cycle type, and task category(ies). 43% of parameters

were reported by less than half of the challenge tasks, from which 9% of parameters have been reported by less than 10%. Those rarely reported parameters were the operator(s) and uncertainty handling (7% of tasks), the organizer participation policy and statistical tests (6% of tasks), and the pre-evaluation method (5% of tasks). In the following, we provide examples of incomplete reporting in the context of training data and pre-processing, annotation details, and metric-based ranking.

58% of challenge tasks made no mention of whether the organizers' training data may have been supplemented with data from other public or private sources. In 75% of tasks, it was unclear whether data pre-processing was performed and for those tasks that reported data pre-processing, this information was mostly given on a very basic level (82%). The generation of the reference annotations was not explained in 66% of challenge tasks and 62% did not mention how many annotators produced the reference annotation. The expertise and number of the annotators were not contributed in 19% of tasks and for the tasks, for which more than one person annotated the same cases of the data set, 45% did not emphasize how the results were aggregated or merged (e.g. majority vote). Inter- and intra-rater-variability was reported for 27% of challenge tasks. Sources of errors in the annotation were not reported in 84% of tasks.

8% of challenge tasks using multiple metrics for the final ranking did not explain how the results were aggregated. The ranking scheme itself was not published for 20% of the tasks. Finally, we asked all Medical Image Computing and Computer Assisted Interventions (MICCAI) 2018 challenge organizers whether they published the full ranking scheme before the challenge took place and 40% of them did not. In the ranking calculations, 82% of challenge tasks did not announce how to deal with missing values in the submissions. Ranking uncertainty analyses were not reported in 94% of challenge tasks. Finally, the number of allowed (re-)submissions per algorithm was not reported in 67% of tasks.

Reproduction of challenge results given submitted reports

Based on the experimental design described in Section 4.4.2, we reimplemented all participating teams of the RobustMIS challenge, trained them on the challenge training data set, and validated them similarly to the original challenge. In the following, we provide an analysis of the missing information and the results of the reimplementation compared to the original challenge results.

Analysis of missing information Based on the missing information and ambiguities described in the above paragraphs, we summarized the shortcomings of the method descriptions for a qualitative summary, as depicted in Table 4.6 and Figure 4.25. The colors indicate how severely the specific issue may impact the final ranking outcome and the confidence with which assumptions have been taken in the reimplementation[2]. A minor deficiency refers to an assumption that had to be made due to missing or blatantly incorrect information, but this assumption was expected to have a minor impact on model performance or there was high confidence that the correct assumption was derived given context. On the other hand, major deficiencies were described as missing design decisions that were either assumed to have a significant impact on final model performance, or where there was little confidence that the proper assumption was made from context, or where the context was completely lost.

[2]Please note that the categorization is based on our subjective indication.

Table 4.5: List of parameters important for reporting a biomedical image analysis challenge, including the coverage of 549 tasks reporting each parameter in percent. The coverage columns are further highlighted by color-coding: Red: 0–19%, light red: 20-49%, light blue: 50–79%, green: 80-100%. Parameters used in the structured challenge submission system are highlighted with an asterisk. Table adapted from [Maier-Hein et al., 2018].

Parameter name	Coverage [%]	Parameter name	Coverage [%]
Challenge organization			
Challenge name*	100	Challenge venue or platform	99
Challenge website*	99	Challenge schedule*	81
Organizing institutions and contact person*	97	Ethics approval*	32
Life cycle type*	100	Data usage agreement	60
Participation conditions			
Interaction level policy*	62	Evaluation software	26
Organizer participation policy*	6	Submission format*	91
Training data policy*	16	Submission instructions	91
Pre-evaluation method	5		
Mission of the challenge			
Field(s) of application*	97	Algorithm target(s)*	99
Task category(ies)*	100	Data origin*	98
Target cohort*	65	Assessment aim(s)*	38
Study conditions			
Study cohort*	88	Acquisition device(s)	25
Context information*	35	Acquisition protocol(s)	72
Center(s)*	44	Operator(s)	7
Imaging modality(ies)*	99		
Challenge data sets			
Distribution of training and test cases*	18	Data pre-processing method(s)	24
Category of training data generation method*	89	Category of test data generation method*	87
Number of training cases*	89	Number of test cases*	77
Characteristics of training cases*	79	Characteristics of test cases*	77
Annotation policy for training cases*	34	Annotation policy for test cases*	34
Annotator(s) of training cases*	81	Annotator(s) of test cases*	78
Annotation aggregation method(s) for training cases*	30	Annotation aggregation method(s) for test cases*	34
Potential sources of reference errors	28		
Assessment method			
Metric(s)*	96	Missing data handling*	18
Justification of metric(s)*	23	Uncertainty handling*	7
Rank computation method*	36	Statistical test(s)*	6
Interaction level handling*	44		
Challenge outcome			
Information on participants	88	Report document	74
Results	87		

During our experiment, we found most of the shortcomings in the model selection and data augmentation, for which we identified deficiencies for six out of seven teams. In addition, we found the data split, model architecture, and data pre-processing subject to be the major issues. The optimizer was the least critical part of the method description.

Rankings and performance analysis The raw metric value distribution of participating teams is shown in Figure 4.26. The original scores are shown in dark blue, whereas the scores from the reimplementation are represented in green. It can be seen that the general performance of the reimplemented methods was substantially worse compared to the original scores. This is especially apparent for team *VIE*. Only the performance for team *Uniandes* was similar to the original metric scores, as this team described their method in sufficient detail.

In the original RobustMIS challenge, two ranking schemes were calculated. Table 4.7 presents the accuracy ranking for the original challenge (a) and the reimplementation (b). It can be seen that the winner and runner-up algorithms changed their ranks after the reimplementation. Generally, only one algorithm stayed robust (team *www*), while all other methods changed their ranks compared to the original ranking with a Kendall's τ of 0.59. The mean change in rank was one rank for this ranking scheme.

The results of the robustness ranking are provided in Table 4.8 for the original challenge (a) and the reimplementation (b). The winner also changed for this ranking scheme. With Kendall's τ between the two rankings of 0.40, the changes are even more pronounced than for the accuracy ranking. Specifically, it should be noted that two algorithms shared the last rank in the official ranking with a 5% percentile of 0.00. For the reimplementation, the number of algorithms sharing the last rank was doubled. The mean change in rank was 1.3 ranks for this ranking scheme and increased compared to the accuracy ranking.

Table 4.6: Overview of shortcomings of method descriptions from the participating teams of the Robust Medical Instrument Segmentation (RobustMIS) challenge [Roß/Reinke et al., 2020]. The severity column provides a subjective indication of how severely the issue may impact final leaderboard outcome, and the confidence with which assumptions have been taken in the reimplementation, described in Chapter 4.4. Teams for which no shortcomings were found, were excluded from the respective topic.

Data pre-processing		
Team	**Severity**	**Description of issue**
Casia SRL	Minor	From their main figure, it can be assumed that no modifications were applied to the images. This was not specified.
Uniandes	Major	Not specified.
VIE	Major	The described resizing step is implausible. Input and output formats and the optical flow were not specified.
www	Major	Not specified.
Data augmentation		
Team	**Severity**	**Description of issue**
Casia SRL	Minor	The combination and hyperparameters were not specified.
fisensee	Minor	The used library and types of augmentations were specified, hyperparameters were missing.
Squash	Minor	The horizontal translation was not specified. 5% of width seemed too small, the upscale factor of 1.5 resulted in images that severely differed from the original images.
Uniandes	Minor	The attempt to incorporate temporal data as data augmentation was described as not effective and was not reimplemented.

VIE	Major	Not specified.
www	Minor	Associated parameters and combination of rotations and flips were not specified.
Data split		
Team	**Severity**	**Description of issue**
caresyntax	Major	The data split was done manually, which could not be exactly reproduced.
Casia SRL	Minor	The reported data set split quantities did not match the actual data set size.
fisensee	Minor	The split for the validation folds yielded major imbalance between the folds.
VIE	Major	Not specified.
www	Major	Not specified.
Model architecture		
Team	**Severity**	**Description of issue**
caresyntax	Minor	The tuning of confidence thresholds was not specified.
Casia SRL	Major	Multiple names for the non-standard network architecture were used, prohibiting proper reimplementation. Description of network details was not sufficient.
VIE	Major	It was not specified how the optical flow was supplied as a model input.
www	Major	The Dense Atrous Convolution (DAC) integration was not specified.
Model backbone and pre-training		
Team	**Severity**	**Description of issue**
caresyntax	Minor	Pre-training on Microsoft-Common Objects in COntext (COCO) seemed unlikely given the description. It may have been confused with the library's naming convention.
VIE	Minor	It remained unclear whether pre-training was applied.
www	Minor	Given the modified architecture, it was unclear whether an Feature Pyramid Network (FPN) or standard configuration was chosen.
Optimizer		
Team	**Severity**	**Description of issue**
VIE	Minor	Not specified. The described hyperparameters made Stochastic Gradient Descent (SGD) plausible.

Learning rate		
Team	**Severity**	**Description of issue**
Casia SRL	Minor	It was unclear how the team specified the term "iterations". It could have been interpreted as actual minibatch iterations or full epochs.
fisensee	Major	The initial learning rate was implausible. No hyperparameters to the learning rate scheduler were specified.
Squash	Major	The specified learning rate did not reach convergence in the reimplementation. It may have been a typo.
Training loss function		
caresyntax	Minor	Not specified, but based on the strong overlap to the official Mask R-CNN implementation [FacebookResearch, 2019], the default loss was likely.
Uniandes	Minor	Not specified, but based on the strong overlap to the official Mask R-CNN implementation [FacebookResearch, 2019], the default loss was likely.
Batch size		
Uniandes	Major	Not specified.
VIE	Major	Not specified.
Training duration		
Team	**Severity**	**Description of issue**
Casia SRL	Minor	Not specified, but the final performance was described.
VIE	Minor	Not specified.
www	Minor	Only the absolute training time on the Graphics Processing Unit (GPU) was specified.
Model selection		

Team	Severity	Description of issue
caresyntax	Minor	Not mentioned, but the validation performance was specified.
Casia SRL	Minor	Not mentioned, but as the validation loss was specified, the lowest validation loss was plausible.
fisensee	Major	Not specified.
Uniandes	Minor	Multiple validation criteria were described, but it was not clear how they were employed.
VIE	Major	Not specified.
www	Major	Not specified.
Inference		
Team	**Severity**	**Description of issue**
caresyntax	Major	Not specified and not thresholds were provided, although this step was explicitly mentioned.
Casia SRL	Major	Not specified how instance segmentation results were obtained. A connected component analysis was likely, but not specified.
fisensee	Minor	Connected component analysis not completely specified.
VIE	Minor	Not specified, but based on the strong overlap to the official Mask R-CNN implementation [FacebookResearch, 2019], the default inference behavior was likely.
www	Minor	Not specified, but based on the strong overlap to the official Mask R-CNN implementation [FacebookResearch, 2019], the default inference behavior was likely.
Programming language / framework		
Team	**Severity**	**Description of issue**
caresyntax	Minor	Not specified. Python and PyTorch were plausible, given the identical configuration to the official PyTorch Mask R-CNN implementation [FacebookResearch, 2019].
Casia SRL	Minor	Not specified. Python and PyTorch were plausible, given that the ResNet-34 is available in PyTorch for the described configuration.
VIE	Minor	Not specified. Python was plausible, given the mention of the *OpenCV* Python library [Bradski, 2000].

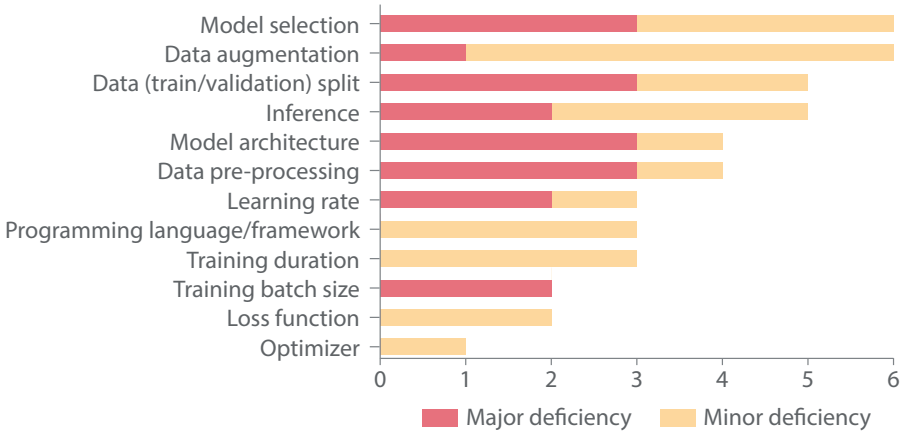


Figure 4.25: Qualitative analysis of deficiencies in the seven method descriptions (x-axis) of the experiment *Reproduction of challenge results given submitted reports* across several different aspects of implementation. A detailed overview is provided in Table 4.6.

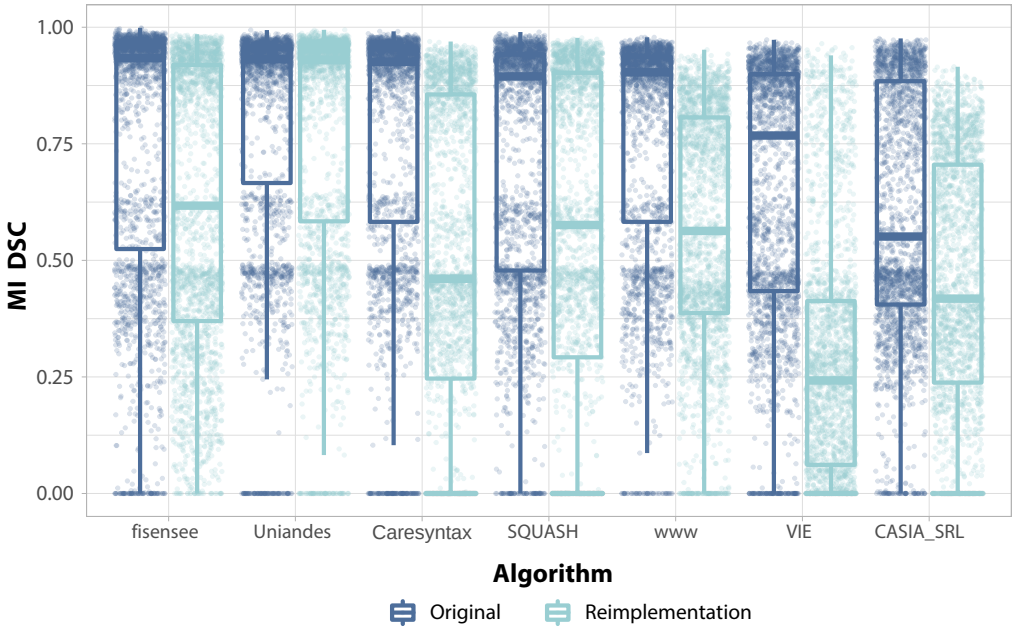


Figure 4.26: Dot- and boxplots showing the individual algorithm performance for Stage 3 of the Robust Medical Instrument Segmentation (RobustMIS) challenge for the Multi-Instance Dice Similarity Coefficient (MI DSC). Results are shown for the original challenge (blue) and the reimplementation (green). A large gap between the original and the reimplemented methods can be observed. The median metric score is represented by the center line of the boxplots and the lower and upper border of the boxes denote the first and third quartiles.

Table 4.7: Accuracy rankings for Stage 3 of the Robust Medical Instrument Segmentation (RobustMIS) challenge of the seven competing teams based on the Multi-Instance Dice Similarity Coefficient (MI DSC). The accuracy is determined by the proportion of how often a team is significantly superior to others divided by the number of teams (Prop. sign). The last column in (b), Δ, refers to the relative rank difference between the original challenge ranking and reimplementation ranking for the respective algorithm. A positive change in ranks is shown by a green arrow ↑, while a rank decrease is shown in red ↓. No rank change is indicated by a gray arrow →.

Team identifier	Prop. sign.	Rank
fisensee	**1.00**	**1**
Uniandes	0.83	2
caresyntax	0.67	3
SQUASH	0.33	4
www	0.33	4
VIE	0.17	6
CASIA SRL	0.00	7

(a) Original ranking as of Roß/Reinke et al. [2020]

Team identifier	Prop. sign.	Rank	Δ
Uniandes	**1.00**	**1**	↑ 1
fisensee	0.83	2	↓ 1
SQUASH	0.67	3	↑ 1
www	0.50	4	→ 0
caresyntax	0.33	5	↓ 2
CASIA SRL	0.17	6	↑ 1
VIE	0.00	7	↓ 1

(b) Ranking based on reimplementation

Table 4.8: Robustness rankings for Stage 3 of the Robust Medical Instrument Segmentation (RobustMIS) challenge of the seven competing teams based on the 5% quantile (Q5) of the Multi-Instance Dice Similarity Coefficient (MI DSC). The last column in (b), Δ, refers to the relative rank difference between the original challenge ranking and reimplementation ranking for the respective algorithm. A positive change in ranks is shown by a green arrow $\uparrow$, while a rank decrease is shown in red $\downarrow$. No rank change is indicated by a gray arrow $\rightarrow$.

Team identifier	Q5 MI DSC	Rank
www	**0.31**	**1**
Uniandes	0.26	2
SQUASH	0.22	3
CASIA SRL	0.19	4
fisensee	0.17	5
caresyntax	0.00	6
VIE	0.00	6

(a) Original ranking as of Roß/Reinke et al. [2020]

Team identifier	Q5 MI DSC	Rank	Δ
Uniandes	**0.28**	**1**	$\uparrow$ 1
fisensee	0.11	2	$\uparrow$ 3
www	0.04	3	$\downarrow$ 2
caresyntax	0.00	4	$\uparrow$ 2
CASIA SRL	0.00	4	$\rightarrow$ 0
SQUASH	0.00	4	$\downarrow$ 1
VIE	0.00	4	$\uparrow$ 2

(b) Ranking based on reimplementation

To investigate ranking uncertainty, we applied bootstrapping to the original challenge and the reimplemented results. For the accuracy ranking, the mean (median, IQR) Kendall's τ between the original ranking and all bootstrap samples was 1.00 (median: 1.00; IQR: (1.00, 1.00)), indicating a very robust ranking. For the reimplementation, the mean (median, IQR) Kendall's τ was only slightly less with 0.98 (median: 0.98; IQR: (0.98, 1.00)). For the robustness ranking, the original challenge ranking was quite less robust with a mean (median, IQR) Kendall's τ of 0.85 (median: 0.98; IQR: (0.98, 1.00)). The robustness ranking was less variable in the reimplementation with a mean (median, IQR) Kendall's τ of 0.97 (median: 1.00; IQR: (1.00, 1.00)).

4.4.4 Discussion

Transparency of validation studies is crucial for being able to interpret and reproduce their results. However, we found that it is not possible to fully understand and replicate challenge results, both in terms of design and participating methods.

Most challenges we analyzed did not provide sufficient information on important design parameters, such as the data annotation process. Given that algorithms rely on the provided labels, it is surprising to see that only little information is provided on the number of annotators, annotation protocols, and aggregation. According to Joskowicz et al. [2019], numerous annotators are needed to capture natural inter- and intra-rater variability. Indeed, as we showed in Section 4.3, the annotators have a crucial effect on the final challenge rankings. Full information on the rankings was not provided for more than two-thirds of the reviewed challenge tasks. This may have critical consequences: As shown in Section 4.1, this may serve as the basis for severe cheating.

Similarly to challenge organizers, participants often fail to transparently describe their methods used for challenge participation. In our attempt to reimplement the seven methods submitted to the RobustMIS challenge, we came up with results substantially different from the original scores. We had to take several assumptions throughout the whole model reimplementation pipeline. For one team, we could not even identify the underlying network architecture. During

our analysis, we found that more complex design choices (such as model selection or data augmentation) are described less thoroughly than simple choices (such as optimizers). This could be because only a small number of optimizers are typically used. Most of the teams successfully described their optimizing method and its hyperparameters in sufficient detail. The model selection, on the other side, was often not directly visible from the method description and often not described at all, although it is a very important step for applying a method to a new application. The RobustMIS challenge employed a rather unusual ranking scheme by using the 5% percentile as an aggregation operator. Model selection would have benefited from including this consideration. However, most of the teams did not specify the model selection step, which may explain why more than half of the teams failed in the reimplemented robustness ranking. Data augmentation is another complex process, which showed major shortcomings in the descriptions. Most augmentation methods come with additional hyperparameters that were often not described. Moreover, augmentations can be applied individually or combined with other augmentations. When combined, the case order and application probabilities additionally need to be specified to ensure reproducibility. Given that the influence of data augmentation has a crucial impact on performance results [Isensee et al., 2021], it is even more critical that it was insufficiently described for almost all methods.

We only provided the reimplementation of challenge results for one task. Nevertheless, we think that our experiments successfully illustrate the lack of transparency, as we were not able to reproduce the results of a single algorithm in the challenge. While the training of Deep Learning (DL) models comes with a high amount of non-determinism [Pham et al., 2020], which additionally contributes to the problem of reproducibility, we think the primary reason for failing to reproduce the participating methods is to the insufficient documentation of model details. As shown in Figure 4.25, many details of the methods remained unclear and were subject to several assumptions from our side. To overcome this problem, challenge participants should make their source codes available.

4.4.5 Conclusion

In this section, we showed that challenge design and algorithms typically cannot be reproduced, although transparency is one of the most important steps of proper validation studies. Lacking transparency may have critical consequences, such as severe changes in the rankings or the possibility of cheating. In addition, reimplementing the winning method is very attractive to many researchers to drive their research. Not being able to compile all information from the descriptions is frustrating and may yield errors due to missing information and misunderstandings propagating through several publications. Given that algorithms should be potentially translated into clinical practice, those errors could lead to severe consequences for patients.

Ideally, a challenge should report all important design parameters on their website and in the resulting publications. Moreover, ideally, challenge algorithms would be publicly available as source codes. As this is not always possible, challenge organizers and participants should make sure that all relevant information is provided, especially for the more complex design choices such as data augmentation and model selection.

4.5 Conclusion related to flaws of biomedical challenges

Throughout this part, we have revealed multiple critical flaws related to several aspects of biomedical image analysis challenges. We found that challenge design is susceptible to manipulation and that weaknesses can be easily exploited by challenge organizers and participants. Challenge metrics form an integral part of the validation of challenge participants' performances and can be misleading if the limitations of metrics are not considered while designing a challenge. In addition,the choice of metrics and aggregation operators, as well as other challenge design aspects, have a major impact on challenge rankings. Finally, we showed that challenge results and participants' algorithms are often not reproducible, decreasing challenge interpretability.

As a result of those issues, we founded the Medical Image Computing and Computer Assisted Interventions (MICCAI) Board Challenge Working Group (which is now the MICCAI Special Interest Group (SIG) for Challenges) and the Biomedical Image Analysis ChallengeS (BIAS) initiative, both aiming for enhancing the quality of biomedical challenges. We worked on solutions for all of the above-mentioned issues and beyond, which will be presented in the following part of this thesis.

5 Improving common practice of Biomedical Image Analysis Validation

5.1 Improving common practice of challenge design

Disclosure

This work has been supervised by Lena Maier-Hein. It has been conducted together with several co-authors, especially Sinan Onogur and Matthias Eisenmann. Please refer to Chapter A.1 for full disclosure.

5.1.1 Introduction

We discovered that the challenge design appears to be ill-conceived in the different sections of Chapter 4. Challenges have a high impact on biomedical image analysis research and often drive future research, given the published data sets or winning algorithms. We should therefore make sure that the results are meaningful, transparent, and of high quality. For this purpose, the focus of this section is on how we may improve challenge design security.

Specific findings in Part 4

Security holes in challenge design can be easily exploited.

Research question investigated in this chapter

How can a challenge design be made more secure?

In the past, challenges were often accepted at conferences based on a very brief description that only included high-level parameters like an abstract, a description of the data collection, and performance metrics. As a result, conference chairs were unable to identify problematic design choices or questionable practices, which may have led to the acceptance of poorly designed challenges. Furthermore, individual researchers, e.g. challenge participants, could not provide additional feedback on the design, as details were not accessible.

As a consequence, we introduce a structured challenge submission system for conferences that was first used in 2018 for the largest conference in the biomedical image analysis domain, the Medical Image Computing and Computer Assisted Interventions (MICCAI) conference. This submission system included all parameters that were found to be important for a challenge (see Table 4.5) as a comprehensive form.

5.1.2 Methods

We introduced a structured challenge submission process for conferences hosting challenges in the biomedical image analysis domain. For the implementation of this structured challenge submission system [Reinke et al., 2018b], we employed the web development framework *Vaadin*[1] (version 8). Vaadin comes with an integrated `Java` backend and `Java` Graphical User Interface (GUI). The framework provided several GUI components, especially useful to implement forms (for example the components `ComboBox`, `TextArea`, and `ProgressBar`), which were connected to the data source and reacted to user events. Since the GUI was run on a Java Virtual Machine (JVM), it did not require Representational State Transfer (REST) services or other methods to access data. In 2020, we updated Vaadin to version 10 (Vaadin Flow) to provide a more modern GUI.

40 out of 53 parameters of the challenge parameter list (see parameters in Table 4.5 that are marked with an asterisk) were implemented in the structured challenge submission system. The selection was based on the parameter ratings of the survey described in Section 4.4.2. Moreover, we added additional parameters that were specific to the concrete planning of the associated conference, asking for the challenge duration (full or half day), whether the challenge would be part of another conference workshop, the expected number of participants, publication and future plans as well as space and hardware requirements. Each parameter was supplemented with a definition and examples to ensure the clarity of the parameter. In 2019, we updated the form according to the newly designed reporting guideline, which is presented in Section 5.4 of this thesis. The `Java` objects were bound to the parameter forms using data binding either with Bean Validation or custom validators. The collected data was stored in an Object Database (ODB) management system and backups were stored multiple times a day on the server. In the first iteration of the submission system, most parameters were provided as a list of enumerations to ensure a standardized challenge collection. However, challenge organizers chose the 'Other' option for most of the parameters, preferring to write a more comprehensive description to avoid ambiguities over a predefined value. Therefore, in later iterations of the submission system, we implemented more free-text areas to give challenge organizers the chance to provide more detailed descriptions.

We implemented two types of user roles for the structured challenge submission system, namely *challenge organizers* and *administrators*. The administrator view provided an overview of all challenges and challenge creators, independent of their status (draft (submitted), accepted, revision required or submitted, rejected, declined, registered (and modified)), including the opportunity to download either a single or multiple challenge design PDF document(s) including all parameters of the respective challenge(s). In addition, we allowed the administrators to change the deadlines, news and updates text paragraphs, and to send emails to all users.

The standard challenge organizer user role mainly featured the creation of challenge proposals. A challenge proposal consisted of a form including all parameters presented above. Parameters needed to be inserted for each challenge task. For the submission of a challenge to a conference, we required challenge organizers to fill in at least 90% of the parameters. We further offered the opportunity for challenge organizer users to duplicate challenge proposals and to download an overview PDF file. In the next step, we extended the website to contain useful information

[1] vaadin.com

for users, including an overview of previously accepted challenges, statistics of those, and best practices for organizing and designing challenges based on community feedback (Section 3.1.3).

In addition to the general standardized submission, we introduced a review process for biomedical challenges. Submitted challenge proposals were reviewed by two to three independent reviewers. Since 2021, we required to have each proposal reviewed by an additional clinical reviewer, assessing the clinical relevance of the design. Challenge proposals requiring a revision were updated by the organizers within the submission system and re-submitted for a follow-up review.

5.1.3 Results

To overcome the lack of reporting, the developed structured challenge submission system [Reinke et al., 2018b] was implemented for the MICCAI 2018 conference. In this system, we collected 138 challenges and 255 challenge tasks so far[2]. Since 2018, it has been used primarily for MICCAI challenges, which account for 88% of all challenges. Furthermore, the submission system was used for IEEE International Symposium on Biomedical Imaging (ISBI) challenges in 2020 and 2021 (9%) and Medical Imaging with Deep Learning conference (MIDL) 2020 challenges (3%), which was the first year this conference hosted challenges.

In Figure 5.1, we compare the challenges captured from 2004–2016, as described in Chapter 3, with the recent trends from challenges that were collected via the structured challenge submission system. It can be seen that some continue, however, some changes can be observed, as we highlight in the following paragraphs.

Compared to the retrospective analysis, the number of challenges per year has been rising with the exception of 2015, in which two challenges with an extremely high amount of tasks were organized (cf. Figure 5.1(a)), with a median of 26 challenges (Interquartile Range (IQR): (22, 35); maximum: 40) and 47 tasks (IQR: (36, 63); maximum: 73). Recent challenges show the same lifecycle type trend compared to previous years, such that they were primarily organized as one-time events with a fixed submission deadline (39%), followed by repeated events with a deadline (21%). Challenges opening for new submissions after the conference deadline were less common, with 21% being repeated and 19% organized as one-time events.

Semantic or instance segmentation remained the most commonly used problem category for challenge tasks (37%), although the percentage halved in recent years. Segmentation is followed by image-level classification (17%) and object detection (11%). The further focus of recent challenges changed to registration and prediction problems rather than retrieval and tracking (cf. Figure 5.1(b)).

Magnetic Resonance Imaging (MRI) (37%) and Computed Tomography (CT) (22%) remained the most commonly applied imaging techniques. However, the fields of application shifted towards research (15%), diagnosis (14%), and decision support (9%) purposes rather than assistance, as prominently during the 2004–2016 challenges. In recent years, the submission method changed to most challenge organizers operating on a specific challenge web platform (37%) or using docker containers (35%), which were less common in the past.

[2] As of September 2022.

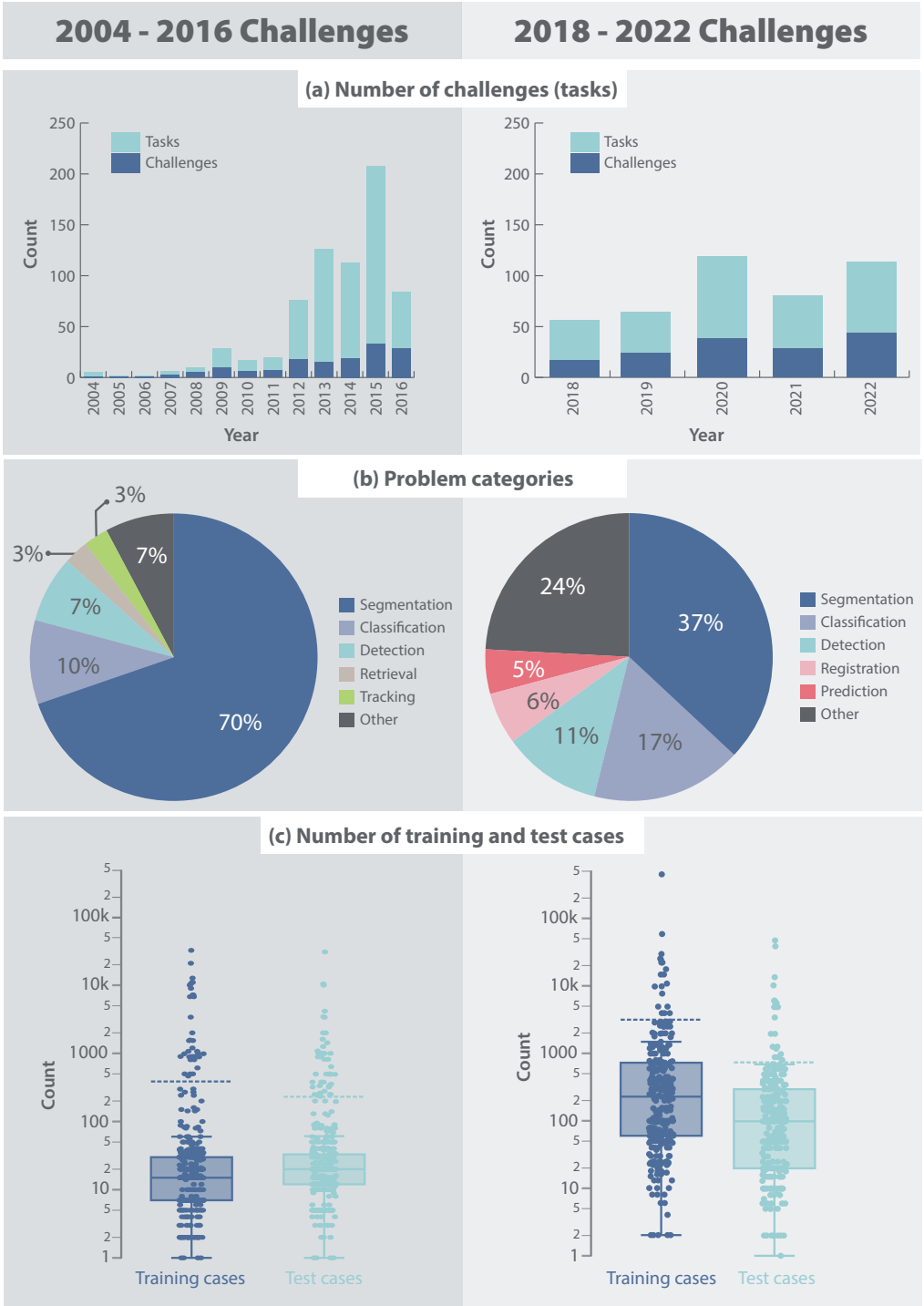


Figure 5.1: Overview and statistics of biomedical image analysis challenges from **2004–2016** (left; inclusion criteria as described in Chapter 3) and from **2018–2022** (right; challenges from the structured challenge submission system as described in Section 4.4). **(a)** Number of challenges and tasks per year. **(b)** Percentage of problem categories assessed in challenge tasks. **(d)** Dots- and boxplots of the number of training and test cases used in challenge data sets.

Challenge data was usually acquired from a median of one center and two devices, in contrast to only one device in the past. Much more challenges reported the ethical approval ID since the challenge submission system required organizers to fill out at least 90% of the parameters. Therefore, a total of 66% of tasks reported the ethical approval ID or mentioned that it will be available at the time of challenge execution. For 34% of recent tasks, no ethical approval was needed.

The size of training data sets increased in recent years, as shown in Figure 3.1(c), with a median of 229 cases, an IQR of (60, 730), and a maximum number of cases of 463,424. The number of test cases was usually lower in present years, with a median of 100 cases, an IQR of (20, 300), and a maximum number of cases of 48,914 from 2018–2022. The overall median ratio of training and test cases raised from 0.75 to 2.33. The data has been annotated by a median of two annotators (IQR: (2, 4); maximum: 20) and an additional 23% of tasks reported that more than one annotator generated the reference data without further specifying the concrete number.

The trend of heterogeneous challenge validation continued in recent years, where the Dice Similarity Coefficient (DSC) (50% of all tasks; 79% of the segmentation tasks) and the Hausdorff Distance (HD)/Hausdorff Distance 95 Percentile (HD95) (26% of tasks; 48% of segmentation tasks) were the most frequently used metrics from a total of 105 metrics. 58% of the metrics were only used in a single task in the past. In recent years, only 16% of challenge tasks did not justify the metric choice at all. Notably 24% justified their choice by stating that the metric was commonly used without providing further context for the metric choice.

A challenge winner was announced in all recent tasks. 45% of tasks calculated the ranking only based on a single metric. Similar to past challenges, we identified ten different ranking schemes and metric-based aggregation was the most frequently used ranking scheme, although the frequency is much lower in comparison (35% *vs.* 76%). Statistical testing increased from 6% of tasks to 48% in recent years. Nonetheless, the t-test (22%) and Wilcoxon signed rank test (18%) remained the most frequently used test strategies.

5.1.4 Discussion

While challenge acceptance in the past relied on only a very brief summary of the design, the structured challenge submission system introduced a comprehensive online tool to collect challenge information and to make challenges more secure. It enables a direct and complete comparison of recent challenges with the retrospective analysis presented in Chapter 3.

We could observe that the number of challenges organized per year is still increasing. Constraints by conferences, such as limited space, could be the reason that the number of challenges decreased in 2016. Similar to previous years, segmentation problems were still the most frequent category, but the number of segmentation problems nearly halved. This enabled other, less common, problem categories, to be assessed. Recent challenges further use larger data sets and more devices, with a trend towards more realistic data sets. Finally, the number of tasks using statistical tests for their analysis has increased by a factor of eight, implying that more and more challenge organizers put emphasis on the analysis of challenge results.

The introduction of the structured challenge submission system allowed us to overcome some of the security holes we presented in Section 4.1. Forcing challenge organizers to provide information for at least 90% of challenge parameters yielded comprehensive design documents. Those were especially helpful for the challenge reviews, in which reviewers were able to comment on all facets of the proposed designs, capturing problems that may not have been visible before introducing the peer review process. Furthermore, the organizers' awareness of certain problems of the challenge design could already have been raised at an early stage. In addition, the submission system also served the purpose of educating organizers on the importance of specific parameters and how they were defined to avoid ambiguities.

However, although the complete challenge design documents were provided to the challenge organizers, we did not have control over the actual execution of the challenges. Thus, we further analyze the concrete implementation of challenges compared to their accepted design documents in Section 5.4.

5.1.5 Conclusion

The structured challenge submission system and review process introduced a specific quality control mechanism for challenges. The process makes sure that only challenges whose *complete* design is sound are accepted for conferences. This represents a novel, long-lasting structural investment, which, through its introduction of true standardization and peer review, can elevate the scientific quality of challenges. Nevertheless, this practice is still not sufficient to ensure transparent reporting and may need further refinement, as we demonstrate in Section 5.4.

5.2 Improving common practice of metrics

Disclosure

This work has been supervised by Lena Maier-Hein. It has been conducted together with several co-authors, especially Paul Jäger. The work has been *submitted to Nature Methods* [Maier-Hein et al., 2022] and accepted as short papers/abstracts at the *Medical Imaging with Deep Learning conference (MIDL) conference* [Reinke et al., 2022a] and for the *Medical Imaging Meets Neural Information Processing Systems (NeurIPS) workshop* [Reinke et al., 2022b]. Please refer to Chapter A.1 for full disclosure.

5.2.1 Introduction

The choice of validation metrics is crucial for designing a biomedical image analysis challenge and validating algorithm performance in general. The metric values provide the researchers with information on how well or poorly the algorithms performed on particular images or the entire data set. As a result, metrics are expected to represent the application-specific validation goal. However, 24% of the challenges captured in the structured challenge submission system (see Section 5.1.3) justified their metric selection by utilizing widely recognized and prominent standard metrics. Yet, in Section 4.2, we demonstrated that most metrics come with several limitations.

Metric selection is one of the most important steps in the validation pipeline. The ranking scheme is largely irrelevant if the metric is flawed because the wrong properties are assessed. If metrics are inadequate, an algorithm could win a challenge for the wrong reasons. Moreover, challenges are often seen as a role model for benchmarking a specific research question. Thus, researchers tend to use the same metrics and validation strategies proposed by a challenge to be able to compare against the original challenge participants. We should therefore make sure that challenges choose metrics in the right way to avoid mistakes being propagated and integrated into research practice.

Therefore, choosing a performance metric because of its frequency or popularity is not valid reasoning. In this section, we thus aim to answer the following research question.

Specific findings in Part 4

Information on metric pitfalls is currently inaccessible.

Metrics are widely chosen based on popularity rather than their clinical suitability, ignoring their pitfalls.

Research question investigated in this chapter

How can we ensure that researchers choose metrics that reflect biomedical needs?

In this section, we propose a problem-aware metric recommendation framework. The recommendations are based on a consensus decision of an international expert consortium. According to Lennerz et al. [2022], we "lack (...) a common language between Artificial Intelligence (AI) and medicine", this is why we the consortium consists of experts from varying domains, including biomedical image analysis AI practitioners, clinicians, biologists, statistitians, and others. The recommendation framework relies on the generation of a problem fingerprint, which systematically captures the properties of the underlying research question and objective of a problem statement. Based on this characterization, metrics can be selected from several metric mappings for different metric families, such as counting or boundary-based metrics. Finally, we present recommendations for common biomedical use cases and show that several popular challenges did not fulfill our requirements of a complete and problem-aware metric selection.

5.2.2 Methods

We once more opted for a multi-stage Delphi process [Brown, 1968], in which the knowledge of multiple experts was merged to create a consensus on the recommendations through a series of surveys, and develop a problem-aware metric recommendation framework. We built upon the consortium presented in Section 4.2.2. The expert consortium was initialized with members from the Medical Image Computing and Computer Assisted Interventions (MICCAI) Board Challenge Working Group, the Biomedical Image Analysis ChallengeS (BIAS) initiative, and the Medical Open Network for Artificial Intelligence (MONAI) Working Group for Evaluation, Reproducibility, and Benchmarks. In addition, we invited experts from the fields of validation, metrics, challenges, and other related meta-research topics. Thus, the consortium evolved over time. The final expert consortium comprised 75 experts (20 female, 55 male). Most of the experts were active in Europe (73%; most of them from Germany (35%) or the UK (12%)), followed by researchers from North America (21%; mostly from the USA (13%) or Canada (8%)), South America (3%), Australia (2%), and Asia (1%). Within the consortium, we defined different groups:

The core team comprised of myself and two others, coordinated the Delphi process, devised the framework based on the expert feedback, prepared and evaluated the surveys, and organized the workshops. The core team generally did not participate in the voting process.

The extended core team assisted the core team in the Delphi process coordination, the preparation of the surveys, and the workshop organization.

The experts participated provided feedback in the surveys. They were further divided into several expert groups, working on dedicated parts of the framework:

The image-level classification expert group worked on the recommendations for image-level classification problems.

The semantic segmentation expert group worked on the recommendations for semantic segmentation problems.

The object detection and instance segmentation expert group worked on the recommendations for object detection and instance segmentation problems.

The biomedical expert group provided input from the application-specific perspective and was composed of medical, clinical, and biological experts.

The cross-cutting concerns expert group worked on recommendations beyond the actual metric selection, such as metric aggregation.

The Delphi process was initialized with a kick-off workshop, in which the scope of the recommendation framework was discussed. The first round of surveys served the purpose of defining the problem categories for which the recommendations should be provided, and the first selection of metrics which should be added as candidates. The initial list of metrics was based on the metrics that were used in challenges (see Chapter 3), complemented by those found in previous work, such as those listed in Taha and Hanbury [2015], Nai et al. [2021], and Hicks et al. [2022]. In this first round, 23 experts participated in the voting process.

The second round aimed for defining inclusion criteria of image analysis problems for the framework. In addition, the experts voted on the first suggestion of the general mapping strategy and the first iteration of the problem category mapping. The round was completed by 26 experts. The refinement of the inclusion criteria and the problem category-specific metric pools were the topics of the third round. They were complemented by defining the framework-specific terminology and the definition of the problem fingerprint. 30 experts participated in round three. After the evaluation of round three, we organized a second workshop, in which the expert groups were presented and group members were assigned for topic-specific discussions.

Round four was divided among the five expert groups, which voted on the problem-specific metric mappings and fingerprints or on questions for their respective topic. Furthermore, we collected relevant biomedical use cases, for which we aimed to instantiate the framework. A total of 37 experts responded to the surveys. Three follow-up workshops were organized after this round to discuss and build the metric mappings. In addition, the pre-final metric mappings were tested several times by researchers in the labs of the experts for a range of different problems and applications. Before the final Delphi round started, we asked the general community to provide feedback on the metric mappings and problem fingerprints. The mappings were refined based on the feedback of 53 participants that were not part of the expert consortium. The final round five aimed for the final consensus building on the problem fingerprints, metric mappings, and decision guides. 32 experts voted in this round.

5.2.3 Results

Our problem-aware metric recommendation framework was built for problems that can be interpreted as classification tasks at various levels. We therefore focused on the four problems of image-level classification, semantic segmentation, object detection, and instance segmentation. Their common validation metrics can be categorized into counting metrics, multi-threshold

metrics, calibration metrics, and distance-based metrics. An overview of the recommendation framework is provided in Figure 5.2.

Prerequisites

Handling of intermediate tasks The framework is intended to suggest metrics for one specific biomedical question. However, many biomedical questions can be split into several intermediate tasks and we recommend using the framework for all of them individually. For example, a clinician may be interested in diagnosing whether a patient suffers from cancer. The task itself is often accompanied by the delineation of the tumor, which would be an intermediate task. The framework would be applied separately for the segmentation of the tumor and the classification of the image. This also holds true for challenges with multiple tasks, which might focus on different aspects. Each task would then be composed of different properties, objectives, and perhaps even different data. Metrics should be chosen separately based on task-specific properties. In addition, metric selection is typically done for every class (of interest[3]) to ensure easily interpretable results. If desired, class-wise results may be aggregated (see metric application).

Handling of multiple classes While some biomedical questions are binary (e.g. cancer versus no cancer), many problems deal with multiple classes with different properties. In this case, some of the metric mappings should be repeated for every class ensuring to reflect the class-specific properties. Below, we indicate which processes should be repeated for every class. Furthermore, we emphasize that the metric selection is solely based on the driving biomedical research question and is independent of the actual algorithm. In our framework, we only focus on the algorithm outputs and we neglect the training process and other model-specific subjects.

Metric pool The metric pool used for the recommendation framework is based on the consensus from the Delphi expert consortium. It includes both popular standard metrics (such as Dice Similarity Coefficient (DSC) and Accuracy) and rather uncommon metrics (such as Expected Cost (EC) or Net Benefit (NB)). Generally, all recommendations address the metric pitfalls presented in Section 4.2. Overall, we recommend the usage of multiple performance metrics to validate algorithm predictions for different properties, thus compensating for limitations of single metrics. Decisions on metric selection are typically application-specific.In situations when the trade-offs between several metric candidates must be taken into account, we provide decision guides that assist in the metric selection while respecting individual preferences. A list of metrics used in the framework is provided in Appendix A.3.

Inderpendency of application domain and imaging modality Finally, it should be noted that the framework is independent from the application domain and imaging modality. The problem fingerprint (see next paragraph) was designed in a way such that it can be used for any biomedical image analysis question related to image-level classification, semantic and instance segmentation, or object detection.

[3] In segmentation problems, for example, the background class may not be of interest and may thus be excluded from the metric selection.

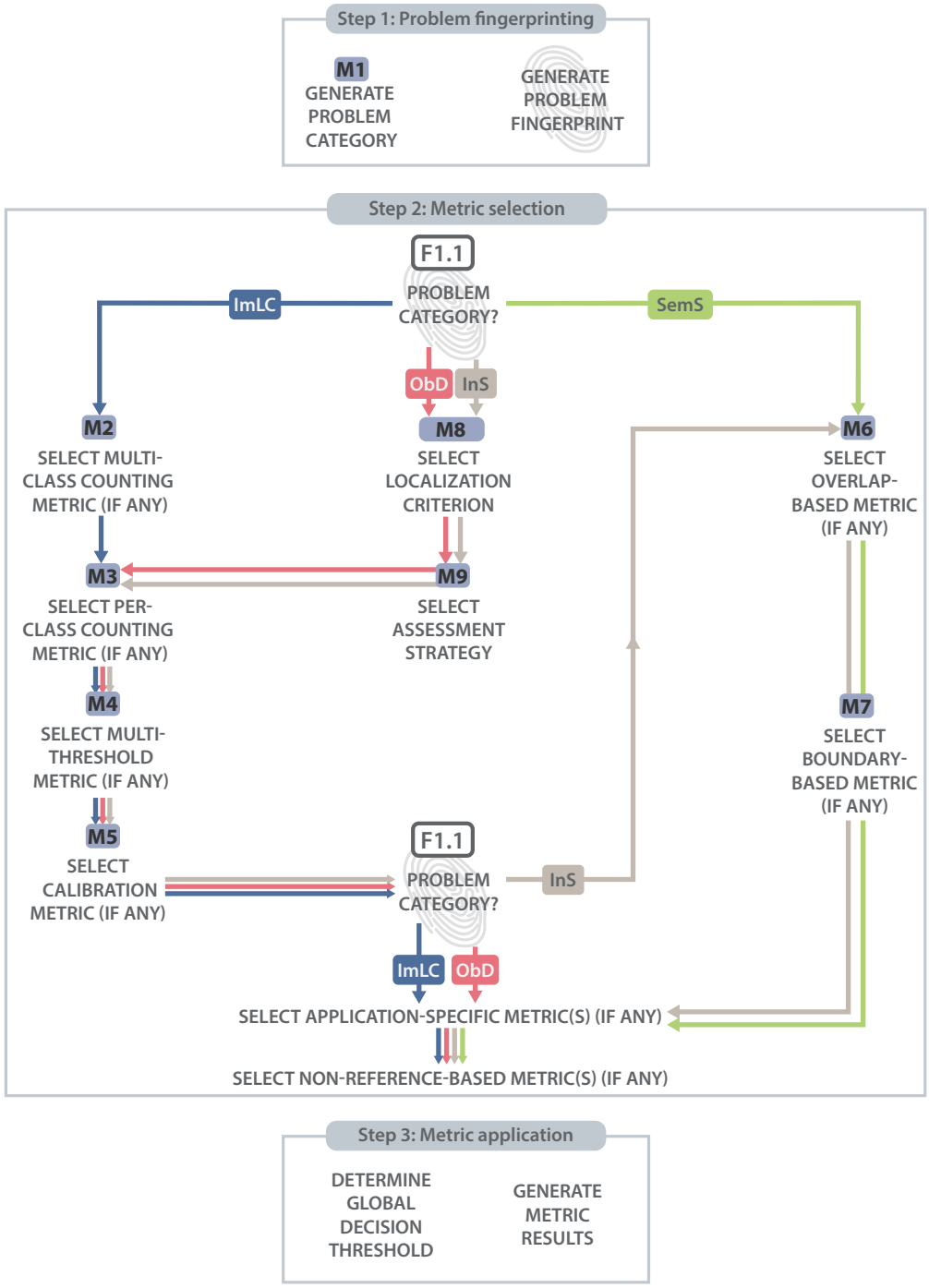


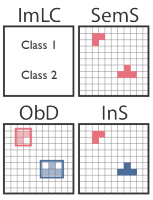
Figure 5.2: Overview of the metrics recommendation framework. In the first step, the **problem fingerprint** is generated for the appropriate problem category (image-level classification (ImLC), semantic segmentation (SemS), object detection (ObD), or instance segmentation(InS)) based on mapping M1 in Figure 5.3. Based on the fingerprint, the metric selection is achieved via the mappings M2 - M9 (see Figures 5.4 - 5.12). Finally, the selected metrics are applied to the data set.

Problem fingerprint

The creation of a **problem fingerprint**, i.e. the structured collection of problem characteristics unique to the underlying biomedical question or problem, is the most crucial stage of the recommendation framework. The problem fingerprint differs across tasks and data sets, this is the reason why it is important to instantiate the fingerprint for every intermediate problem or challenge task. The problem fingerprint is composed of five categories, which were built from the metric pitfall taxonomy presented in Section 4.2. In the first step, the *problem category* (fingerprint F1.1) for the driving biomedical question is observed. The problem category mapping M1 can be found in Figure 5.3. This mapping, which is assembled as a decision tree, asks for global or local interest, where global interest refers to reference annotations on image level and local interest refers to reference annotations on object or pixel level.

The *domain interest problem characteristics* cover the actual interest behind the underlying biomedical question. *Target structure problem characteristics* capture the appearance of the objects or structures in an image and are especially important for object detection and semantic and instance segmentation problems independent from the interest. The *data set problem characteristics* describe the data set on which the metrics should be calculated, i.e. the test data set in a challenge. Finally, the *algorithm output problem characteristics* describe which algorithm output is expected. The complete fingerprint is provided in Tables 5.1 - 5.5. The tables indicate for which problem category the problem fingerprint is relevant. In addition, they provide a description and illustration of every fingerprint.

Table 5.1: Problem fingerprint for the **general problem category** (Image-level Classification (ImLC), Semantic Segmentation (SemS), Object Detection (ObD), Instance Segmentation (InS)). Each problem fingerprint comes with an identifier (ID), an illustration, and a description. Table adapted from Maier-Hein et al. [2022].

ID	Fingerprint	Illustration	Description
F1.1	Image processing category identified by category mapping		The driving biomedical problem is assigned to one of the following problem categories via the category mapping. **Image level classification (ImLC)**: assignment of one or multiple category labels to the entire image. *Example: disease screening; deciding on the presence or absence of a certain condition/pathology without localizing the phenomenon.* **Semantic segmentation (SemS):** assignment of one or multiple category labels to each pixel. *Example: surgical scene segmentation for autonomous robotics; assigning each pixel the corresponding structure/organ/pathology label.* **Object detection (ObD):** detection and localization of structures of one or multiple categories. *Example: detection and bounding box-based localization of polyps in colonoscopy sequences.* **Instance segmentation (InS):** detection and delineation of each distinct object of a particular class. It can be viewed as simultaneously performing the tasks of semantic segmentation and object detection. In contrast to object detection, instance segmentation also involves the accurate marking of the object boundary. In contrast to semantic segmentation, it distinguishes different instances of the same class. *Example: cell segmentation with a subsequent goal of cell counting. Detection and localization of structures of one or multiple categories.*

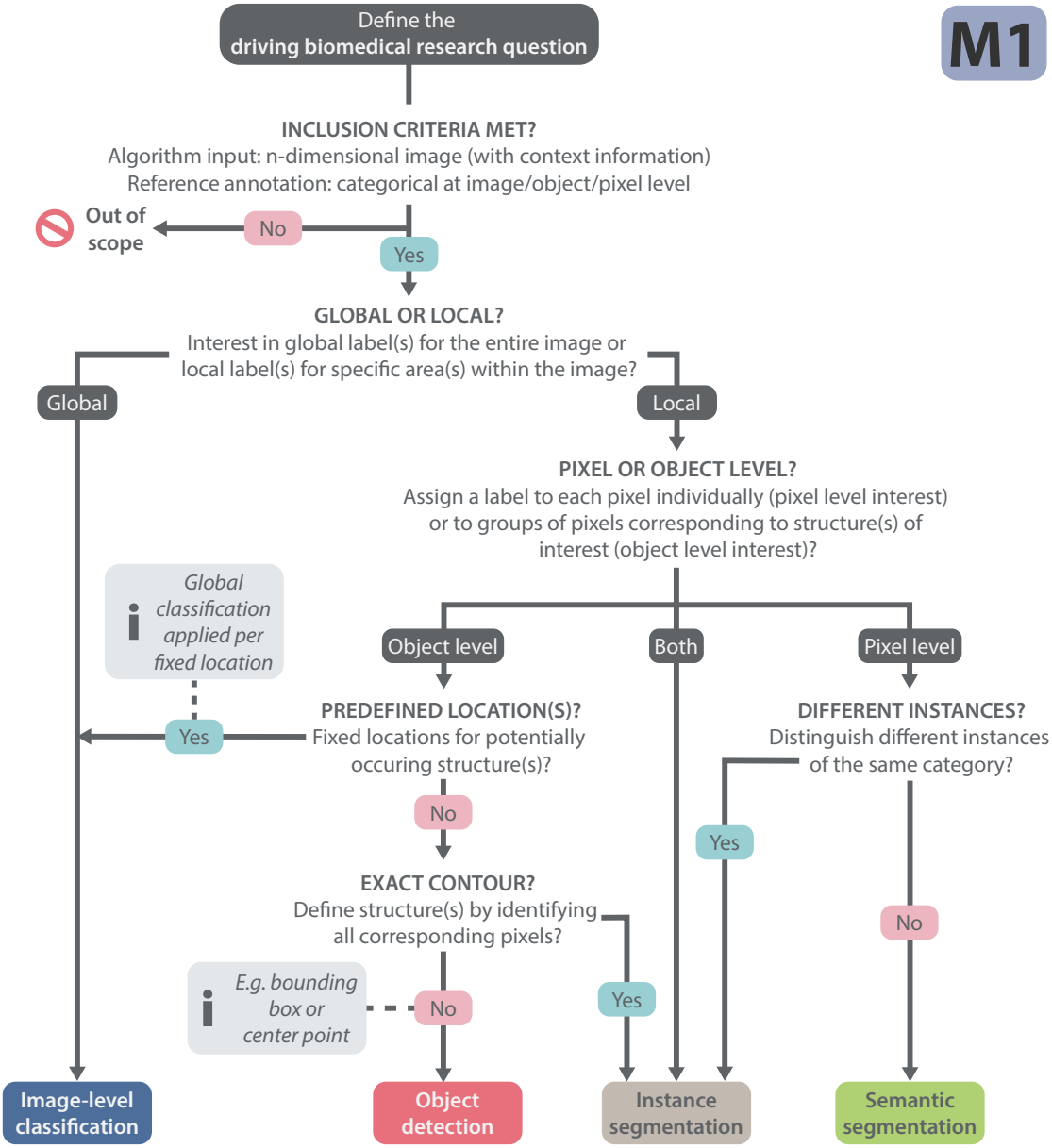
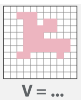
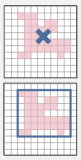
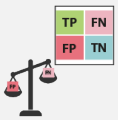
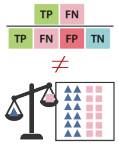
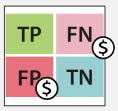
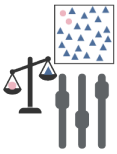
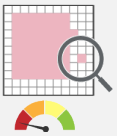
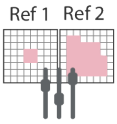
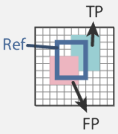


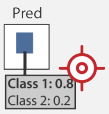
Figure 5.3: **Mapping M1:** Problem category mapping to identify the problem fingerprint F1.1. The mapping starts with the driving biomedical question and moves through multiple questions before arriving at the appropriate problem category (image-level classification, semantic segmentation, object detection, or instance segmentation).

Table 5.2: Problem fingerprint for **domain interest problem characteristics**. Each problem fingerprint comes with an identifier (ID), an illustration, and a description. In addition, every fingerprint is assigned to one or multiple problem categories (Image-level Classification (ImLC), Semantic Segmentation (SemS), Object Detection (ObD), Instance Segmentation (InS)). Table adapted from Maier-Hein et al. [2022].

ID	Fingerprint	ImLC	SemS	ObD	InS	Illustration	Description
F2.1	Particular importance of structure boundaries		✓		✓		The biomedical application requires exact structure boundaries. *Example: segmentation for radiotherapy planning; knowledge of exact structure boundaries is crucial to destroy the tumor while sparing healthy tissue.* Important: Overlap-based metrics do not measure shape agreement. In the cases of complex shapes (high boundary-to-volume ratio), it is therefore typically advisable to set this property to true.
F2.2	Particular importance of structure volume		✓		✓		The biomedical application requires accurate knowledge of structure volumes. *Example: liver segmentation as the basis for remnant liver volume computation in surgical resection planning.*
F2.3	Particular importance of structure center (e.g. in cells, vessels)		✓	✓	✓		The biomedical application requires accurate knowledge of structure centers. *Example: cell centers are subsequently used for cell tracking and cell motion characterization, so false center movement should be suppressed.*
F2.4	Desired granularity of localization			✓			The granularity of required localization can vary in object detection tasks. We distinguish two main categories: **Only position:** given an n-dimensional image, the object is represented by its position, encoded in n degrees of freedom (e.g. xy/xyz coordinates of the center point) **Rough outline:** a rough outline of the object is provided, typically given by simple geometric approximations such as bounding boxes or ellipsoids. It should be noted that if a substantial fraction of objects is tiny (F3.1), any outline-based localization becomes very noisy. In such cases, users might want to consider alternative localization strategies, such as a center-point-based localization.
F2.5.1	*Penalization of errors:* Unequal interest across classes	✓	✓	✓	✓		There is a preference for one or several of the classes. This has implications for both the metric selection and the metric aggregation. *Example 1: In cell classification scenarios, it may be more important to correctly classify tumor cells compared to correctly classifying muscle cells or connective tissue.* *Example 2: in full surgical scene segmentation for autonomous robotics, critical structures, such as nerves or vessels, should be localized more accurately compared to fatty tissue*

ID	Fingerprint	ImLC	SemS	ObD	InS	Illustration	Description
F2.5.2	*Penalization of errors:* Unequal severity of class confusions	✓	✓	✓	✓		Any class can be confused with another, but certain mismatches are more severe than others, from a domain point of view. *Example 1 (binary, ObD): polyp detection; a False Negative (FN) (missed polyp) is clinically much more severe than a False Positive (FP).* *Example 2 (multi-class): Depending on the application, confusing different kinds of immune cells is more problematic compared to confusing an immune cell with a tumor epithelial cell.* *Example 3 (multi-class, ImLC): lung tumor categorization T1-T5 depends largely on structure size, implying an ordinal scale of classes. Thus, penalization of class confusions should reflect this ordinal scale.* Specifically, in InS problems, the property needs to be set separately for the validation of the (a) detection (relevant decision guide: 3.5) and (b) segmentation performance (relevant mapping: M6. At object level, FNs (missed instances) are sometimes more severe than FPs, while FNs (e.g. undersegmentation) and FPs (e.g. oversegmentation) may be equally important at pixel level.
F2.5.3	*Penalization of errors:* Mismatch between class prevalences and class importance	✓					The class prevalences do not reflect the class importance. There are three scenarios for mismatches between class prevalences and class importance: Class prevalences are balanced (F4.1 = FALSE), but there is an unequal interest across classes (F2.5.1 = TRUE). Class imbalance is present (F4.1 = TRUE), but there is an equal interest across classes (F2.5.1 = TRUE). Class imbalance is present (F4.1 = TRUE) and there is an unequal interest across classes (F2.5.1 = TRUE), but the way in which classes are balanced does not match the "imbalance of interest". Importantly, while scenarios 1 and 2 can be expressed with other fingerprints, scenario 3 represents a new set of use cases.
F2.5.4	*Penalization of errors:* Costs for class confusions available	✓					In the case of an unequal severity of class confusions (F2.5.2 = TRUE), these unequal severities might be explicitly defined in the form of costs values associated with each confusion. For example, a cost analysis may lead to the result that FP errors are five times more costly than FN errors. In case such costs are defined or can be estimated with adequate accuracy for the use case, it is possible to apply certain metrics which explicitly consider these costs in validation (e.g. Weighted Cohen's Kappa (WCK) and Expected Cost (EC)). If costs are not provided and cannot be estimated, we recommend to proceed with validation separately for individual classes.

ID	Fingerprint	ImLC	SemS	ObD	InS	Illustration	Description
F2.5.5	*Penalization of errors:* Compensation for class imbalances requested	✓	✓	✓	✓		Severe class imbalances might impede interpretability and objective assessment of method validation and e.g. lead to overly optimistic conclusions. Some metrics compensate for such effects. *Example 1 (ImLC; multi-class classification with one dominant class): Accuracy would reflect this imbalance and allow for high scores of an uninformed classifier, which impedes the interpretability of scores. Balanced Accuracy (BA) and Matthews Correlation Coefficient (MCC), on the other hand, compensate for this effect as an uninformed classifier would have a known fixed metric value (BA: 1 divided by the number of classes, MCC: 0) irrespective of the class imbalance.* *Example 2 (ObD, InS, ImLC; binary classification with the negative class being overrepresented in the form of easy-to-classify TN): Metric scores considering balanced discrimination of two classes (e.g. LR+ or AUROC) might be dominated by TN and thus give an overly optimistic picture of the performance, especially if practical interest lies with the positive class. This is especially true in screening tasks (e.g. cancer detection in a cohort of mostly healthy patients). Metrics not considering TN (e.g. F_β Score or AP) compensate for this effect and enable focusing on the discrimination of the positive class.*
F2.5.6	*Penalization of errors:* Handling of spatial outliers		✓		✓		Spatial outliers are FP predictions that feature a large distance to the reference. They can be handled in three different ways: **Distance-based penalization with outlier focus:** Outliers should be heavily penalized as a function of the distance to the reference contour. **Distance-based penalization with contour focus:** Outliers should be penalized as a function of the distance to the reference, but the assessment should focus on the general contour agreement rather than individual outliers. **Existence-based penalization:** The existence of spatial outliers should be penalized irrespective of their distance to the reference contour. Note that distance-based penalization is not possible when either the reference or the prediction is empty. In applications in which many of such cases potentially occur, we therefore recommend an existence-based penalization.
F2.5.7	*Penalization of errors:* Compensation for annotation imprecisions requested		✓		✓		The reference annotation is typically only an approximation of the (forever unknown) ground truth. It may be desirable to compensate for known uncertainties, such as intra-rater or inter-rater variability, by configuring the metric accordingly. This is only possible for some metrics.
F2.5.8	*Penalization of errors:* Penalization of multiple predictions assigned to the same reference object requested			✓	✓		ObD and InS algorithms involve the step of assigning predicted objects to reference objects. This may result in more than one prediction being assigned to the same reference. This fingerprint property should be set to true if all but one prediction of such an assignment should be penalized as FP, and set to false if these spare predictions should be ignored during validation.

ID	Fingerprint	ImLC	SemS	ObD	InS	Illustration	Description
F2.6	Cutoff on predicted class scores	✓		✓	✓		Modern algorithms output continuous class scores. Making a classification decision requires setting a cutoff value on the scores, thereby generating a (cutoff-specific) confusion matrix. This matrix enables the computation of popular single-threshold counting metrics, such as sensitivity, Positive Predictive Value (PPV) and F_1 Score. Depending on domain interest the cutoff can be set in multiple ways: **Target value-based cutoff:** The cutoff represents the threshold for which a specific target metric value (e.g. Sensitivity = 0.95) is achieved. Other metric values (e.g. Specificity) are then reported for this specific threshold. We use the notation Metric1@Metric2 (e.g. Specificity@Sensitivity = 0.95) in this case. **Optimization-based cutoff:** The cutoff is inferred by optimizing a target metric, such as the F_1 Score, on a validation data set. **Argmax-based cutoff:** If no target value is defined, no separate data split for optimization is available, or there are concerns w.r.t generalization of data-based cutoff optimization, a common option is to follow the principle of a Bayes classifier and pick the class with the highest predicted class score. **Benefit-cost-based cutoff:** In case the predicted class scores express the risk associated with a case belonging to a certain class (see F2.7), and there is a task-related risk-threshold provided (e.g., only treat patients with cancer risk >10%), one can apply this threshold directly to the scores without data-driven optimization. Notably, provided risk thresholds correspond to a cost ratio of TP versus FP (e.g., not more than 10 FP per 1 TP should be treated). The Net Benefit metric considers this cost ratio as an inherent part of performance measure. **No cutoff:** Examples for no interest in validating a method at a certain cutoff are (1) focus on general methodological performance across many tasks and data sets without application interest, or (2) concerns regarding the comparability of results based on a single cutoff that is fixed across varying study cohorts (see also F4.2).
F2.7.1	*Calibration of predicted class scores :* Calibration assessment requested	✓		✓	✓		This property should be set to true if the predicted class scores should match the true probability of cases belonging to the predicted class (e.g. the probability of a patient to develop a certain disease in prediction problems). Methods subject to validation in this context are either classification models (testing their inherent calibration quality) or so called re-calibration methods, i.e. accuracy-preserving (bijective) transformation on the classifier outputs aiming to improve calibration quality.

ID	Fingerprint	ImLC	SemS	ObD	InS	Illustration	Description
F2.7.2	*Calibration of predicted class scores:* Comparative calibration assessment requested	✓		✓	✓		**Comparison of re-calibration methods on the same classifier:** The potential benefit of one or more re-calibration methods is to be assessed and compared. The desired validation output is a ranking of re-calibration methods (including the performance of "no re-calibration") from which the best method can be selected. **Comparison of calibration performance across classifiers:** This comparison of classification models potentially includes (re-) calibration methods applied on their outputs. The desired validation output is a ranking of methods according to calibration quality. **Comparison of overall performance across classifiers:** Overall performance refers to the joint assessment of discrimination performance and calibration quality. This comparison of classification models potentially includes re-calibration methods applied on their outputs. The desired validation output is a single ranking naturally weighting both aspects. **No comparative assessment:** If the interest lies in understanding the reliability of predicted class scores for one given model, no metrics for comparative assessment are required.
F2.7.3	*Calibration of predicted class scores:* Interpretable estimate of calibration error requested	✓		✓	✓		There is an interest in understanding the reliability of predicted class scores for a given model as a basis for interpreting and communicating results. The desired validation output is a single score which provides an insight into how well the model is calibrated.

Table 5.3: Problem fingerprint for **target structure problem characteristics**. Each problem fingerprint comes with an identifier (ID), an illustration and a description. In addition, every fingerprint is assigned to one or multiple problem categories (Image-level Classification (ImLC), Semantic Segmentation (SemS), Object Detection (ObD), Instance Segmentation (InS)). Table adapted from Maier-Hein et al. [2022].

ID	Fingerprint	ImLC	SemS	ObD	InS	Illustration	Description
F3.1	Small size of structures relative to pixel size		✓	✓	✓		Structures of the provided class are only a few pixels in size. *Example: multiple sclerosis lesions in magnetic resonance imaging (MRI) scans.*
F3.2	High variability of structure sizes (within one image, across images)		✓	✓	✓		The target structures vary substantially in size, such that some structures are several times the size of others. *Example: polyps in colonoscopy screening, where some polys are several times the size of others.* *Counterexample: large organs, such as the liver or the kidneys, which are relatively comparable in size across individuals.*
F3.3	Target structures feature tubular shape		✓	✓	✓		The target structures feature a tubular shape. *Examples: vessels, fibers.*
F3.4	Possibility of multiple labels per unit (pixel or image)	✓	✓		✓		**ImLC:** Multiple categories may be assigned to one image. *Example: image classified with multiple pathologies during a screening process.* **SemS/InS:** Multiple categories may be assigned to one pixel. *Example: labels 'tumor core' and 'tumor' assigned to the same pixel.*
F3.5	Possibility of overlapping or touching target structures (e.g. medical instruments or cells)		✓	✓	✓		Different instances of a class can overlap or touch each other. *Examples: overlapping cells or organisms, such as BBBC010 (worms in a dish); overlapping medical instruments in laparoscopy.*
F3.6	Possibility of disconnected target structure(s)			✓	✓		A given structure appears disconnected in the given image. *Examples: single tomographic image slice depicting complex vessel, partially occluded by a medical instrument.*

Table 5.4: Problem fingerprint for **data set problem characteristics**. Each problem fingerprint comes with an identifier (ID), an illustration, and a description. In addition, every fingerprint is assigned to one or multiple problem categories (Image-level Classification (ImLC), Semantic Segmentation (SemS), Object Detection (ObD), Instance Segmentation (InS)). Table adapted from Maier-Hein et al. [2022].

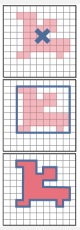
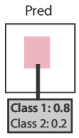
ID	Fingerprint	ImLC	SemS	ObD	InS	Illustration	Description
F4.1	Presence of class imbalance	✓	✓	✓	✓		The class prevalences differ substantially. Further reading Mahani and Ali [2019].
F4.2	Provided class prevalences reflect the population of interest	✓					The class prevalences are representative of the prevalences to be expected in the population of interest. **This property should be set to true** if either the validation interest is constrained to the data set at hand or no variation of prevalences is expected in other cohorts and upon application of the method. **The property should be set to false** if the variation of prevalences is expected to occur beyond the current data set and, at the same time, comparability across study cohorts or estimation of method performance upon future application are requested. In this case, only prevalence-independent metrics will be recommended.
F4.3.1	*Uncertainties in the reference:* High inter-rater variability		✓	✓	✓		The reference can be assumed to be noisy due to high inter-rater variability.
F4.3.2	*Uncertainties in the reference:* Possibility of spatial outliers in reference annotation		✓		✓		The reference may feature spatial outliers that are distant from the (unknown) ground truth.
F4.4	Granularity of provided reference annotations			✓			The granularity of the reference can vary in object detection problems. We distinguish three main categories: **Only position:** Given an n-dimensional image, the object is represented by its position, encoded in n degrees of freedom (e.g. xy/xyz coordinates of the center point) **Rough outline:** A rough outline of the object is provided, typically given by simple geometric objects such as bounding boxes or ellipsoids. **Exact outline:** The object is outlined exactly.
F4.5	Non-independence of test cases	✓	✓	✓	✓		The test cases are hierarchically structured, indicating non-independence of test cases. *Examples: multiple images of the same patient, hospital, or video.*
F4.6	Possibility of reference without target structure(s)		✓	✓	✓		There are test cases in which the reference for at least one class is empty.

Table 5.5: Problem fingerprint for **algorithm output problem characteristics**. Each problem fingerprint comes with an identifier (ID), an illustration, and a description. In addition, every fingerprint is assigned to one or multiple problem categories (Image-level Classification (ImLC), Semantic Segmentation (SemS), Object Detection (ObD), Instance Segmentation (InS)). Table adapted from Maier-Hein et al. [2022].

ID	Fingerprint	ImLC	SemS	ObD	InS	Illustration	Description
F5.1	Availability of predicted class scores	✓		✓	✓		Modern algorithms output continuous class scores, which are often interpreted as predicted class probabilities. These scores contain relevant information about the performance of a model and are thus crucial for comprehensive and meaningful validation. **ObD:** In object detection, predicted class probabilities are typically available for each detected object. **InS:** Instance segmentation problems in the biomedical domain are often approached by adding a post-processing step (e.g. connected component analysis) to a semantic segmentation algorithm. In this process, predicted class probabilities often get lost. If no predicted class probabilities are available, this property is set to false.
F5.2	Possibility of algorithm output not containing the target structure(s)		✓	✓	✓		The algorithm may yield output images in which not all classes are present.
F5.3	Possibility of invalid algorithm output (e.g. Prediction is NaN)	✓	✓	✓	✓		The files representing the algorithm output can contain invalid output.
F5.4	Possibility of overlapping predictions			✓	✓		Predictions of the algorithm can potentially overlap.

Metric selection

Based on the appropriate problem category and the problem fingerprint, metrics are selected from a problem category-specific metric pool. The selection is made via several metric mappings, as shown in Figure 5.2.

For **image-level classification** problems, we recommend selecting a *multi-class counting metric* (mapping M2; Figure 5.4), but only if one is interested in a cutoff on the predicted class scores, and therefore in building a confusion matrix. If a metric of this family should be used, the researcher is guided through several questions based on the problem fingerprint. Potential candidates are Accuracy, BA, Expected Cost (EC), MCC, and Weighted Cohen's Kappa (WCK). In the case of a node requesting to select between two metrics (e.g. "Select from EC, WCK"), a decision guide helps in the decision making process. It should be noted that the EC can be adjusted by specific choice of the costs and priors to serve as a generalization of BA and Accuracy (see respective information boxes in Figure 5.4 and Equation 2.40).

To measure the performance per class, a *per-class counting metric* (mapping M3; Figure 5.5) should be selected if a confusion matrix should be created, based on which the metric is calculated. We suggest different candidates depending on the kind of cutoff for the predicted class scores. For a cutoff based on a pre-defined target value, the confusion matrix and respective metric

scores should be chosen according to the pre-defined target value (e.g. Sensitivity@Specificity). In this case, the metric should be only selected for the class with the target value. In the case of optimization- or `argmax`-based cutoff value and based on the class prevalences, we suggest the Positive Likelihood Ratio (LR+), Sensitivity, and F_β Score as candidates. If a risk-based cutoff is preferred, our recommended metrics are the NB or EC. The metric candidate(s) should be selected for each class.

If predicted class scores are available, we further recommend selecting a *multi-threshold metric* (mapping M4; Figure 5.7). For image-level classification problems, we suggest the Area under the Receiver Operating Characteristic Curve (AUROC) and Average Precision (AP) as candidates, which are chosen based on whether the class prevalences reflect the population of interest.

If calibration assessment is required, we further recommend selecting a *calibration metric* (mapping M5; Figure 5.8). For comparative calibration assessment, we recommend selecting an overall performance measure (Brier Score (BS) or Negative Log Likelihood (NLL)) capturing both discrimination and calibration for the comparison of (re-) calibration methods of a fixed classifier (use case U1 in Figure 2.23) or for comparison of the overall performance of multiple classifiers (use case U3 in Figure 2.23). For the comparison of the calibration performance across classifiers (use case U2 in Figure 2.23), we recommend either the Class-wise Calibration Error (CWCE) if there is no mismatch between the class prevalences and class importances, and Kernel Calibration Error (KCE) (canonical calibration) otherwise. For an interpretable estimate of the calibration (use case U4 in Figure 2.23), we recommend reporting the CWCE per class in the case of a mismatch between class prevalences and class importances. Otherwise, we recommend selecting between either top-label calibration (Expected Calibration Error (ECE)) or canonical calibration (Expected Calibration Error Kernel Density Estimate (ECE^{KDE})), the latter with additionally reporting the CWCE per class. The decision between both strategies is facilitated by the respective decision guide. Since ECE, ECE^{KDE}, and CWCE are known to potentially underestimate the true calibration error [Gruber and Buettner, 2022], we advise reporting the Root Brier Score (RBS) in addition to either metrics as a guaranteed lower bound of the true calibration error [Gruber and Buettner, 2022]. This guarantee provides additional information, especially in safety-critical applications where the calibration error must not be underestimate.

For **semantic segmentation** problems, we recommend selecting an *overlap-based* and a *boundary-based metric*. An overlap-based metric (mapping M6; Figure 5.9) should always be chosen unless the target structures are all consistently small and the reference annotation is very noisy. In this case, we recommend reformulating the problem as an object detection task. We recommend the Centerline Dice Similarity Coefficient (clDice) if the centerline of structures is the exclusive interest. If not, based on potentially unequal penalization of over- and undersegmentation, we recommend either selecting the F_β Score or the DSC or Intersection over Union (IoU) based on the respective decision guide. The mapping should be repeated for every class.

In most cases, the overlap-based metrics should be accompanied by a boundary-based metric (mapping M7; Figure 5.10). More specifically, a boundary-based metric should be selected if one is particularly interested in the structure boundary, if a specific handling strategy for outliers is requested, if imprecisions in the annotation should be compensated or if the target structures are of variable sizes. If the compensation for imprecisions in the annotation is required, we recommend selecting the Normalized Surface Distance (NSD). If not, based on the outlier

strategy, we suggest the Average Symmetric Surface Distance (ASSD), Boundary Intersection over Union (Boundary IoU), Hausdorff Distance (HD), Mean Absolute Surface Distance (MASD), NSD, or the X^{th} Percentile HD (e.g. HD95) as candidates. Boundary-based metrics should not be used if overlapping or touching structures are possible to avoid comparing a structure to the wrong boundary. The mapping should be repeated for every class.

For **object detection** problems, we first recommend choosing an appropriate *localization criterion* (mapping M8; Figure 5.11). The recommendation is based on the granularity of the localization and the reference annotation. Potential candidates include the Center Distance (if only the position is of importance), the Point inside the Mask/Box/Approximation or the Mask Intersection over Union (Mask IoU) that is greater than 0 (if only the position is of importance although an exact or rough outline is available), or the Box or Approximation IoU for interest in the rough outline. The mapping should be repeated for every class.

Afterwards, *the assignment strategy* (mapping M9; Figure 5.12) is selected. Based on the availability of the predicted class scores, we recommend either selecting the Greedy (by Score) Matching (if available), or choosing between the Greedy (by "localization criterion"[4]) Matching, the Optimal (Hungarian) Matching and the Matching via the "localization criterion"[4] > 0.5. In addition, one should decide whether multiple predictions assigned to the same reference structure should be penalized as FPs or not. The mapping should be repeated for every class.

Moreover, metrics for assessing the detection performance should be selected. Here, we recommend following a similar branch compared to the image-level classification metrics, namely in choosing a *per-class counting metric* (mapping M3; Figure 5.6) and a *multi-threshold metric* (mapping M4; Figure 5.7). For the per-class counting metrics, the decision is once more dependent on the interest in a cutoff value. If so, along with a pre-defined target, we recommend choosing a metric at a specific target value, e.g. PPV@Sensitivity. If there is no pre-defined target value, we recommend the F_β Score. For the multi-threshold metric, we suggest the AP and Free-Response Receiver Operating Characteristic (FROC) Score as candidates. The decision should be based on the respective decision guide. If the predicted class scores should be interpretable, we recommend selecting a *calibration metric* (mapping M5; Figure 5.8), such as the BS or the ECE. In object detection problems, we do not recommend the NLL (see Section 2.4.5).

For **instance segmentation** problems, we recommend combining the object detection and semantic segmentation pipeline. In order to assess the detection quality, we suggest selecting the *localization criterion* (mapping M8; Figure 5.11; here we recommend selecting between Boundary IoU, Mask IoU and Intersection over Reference (IoR)), and the *assignment strategy* (mapping M9; Figure 5.12). As detection metrics, we analogously select a *per-class counting metric* (mapping M3; Figure 5.6), and a *multi-threshold metric* (mapping M4; Figure 5.7). For the latter, we further add the Panoptic Quality (PQ) to the candidates. If the predicted class scores should be interpretable, we recommend selecting a *calibration metric* (mapping M5; Figure 5.8). In instance segmentation problems, we do not recommend the NLL (see Section 2.4.5). Finally, to assess the segmentation quality, we recommend selecting an *overlap-based metric* (mapping M6; Figure 5.9) and a *boundary-based metric* (mapping M7; Figure 5.10). Note that the segmentation metrics are only applied to the identified True Positive (TP) instances and are computed per instance.

[4]"Localization criterion" refers to the overlap-based localization criterion selected in mapping M8 (Figure 5.11).

Metric mappings

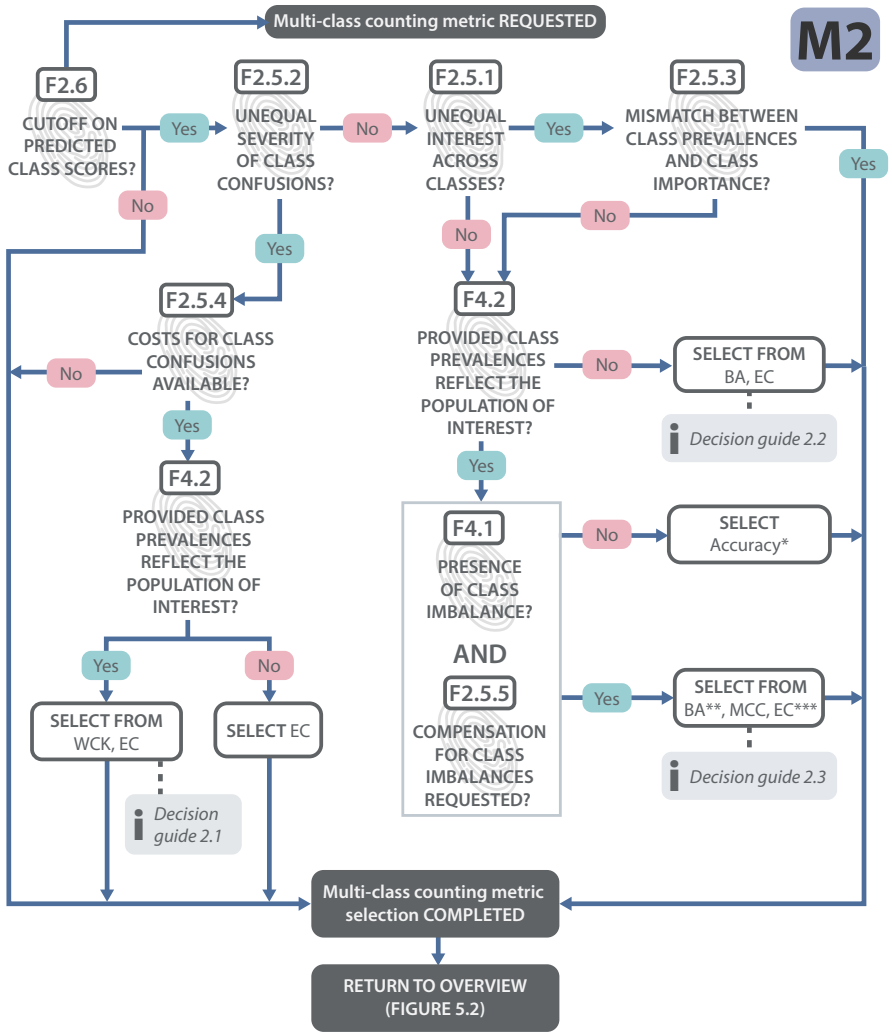


Figure 5.4: **Mapping M2: Multi-class counting metrics.** The mapping is based on the generated problem fingerprint, which is referred to as "FX.Y" for image-level classification problems. The decision guides can be found in Tables 5.6 - 5.8. An overview of relevant limitations of the metric candidates is provided in Table 4.1. Used abbreviations: Balanced Accuracy (BA), Expected Cost (EC), and Matthews Correlation Coefficient (MCC). Figure adapted from [Maier-Hein et al., 2022].

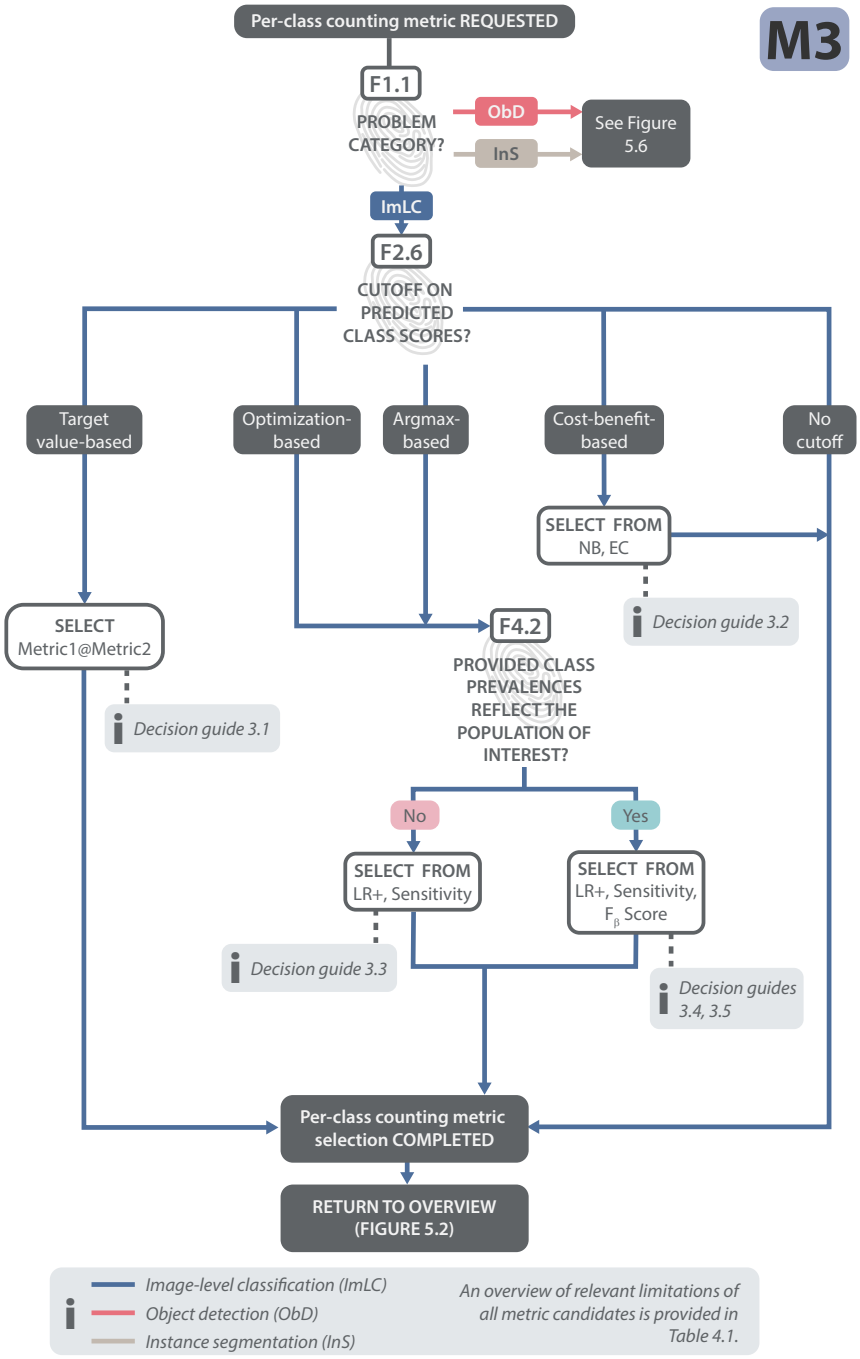


Figure 5.5: **Mapping M3 (Part 1): Per-class counting metrics.** The mapping is based on the generated problem fingerprint, which is referred to as "FX.Y" for image-level classification (ImLC), object detection (ObD), and instance segmentation (InS) problems. The path through M5 for ObD and InS problems is shown in Figure 5.6 to enhance the readability. The decision guides can be found in Tables 5.9 - 5.13. An overview of relevant limitations of the metric candidates is provided in Table 4.1. Further abbreviation: Positive Likelihood Ratio (LR+). Figure adapted from [Maier-Hein et al., 2022].

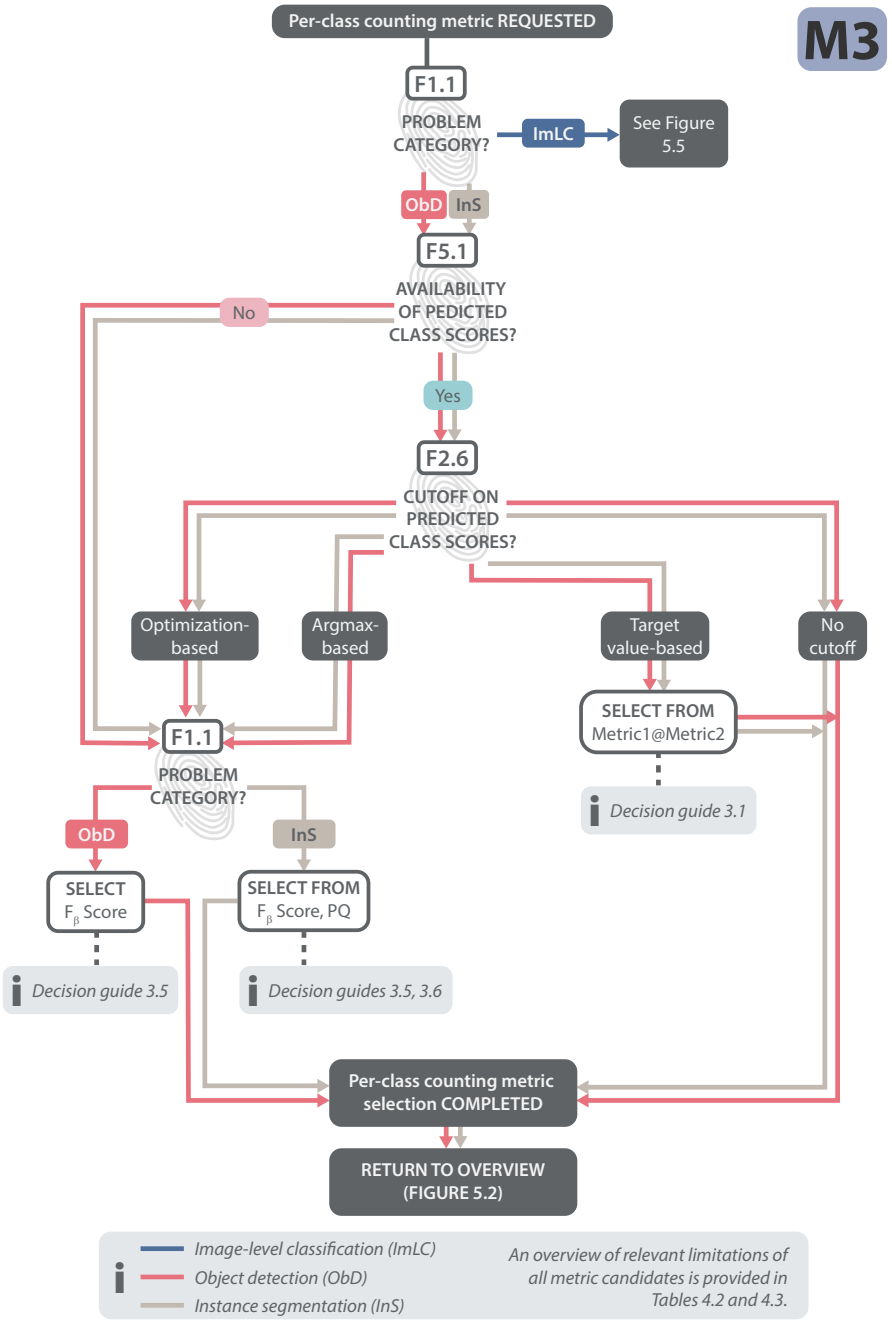


Figure 5.6: **Mapping M3 (Part 2): Per-class counting metrics.** The mapping is based on the generated problem fingerprint, which is referred to as "FX.Y" for image-level classification (ImLC), object detection (ObD), and instance segmentation (InS) problems. The path through M5 for ImLC problems is shown in Figure 5.5 to enhance the readability. The decision guides can be found in Tables 5.10 - 5.13. An overview of relevant limitations of the metric candidates is provided in Tables 4.1 and 4.3. Further abbreviation: Panoptic Quality (PQ). Figure adapted from [Maier-Hein et al., 2022].

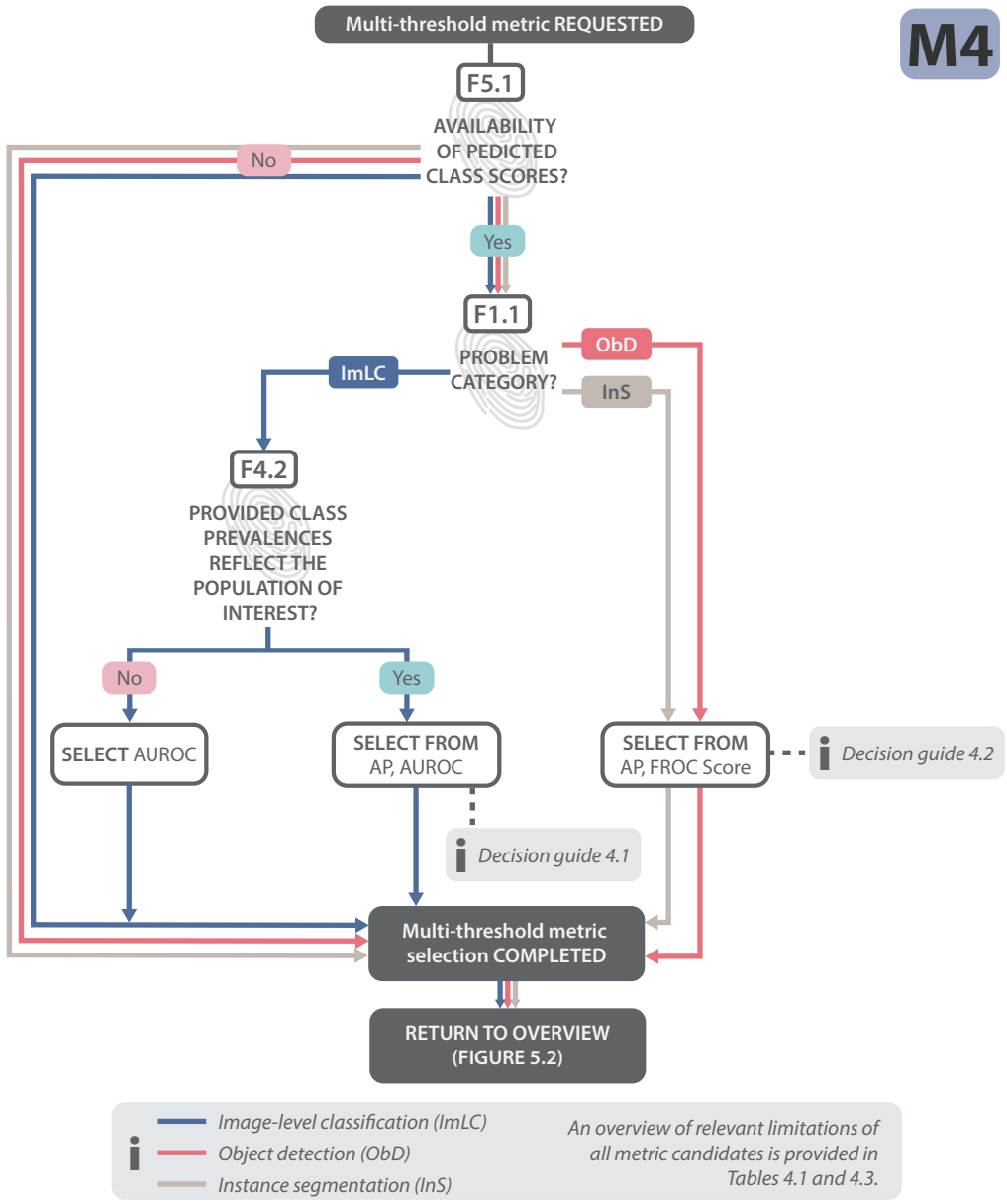


Figure 5.7: **Mapping M4: Multi-threshold metrics.** The mapping is based on the generated problem fingerprint, which is referred to as "FX.Y" for image-level classification (ImLC), object detection (ObD), and instance segmentation (InS) problems. The decision guide can be found in Table 5.14. An overview of relevant limitations of the metric candidates is provided in Tables 4.1 and 4.3. Further abbreviations: Area under the Receiver Operating Characteristic Curve (AUROC), Average Precision (AP), and Free-Response Receiver Operating Characteristic (FROC) Score. Figure adapted from [Maier-Hein et al., 2022].

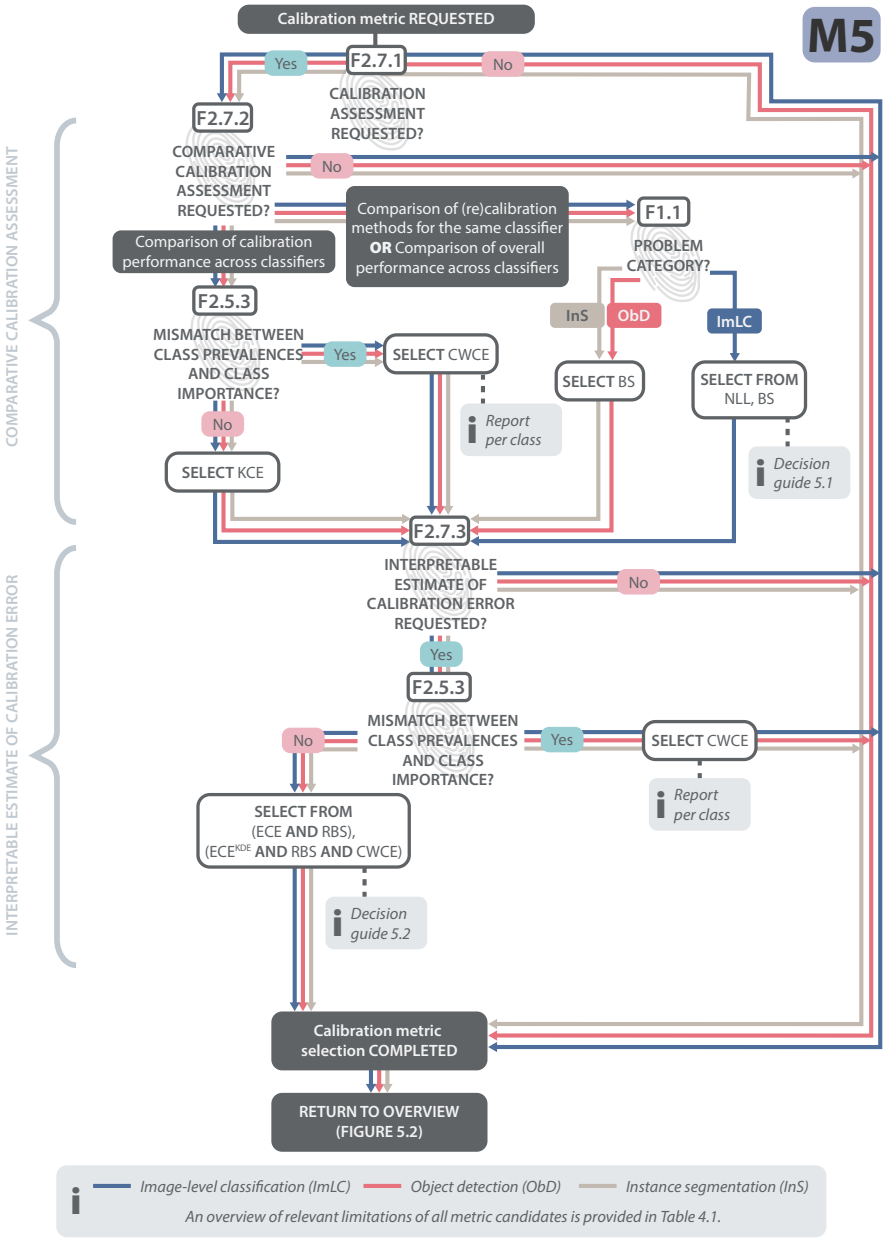


Figure 5.8: **Mapping M5: Calibration metrics.** The mapping is based on the generated problem fingerprint, which is referred to as "FX.Y" for image-level classification (ImLC), object detection (ObD), and instance segmentation (InS) problems. The decision guide can be found in Tables 5.16 - 5.17. An overview of relevant limitations of the metric candidates is provided in Table 4.1. Further abbreviation: Expected Calibration Error (ECE). Figure adapted from [Maier-Hein et al., 2022].

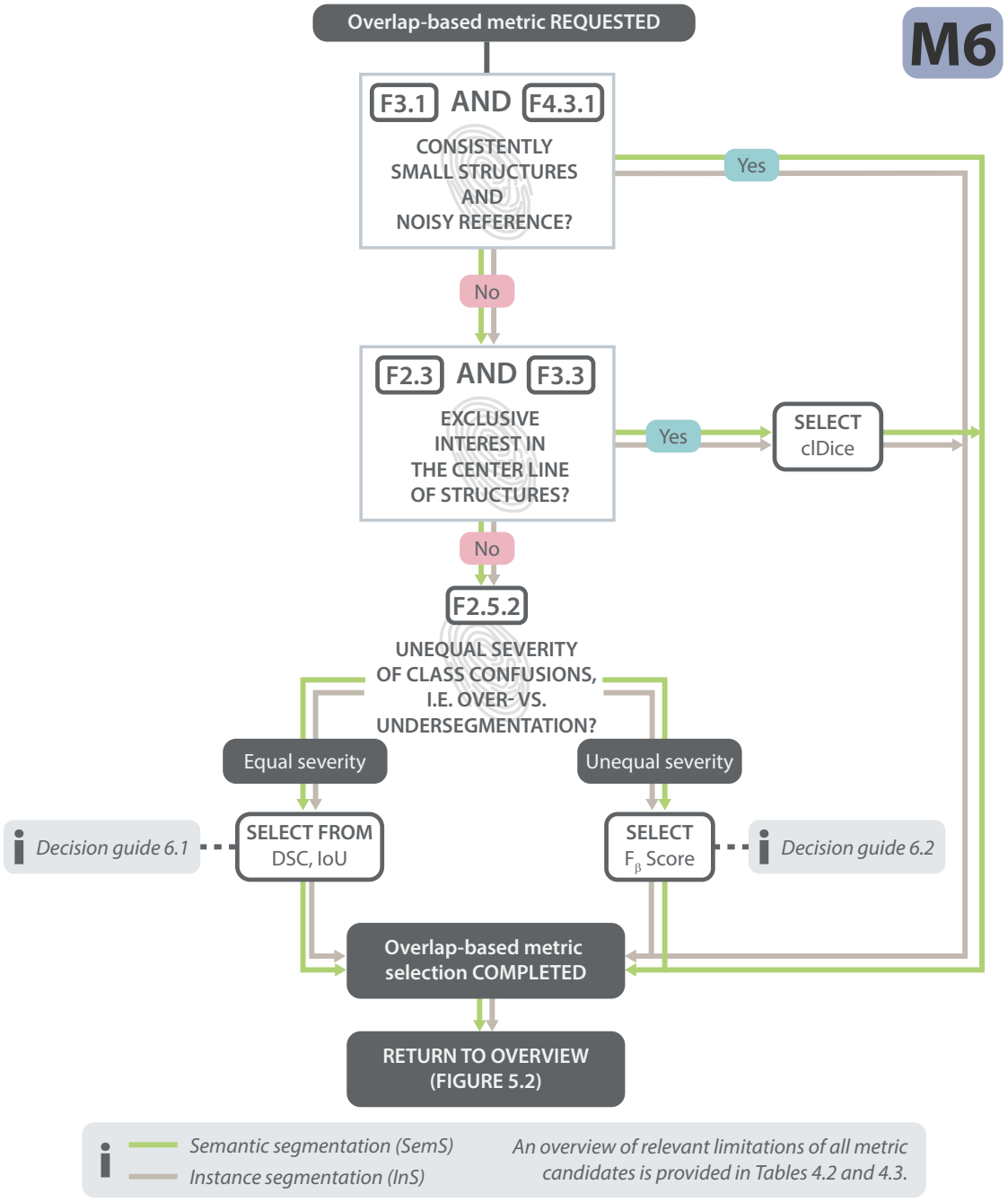


Figure 5.9: **Mapping M6: Overlap-based metrics.** The mapping is based on the generated problem fingerprint, which is referred to as "FX.Y" for semantic segmentation (SemS) and instance segmentation (InS) problems. The decision guides can be found in Tables 5.18 - 5.19. An overview of relevant limitations of the metric candidates is provided in Tables 4.2 and 4.3. Further abbreviations: Centerline Dice Similarity Coefficient (clDice), Dice Similarity Coefficient (DSC) and Intersection over Union (IoU). Figure adapted from [Maier-Hein et al., 2022].

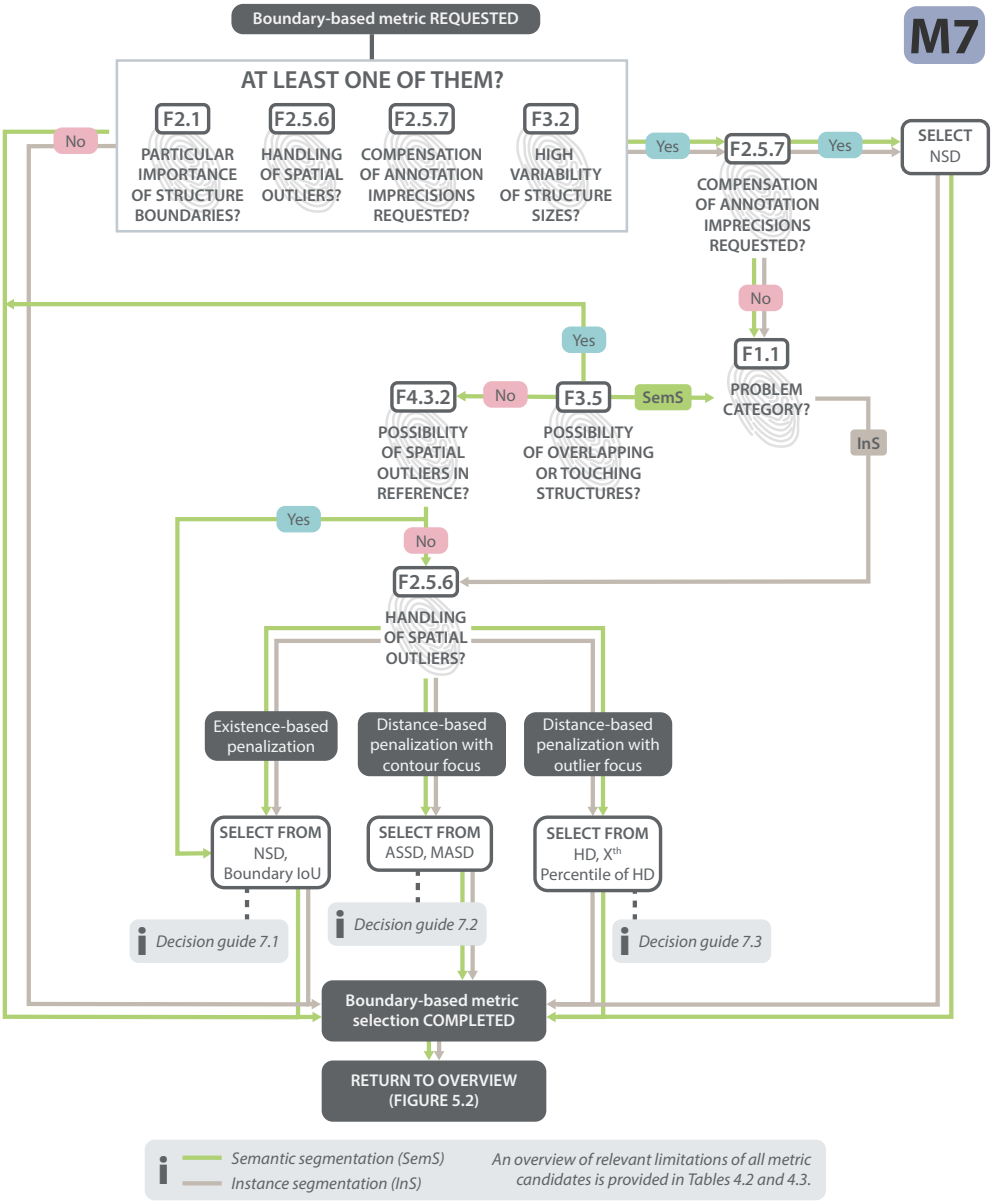


Figure 5.10: **Mapping M7: Boundary-based metrics.** The mapping is based on the generated problem fingerprint, which is referred to as "FX.Y" for semantic segmentation (SemS) and instance segmentation (InS) problems. The decision guides can be found in Tables 5.20 - 5.22. An overview of relevant limitations of the metric candidates is provided in Tables 4.2 and 4.3. Further abbreviations: Average Symmetric Surface Distance (ASSD), Boundary Intersection over Union (Boundary IoU), Hausdorff Distance (HD), Mean Absolute Surface Distance (MASD), Normalized Surface Distance (NSD), and Intersection over Union (IoU). Figure adapted from [Maier-Hein et al., 2022].

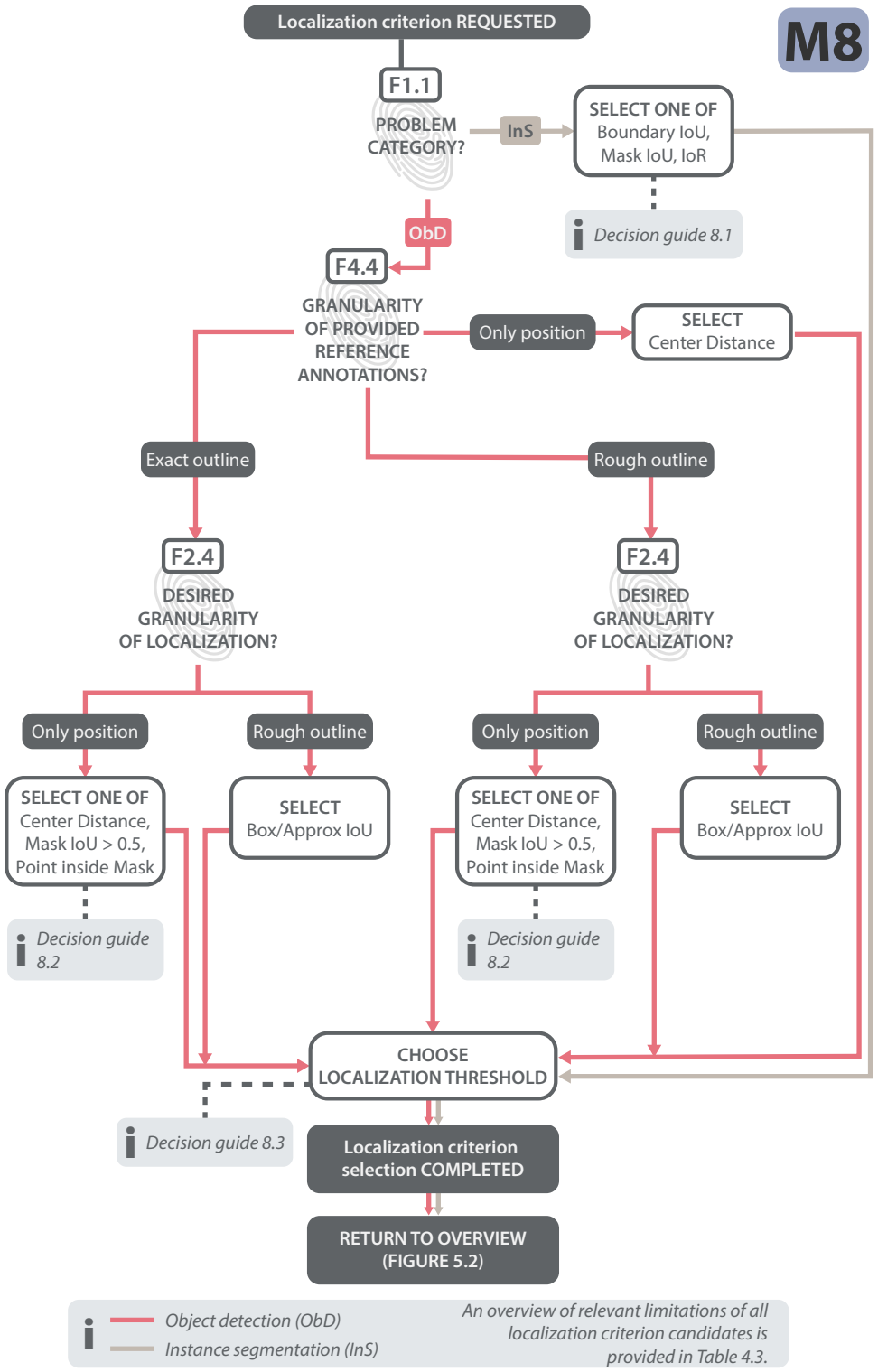


Figure 5.11: **Mapping M8: Localization criteria.** The mapping is based on the generated problem fingerprint, which is referred to as "FX.Y" for object detection (ObD) and instance segmentation (InS) problems. The decision guides can be found in Tables 5.23 - 5.25. An overview of relevant limitations of the metric candidates is provided in Table 4.3. Further abbreviations: Boundary Intersection over Union (Boundary IoU), Box/Approximation Intersection over Union (Box/Approx IoU), Intersection over Reference (IoR), and Mask Intersection over Union (Mask IoU). Figure adapted from [Maier-Hein et al., 2022].

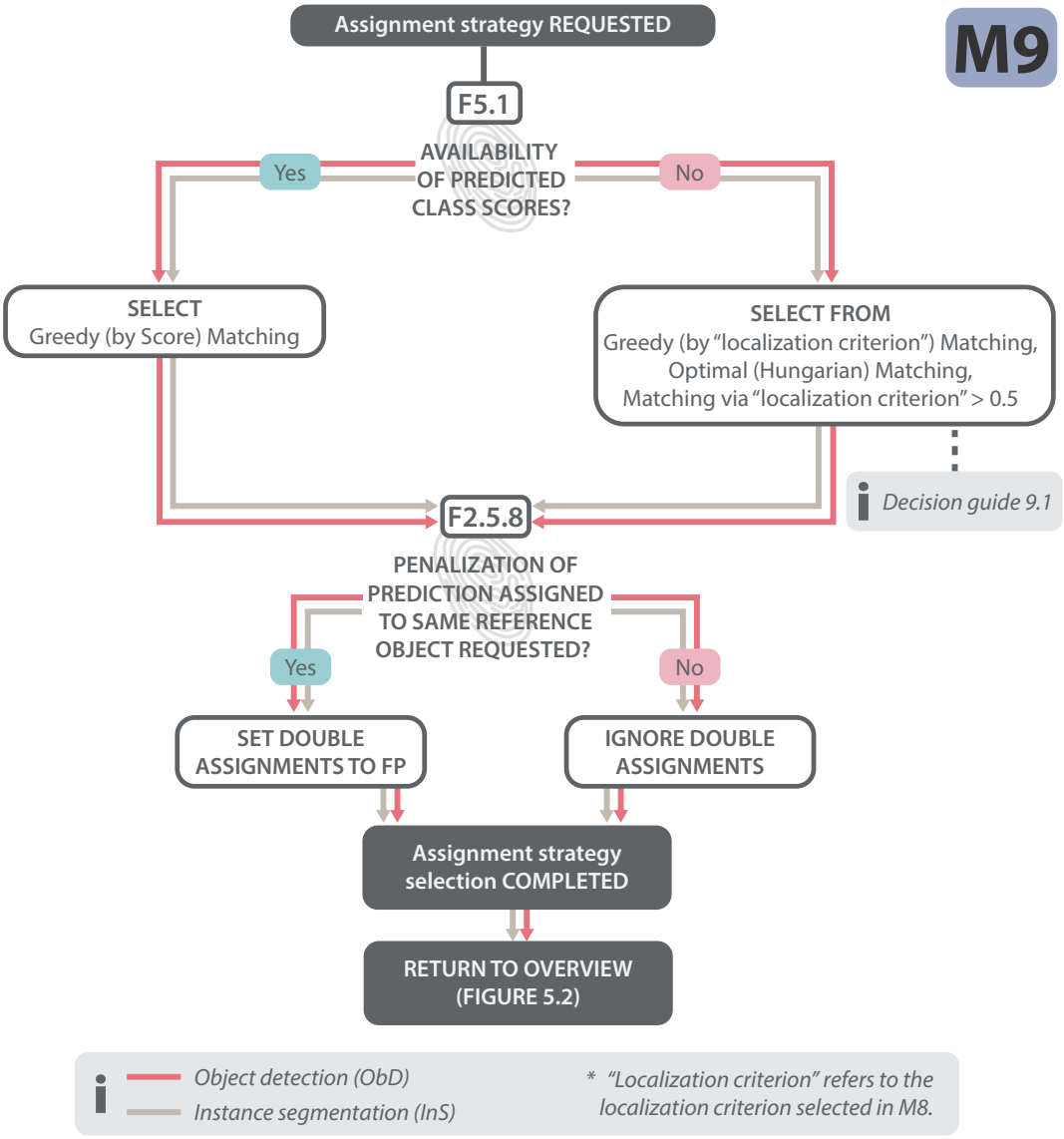


Figure 5.12: **Mapping M9: Assignment strategies.** The mapping is based on the generated problem fingerprint, which is referred to as "FX.Y" for object detection (ObD) and instance segmentation (InS) problems. The decision guide can be found in Table 5.26. Figure adapted from [Maier-Hein et al., 2022].

Decision guides

In the following, we present the decision guides, which should be used for the final step of the problem-aware metric selection. Most of the metric mappings end in selection nodes, in which a decision between two or more metrics is required. The decision guides list the advantages and disadvantages of the metrics. Only the conditions mentioned in the relevant mappings should be used to make the decision, i.e. the decision guides should not be used for a standalone decision;

they already incorporate the traversal through the mappings and therefore, the decisions based on the problem fingerprint. It should be noted that some properties of the metric candidates can be both an advantage and a disadvantage based on the actual interest and problem.

Decision guides for selecting multi-class counting metrics (M2)

Decision guide 2.1 This decision guide refers to mapping M2 (Figure 5.4) for choosing between the WCK and the EC as a multi-class counting metric. Table 5.6 lists the advantages and disadvantages of both metrics.

Table 5.6: Decision guide 2.1 for choosing between Weighted Cohen's Kappa (WCK) and the Expected Cost (EC) as a multi-class counting metric (mapping M2 in Figure 5.4). Context: unequal interest across classes (F2.5.1 ✔) and/or unequal severity of class confusions (F2.5.2 ✔) and provided class prevalences reflect the population of interest (F4.2 ✔). Advantages of a metric are indicated by 👍, while disadvantages are shown by 👎. Table adapted from Maier-Hein et al. [2022].

WCK	EC
👍 Allows for problem-specific penalties	👍 Allows for problem-specific penalties
👍 Commonly used in the biomedical community	👍 Thorough theoretical grounding
👎 Difficult to interpret	👍 Intuitive interpretation
👎 Symmetric measure for an asymmetric situation (comparing reference and prediction rather than multiple raters)	👍 Asymmetric measure
👎 Quadratic weighted version may yield "paradoxical results" [Warrens, 2012]	👍 Simultaneous assessment of discrimination and calibration
	👎 Uncommon in the biomedical community

Decision guide 2.2 This decision guide refers to mapping M2 (Figure 5.4) for choosing between the BA and the EC as a multi-class counting metric. Table 5.7 lists the advantages and disadvantages of both metrics.

Table 5.7: Decision guide 2.2 for choosing between the Balanced Accuracy (BA) and the Expected Cost (EC) as a per-class counting metric (mapping M2 in Figure 5.4). Context: Equal interest across classes (F2.5.1 ✘), equal severity of class confusions (F2.5.2 ✘), and provided class prevalences do not reflect the population of interest (F4.2 ✘). Advantages of a metric are indicated by 👍, while disadvantages are shown by 👎. Table adapted from Maier-Hein et al. [2022].

BA	EC
👍 Intuitive interpretation	👍 Enables modification of the class priors for the intended use
👎 Assumes that class priors are uniform and not modifiable	👎 Difficult to interpret; may require usage of the normalized variant of the metric

Decision guide 2.3 This decision guide refers to mapping M2 (Figure 5.4) for choosing between the BA, the MCC, and the normalized variant of the EC as a multi-class counting metric. In class-balanced scenarios, Accuracy is the default metric (see M2; Figure 5.4). However, class imbalance introduces three effects on this metric that can be compensated: (1) lost random reference, (2) lost equal importance of classes, and (3) lost consideration of predictive values. BA, MCC, and the normalized variant of the EC can compensate for them in different aspects, as indicated in Table 5.8.

Table 5.8: Decision guide 2.3 for choosing between Balanced Accuracy (BA), Matthews Correlation Coefficient (MCC), and the normalized variant of the Expected Cost (EC) (EC_{norm}) as a multi-class counting metric (mapping M2 in Figure 5.4). Context: class prevalences reflect the population of interest (F4.2 ✔) but compensation for imbalanced classes is required (F4.1 ✔ and F2.5.5 ✔). Advantages of a metric are indicated by 👍, while disadvantages are shown by 👎. Further abbreviations: Negative Predictive Value (NPV), Positive Predictive Value (PPV). Table adapted from Maier-Hein et al. [2022].

BA	**MCC**	**EC_{norm}**
👍 Restores random reference ($1/n_{classes}$) 👍 Restores equal importance of classes: errors are penalized with respect to respective class Sensitivities, thus ensuring equal importance across class Sensitivities 👍 Intuitive interpretation 👎 Does not restore lost considerations of predictive values, i.e. BA can yield near-perfect scores for predictions with low predictive value	👍 Restores random reference (0) 👍 Restores equal importance of classes: errors are penalized in reference to respective Sensitivities and predictive values giving equal weight across Sensitivities and across predictive values irrespective of prevalence 👍 Restores consideration of predictive values (all four basic rates need to be high to get a high MCC score) 👎 No interpretable scale for non-random values, which makes it harder to interpret	👍 Restores random reference (1) 👍 Simple rescaling of Accuracy, but ranking remains unchanged 👍 Rescaling ensures that above-random performance implies good predictive values 👎 Does not restore equal importance of classes: even errors in the dominant class are counted as if they occurred in the rare classes, thereby not giving equal weight to individual class Sensitivities (analogously to Accuracy). This makes EC_{norm} the strictest of the three metrics in terms of interpreting systems to perform better than random (the total number of errors, irrespective of the associated class, needs to be lower than the number of events in the rare class)

It should be noted that in the specific case of this decision, BA and EC are equivalent if the costs should compensate for class imbalance, i.e. the costs of EC are chosen as $w_{ii} = 0$ and $w_{ij} = \frac{1}{NP_i}$ with N referring to the number of classes and P_i being the class prevalence of class i (see Equation 2.45). However, for this decision guide, we explicitly use the normalized variant of EC as an additional metric candidate.

Decision guides for selecting per-class counting metrics (M3)

Decision guide 3.1 This decision guide refers to mapping M3 (Figures 5.5 and 5.6) for choosing a metric according to a specified target value (Metric1@Metric2). In such a situation, the cutoff is the threshold at which a particular pre-defined target metric value is attained. For example, an application might require a Sensitivity of 0.98. The chosen thresholds are then optimized for this target value and other metrics can be reported for this target value. By drawing the AUROC, for example, the corresponding Specificity value can be easily obtained. Based on the general situation, the exact metrics can be chosen. The following points help to select a suitable metric:

Image-level classification: If class prevalences do **not** reflect the population of interest (F4.2 ✖), prevalence-dependent metrics should be avoided. Instead, the Sensitivity@Specificity and Specificity@Sensitivity are proper metrics, depending on the metric for which a target-value was defined. If prevalences reflect the population of interest (F4.2 ✔), and class imbalances should be compensated (F4.1 ✔ and F2.5.5 ✔), PPV/NPV@Sensitivity (or vice versa) can be selected instead.

Object detection/instance segmentation: Since True Negative (TN) are typically undefined in object detection and instance segmentation tasks, Specificity and NPV should usually not be selected. Instead, metrics that can be derived from the Precision-Recall (PR) or FROC curves should be prioritized (ideally, according to the selected multi-threshold metric in M4), namely PPV/False Positives per Image (FPPI)@Sensitivity or vice versa. While the AP metric (PPV@Sensitivity/Sensitivity@PPV) is broadly known in the computer vision community and has a standardized definition, the FROC Score (FPPI@Sensitivity/Sensitivity@FPPI) is more commonly used by clinicians given its easy interpretation. However, the range of the FPPI is not standardized. Popular benchmarks [Van Ginneken et al., 2010; Setio et al., 2017] used the following values: 1/8, 1/4, 1/2, 1, 2, 4, 8..

Decision guide 3.2 This decision guide refers to mapping M3 (Figure 5.5) for choosing between the NB and the EC as a per-class counting metric. Both metrics are linked to cost-benefit analysis and are technically very similar since both are computed by multiplying costs and (variations of) error rates. Table 5.9 lists the advantages and disadvantages of both metrics.

Table 5.9: Decision guide 3.2 for choosing between theNet Benefit (NB) and the Expected Cost (EC) as a per-class counting metric (mapping M3 in Figure 5.5). Context: benefit-cost-based cutoff on predicted class scores requested (F2.6). Advantages of a metric are indicated by 👍, disadvantages are shown by 👎, and 👉 refers to neutral comments. Table adapted from Maier-Hein et al. [2022].

NB	EC
👉 Costs are implicitly defined via a risk expressing all cost-benefit considerations in one intuitive notion	👉 Costs are explicitly defined
👍 Comes with individually definable exchange rate parameter for benefit-cost tradeoff	👍 Comes with individually definable cost parameters
👍 Allow to determine a cutoff on predicted class scores based on risks	👍 Allows to determine a cutoff on predicted class scores based on costs or to alternatively determine an empirical cutoff by being minimized based on a dedicated data split
👍 Intuitive interpretation	👍 Enables modification of the class priors for the intended use and can thus be made prevalence-independent
👍/👎 Should be assessed for a range of clinically relevant threshold	👍/👎 Can be assessed at a single threshold
👎 Assumes that class priors are uniform and not modifiable	👍/👎 Normalized variant enables good interpretability
👎 Rather unknown in the computer vision community, while popular in clinical studies	👎 Rather unknown in the computer vision community

Decision guide 3.3 This decision guide refers to mapping M3 (Figures 5.5 and 5.6) for choosing between the LR+ and Sensitivity as a per-class counting metric. Table 5.10 lists the advantages and disadvantages of both metrics.

Table 5.10: Decision guide 3.3 for choosing between the LR+ and Sensitivity as a per-class counting metric (mapping M3 in Figures 5.5 and 5.6). Context: Optimization- or argmax-based cutoff on predicted class scores (F2.6) and provided class prevalences not reflecting the population of interest (F4.2 ✗). Advantages of a metric are indicated by 👍, while disadvantages are shown by 👎. Table adapted from Maier-Hein et al. [2022].

Sensitivity	**LR+**
👍 **Sensitivity and LR+ convey similar information when reported for all classes individually**	
👍 Easy to interpret in multi-class settings (e.g. "one versus rest") 👎 If the cutoff is to be determined based on optimization on the target class of a dedicated data split (F2.6), special consideration is required when selecting Sensitivities per class: The cutoff can not be optimized based on a single Sensitivity in the target class. Possible workarounds include re-considering F2.6 to opt for an argmax-based cutoff or optimizing a weighted average over sensitivity for all classes instead. The latter option should only be considered if meaningful weights across classes can be defined (e.g. based on class importance)	👍 Binary classification: LR+ of the target class conveys Sensitivities of both classes in a single score and is often reported in clinical studies due to its intuitive interpretation

Decision guide 3.4 This decision guide refers to mapping M3 (Figures 5.5 and 5.6) for choosing between the LR+, Sensitivity, and the F_β as a per-class counting metric. For this decision guide, it should be noted that per-class validation is commonly performed in a "one versus rest" fashion that naturally introduces class imbalance in the validation. Exceptions are binary scenarios with two balanced classes. For this exception, no compensation for class imbalance is needed (F2.5.5 ✗) and the choice between the three metrics becomes less relevant, i.e. no relevant pitfalls occur for either of the three. Thus, the decision can be made based on which metric is easier to interpret in a given task.

If deciding whether compensation for class imbalance is required (F2.5.5 ✓), it should be noted that – because metrics are reported individually per class – effects of restoring a random reference and equal class balance are not relevant in this context. Table 5.11 lists the advantages and disadvantages of the three metrics.

Table 5.11: Decision guide 3.4 for choosing between the LR+, Sensitivity, and the F_β as a per-class counting metric (mapping M3 in Figures 5.5 and 5.6). Context: Optimization- or argmax-based cutoff on predicted class scores (F2.6) and provided class prevalences reflecting the population of interest (F4.2 ✔). Advantages of a metric are indicated by 👍, while disadvantages are shown by 👎. Table adapted from Maier-Hein et al. [2022].

Sensitivity	**LR+**	**F_β Score**
👍 Straightforward interpretation 👎 Issue with data-driven optimization of the cutoff (see decision guide 3.3 in Table 5.10) 👎 Does not restore lost considerations of predictive values. Pitfall: BA (averaged Sensitivities per class) may yield near-perfect scores for systems with low predictive value	👍 Binary classification: LR+ of the target class conveys Sensitivities of both classes in a single score and is often reported in clinical studies due to its intuitive interpretation 👎 More complex interpretation in multi-class settings 👎 Does not restore lost considerations of predictive values. Pitfall: BA can yield near-perfect scores for systems with low predictive value	👍 Restores consideration of predictive values, protecting against associated failure modes 👍/👎 Compared to Sensitivities per class, this adds a new layer of complexity w.r.t interpretation, although the score can be interpreted as the harmonic mean of sensitivity and positive predictive value per class

Decision guide 3.5 This decision guide refers to mapping M3 (Figures 5.5 and 5.6) for determining the hyperparameter β of the F_β Score. Table 5.12 lists the implications of different values.

Table 5.12: Decision guide 3.5 for determining the hyperparameter β of the F_β Score as a per-class counting metric (mapping M3 in Figures 5.5 and 5.6). Context (Image-level Classification (ImLC)): Optimization- or argmax-based cutoff on predicted class scores (F2.6) and provided prevalences reflect the population of interest (F4.2 ✔). Context (Object Detection (ObD)/Instance Segmentation (InS)): Either no predicted class scores available (F5.1 ✘) or optimization- or argmax-based cutoff (F2.6). Advantages of a metric are indicated by 👍, while disadvantages are shown by 👎. Used abbreviations: False Positive (FP), False Negative (FN), Positive Predictive Value (PPV). Table adapted from Maier-Hein et al. [2022].

$\beta < 1$ (e.g. 0.5)	$\beta = 1$	$\beta > 1$ (e.g. 2)
Higher penalization of PPV, i.e. FP	Equal treatment of PPV and Sensitivity, i.e. FP and FN	Higher penalization of Sensitivity, i.e. FN

Decision guide 3.6 This decision guide refers to mapping M3 (Figures 5.5 and 5.6) for choosing between the F_β Score and the PQ as a per-class counting metric. Table 5.13 lists the advantages and disadvantages of both metrics.

Table 5.13: Decision guide 3.6 for choosing between the F_β Score and the Panoptic Quality (PQ) as a per-class counting metric (mapping M3 in Figures 5.5 and 5.6). Context: Either no predicted class scores available (F5.1 ✗) or optimization- or argmax-based cutoff on predicted class scores (F2.6) and instance segmentation (F1.1 = IS). Table adapted from Maier-Hein et al. [2022].

F_β Score	**PQ**
👍 / 👎 Pure detection metric	👍 / 👎 Simultaneous assessment of detection and segmentation performance in one score

Decision guides for selecting multi-threshold metrics (M4)

Decision guide 4.1 This decision guide refers to mapping M4 (Figure 5.7) for choosing between the AP and the AUROC as a multi-threshold metric. Table 5.14 lists the advantages and disadvantages of both metrics.

Table 5.14: Decision guide 4.1 for choosing between Average Precision (AP) and the Area under the Receiver Operating Characteristic Curve (AUROC) as a multi-threshold metric (mapping M4 in Figure 5.7). Context: availability of predicted class scores (F5.1 ✔), Image-level Classification (ImLC) (F1.1), and provided class prevalences reflect the population of interest (F4.2). Advantages of a metric are indicated by 👍, while disadvantages are shown by 👎. Used abbreviations: False Positive (FP), True Negative (TN). Table adapted from Maier-Hein et al. [2022].

AP	**AUROC**
👍 In the case of class imbalance, AP restores the consideration of predictive values and prevents associated failure cases of AUROC 👎 Features no similar interpretation to AUROC	👍 Features interpretability as the probability of a randomly sampled positive case having a higher score than a randomly sampled negative case and a fixed random reference score at 0.5 👎 Class imbalance: may produce overly optimistic scores, since FP penalization is suppressed by a large number of TN

Decision guide 4.2 This decision guide refers to mapping M4 (Figure 5.7) for choosing between the AP and the FROC Score as a multi-threshold metric. Table 5.15 lists the advantages and disadvantages of both metrics.

Table 5.15: Decision guide 4.2 for choosing between Average Precision (AP) and the Free-Response Receiver Operating Characteristic (FROC) Score as a multi-threshold metric (mapping M4 in Figure 5.7). Context: availability of predicted class scores (F5.1 ✔) and Object Detection (ObD) or Instance Segmentation (InS) (F1.1). Further abbreviations: False Positives per Image (FPPI). Advantages of a metric are indicated by 👍, while disadvantages are shown by 👎. Table adapted from Maier-Hein et al. [2022].

AP	FROC Score
👍 Commonly used in computer vision community 👍 Standardized definition 👎 Rejecting predictions of low confidence requires setting a cutoff value for the confidence scores 👎 Neglects the number of images	👍 Commonly used in the clinical context 👍 Incorporates the number of images 👍 Rejecting predictions of low confidence is naturally given 👎 Definition of FPPI not standardized. Popular benchmarks (Van Ginneken et al. [2010]; Setio et al. [2017]) used the following values: 1/8, 1/4, 1/2, 1, 2, 4, 8.

Decision guides for selecting calibration metrics (M5)

Decision guide 5.1 This decision guide refers to mapping M5 (Figure 5.8) for choosing between the BS and the NLL) as a calibration measure. Table 5.16 lists the advantages and disadvantages of both metrics.

Table 5.16: Decision guide 5.1 for choosing between Brier Score (BS) and Negative Log Likelihood (NLL) as a calibration measure (mapping M5 in Figure 5.8). Context: Comparison of re-calibration methods for the same classifier or comparison of overall performance across classifiers requested (F2.7.2). Advantages of a metric are indicated by 👍, while disadvantages are shown by 👎. Further abbreviation: Brier Skill Score (BSS). Table adapted from Maier-Hein et al. [2022].

BS	NLL
👎 Treats errors of all events the same irrespective of the class prevalence, i.e. scores may drastically change when the prevalence changes and thus makes it highly prevalence-dependent 👍/👎 Prevalence dependency can be compensated by normalizing using BSS instead (normalized variant of BS). Using BSS instead is a rescaling of scores for interpretability, for which errors are still treated equally (missing a frequent event is as bad as missing a rare event) resulting in a strict interpretation of above-random performance where the total number of errors has to be lower than the number of events in the rare class	👎 The logarithm introduces a stronger penalization of tail probabilities 👍/👎 Naturally higher penalization of naive systems. More conservative predictions are favored. For instance, in class imbalanced scenarios, naive systems are prone to missing rare events which results in strong penalization

Decision guide 5.2 This decision guide refers to mapping M5 (Figure 5.8) for choosing between two sets of metrics as interpretable calibration measures: (1) top-label ECE and RBS and (2) canonical ECE^{KDE}, per-class CWCE, and RBS. The decision between these two sets of metrics boils down to determining whether predicted class scores should be tested for top-label calibration, i.e. whether the interest is limited to the predicted scores that lead to the classification decision, or a more comprehensive calibration condition, i.e. requesting that all predicted scores to be calibrated.

In both cases, we recommend to additionally report RBS as a guaranteed upper bound on the calibration error. Since calibration error estimates such as ECE, CWCE, or ECE^{KDE} are know to over- or underestimate the error [Gruber and Buettner, 2022], this guarantee provides additional information, especially in safety-critical applications where the calibration error must not be underestimate. Table 5.17 lists the reasons for selecting either selecting top-label assessment (ECE) or focus on all predicted class scores (CWCE and ECE^{KDE}).

Table 5.17: Decision guide 5.2 for choosing between top-label assessment (Expected Calibration Error (ECE)) or focus on all predicted class scores (Class-wise Calibration Error (CWCE) and Expected Calibration Error Kernel Density Estimate (ECE^{KDE})) for calibration assessment (mapping M5 in Figure 5.8). Context: Interpretable estimate of calibration error requested (F2.7.3 ✔) and no mismatch between class prevalences and class importance (F2.5.3 ✖). Table adapted from Maier-Hein et al. [2022].

Top-label assessment (ECE)	Focus on all predicted class scores (CWCE and ECE^{KDE})
👉 The defined problem is linked to one particular decision process (predicting the correct class). If the underlying biomedical question refers to validating this exact decision process, top-label error might be the right choice, because it directly reflects a focus on the decisions made by the classifier. Conflating the calibration of decisions with other probabilities might be interpreted as washing out the task focus in this case	👉 Often, the interest of a problem goes beyond the core decision process of the classifier (predicting the correct class). In clinical context, for instance, several potential outcomes might trigger their individual treatment decisions given a certain probability for the outcome. In such scenarios, calibration of all probabilities (i.e. marginal and or canonical calibration) might be of interest 👉 Another reason might be manual correction of decision making: In multi-class settings, cutoffs are often based on a simple argmax operation, but in the case of unequal severity of class confusions (F2.5.2 ✔), the default cutoff might not perfectly reflect the task interest. Consider, for instance, a multi-way classification of tumor categories, where some are more aggressive and thus relevant than others. It might be desired to lower the probability thresholds at which the worst-case scenarios are considered. For such interventions, calibrated probabilities across all classes are required

Decision guides for selecting overlap-based metrics (M6)

Decision guide 6.1 This decision guide refers to mapping M6 (Figure 5.9) for choosing between the DSC and IoU as an overlap-based metric. Table 5.18 lists the advantages and disadvantages of both metrics.

Table 5.18: Decision guide 6.1 for choosing between the Dice Similarity Coefficient (DSC) and Intersection over Union (IoU) as an overlap-based metric (mapping M6 in Figure 5.9). Context: no exclusive interest in the centerline of structures (F2.3 ✖, F3.3 ✖) and equal severity of class confusions (F2.5.2 ✖). Advantages of a metric are indicated by 👍, while 👉 refers to neutral comments. Table adapted from Maier-Hein et al. [2022].

DSC	IoU
👉 Closely related to IoU, see Equation 2.50 👍 Commonly used in the medical community	👉 Closely related to DSC, see Equation 2.50 👍 Commonly used in the biological and computer vision communities

Decision guide 6.2 This decision guide refers to mapping M6 (Figure 5.9) for determining the hyperparameter β of the F_β Score. Table 5.19 lists the implications of different values.

Table 5.19: Decision guide 6.2 for determining the hyperparameter β of the F_β Score as an overlap-based metric (mapping M6 (Figure 5.9)). Used abbreviations: False Positive (FP), False Negative (FN), Positive Predictive Value (PPV). Context: no exclusive interest in the centerline of structures (F2.3 ✗, F3.3 ✗) and unequal severity of class confusions (F2.5.2 ✓). Table adapted from Maier-Hein et al. [2022].

$\beta < 1$ (e.g. 0.5)	$\beta = 1$	$\beta > 1$ (e.g. 2)
Higher penalization of PPV, i.e. FP (oversegmentation)	Equal treatment of PPV and Sensitivity, i.e. FP and FN (over- and undersegmentation)	Higher penalization of Sensitivity, i.e. FN (undersegmentation)

Decision guides for selecting boundary-based metrics (M7)

Decision guide 7.1 This decision guide refers to mapping M7 (Figure 5.10) for choosing between the Boundary IoU and NSD as a boundary-based metric. Table 5.20 lists the advantages and disadvantages of both metrics.

Table 5.20: Decision guide 7.1 for choosing between the Boundary Intersection over Union (Boundary IoU) and Normalized Surface Distance (NSD) as a boundary-based metric (mapping M7 in Figure 5.10). Context: possibility of spatial outliers in the reference annotation (F4.3.2 ✓) or, if no possibility of spatial outliers, in the case of existence-based penalization of outliers (F2.5.6). Advantages of a metric are indicated by 👍, while disadvantages are shown by 👎. 👉 refers to neutral comments. Table adapted from Maier-Hein et al. [2022].

Boundary IoU	**NSD**
👉 Hyperparameter can be selected based on inter-rater variability to capture noise and imprecisions	👉 Hyperparameter can be selected based on inter-rater variability to capture severe inconsistencies
👍 Able to compensate for annotation imprecisions	👍 Assesses errors in the boundary as severe inconsistencies
👎 Selection of hyperparameter influences the metric scores	👎 Selection of hyperparameter influences the metric scores

Decision guide 7.2 This decision guide refers to mapping M7 (Figure 5.10) for choosing between the ASSD and MASD as a boundary-based metric. Table 5.21 lists the advantages and disadvantages of both metrics.

Table 5.21: Decision guide 7.2 for choosing between the Average Symmetric Surface Distance (ASSD) and Mean Absolute Surface Distance (MASD) as a boundary-based metric (mapping M7 in Figure 5.10). Context: distance-based outlier penalization with contour focus (F2.5.6). Advantages of a metric are indicated by 👍, while disadvantages are shown by 👎. Table adapted from Maier-Hein et al. [2022].

ASSD	**MASD**
👎 If one object is much larger than the other, its boundary affects the score much more	👍 Both boundaries are equally considered

Decision guide 7.3 This decision guide refers to mapping M7 (Figure 5.10) for choosing between the HD and the X^{th} Percentile Hausdorff Distance (X^{th} Percentile HD) (e.g. the 95% Percentile (HD95)) as a boundary-based metric. Table 5.22 lists the advantages and disadvantages of both metrics.

Table 5.22: Decision guide 7.3 for choosing between the Hausdorff Distance (HD) and the X^{th} Percentile HD as a boundary-based metric (mapping M7 in Figure 5.10). Context: distance-based outlier penalization with outlier focus (F2.5.6). Advantages of a metric are indicated by 👍, while disadvantages are shown by 👎. Table adapted from Maier-Hein et al. [2022].

HD	**X^{th} Percentile HD**
👎 Spatial outliers substantially affect the metric score	👍 Spatial outliers are compensated for

Decision guides for selecting localization criteria (M8)

Decision guide 8.1 This decision guide refers to mapping M8 (Figure 5.11) for choosing between the Boundary IoU, Mask IoU, and IoR as a localization criterion. Table 5.23 lists the advantages and disadvantages of the metrics.

Table 5.23: Decision guide 8.1 for choosing between the Boundary IoU, Mask IoU, and IoR as a localization criterion (mapping M8 in Figure 5.11). Context: Instance Segmentation (InS) (F1.1). Advantages of a metric are indicated by 👍, while disadvantages are shown by 👎. Table adapted from Maier-Hein et al. [2022].

Boundary IoU	Mask IoU	IoR
👍 Less strict penalization of small structures, depending on hyperparameter d	👍 Established and commonly used criterion	👍 Less strict penalization of scenarios with touching structures, in which several predictions may overlap a single reference
👍/👎 Focus on object boundaries	👍/👎 No focus on object boundaries	👍/👎 No focus on object boundaries
👎 Selection of hyperparameter influences the metric scores	👍/👎 No hyperparameter	👍/👎 No hyperparameter
👎 Newly proposed metric, thus not (yet) established	👎 Over-penalization of small structures	👎 Very uncommon criterion
👎 Might over-penalize scenarios with touching structures, in which several predictions may overlap a single reference	👎 Might over-penalize scenarios with touching structures, in which several predictions may overlap a single reference	

Decision guide 8.2 This decision guide refers to mapping M8 (Figure 5.11) for choosing between the Center Distance, the Point inside Mask/Box/Approximation, and the Mask IoU > 0.5 as a localization criterion. Table 5.24 lists the advantages and disadvantages of the metrics.

Table 5.24: Decision guide 8.2 for choosing between the Center Distance, the Point inside Mask/Box/Approximation (Approx), and the Mask Intersection over Union (Mask IoU) > 0.5. as a localization criterion (mapping M8 in Figure 5.11). Context: Object Detection (ObD) problems (F1.1) in the case of either (1) reference annotations provided as exact outline (F4.4) and a desired localization as only position (F2.4) or (2) reference annotations provided as rough outline (F4.4) and a desired localization as only position (F2.4). Advantages of a metric are indicated by 👍, while disadvantages are shown by 👎. Table adapted from Maier-Hein et al. [2022].

Center Distance	**Point inside Mask/Box/Approx**	**Mask IoU > 0.5**
👍 Adjustment of localization strictness possible (by adjusting the threshold value for the distance to the center) 👍/👎 Rough estimate of object location 👎 Bad approximation of tubular or disconnected structures	👍 Better approximation of tubular or disconnected structures 👍/👎 Rough estimate of object location 👎 No adjustment of localization strictness possible	👍 Adjustment of localization strictness possible (by adjusting the threshold value for an overlap of 0.5) 👍 More precise estimate of object location than point-based estimates 👎 Very loose criterion (e.g. Mask IoU > 0) may favor very large predictions irrespective of the location

Decision guide 8.3 This decision guide refers to the choice of a localization threshold, which is needed for most of the suggested localization criteria (only for the Point inside Mask/Box/Approximation, no threshold is needed). For the other localization metrics, the threshold decides whether a predicted object is considered a hit or not. Thus, the choice of the threshold is highly important and impacts the final metric scores. In the general computer vision domain, it is common to compute several thresholds and average the metric scores for all thresholds. The higher the threshold, the more restrictive is the choice of TP objects. Under certain conditions, a lower threshold may be considered:

Table 5.25: Decision guide 8.3 for determining the localization threshold (mapping M8 in Figures 5.11). 👉 refers to neutral comments. Table adapted from Maier-Hein et al. [2022].

Conditions for low threshold	Conditions for high threshold
👉 No interest in localizing objects precisely 👉 3D images since the overlap ratio may need to be refined given the cubical increasing volume 👉 High variability of structure sizes across or within images (F3.2) 👉 High inter-rater variability (F4.3.1) 👉 Consistently small structure sizes (F3.1)	👉 Interest in localizing objects precisely 👉 Possibility of overlapping or touching structures (F3.5)

Decision guides for selecting assignment strategies (M9)

Decision guide 9.1 This decision guide refers to mapping M9 (Figure 5.12) for choosing between the Greedy (by "localization criterion"[5]) Matching, the Optimal (Hungarian) Matching and the Matching via "localization criterion"[5] > 0.5 as an assignment strategy. Table 5.26 lists the advantages and disadvantages of the strategies.

Table 5.26: Decision guide 9.1 for choosing between the Greedy (by "localization criterion"[5] Matching, the Optimal (Hungarian) Matching and the Matching via "localization criterion"[5] > 0.5 as an assignment strategy (mapping M9 in Figure 5.12). Context: lack of predicted class scores (F5.1 ✖). Advantages of a metric are indicated by 👍, while disadvantages are shown by 👎. Table adapted from Maier-Hein et al. [2022].

Greedy (by "localization criterion") Matching	Optimal (Hungarian) Matching	Matching via "localization criterion" > 0.5
👍/👎 Avoids elaborate matching strategies 👎 Only usable if predictions are not overlapping	👍/👎 Elaborated matching strategy 👎 Tends to overly optimistic validation of ambiguous predictions	👍 Alternative to the Greedy (by Score) Matching if predicted class scores are unavailable 👍/👎 Elaborated matching strategy

[5] "Localization criterion" refers to the localization criterion selected in mapping M8 (Figure 5.11).

Addressing metric pitfalls

In Section 4.2, we presented a comprehensive overview of metric pitfalls, which were categorized into a pitfall taxonomy. In Table 5.27, we show how we addressed the pitfalls in the recommendation framework.

Table 5.27: Overview of how the pitfalls described in Section 4.2 were addressed in metrics recommendation framework. Table adapted from Maier-Hein et al. [2022].

Pitfall	***Addressed by***
	Incorrect problem categorization
Wrong choice of problem category	Introduction of problem category mapping M1 (Figure 5.3), which is traversed in the first step of the framework.
	Disregard of the domain interest
Importance of benefit-cost-analysis	Introduction of fingerprint item F2.6 (cutoff on predicted class scores), which includes a benefit-cost-based cutoff analysis. In addition, we recommend the NB and EC metrics for such a situation.
Importance of structure boundaries	Introduction of fingerprint item F2.1 (particular importance of structure boundaries). In addition, we recommend complementing overlap-based metrics (M6, Figure 5.9) with boundary-based metrics (M7, Figure 5.10).
Importance of structure volume	Introduction of fingerprint item F2.2 (particular importance of structure volume). In addition, we recommend complementing the metric selection with application-specific volume-based metrics.
Importance of structure center(line)	Introduction of fingerprint item F2.3 (particular importance of structure center(line)). In addition, we recommend selecting the clDice metric as an overlap-based metric for semantic or instance segmentation problems. In object detection problems, we recommend using the Center Distance as a localization criterion.
Importance of confidence awareness	Introduction of fingerprint item F2.7.1 (calibration assessment requested; dedicated recommendations on calibration) and introduce a calibration metric mapping (M5, Figure 5.8).
Unequal severity of class confusions	Introduction of fingerprint item F2.5 (penalization of errors). In addition, we recommend selecting the EC metric for image-level classification problems and adjusting the hyperparameter of the F_β Score for penalizing either oversegmentation or undersegmentation (see Table 5.19).
Importance of comparability across data sets	Introduction of fingerprint item F4.2 (provided class prevalences reflect the population of interest). In addition, we make sure to recommend only prevalence-independent metrics in cases of prevalences not reflecting the population of interest.
	Disregard of the properties of the target structure
Small structure sizes	Introduction of fingerprint item F3.1 (small size of structures relative to pixel size). In addition, we recommend to rephrase the problem as object detection (see M1, Figure 5.3). In object detection problems, we recommend using a lower localization threshold (see Table 5.25).
High variability of structure sizes	Introduction of fingerprint item F3.2 (high variability of structure sizes). In addition, we recommend using a lower localization threshold in object detection problems (see Table 5.25). We further recommend applying size stratification (see next paragraph).
Complex structure shapes	Introduction of fingerprint item F3.3 (target structures feature tubular shape). In addition, we recommend selecting the clDice in segmentation problems with objects of tubular shapes. In object detection problems, we recommend using approximations instead of bounding boxes, and using a Point inside Approximation localization criterion.
Occurrence of overlapping or touching structures	Introduction of fingerprint item F3.5 (possibility of overlapping or touching target structures). In addition, we recommend phrasing the problem as an instance rather than semantic segmentation problem (see M1, Figure 5.3). In object detection problems, we recommend using a higher localization threshold in object detection problems (see Table 5.25).

Pitfall	*Addressed by*
Occurrence of disconnected structures	Introduction of fingerprint item F3.6 (possibility of disconnected target structure(s)). In addition, we recommend using approximations instead of bounding boxes. Furthermore, as localization criterion, we recommend the Point inside Mask/Approximation criterion.
Disregard of the properties of the data set	
High class imbalance	Introduction of fingerprint items F4.1 (presence of class imbalance) and F2.5.5 (compensation for class imbalances requested). In addition, we recommend compensation of high class imbalance by selecting metrics that are independent from the prevalence (e.g. BA).
Small test set size	We recommend reporting confidence intervals for all metrics (see Metric Application).
Noisy reference standard	Introduction of fingerprint items F4.3.1 (high inter-rater variability) and F2.5.7 (compensation for annotation imprecisions requested). In addition, we recommend selecting the NSD as a boundary-based metric and selecting its hyperparameter according to the inter-rater variability.
Spatial outliers in reference	Introduction of fingerprint items F4.3.2 (possibility of spatial outliers in reference annotation) and F2.5.6 (handling of spatial outliers). In addition, we recommend selecting metrics that are not sensitive to outliers, such as NSD or Hausdorff Distance 95 Percentile (HD95).
Occurrence of cases with an empty reference	Introduction of fingerprint item F4.6 (possibility of reference without target structure(s)). In addition, we introduce further aggregation-related recommendations in the next paragraph.
Disregard of the properties of the algorithm output	
Possibility of empty prediction	Introduction of fingerprint item F5.2 (possibility of algorithm output not containing the target structure(s)). In addition, we introduce further aggregation-related recommendations in the next paragraph.
Possibility of overlapping predictions	We recommend phrasing the problem as an instance rather than semantic segmentation problem. In addition, we recommend an assignment strategy based on IoU > 0.5 if overlapping predictions are not possible and no predicted class scores are available (F5.1).
Lack of predicted class scores	Introduction of fingerprint item F5.1 (availability of predicted class scores). In addition, we recommend using predicted class scores and selecting an appropriate cutoff strategy (F2.6) in addition to calibration assessment (F2.7.1).

Metric application

Finally, after the metric selection based on the mappings M2 to M9, the metric candidates need to be applied to the data set. First, in the case of multiple classes, a *global decision threshold* should be determined that can be used for all classes. To ensure wide applicability and prevent an overestimation of particular classes, this step is crucial. Table 5.28 contains our recommendations related to the *metric implementation, aggregation, ranking, and reporting*.

Table 5.28: Overview of metric application recommendations. Table adapted from Maier-Hein et al. [2022].

Pitfall	***Recommendation***
	Metric implementation
Non-standardized metric definition	Publicly available metric libraries should be carefully inspected before using. Ideally, reference implementations should be used.
Discretization issues	For metrics like the ECE, unbiased estimates should be used whenever possible. An example is the ECE^{KDE} (see Section 2.4.5).
	Metric aggregation
Presence of multiple classes	The generation of metric results should be done *per class*. While multi-class counting metrics directly measure the performance over all classes, a per-class validation is typically beneficial. For example, some classes might be more important than others (F2.5.1) or the algorithm may perform much worse on one class compared to others (see Figure 4.13(b)).
Non-independence of test data (F4.5)	Biomedical data is often hierarchically. structured. This structure should be respected. For example, as shown in Figure 4.14(a), the metric scores should first be aggregated per patient to avoid biases towards one patient. Afterwards, the aggregates should be aggregated into a single score.
Hierarchical label structure (F4.5)	Similarly, correlation between individual classes should be respected and results should be aggregated hierarchically.
High variability of structure sizes (F3.2)	The results should be stratified by structure size.
Possibility of invalid algorithm output (F5.3)	In the case of challenges, an invalid submission may occur, for example, if the results for some data points were not submitted. If such a setting is possible, one should define a strategy on how to handle these missing values. For instance, they could be set to the worst possible metric value or handled via a case-based ranking scheme (see Section 2.1.2). The missing value strategy may highly impact the aggregated results, as demonstrated in Figures 4.14(b) and (c).
Availability of meta-information	If meta-information is available, the information should be used for performance validation. This could reveal important information. For example, the prediction in Figure 4.15(b) performs differently for men and women.
	Rankings (if any)
Metric relationships	In the case of highly related metrics (e.g. DSC and IoU), only one of them should be selected for a ranking, since the other would not add any additional information. However, reporting related metrics may still be chosen to provide a value comparable across communities.
Ranking uncertainties	As we show in Section 5.3, reporting rankings should go beyond simple ranking tables. In that Section, we propose several advanced analysis and visualization techniques for addressing ranking uncertainty.

Pitfall	*Recommendation*
	Metric reporting and interpretation of values
Raw metric values	The raw metric values should be visualized in addition to reporting the aggregated scores. For example, violin plots or color-coded dots- and boxplots could be used for visualization (see Figure 4.21(b)). Moreover, we recommend reporting confidence intervals for all metric values in addition to descriptive statistics such as mean, median, Interquartile Range (IQR) and similar.
Multiple test runs	Deep Learning (DL) algorithms are subject to non-determinism, i.e. the results slightly differ for different runs. To account for this problem, multiple runs can be performed and the respective variance could be reported or ensembling techniques could be leveraged [Pham et al., 2020; Summers and Dinneen, 2021].
Number of decimal places	The number of reported decimal places of the (aggregated) metric scores should reflect the inter-rater variability (uncertainty) and the relevance of the reference.
Reporting guidelines	The general reporting should ideally follow respective reporting guidelines. In the case of challenges, we present the specified BIAS guideline [Maier-Hein et al., 2020] in Section 5.4. Depending on the application, the respective guideline can be found via the Enhancing the QUAlity and Transparency Of health Research (EQUATOR) network [Altman et al., 2008].

Recommendations for popular biomedical challenges

We applied the recommendation framework to prominent biomedical image analysis challenges to analyze their metric choices. One of the most well-known MICCAI challenges is the repeated **Multimodal Brain Tumor Image Segmentation (BRATS) challenge** [Menze et al., 2014], aiming for the segmentation of glioma from the human brain. The semantic segmentation is performed for the tumor regions whole tumor, tumor core, and active tumor. The challenge uses the DSC and HD as primary metrics and the Sensitivity and Specificity as additional metrics. The metrics were individually calculated for every tumor region. Based on the problem-specific fingerprint and metric mappings M6 (Figure 5.9) and M7 (Figure 5.10), we agree with the per-region validation and with the DSC metric. However, we would not recommend using Sensitivity and Specificity for segmentation tasks. Sensitivity is already assessed with the DSC score. Specificity relies on the TN. For pixel-level validation, the TN would refer to the background of the image. As the background may be very large, we do not recommend using metrics like Specificity, as the actual interest may be overruled by a high amount of background pixels. Given the inter-rater variability described in their work [Menze et al., 2014], we would recommend using NSD rather than HD as a boundary-based metric, since NSD is able to compensate for annotation imprecisions. Furthermore, HD is very sensitive to spatial outliers and the actual value may be misleading if only a few outliers were obtained in the reference annotations. However, it should be noted that NSD had not yet been introduced in literature at the time of the first BRATS challenge. Nevertheless, it was repeated every year and the metrics could have been adjusted to better choices.

The **International Skin Imaging Collaboration (ISIC) challenge** [Codella et al., 2018] is a popular competition aiming for the segmentation of skin lesions and the classification of skin diseases. The challenge used the DSC, IoU, and pixel-wise Accuracy metrics for the validation of the *skin lesion segmentation task*. Given that DSC and IoU are mathematically closely related, we would highly recommend not using them together (see Table 5.18 and Figure 4.16(a)). In addition, the Accuracy is dependent on the TN. Similar to the Specificity used in the BRATS challenge, we would not recommend using Accuracy for the ISIC segmentation challenge. Moreover, the challenge did

not assess the quality of the structure boundaries. Given that the DSC and IoU come with several limitations and given the high variability of structure sizes in the skin lesion data set, we would recommend using a boundary metric. From the challenge publication, it is unclear whether to expect a high inter-rater variability. Based on mapping M7 (Figure 5.10), we would thus either recommend using the NSD to compensate for a potentially noisy reference or the MASD if focus on contours is of greater interest. For the *disease classification task*, the challenge proposed the AUROC and the Specificity at several Sensitivity target values as validation metrics. Based on the problem description, both metrics are in line with our recommendation framework. We would additionally suggest BA as a multi-class counting metric (mapping M2, Figure 5.4) and a Proper Scoring Rule such as BS to assess the interpretability of the predicted class scores (mapping M5, Figure 5.8).

The **Cancer Metastases in Lymph Nodes Challenge (CAMELYON)** [Bejnordi et al., 2017] was composed of two tasks, the detection of lymph node metastasis from Whole Slide Imaging (WSI) and the classification of images into cancer versus no cancer. They used the FROC Score for the *lymph node detection task*, which is in line with the recommendation framework (mapping M4, Figure 5.7). However, this was the only metric for this task and neither the localization criterion (mapping M8, Figure 5.11) nor the assignment strategy (mapping M9, Figure 5.12) was defined. Based on their description of the data, we would recommend the Box/Approx IoU localization criterion and the Greedy (by Score) matching strategy. In addition, we would suggest using the F_β Score to assess the overall detection quality at an argmax-based cutoff value (mapping M3, Figure 5.6). Similarly, the *binary metastasis classification task* only used the AUROC as a metric. We would again complement it with a per-class counting metric for a default cutoff value. Given the rather imbalanced data set, we would suggest the F_β Score.

Finally, the **Gland Segmentation (GlaS) challenge** [Sirinukunwattana et al., 2017] was organized with the purpose of instance segmentation of glands in colon histology images. The challenge suggested the pixel-wise F_1 Score to assess the detection quality of algorithms. This measure was further used as a localization criterion with at least 50% of overlap necessary for being counted as TP. In addition, they proposed the per-instance DSC, Adjusted Rand Index, and HD for assessing the segmentation quality. Based on the problem description, we agree with the pixel-wise F_1 Score being used as the localization criterion (mapping M8, Figure 5.11). However, we would not recommend using the pixel-wise F_1 Score version also to assess the detection quality, since this may cause problem category-metric mismatch problems (see Figure 4.3). Instead, we would recommend calculating the object-level F_1 Score instead (mapping M3, Figure 5.6). In addition, there are no details present on whether predicted class scores are available, but we assume that they were available based on the description of the individual algorithms. We therefore would suggest using the Greedy (by Score) assignment strategy (mapping M9, Figure 5.12) and adding a multi-threshold metric (mapping M4, Figure 5.7) to the metrics. Based on the clinical importance of the objective, we would recommend using the FROC Score. We further agree with the DSC as an overlap-based metric (mapping M6, Figure 5.9). The Adjusted Rand Index is not part of the framework but may be added as a complement to the DSC score. No information with respect to inter-rater variability or spatial outliers was provided, we could therefore not estimate whether the HD is an appropriate boundary metric. If spatial outliers were present, we would rather recommend the HD95 (mapping M7, Figure 5.10).

The framework can not only be applied to challenges, but to any classification-related biomedical

image analysis scenario. From the resulting recommendations for common biomedical use cases, we could observe that scenarios concerned with assessing similar properties yield the same metric recommendations. For example, the semantic segmentation of large objects, such as use cases "lung cancer cell segmentation from microscopy images" [Castilla et al., 2018] and "liver segmentation in Computed Tomography (CT) images" [Antonelli/Reinke et al., 2022; Simpson et al., 2019] yield a similar problem fingerprint for very different problems from different domains and modalities. In Appendix A.4, we present the metric recommendations for four common biomedical use cases per problem category.

5.2.4 Discussion

In previous sections, we showed that rankings in challenges and benchmarking experiments heavily rely on the choice of validation metrics. Every metric assesses different properties of a prediction. For example, an algorithm that perfectly predicts the boundary of a structure receives a perfect boundary-based metric score and is thus ranked highly. However, this prediction may have only predicted the outline, meaning that it could have missed a huge hole inside of the object. In this case, the overlap-based metric scores would thus be very low, yielding a low rank. To sum up, the choice of the metric is extremely relevant, not only in challenge rankings but also in decision-making in clinical practice.

One of the most important steps in a challenge is to establish a meaningful, precise, and feasible objective. The problem category and metric selection should be solely based on this biomedical objective and research problem, not based on popularity. Nonetheless, 24% of recent challenges select metrics for their eminence, not for their suitability.

With our metric recommendation framework, we offer resources to researchers such that they can choose metrics while being cognizant of potential problems. The framework was created based on the agreement of multiple specialists with various levels of experience. While we invited multiple technical experts from the image analysis domain, we made sure to include a substantial amount of clinical and biological experts, as well as experts from other domains such as statistics. Based on their specific expertise, they made sure that the recommendations are suitable for a range of different applications in the medical, clinical, and biological domains. The framework and metric candidates were tested several times and iteratively refined.

Although we included popular metrics like the DSC in the metric candidate list, we also made sure that those metrics would only be selected under the correct circumstances and would be complemented with other metrics that can compensate for their limitations. The focus of the framework was the suitability and correctness of the metric selection, which also may yield recommendations beyond common practice. Metrics are often historically motivated and remain as an unquestioned default although they may not be suitable for recent developments in models or applications. To show that common practice does not always follow principles of best practices, we revisited four popular biomedical image analysis challenges and reviewed their metric choices. While most of them select some suitable candidates, none performed a comprehensive selection of appropriate metrics. This is especially critical given the fact that researchers often rely on challenges and use the same validation methods for a comparison of their own algorithms to those of challenge participants.

A limitation of the framework is certainly that we only cover four problem categories. However, as we have shown in Chapter 3 of this thesis, they represent the largest part of biomedical challenges (87% in the period from 2004 to 2016 and 65% in recent years). We are therefore already covering a large range of problems. Nonetheless, we are planning an extension to further problems and applications in the future.

The complexity of the framework is another limitation of the recommendations. To make the recommendations more easily accessible and to avoid biased decisions, we further plan to implement a web-based metric recommendation toolkit. The presented problem fingerprint and metric mappings will serve as a basis for the implementation. Only relevant fingerprints will be answered by the user, thus, the complexity of the framework will be substantially reduced. The toolkit will be written in Python and will be constructed as an interactive form, which guides the user through the different mappings. We plan to hire a designer to optimize the user experience.

Finally, we present metric profiles of all metric candidates in Appendix A.3 of this thesis. They summarize relevant information about each metric, including a description, relevant limitations, and recommendations. Furthermore, it should be noted that metric implementation is not standardized. Even for common metrics, the concrete implementations often differ across frameworks and packages. This is why we are currently working on a dedicated package including the implementation of all metrics that we are recommending in our metric framework. Specifically, we also address common shortcomings in standard implementation, for example, by adding the prevalence as a parameter for the implementation of PPV and NPV to ensure prevalence correction. We plan to integrate this repository as a side package in the MONAI framework.

5.2.5 Conclusion

Problem-aware metric selection is a crucial step in the validation pipeline. An inappropriate metric choice is carried through the entire assessment of an algorithm, impacting the final rankings, challenge winners, and possibly the decision on which algorithm may be translated into clinical practice.

Our metric recommendation framework can be used for several biological, medical, and clinical use cases, even beyond challenges. Metric choice is also of utmost importance for the individual validation of a new algorithm, and can even be used for general computer vision problems. A problem-aware metric selection should be the foundation for every challenge and algorithm validation procedure, upon which all subsequent steps, such as the ranking analysis, the reporting, and other further analyses stand. With our metric recommendation framework, we hope to guide the research community toward a more problem-aware validation.

5.3 Improving common practice of rankings

Disclosure

This work has been supervised by Lena Maier-Hein. It has been conducted together with several co-authors, especially Manuel Wiesenfarth, Matthias Eisenmann, and Annette Kopp-Schneider. Parts of the work have been published in *Nature scientific reports* [Wiesenfarth et al., 2021], *Nature communications* [Antonelli/Reinke et al., 2022] and *Medical Image Analysis* [Roß/Reinke et al., 2020]. Please refer to Chapter A.1 for full disclosure.

5.3.1 Introduction

Challenge rankings are an important step to determine the best-performing algorithms for a dedicated research problem. However, in Chapter 4 of this thesis, we showed that the concrete ranking schemes are often not transparently reported. Theoretically, it would be simple to manipulate challenge rankings by exploiting this bad practice.

This is due to the fact that challenge rankings are extremely sensitive to their design and the underlying data. In Section 4.3, we demonstrated that challenge rankings may change substantially when altering the ranking method, aggregation, annotator, or data set. Thus, challenge winners are frequently unstable, and it may be unfair to choose one algorithm if a small adjustment to the ranking scheme would have caused another participant to win.

In this section, we first analyze how rankings are presented in challenge reports. We hypothesize that many challenges base their analysis of results solely on ranking tables without further visualization. We then explore ranking uncertainty analysis methods and present advanced visualization techniques that may help interpret the performances of the challenge participants.

Specific findings in Part 4

Rankings are highly sensitive to various design parameters.

Research questions investigated in this chapter

Is the interpretability of challenge results enhanced by advanced result visualization techniques?

Can advanced visualization techniques easily be applied to several challenges?

We apply those visualization techniques to three different challenges to explore whether they increase the interpretability over ranking tables and whether they are easy to apply. The first challenge simulates extreme cases to illustrate the advantages of the visualization methods. In addition, we present the results for the Robust Medical Instrument Segmentation (RobustMIS)

and the Medical Segmentation Decathlon (MSD) challenges. While both challenges focus on segmentation, they have different goals and assessment methods. Given their differences, we show that the ranking uncertainty analyses and visualization techniques can be easily applied to challenges, independent of their manifestation.

5.3.2 Methods

Analysis of common practice

For the analysis of common visualization practice, we utilized the same challenge database that was introduced in Chapter 3. Our analysis was based on all challenges whose results were published in journals or conference proceedings, yielding 82 challenges in the period 2004 to 2016 [Maier-Hein et al., 2018]. We analyzed all 82 challenge reports and their result visualization techniques, which were grouped into one or multiple of the following categories:

- No visualization
- Visualization of metric values (e.g. as scatterplots or curve visualization such as plotting the Receiver Operating Characteristic (ROC) curve)
- Boxplots
- Visualization of qualitative results of algorithms (e.g. segmentation on top of an image)
- Visualization of ranking variability
- Other visualization

Visualization techniques

In the following, we explain the employed visualization and analyzing techniques. An overview of all plots with examples is provided in Figure 5.13.

Visualization of raw assessment data: While challenge results are often directly provided in aggregated ranking tables (see Section 5.3.3), the visualization of the raw assessment data helps to reveal the underlying distribution of performance results per algorithm. An easy way to visualize the metrics per challenge participant are ***dots- and boxplots*** (see Figure 5.13, assessment data, left). The boxplots include information on the first and third quartile as lower and upper borders of the box, the median as a horizontal line within the boxes, and the 1.5 Interquartile Range (IQR) range illustrated by vertical lines. The individual metric scores are shown jittered on top to visualize the distribution of values. Dots- and boxplots are easy to interpret and directly visualize descriptive statistics such as the median and quantiles of the data distribution. They further help to identify implausible values or inconsistencies in the assessment data. For settings, in which algorithms are easily distinguishable, the rankings can be directly deviated from dots- and boxplots. This is more complicated for situations, in which performances are more diverse (as for the RobustMIS challenge introduced below). However, there is no connection between the individual metric values corresponding to the same test cases, therefore not revealing cases that were especially easy or hard to process across challenge participants.

An alternative visualization technique are ***podium plots*** (see Figure 5.13, assessment data, middle). Podium plots are also known as *benchmark experiment plots* and were first presented in Eugster et al. [2008]. In contrast to dots- and boxplots, podium plots connect the individual metric

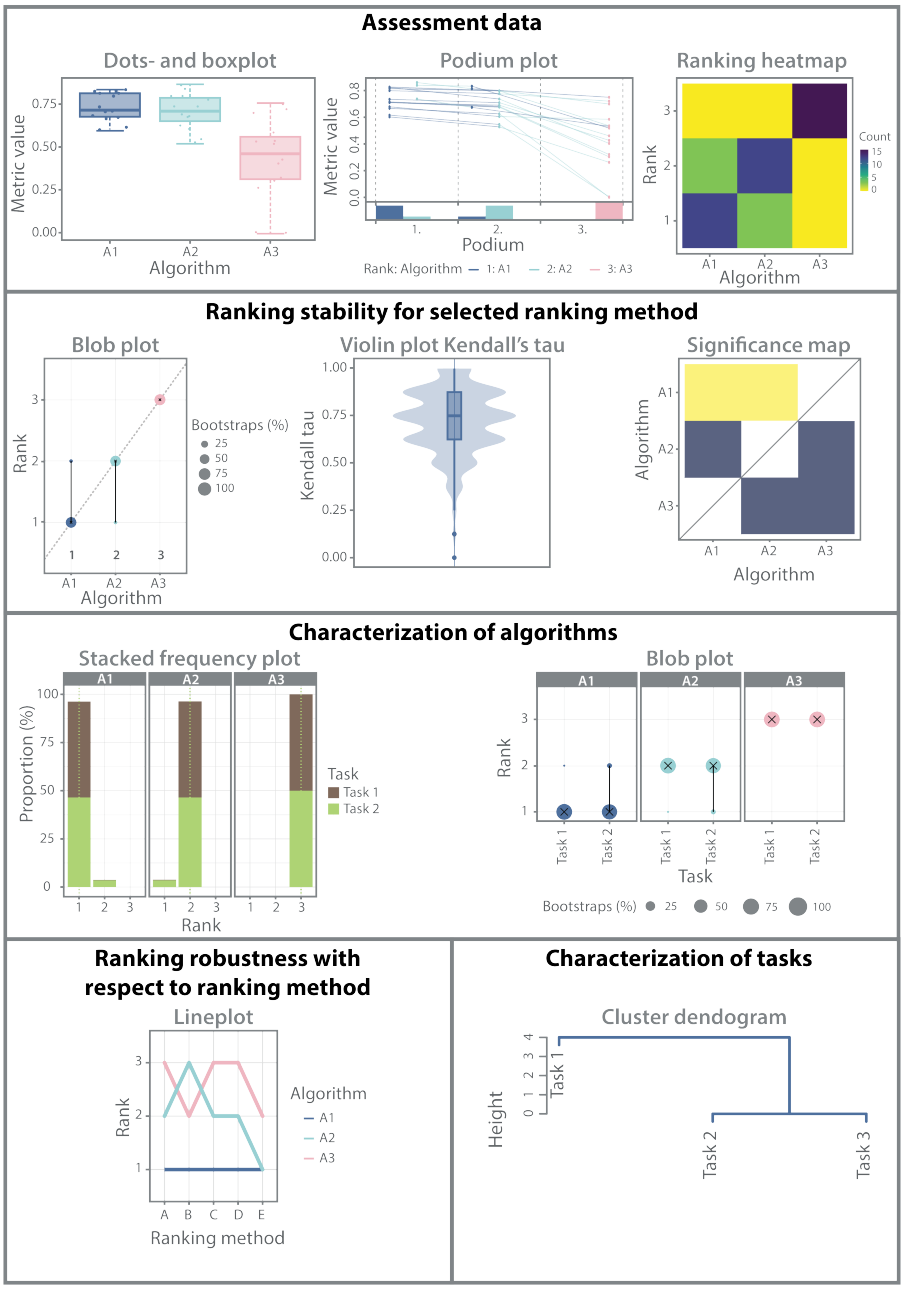


Figure 5.13: Overview of the employed visualization techniques. The *raw assessment data* is visualized with dots- and boxplots, podium plots, or ranking heatmaps. The *ranking stability for a selected ranking method* can be visualized by blob plots, violin plots of Kendall's τ or significance maps. For the *characterization of algorithms*, stacked frequency plots or blob plots per algorithm can be utilized. To visualize the *ranking robustness with respect to the ranking method*, lineplots are employed. For multi-task challenges, the *characterization of tasks* can be achieved via cluster dendrograms.

scores from the same test case for the different participating teams. In that way, one may easily spot influential test cases from the data set. The upper part of a podium plot contains dots representing the metric values color-coded by the algorithms that produced this score. A line connects metric values that represent the same cases. The plot is split into podiums representing possible ranks. The dots, i.e. the metric values, are assigned a rank (podium) based on the rank that the respective participating team would achieve for the particular test case. Per podium, dots are presented in columns representing every algorithm. The lower part of the podium plot shows a frequency plot of how often a challenge participant achieves a specific podium rank over all test cases. Although podium plots help to overcome problems of dots- and boxplots, there are relatively complex and hard to read for many test cases and algorithms. It is important to note that the podium plot assigns random ranks in the event of a tie.

Ranking heatmaps (see Figure 5.13, assessment data, right) do not visualize the metric distribution per algorithm but rather show how often a challenge participant receives a specific rank when ranked per test case, i.e. the cell (i, A_j) corresponds to the absolute frequency of cases where the algorithm A_j achieved the rank i. Heatmaps with high contrast indicate better separation of participating teams. In particular, a high count on the diagonal of the heatmap implies a high degree of concordance in the ranking. Ranking heatmaps are especially useful in the case of many algorithms or test cases, which would lead to low readability of podium plots.

Ranking stability for a selected ranking method: As described in Section 4.3, bootstrapping approaches can be used to investigate ranking stability for a given ranking method. The results of the bootstrapping analysis can be visualized by ***blob plots***, in which blobs indicate the relative frequency of an algorithm achieving a particular rank. The radius of the blobs increases proportionally with the relative frequency of ranks. The median rank is provided by a black cross while 95% of the distribution of bootstrap ranks is shown by a black line (2.5th-97.5th percentile).

To directly visualize the ranking variability across bootstrap samples, Kendall's τ can be computed for a comparison between the original ranking and the bootstrap rankings. The distribution of Kendall's τ scores is then visualized in form of a ***violin plot***. Violin plots combine a boxplot and a density plot to visualize the distribution of Kendall's τ values.

In contrast to bootstrapping, a test-based approach can be employed to investigate ranking stability. A ***significance map*** visualizes whether a specific algorithm is significantly superior to others. We visualize the pairwise significant test results as an incidence matrix. An algorithm on the abscissa that significantly performed better than another algorithm on the ordinate is shown by a yellow cell. In contrast, blue cells indicate that the algorithm is not significantly superior. As a default, we use a Wilcoxon signed rank test with an alpha level of 5% with adjustment for multiple testing according to Holm [Holm, 1979].

Characterization of algorithms: For challenges with multiple tasks, further plots can be used to characterize the performance of algorithms across tasks. Please note that these visualization techniques can only be used if the same algorithms compete for all tasks. The ordering of algorithms in the plots is based on a consensus ranking which refers to an aggregated ranking over all tasks. We compute the consensus ranking as the average of ranks across tasks [Lin, 2010; Hornik and Meyer, 2007]. For every participating team, ***blob plots*** can again be utilized to visualize the frequency of achieved ranks over the bootstrap samples. In this case, we won't show

the participating teams on the x-axis but replace them with the tasks. A blob plot is shown for every participating algorithm. This allows us to directly spot ranking differences between algorithms.

We provide an alternative representation in the form of ***stacked frequency plots*** of the observed ranks across bootstrap samples. The proportion of achieved ranks, given on the x-axis, for all bootstrap samples is provided on the y-axis. The plot is color-coded by the task and one plot is shown for every algorithm. In the case of algorithms that achieved the same ranking in multiple tasks for the whole assessment set, vertical lines appear on top of each other. By plotting vertical lines, we can compare each algorithm's performance over the different tasks.

Characterization of tasks: Similarly, it may be interesting to characterize the tasks of a challenge and check for similarities. ***Blob plots*** can be generated for every task to compare the ranking stability across tasks. Another way of analyzing results is to investigate similarities between tasks. For this approach, we propose a ***cluster dendrogram*** (see also [Eugster et al., 2008]), clustering tasks in which the rankings of participating teams are similar. The cluster dendrogram is based on hierarchical clustering [Hastie et al., 2009]. Clusters are generated based on the distance measure Spearman's footrule (see Section 2.2) and complete agglomeration [Landau et al., 2011].

Robustness of the ranking based on the ranking method: As shown in Section 4.3, challenge results heavily depend on the chosen ranking method. An easy way to visualize the ranks of challenge participants across different ranking methods are **lineplots**. The rank of a challenge participant is indicated by a point on the line so that changes across ranking methods can be easily spotted.

Open-source ranking toolkit

We implemented all of the described visualization and analysis techniques in an R framework called *challengeR* [Wiesenfarth et al., 2019]. In the toolkit, the user can generate a full PDF report including all of the presented figures and analyses. The *challengeR* package offers the user to easily calculate the rankings and perform bootstrapping. Based on the results, the report is automatically generated. The user only needs to load the package and the challenge assessment data as a table containing the following columns:

- A *test case identifier* specifies the test image for which the results are calculated
- An *algorithm identifier* specifies the algorithm for which the results are calculated
- A *metric value* is provided for every test case and algorithm
- For multi-task challenges, a *task identifier* specifies the given tasks

Currently, only one metric at a time can be processed in *challengeR*. Challenges with multiple metrics can be interpreted as multi-task challenges, with the metric name being the task identifier.

In the readme of *challengeR*, we provide several examples on how to use the toolkit in addition to a troubleshooting section. The toolkit has so far been employed by a number of users for algorithm benchmarking and challenge validation. *challengeR* offers a customizable challenge analysis. Users may, for example, change color schemes or the ranking correlation coefficient.

Challenges used for application of visualization techniques

We applied the visualization techniques presented above to one simulated challenge and two challenges organized by ourselves: the RobustMIS challenge and the MSD challenge, which will be presented in the following paragraphs.

Simulated challenge
We applied the visualization techniques presented above to one simulated challenge and two challenges organized by ourselves: the RobustMIS challenge and the MSD challenge (see below). The simulated challenge was designed such that it best shows the benefits of the presented visualization and analysis techniques. The challenge utilizes a metric bounded between 0 and 1 with 1 referring to a perfect agreement with the reference annotation and 0 meaning no agreement with the reference annotation (such as the Dice Similarity Coefficient (DSC) metric). Synthetically, we generated challenge results for five algorithms *Algorithm 1* to *Algorithm 5*. The challenge comprised three tasks with 120 simulated test cases each:

- For the *worst-case simulation*, we obtained and assigned all permutations of {0.80, 0.85, 0.90, 0.95, 1.00} to *Algorithm 1* to *Algorithm 5*, such that every algorithm received the same amount of the exact same metric values. In this task, the algorithm performances were indistinguishable by design.
- For the *best-case simulation*, we randomly draw the metric scores from a uniform distribution in the range of [0.9, 1.0) for *Algorithm 1*, [0.8, 0.9) for *Algorithm 2*, ..., and [0.6, 7) for *Algorithm 5*. In this task, the algorithm performances were perfectly distinguishable and introduced a natural ranking by design.
- For the *random simulation*, we randomly assigned metric scores to *Algorithm 1* to *Algorithm 5* ranging between 0 and 1. This task introduced differences between the algorithms due to chance alone.

A metric-based aggregation with the mean was chosen as the primary ranking scheme. We used the presented visualization techniques and the *challengeR* toolkit to investigate ranking uncertainty.

The Robust Medical Instrument Segmentation Challenge
The RobustMIS challenge was part of the Medical Image Computing and Computer Assisted Interventions (MICCAI) conference 2019 [Roß/Reinke et al., 2020]. The comprehensive challenge design document is provided in Appendix A.6 and an overview is provided in Figure 5.14. Our goal was the benchmarking of methods designed for the segmentation and detection of medical instruments in videos of minimally invasive surgery, specifically, for laparoscopic videos. In this challenge, we assessed two primary goals: (1) finding *robust* algorithms and (2) finding algorithms that *generalize* across different degrees of difficulty. The challenge was organized into three tasks, binary instrument segmentation, instrument detection, and instrument instance segmentation. For the binary segmentation task, the requested output was a binary mask of medical instruments with 1 indicating pixels of the instrument class and 0 indicating pixels of the background. Similarly, for the instrument instance segmentation task, the requested output were image masks with numbers 1, 2, etc. representing pixels of instrument instances and 0 representing the background. For the instrument detection task, only a rough localization was required in the form of bounding boxes.

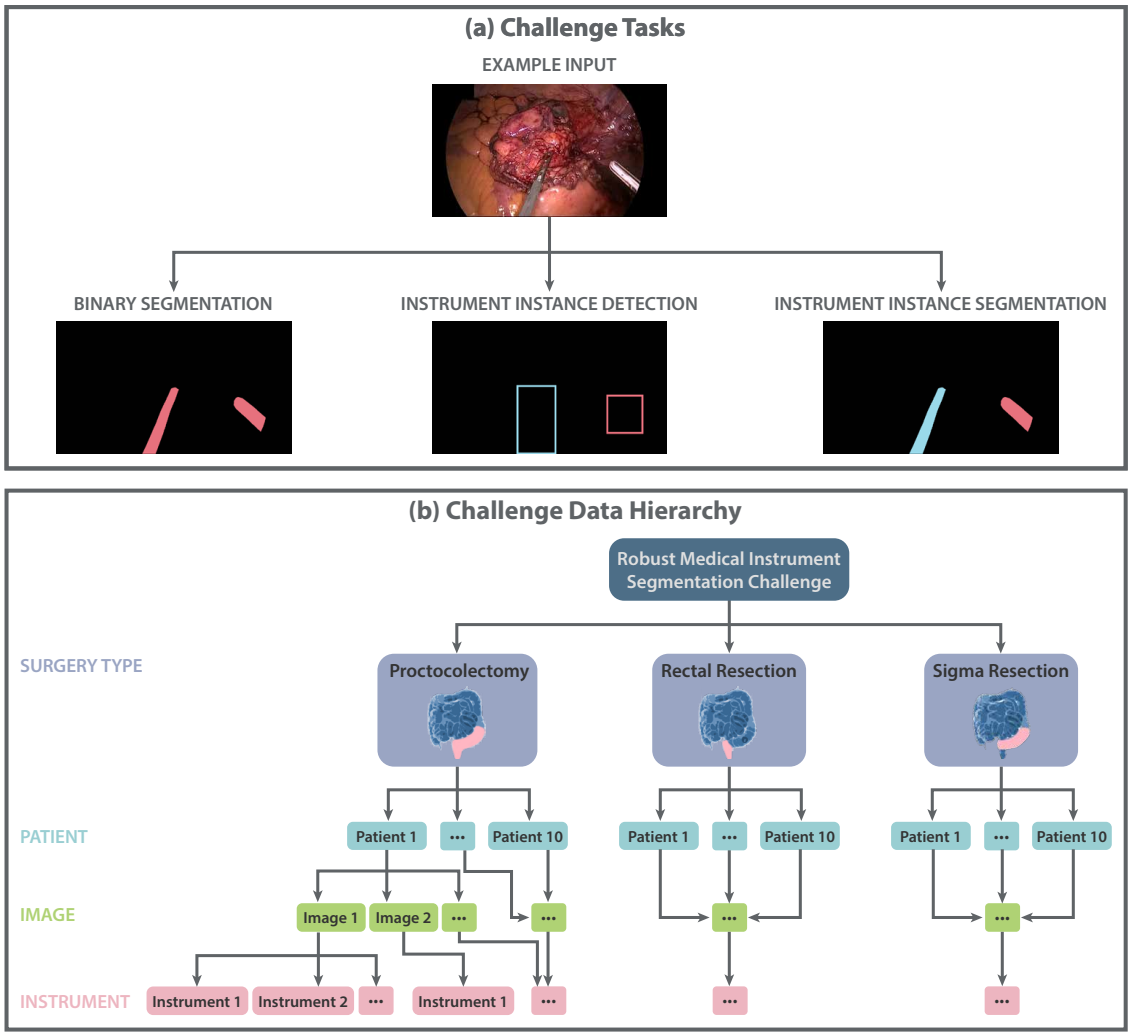


Figure 5.14: Overview of the Robust Medical Instrument Segmentation (RobustMIS) challenge. **(a) Overview of challenge tasks.** The challenge was organized in three tasks: Binary segmentation, instrument instance detection, and instrument instance segmentation. **(b) Hierarchical scheme of the challenge data.** The challenge data was taken from three surgery types, from which images were taken from ten different patients each. Images contained a different number of instruments.

For the challenge data, we used the segmentation data of the Heidelberg Colorectal (HeiCo) data set [Maier-Hein et al., 2021]. The data comprised image frames from 30 video procedures from proctocolectomy, rectal resection, and sigmoid resection (ten videos/patients for each type of surgery). A total of 10,040 frames were extracted. At the time of challenge organization, this was the largest annotated data set in the field. For the training data, we only used data from proctocolectomy (2,943 frames) and rectal resection (3,040 frames), with a total of 5,983 training cases. To assess the generalization capabilities of the participating algorithms, we split our test data set (4,057 cases) into three stages with increasing difficulty:

Stage 1: We chose the test data for Stage 1 from patients from proctocolectomy (325 cases) or rectal resection (338 cases) surgeries that were visible in the training data. Stage 1 comprised 663 image frames.

Stage 2: We chose the test data for Stage 2 from the same types of surgery (proctocolectomy with 225 cases or rectal resection with 289 cases) that were visible in the training data, but from patients that did not show up in the training data set. Stage 2 comprised 415 image frames.

Stage 3: We chose the test data for Stage 3 from a different (but similar) type of surgery that was not part of the training data set, namely sigma resection. Challenge participants did not know which kind of surgery was used for this stage. Stage 3 comprised 2,880 image frames.

We defined a training case as a ten-second video snippet in the form of 250 consecutive image frames as well as a reference annotation for the last frame of each video. Participants could choose whether they wanted to incorporate the temporal information of the video snippet in their methods. The test cases were similarly defined but without a reference annotation. The test data was kept hidden from the challenge participants. We required challenge participants to submit their methods as docker containers [Docker, 2022] to avoid cheating.

The data set was initially annotated by the company UnderstandAI. In a follow-up step, we identified inconsistencies in the annotation and agreed on a labeling instruction, which can be found in the complete challenge design document in Appendix A.6. The annotations were then refined by 14 engineers and four medical students following the labeling instructions. In the case of disagreement or ambiguities, a team of one medical student and two engineers agreed on a consensus decision. In the final step, a medical expert reviewed all annotations and potential errors were discussed and refined by an engineer and a medical expert.

The DSC and Normalized Surface Distance (NSD) metrics (see Section 2.4) were used for the performance assessment of the binary segmentation task. A threshold of $\tau = 13$ was chosen for the NSD based on inter-rater variability. For the instrument detection and instance segmentation tasks, we used the Intersection over Union (IoU) as a localization criterion with a threshold of 0.3 and the Optimal Hungarian Matching (see Section 2.4) to assign predicted instances to reference instances. We calculated the DSC and NSD metrics for every instance for the instrument instance segmentation task and aggregated the results with the mean to calculate the Multi-Instance Dice Similarity Coefficient (MI DSC) and Multi-Instance Normalized Surface Distance (MI NSD), which we used as performance metrics. The object-level F_1 Score served as the per-class counting metric for the instrument detection task. Since we did not request the predicted class scores of the submitted algorithms, we did not calculate a multi-threshold metric for performance assessment.

To assess the robustness and accuracy of algorithms, we defined two ranking schemes for the binary and instance segmentation tasks. We used a test-based ranking scheme (see Section 2.1.2) based on a Wilcoxon signed rank test with a significance level of $\alpha = 0.05$ as our *accuracy ranking*. As we were particularly interested in the robustness of algorithms, we focused on the worst-case performance in our *robustness ranking*, for which we applied a metric-based ranking (see Section 2.1.2) with the 5% percentile serving as aggregation operator. The rankings were computed separately for the DSC/MI DSC and the NSD/MI NSD. The object-level F_1 Score was calculated over the whole data set and yielded a single score per participant, which naturally defined the ranking for the instrument detection task. Potentially missing values were set to 0 for all metrics. We used the presented visualization techniques and the *challengeR* toolkit to investigate ranking uncertainty.

The Medical Segmentation Decathlon
We organized the MSD challenge as part of the MICCAI conference 2018 [Antonelli/Reinke et al., 2022]. The complete challenge design document can be found in Appendix A.7. The major goal of the challenge was to find a semantic segmentation algorithm that worked for a wide range of various tasks and data without requiring the model to be changed or any human interaction to accomplish the specific task. For this purpose, the challenge included ten separate data sets, each representing a distinct anatomical structure, with a total of 17 Region of Interests (ROIs). Data was collected by several institutions worldwide, each following its own data acquisition and labeling standards. Seven data sets were used for the training phase, referred to as the *development phase*, with the goal of developing the general-purpose algorithms:

- **Brain:** 750 4D Multiparametric Magnetic Resonance Imaging (mp-MRI) volumes (484 training cases, 266 test cases), three ROIs (edema, enhancing and non-enhancing tumor)
- **Heart:** 30 3D Magnetic Resonance Imaging (MRI) volumes (20 training cases, 10 test cases), one ROI (left atrium)
- **Hippocampus:** 394 3D MRI volumes (263 training cases, 131 test cases), two ROIs (anterior and posterior of hippocampus)
- **Liver:** 210 3D Computed Tomography (CT) volumes (131 training cases, 70 test cases), two ROIs (liver and liver tumor)
- **Lung:** 96 3D CT volumes (64 training cases, 32 test cases), one ROI (lung tumor)
- **Pancreas:** 420 3D CT volumes (282 training cases, 139 test cases), two ROIs (pancreas and pancreatic tumor mass)
- **Prostate:** 48 4D mp-MRI volumes (48 training cases, 32 test cases), two ROIs (prostate Peripheral Zone (PZ) and Transition Zone (TZ))

The algorithms were evaluated on the hidden test sets and participants could follow their scores on a daily basis to avoid model overfitting. The second phase of the challenge, referred to as the *mystery phase*, served as the actual validation and ranking of algorithms. Three additional data sets were used for this phase, only being made accessible to teams who successfully participated in the first phase:

- **Colon:** 190 3D CT volumes (126 training cases, 64 test cases), one ROI (colon cancer primaries)
- **Hepatic vessels:** 443 3D CT volumes (303 training cases, 140 test cases), two ROIs (hepatic vessels and hepatic tumor)

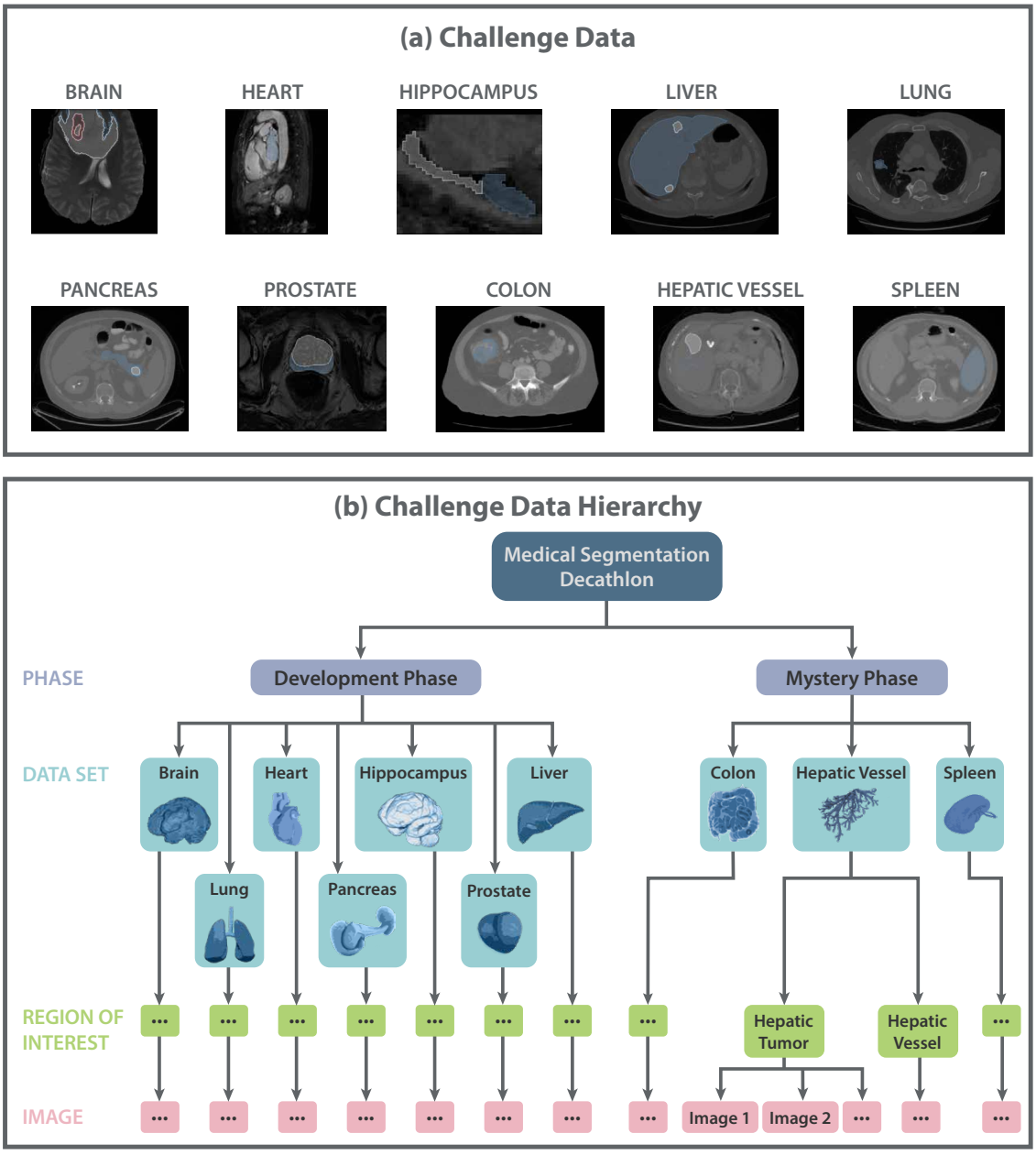


Figure 5.15: Overview of the Medical Segmentation Decathlon (MSD) challenge. **(a) Overview of challenge data.** The challenge was composed of ten data sets. Example images are from Simpson et al. [2019]. **(b) Hierarchical scheme of the challenge data.** Each of the ten data sets was composed of a varying number of regions of interest and for every region, a different number of images was used for the challenge.

- **Spleen:** 61 3D CT volumes (41 training cases, 20 test cases), one ROI (spleen)

Algorithms could be trained on those three additional data sets but could only be submitted once for the final validation.

For performance assessment, we decided to use two validation metrics for all data sets, although they might not be well suited for each of the data sets. However, choosing specific metrics for every data set and ROI would have hindered a statistical comparison across tasks. We therefore picked the DSC and NSD metric, similar to the metrics used for the binary segmentation task of the RobustMIS challenge. For the NSD, the threshold was chosen per data set based on feedback from the clinicians working on the annotations. We chose the following thresholds: 5mm for the brain data, 4mm for the heart, 1mm for the hippocampus, 7mm for the liver, 2mm for the lung, 4mm for the prostate, 4mm for the colon, 3mm for the hepatic vessel and 3mm for the spleen data set.

We chose a test-based ranking scheme with a Wilcoxon signed rank test ($\alpha = 0.05$) as our basic ranking method which was calculated per target ROI. Two final rankings were computed, one for the development and one for the mystery phase. For all data sets and ROIs of each phase, the individual rankings were averaged to achieve the final ranks for every participating algorithm. This procedure was chosen to compensate that every data set had a substantially different sample size. By computing the rankings per ROI and aggregating them in a second step, we made sure that the larger data sets did not overrule the smaller data sets (e.g. heart). We used the presented visualization techniques and the *challengeR* toolkit to investigate ranking uncertainty. Results are mainly shown for the per-ROI or per-data set rankings.

5.3.3 Results

Common practice related to challenge visualization

We analyzed 82 challenges organized in the period 2004 to 2016 that published their results in journals [Maier-Hein et al., 2018]. 27% of those challenge reports were solely based on ranking tables on final ranks without any further visualization. 44% of challenges further employed a visualization of raw metric values in the form of bar- or scatterplots or plotting multi-threshold curves such as the ROC curve and an additional 9% provide a visualization of metric scores plotted against each other. Boxplots without plotting the individual metric values on top were used by 39% of challenges. 40% provided a visualization of qualitative results of participants and 11% used other visualization techniques such as Bland-Altman plots [Cleveland, 1993]. Only a single challenge visualized ranking variability by comparing different ranking methods [Arganda-Carreras et al., 2015].

Since the *challengeR* toolkit for ranking uncertainty analysis was published in 2019, it was used as an analysis tool since then for multiple challenges and benchmarking experiments (e.g. [Oreiller et al., 2022; Roth et al., 2022; Huaulmé et al., 2021]), or algorithm validation in general (e.g. [Baur et al., 2021; Daza et al., 2020; Ayala et al., 2022]), implying that the community was willing to expand their challenge analysis techniques towards ranking uncertainty analysis. The open-source repository was marked with 23 stars on GitHub. Since its first release, we resolved two external and twelve internal issues in four patch releases thanks to community feedback[6].

[6] As of September 2022.

Simulated Challenge

We simulated a challenge with different conditions to emphasize the advantages of the *challengeR* toolkit. The challenge comprised three tasks, a best-case, a worst-case, and a fully random situation. The best-case and fully random simulation yielded the same ranking tables, as shown in Table 5.29, although they were constructed in very different ways. The best-case simulation task was designed as a perfectly distinguishable ranking, while the ranking of the fully random simulation was given by chance. The worst-case scenario yielded the same scores and ranks for all algorithms.

The information of how close algorithms were apart was hidden in the ranking tables but can be seen when inspecting the dots- and boxplots in Figure 5.16, the podium plots in Figure 5.17, or the ranking heatmaps in Figure 5.18. The podium plots and heatmaps reveal a clear rank distribution for every case for the best-case simulation. For the worst-case simulation, it is visible that all algorithms achieve the same ranks for all cases, while the situation is extremely fuzzy for the fully random scenario.

Table 5.29: Rankings for the three tasks of the simulated challenge. Rankings are provided for a generic metric bounded between 0 and 1 for a metric-based aggregation with the mean.

Algorithm	Rank
Algorithm 1	1
Algorithm 2	2
Algorithm 3	3
Algorithm 4	4
Algorithm 5	5

(a) Best-case simulation

Algorithm	Rank
Algorithm 1	1
Algorithm 2	2
Algorithm 3	3
Algorithm 4	4
Algorithm 5	5

(b) Fully random simulation

Algorithm	Rank
Algorithm 1	1
Algorithm 2	1
Algorithm 3	1
Algorithm 4	1
Algorithm 5	1

(c) Worst-case simulation

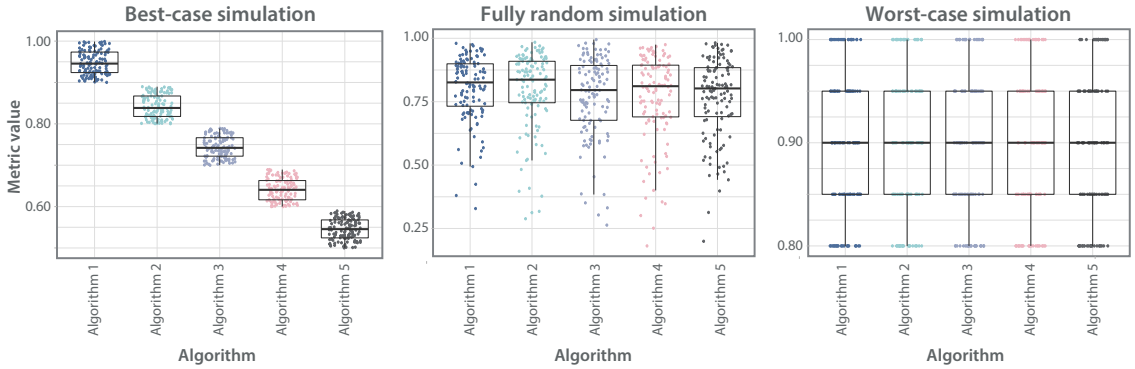


Figure 5.16: Dots- and boxplots illustrating the individual algorithm performances for the simulated challenge for the best-case (left), fully random (middle), and worst-case (right) simulations. Rankings are provided for a generic metric bounded between 0 and 1 for a metric-based aggregation with the mean. Figure adapted from [Wiesenfarth et al., 2021].

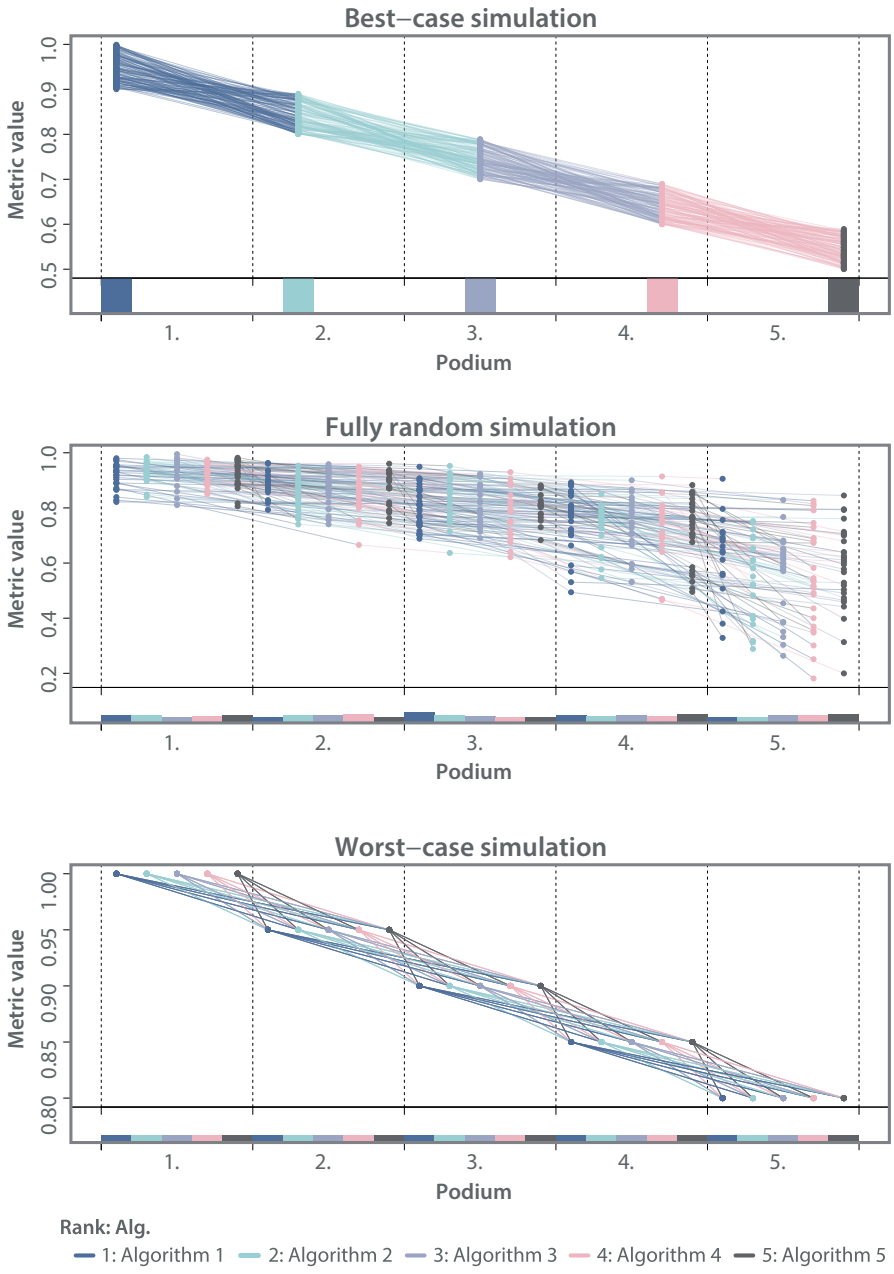


Figure 5.17: Podium plots illustrating the individual algorithm performances for the simulated challenge for the best-case (left), fully random (middle), and worst-case (right) simulations. Rankings are provided for a generic metric bounded between 0 and 1 for a metric-based aggregation with the mean. Top: Dots represent the metric values color-coded by the algorithms that produced this score. Metric values corresponding to identical cases are connected by a line. Bottom: Frequency plot of how often an algorithm achieved a specific rank (podium) over all cases. Per podium, dots are presented in columns representing every algorithm. Figure adapted from [Wiesenfarth et al., 2021].

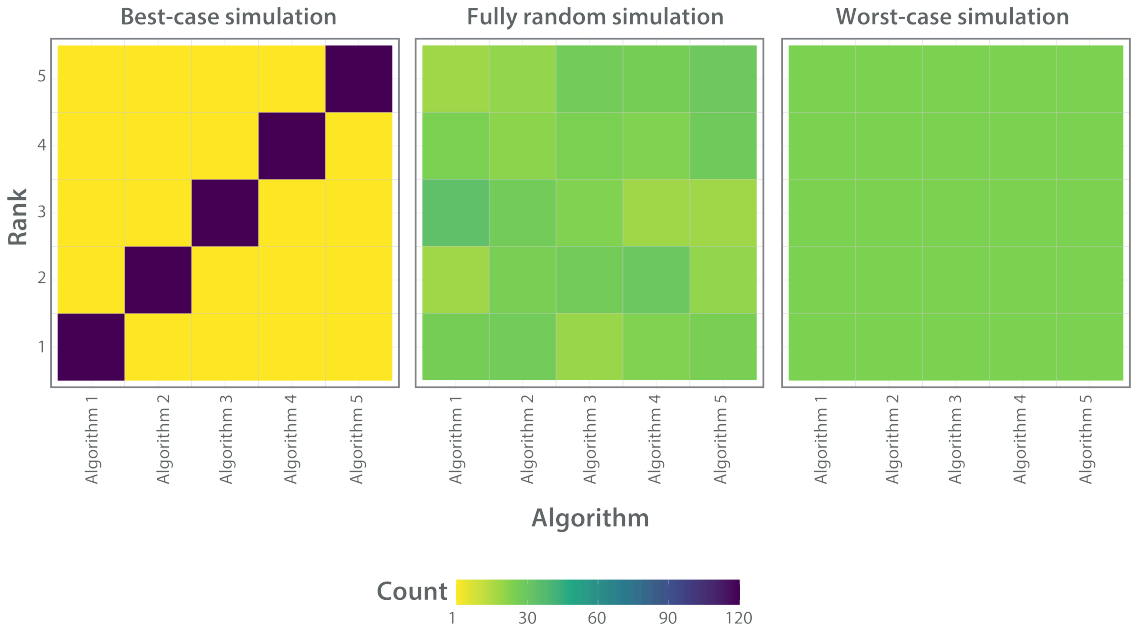


Figure 5.18: Ranking heatmaps for the simulated challenge for the best-case (left), fully random (middle), and worst-case (right) simulations. Rankings are provided for a generic metric bounded between 0 and 1 for a metric-based aggregation with the mean. Each cell (i, A_j) corresponds to the absolute frequency of cases where the algorithm A_j achieved the rank i. Figure adapted from [Wiesenfarth et al., 2021].

This trend is also visible in the blob plots (Figure 5.19), which show the results of the bootstrapping analysis for investigating the ranking stability. The rankings were extremely stable against small perturbations of the data set for the best-case simulation. For the fully random simulation, the ranking was very unstable, meaning that every algorithm appeared at every rank at some point. While *Algorithm 1* and *Algorithm 2* were more similar than the others and slightly superior, the remaining three algorithms were very close together and exchanged ranks at the same frequency. It was even more critical for the worst-case simulation task. Here, all algorithms achieved all ranks with nearly the exact same frequency across bootstrap samples. In this case, a ranking did not make any sense since all algorithms showed the same performance.

The violin plots of Kendall's τ values over the bootstrap rankings (Figure 5.20) further show that there was no ranking variability in the best-case simulation. The mean, median, and IQR were equal to 1.0, implying identical rankings. On the other hand, we could observe very low values for the fully random simulation, yielding a mean Kendall's τ of 0.57 (median: 0.6, IQR: (0.4, 0.8)). For the worst-case simulation, we could not compute Kendall's τ between the rankings, given that all algorithms yielded the same rank for every bootstrap sample.

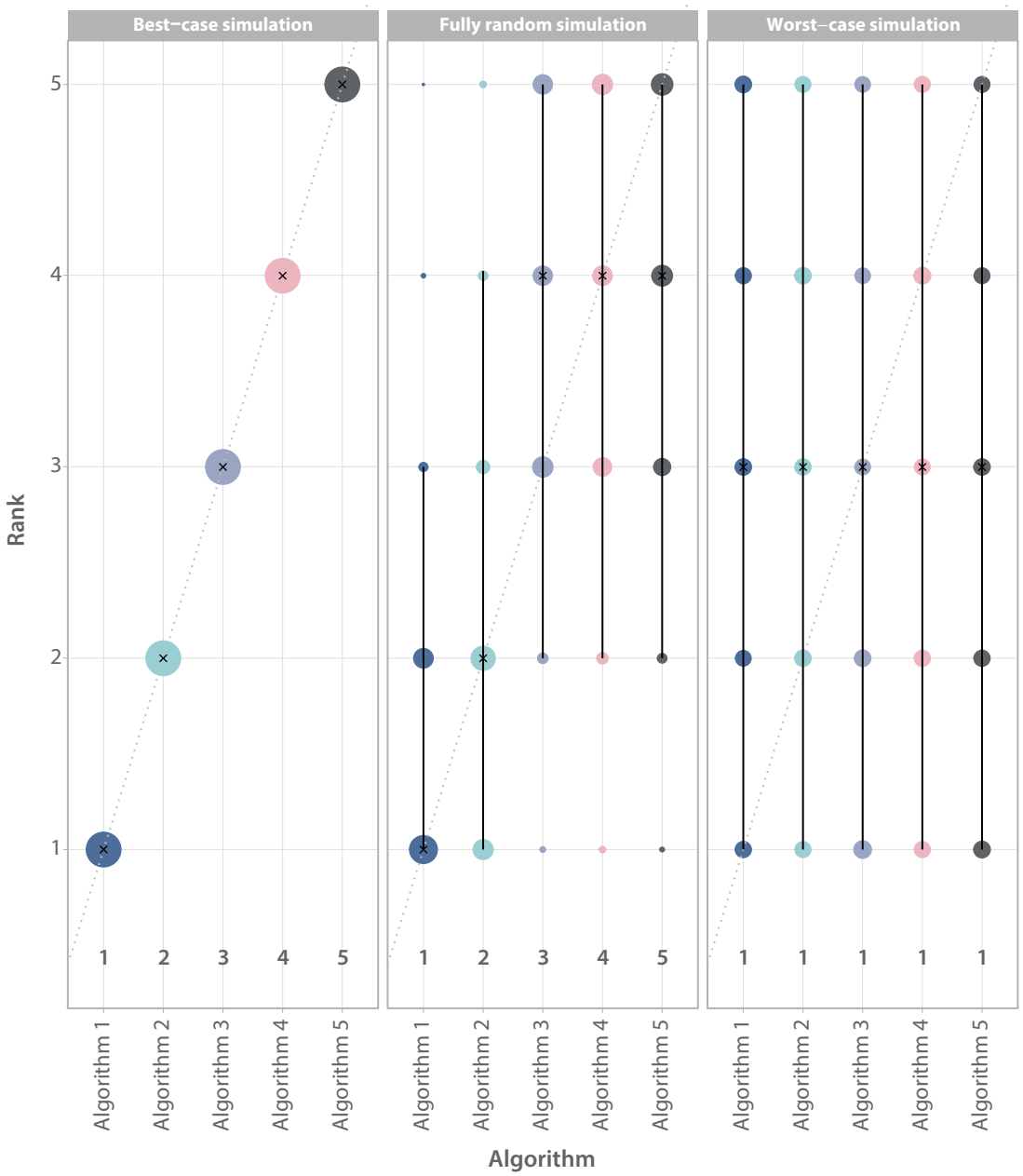


Figure 5.19: Blob plots visualizing ranking uncertainty over 1,000 bootstrap samples for the simulated challenge for the best-case (left), fully random (middle), and worst-case (right) simulations. Rankings are provided for a generic metric bounded between 0 and 1 for a metric-based aggregation with the mean. The radius of the blobs increases with an increasing relative frequency of ranks. The median rank is provided by a black cross while 95% of the distribution of bootstrap ranks is shown by a black line (2.5th-97.5th percentile). Figure adapted from [Wiesenfarth et al., 2021].

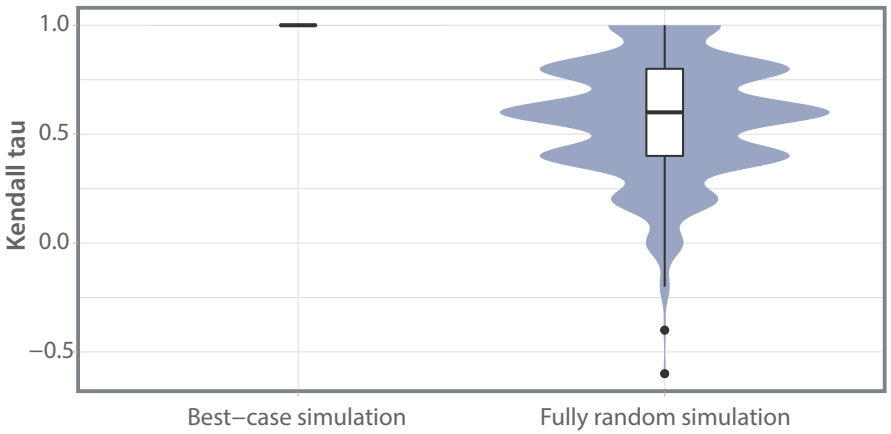


Figure 5.20: Violin plots visualizing Kendall's τ values over 1,000 bootstrap samples for the simulated challenge for the best-case (left), fully random (middle), and worst-case (right) simulations. Rankings are provided for a generic metric bounded between 0 and 1 for a metric-based aggregation with the mean. Figure adapted from [Wiesenfarth et al., 2021].

While the significance maps (Figure 5.21) show that every algorithm was significantly superior to its successors for the best-case simulation, no significance could be found for the other two tasks. When changing the ranking scheme, the rankings for the best-case and worst-case simulations remained stable, while changing completely across ranking schemes for the fully random simulation (Figure 5.22).

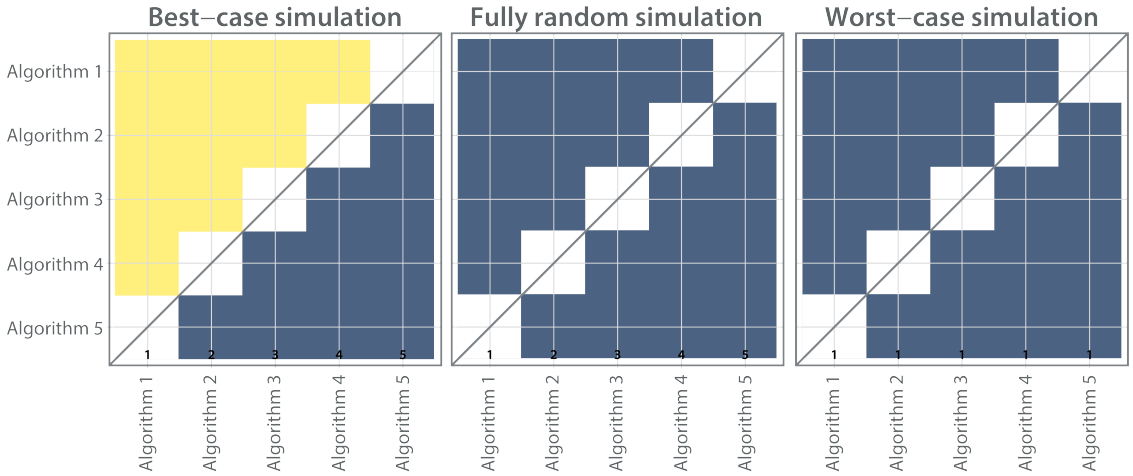


Figure 5.21: Significance maps visualizing pairwise significant test results for a Wilcoxon signed rank test for the simulated challenge for the best-case (left), fully random (middle), and worst-case (right) simulations. Rankings are provided for a generic metric bounded between 0 and 1 for a metric-based aggregation with the mean. Yellow cells indicate that the metric values of the algorithm on the x-axis are significantly superior to those of the algorithm on the y-axis, while blue cells show algorithms that are not significantly superior. Figure adapted from [Wiesenfarth et al., 2021].

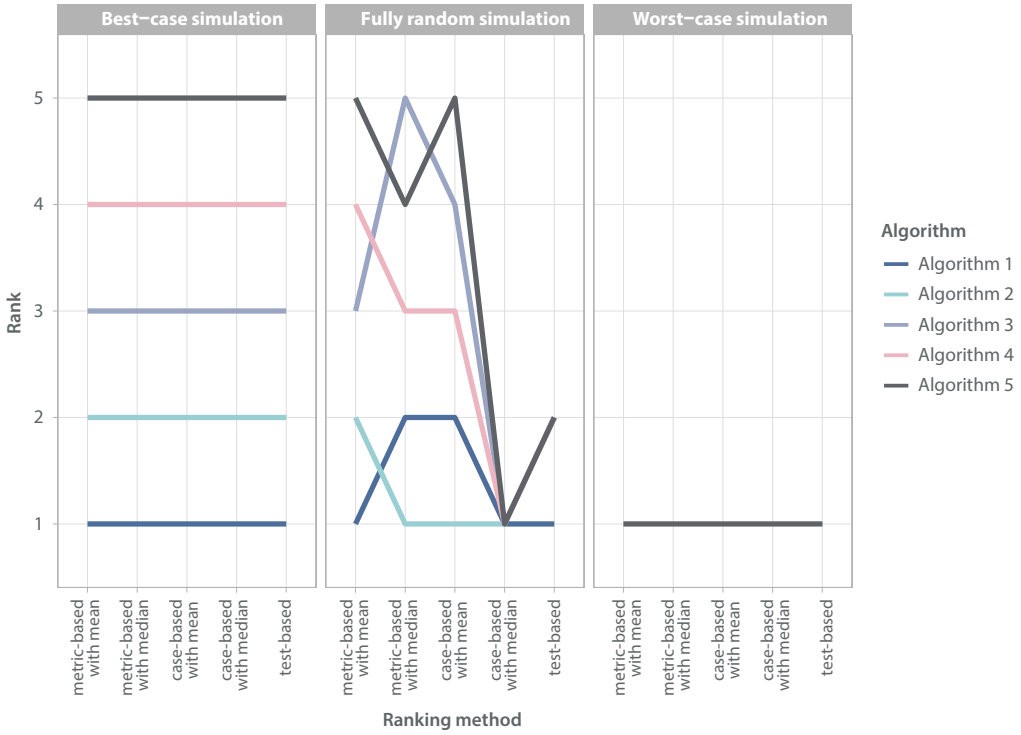


Figure 5.22: Lineplots illustrating the ranking robustness across four different ranking methods for the simulated challenge for the best-case (left), fully random (middle), and worst-case (right) simulations. Figure adapted from [Wiesenfarth et al., 2021].

The Robust Medical Instrument Segmentation Challenge

A total of twelve teams submitted their methods to the RobustMIS challenge, from which ten participated in the binary segmentation task and seven in the instrument detection and instance segmentation tasks. The submitted algorithms were mainly based on Mask R-CNN and U-Net architectures (see Section 2.3). A detailed description of the participating algorithms of the detection and instance segmentation tasks is provided in Section 4.4.

The rankings for Stage 3 for all tasks and metrics are provided in Tables 5.30-5.32. The ranking Table 5.30 shows that the teams *nnU-Net* and *haoyun* seem to have generated the best methods for the binary segmentation problem. For the detection task, team *Uniandes* succeeds over the other teams but only with an increase of 0.01 in the F_1 Score, indicating that the submitted algorithms are quite close in their detection performances (see Table 5.31). For the instrument instance segmentation task, no clear winner for all four rankings can be seen. For example, while team *nnU-Net* was on rank 1 for the MI DSC accuracy ranking, it only achieved rank 5 for the robustness ranking for the same metric. Generally, it can be seen that the chosen ranking scheme and metric had a large influence on the resulting algorithm order.

Table 5.30: Rankings for Stage 3 of the binary segmentation task of the Robust Medical Instrument Segmentation (RobustMIS) challenge. Rankings are provided for the Dice Similarity Coefficient (DSC) (top) and the Normalized Surface Distance (NSD) (bottom) metrics. Accuracy rankings are shown on the left and are based on the proportion of significant tests divided by the number of algorithms (Prop. Sign). The robustness rankings are shown on the right and are based on the 5% quantile (Q5) of the DSC/NSD.

DSC: ACCURACY RANKING			**DSC: ROBUSTNESS RANKING**		
Team identifier	**Prop. Sign.**	**Rank**	**Team identifier**	**Q5 DSC**	**Rank**
fisensee	**1.00**	**1**	***haoyun***	**0.52**	**1**
haoyun	0.89	2	*CASIA_SRL*	0.50	2
CASIA_SRL	0.78	3	*www*	0.49	3
Uniandes	0.67	4	*fisensee*	0.34	4
caresyntax	0.56	5	*Uniandes*	0.28	5
SQUASH	0.44	6	*SQUASH*	0.22	6
www	0.33	7	*caresyntax*	0.00	7
Djh	0.22	8	*Djh*	0.00	7
VIE	0.11	9	*NCT*	0.00	7
NCT	0.00	10	*VIE*	0.00	7

NSD: ACCURACY RANKING			**NSD: ROBUSTNESS RANKING**		
Team identifier	**Prop. Sign.**	**Rank**	**Team identifier**	**Q5 NSD**	**Rank**
haoyun	**0.89**	**1**	***haoyun***	**0.63**	**1**
fisensee	**0.89**	**1**	*CASIA_SRL*	0.62	2
CASIA_SRL	0.67	3	*www*	0.57	3
Uniandes	0.67	3	*fisensee*	0.45	4
caresyntax	0.56	5	*Uniandes*	0.32	5
www	0.44	6	*SQUASH*	0.26	6
SQUASH	0.33	7	*caresyntax*	0.00	7
VIE	0.22	8	*Djh*	0.00	7
NCT	0.11	9	*NCT*	0.00	7
Djh	0.00	10	*VIE*	0.00	7

Table 5.31: Ranking for Stage 3 of the instrument detection task of the Robust Medical Instrument Segmentation (RobustMIS) challenge for the F_1 Score.

Team identifier	F_1 Score	Rank
Uniandes	**0.91**	**1**
www	0.90	2
caresyntax	0.89	3
SQUASH	0.86	4
fisensee	0.86	5
VIE	0.82	6

Table 5.32: Rankings for Stage 3 of the instrument instance segmentation task of the Robust Medical Instrument Segmentation (RobustMIS) challenge. Rankings are provided for the Multi-Instance Dice Similarity Coefficient (MI DSC) (top) and the Multi-Instance Normalized Surface Distance (MI NSD) (bottom) metrics. Accuracy rankings are shown on the left and are based on the proportion of significant tests divided by the number of algorithms (Prop. Sign). The robustness rankings are shown on the right and are based on the 5% quantile (Q5) of the MI DSC/MI NSD.

MI DSC: ACCURACY RANKING			MI DSC: ROBUSTNESS RANKING		
Team identifier	**Prop. Sign.**	**Rank**	**Team identifier**	**Q5 MI DSC**	**Rank**
fisensee	**1.00**	**1**	***www***	**0.31**	**1**
Uniandes	0.83	2	*Uniandes*	0.26	2
caresyntax	0.67	3	*SQUASH*	0.22	3
SQUASH	0.33	4	*CASIA_SRL*	0.19	4
www	0.33	4	*fisensee*	0.17	5
VIE	0.17	6	*caresyntax*	0.00	6
CASIA_SRL	0.00	7	*VIE*	0.00	6

MI NSD: ACCURACY RANKING			MI NSD: ROBUSTNESS RANKING		
Team identifier	**Prop. Sign.**	**Rank**	**Team identifier**	**Q5 MI NSD**	**Rank**
Uniandes	**1.00**	**1**	***www***	**0.35**	**1**
caresyntax	0.67	2	*Uniandes*	0.29	2
fisensee	0.50	3	*CASIA_SRL*	0.27	3
www	0.50	3	*SQUASH*	0.26	4
SQUASH	0.33	5	*fisensee*	0.16	5
VIE	0.17	6	*caresyntax*	0.00	6
CASIA_SRL	0.00	7	*VIE*	0.00	6

To provide deeper insights in the challenge results, we utilized the visualization techniques implemented in the *challengeR* package. We provide the analysis for the segmentation tasks only, as it is mainly developed for validation per case, not per data set (as done for the instrument detection task). In addition, we provide results for the most difficult test data set (Stage 3) only. This stage was the most complex, but most realistic one. A comparison across stages is provided below in Figures 5.27 and 5.28.

First, we visualized the raw metric value distribution of the segmentation tasks with dots- and boxplots as shown in Figure 5.23. It can be seen that the IQR was much smaller for the binary compared to the instance segmentation task. Overall, the metric values were higher for the binary segmentation with the first algorithms being quite close together, especially for the NSD metric. The dots- and boxplots were accompanied by ranking heatmaps in Figure 5.24, which show how often a challenge participant received a specific rank when ranked per test case. A trend of a ranking is visible for the DSC/MI DSC, being more fuzzy for the instrument instance segmentation task. Similar trends could be observed for the NSD/MI NSD metrics. Given the large data set of 2,880 cases for Stage 3, the podium plots were not readable and are omitted for this example.

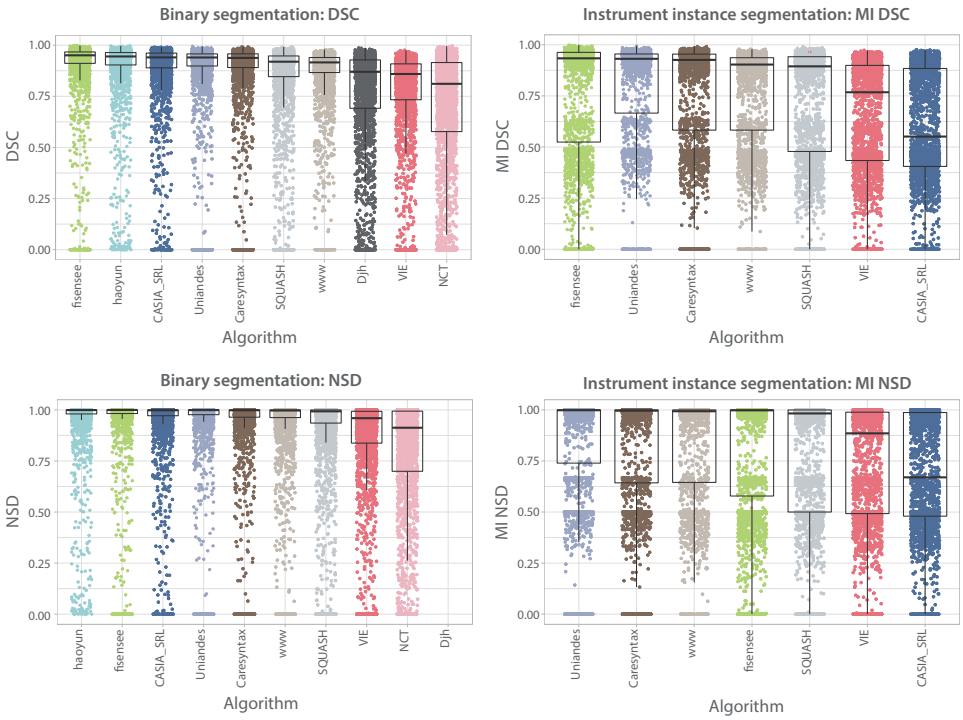


Figure 5.23: Dots- and boxplots illustrating the individual algorithm performances for Stage 3 of the Robust Medical Instrument Segmentation (RobustMIS) challenge for the binary (left) and instrument instance segmentation (right) tasks. Metric values are shown for the Dice Similarity Coefficient (DSC), Multi-Instance Dice Similarity Coefficient (MI DSC), Normalized Surface Distance (NSD), and Multi-Instance Normalized Surface Distance (MI NSD). Figure adapted from [Roß/Reinke et al., 2020].

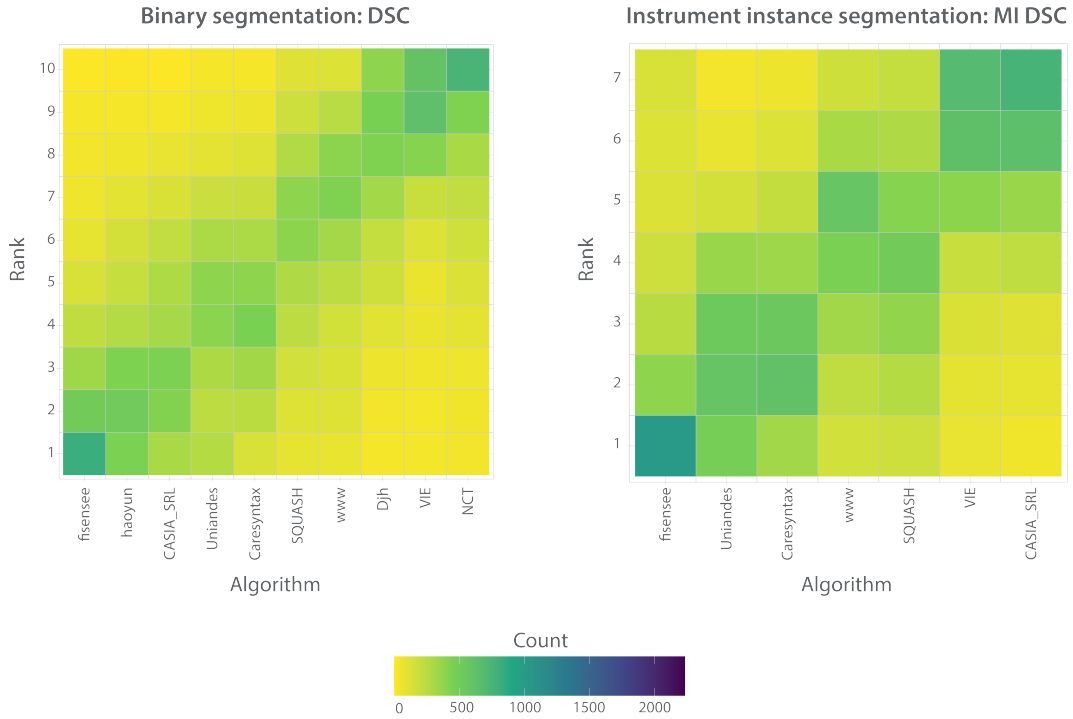


Figure 5.24: Ranking heatmaps for Stage 3 of the Robust Medical Instrument Segmentation (RobustMIS) challenge for the binary (left) and instrument instance segmentation (right) tasks. Metric values are shown for the Dice Similarity Coefficient (DSC) and Multi-Instance Dice Similarity Coefficient (MI DSC). Each cell (i, A_j) corresponds to the absolute frequency of cases where the algorithm A_j achieved the rank i. Figure adapted from [Roß/Reinke et al., 2020].

We performed bootstrap analysis to investigate ranking uncertainty. Descriptive statistics of Kendall's τ values are provided in Table 5.33. In addition, the variability of ranks across tasks is shown in the form of blob plots for the rankings computed with the (MI) DSC in Figure 5.25. It can be easily seen that the accuracy rankings for both the binary and instance segmentation tasks were very stable against small perturbations. Contrarily, the robustness rankings were quite unstable, allowing for many variations in the achieved ranks. The same trends were observed for the rankings based on the (MI) NSD.

Table 5.33: Descriptive statistics of Kendall's τ for comparing the original ranking with 1,000 bootstrapped rankings for Stage 3 of the Robust Medical Instrument Segmentation (RobustMIS) challenge. Mean, Median, 25% Quartile (Q25) and 75% Quartile (Q75) of Kendall's τ are provided for different ranking schemes for the Dice Similarity Coefficient (DSC), Multi-Instance Dice Similarity Coefficient (MI DSC), Normalized Surface Distance (NSD) and Multi-Instance Dice Similarity Coefficient (MI DSC) metrics.

Ranking scheme	**Mean**	**Median**	**Q25**	**Q75**
Binary instrument segmentation, DSC Accuracy	0.99	1.00	1.00	1.00
Binary instrument segmentation, NSD Accuracy	0.95	0.95	0.95	0.96
Binary instrument segmentation, DSC Robustness	0.68	0.69	0.60	0.78
Binary instrument segmentation, NSD Robustness	0.73	0.73	0.64	0.82
Instrument instance segmentation, MI DSC Accuracy	0.99	1.00	1.00	1.00
Instrument instance segmentation, MI NSD Accuracy	0.99	1.00	1.00	1.00
Instrument instance segmentation, MI DSC Robustness	0.77	0.81	0.62	0.90
Instrument instance segmentation, MI NSD Robustness	0.71	0.81	0.52	0.90

Figure 5.26 illustrates the significance maps for the participating teams. They support the results from above that the rankings were quite clear for the accuracy ranking schemes. In most cases, an algorithm was significantly superior to its successors. The robustness rankings were not obvious in terms of statistical significance. For example, the winning algorithm for the MI DSC robustness ranking (team *www*) was only significantly superior to two other methods (at ranks 4 and 7) and no statistical superiority could be measured for the other teams. Similar trends could be observed for the (MI) NSD,

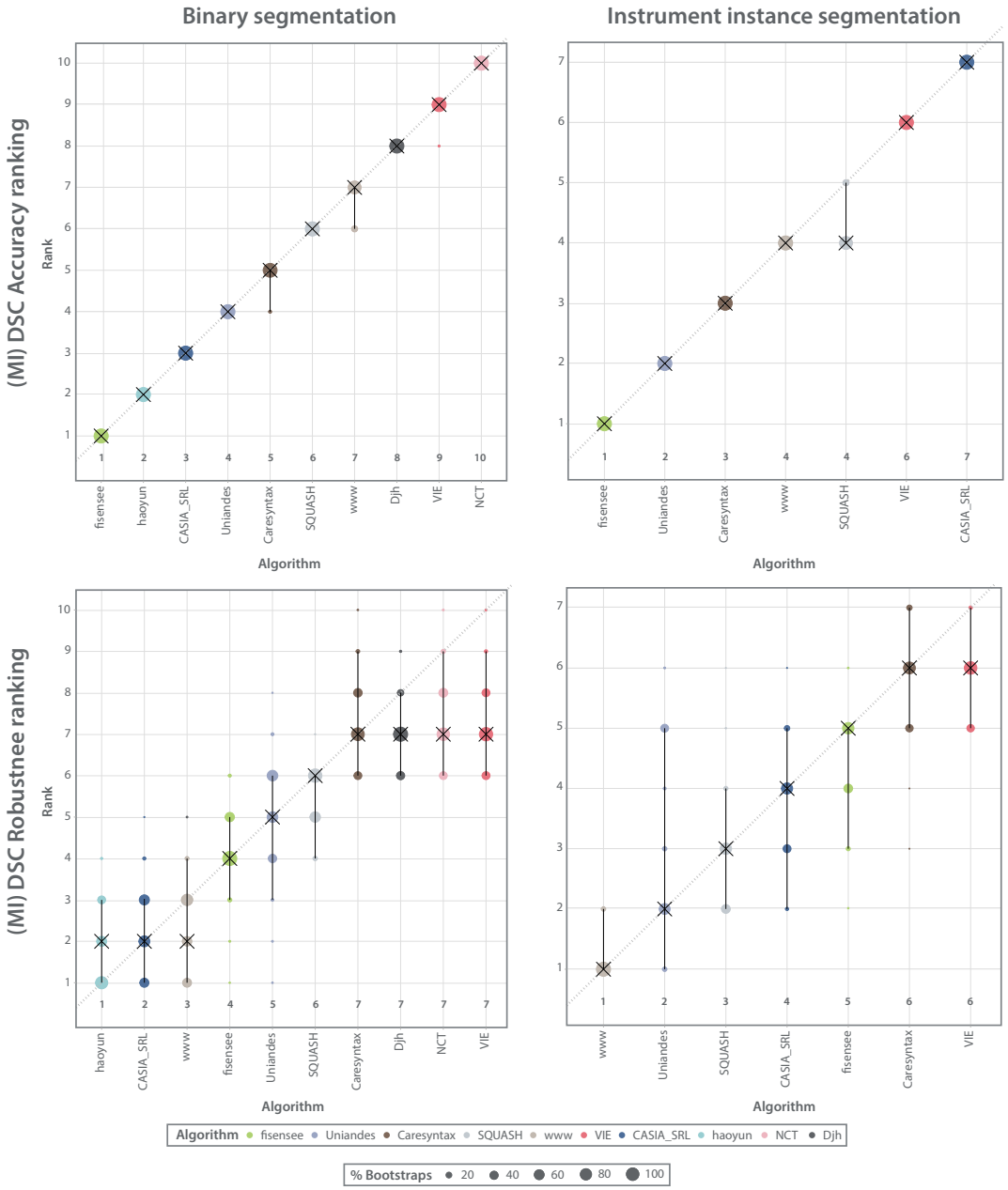


Figure 5.25: Blob plots visualizing ranking uncertainty over 1,000 bootstrap samples for Stage 3 of the Robust Medical Instrument Segmentation (RobustMIS) challenge for the binary (left) and instrument instance segmentation (right) tasks and the accuracy (top) and robustness (bottom) rankings. Metric values are shown for the Dice Similarity Coefficient (DSC) and Multi-Instance Dice Similarity Coefficient (MI DSC). The radius of the blobs increases with an increasing relative frequency of ranks. The median rank is provided by a black cross while 95% of the distribution of bootstrap ranks is shown by a black line (2.5th-97.5th percentile). Figure adapted from [Roß/Reinke et al., 2020].

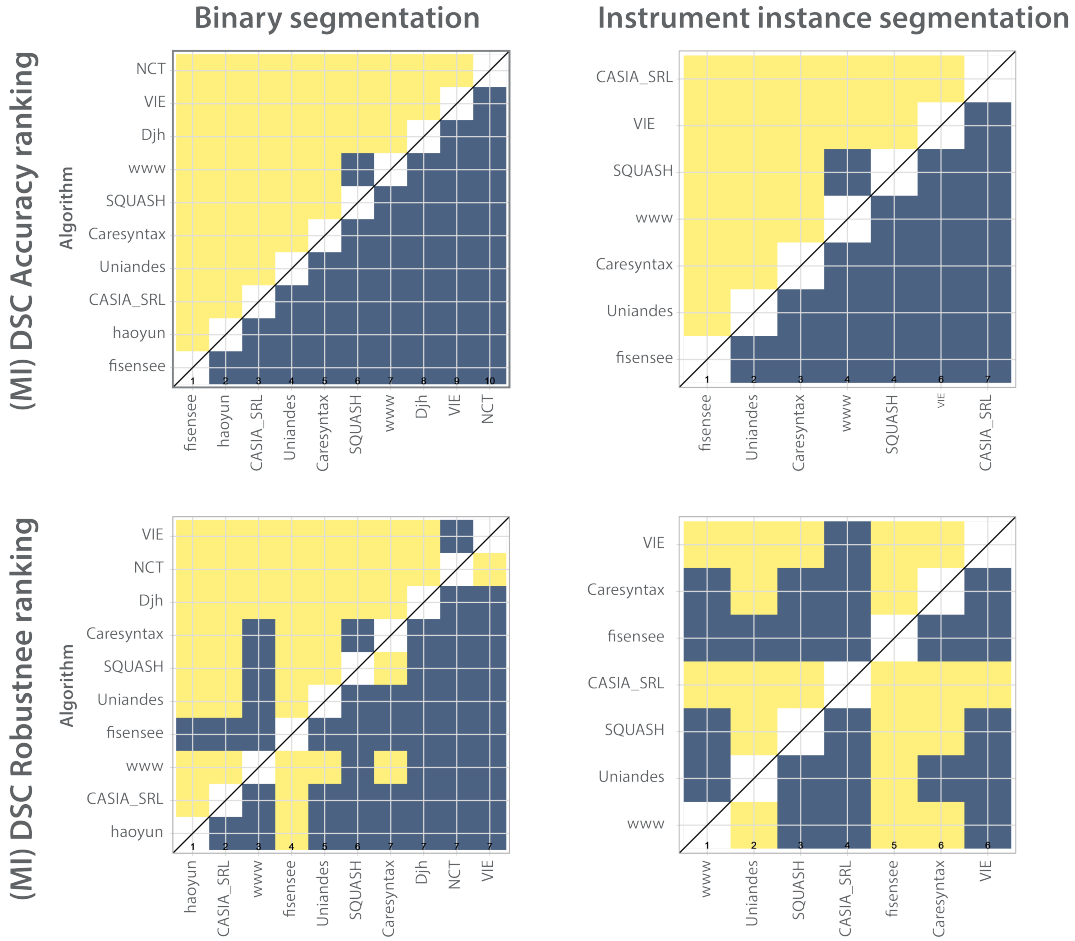


Figure 5.26: Significance maps visualizing pairwise significant test results for a Wilcoxon signed rank test for Stage 3 of the Robust Medical Instrument Segmentation (RobustMIS) challenge for the binary (left) and instrument instance segmentation (right) tasks. Metric values are shown for the Dice Similarity Coefficient (DSC) and Multi-Instance Dice Similarity Coefficient (MI DSC). Yellow cells indicate that the metric values of the algorithm on the x-axis are significantly superior to those of the algorithm on the y-axis.

Figure 5.27 illustrates the rising difficulty of the different stages through the decreasing (MI) DSC values, indicating that generalizing from the training domain was a difficult task for the participating algorithms. The results were comparable for the (MI) NSD scores. The achieved rankings of all teams under bootstrap variability across stages are shown in Figure 5.28 exemplarily for the MI DSC instrument instance segmentation rankings. The ranks for the accuracy rankings were relatively stable across tasks (also for the other ranking schemes), only resulting in lower performance scores for all teams (see Figure 5.27). However, the robustness rankings varied substantially across stages for all ranking schemes.

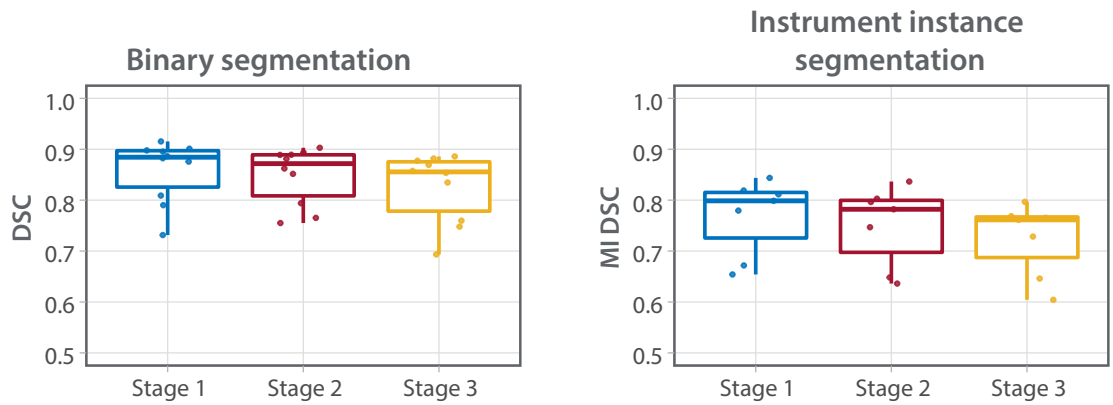


Figure 5.27: Dots- and boxplots for all stages of the mean performance of every participating team of the Robust Medical Instrument Segmentation (RobustMIS) challenge for the binary (left) and instrument instance segmentation (right) tasks. Metric values are shown for the Dice Similarity Coefficient (DSC) and Multi-Instance Dice Similarity Coefficient (MI DSC). Figure adapted from [Roß/Reinke et al., 2020].

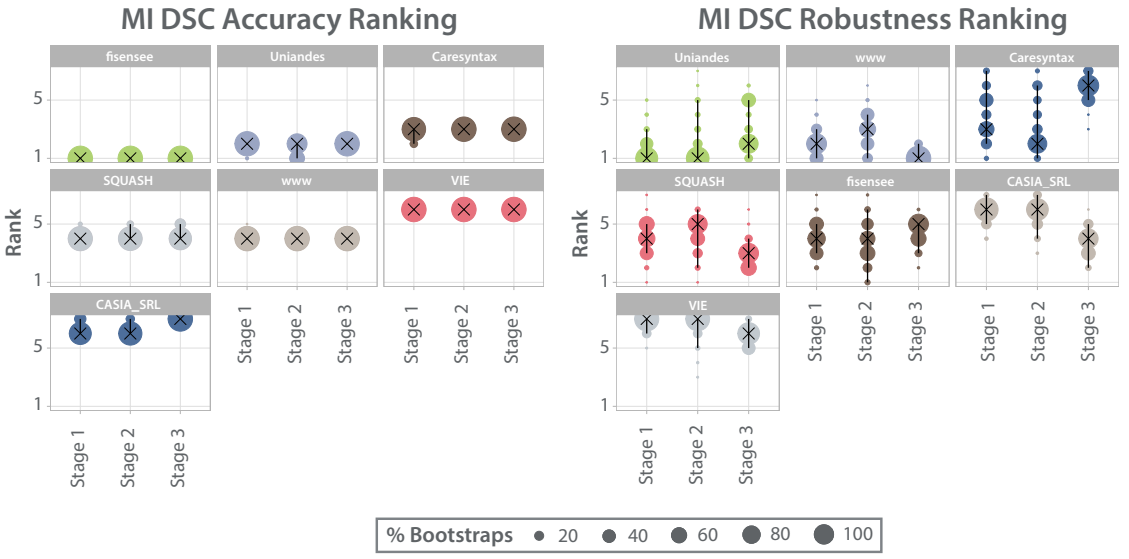


Figure 5.28: Blob plots visualizing ranking uncertainty over 1,000 bootstrap samples for all stages separately for every participating team of the Robust Medical Instrument Segmentation (RobustMIS) challenge for the instrument instance segmentation task. Results are shown for the accuracy (left) and robustness (right) rankings for the Multi-Instance Dice Similarity Coefficient (MI DSC) metric. The radius of the blobs increases with an increasing relative frequency of ranks. The median rank is provided by a black cross while 95% of the distribution of bootstrap ranks is shown by a black line (2.5th-97.5th percentile).

The Medical Segmentation Decathlon

A total of 19 teams submitted their algorithms for the mystery phase. All methods were based on Convolutional Neural Networks (CNNs). The U-Net [Ronneberger et al., 2015] (see Section 2.3) was the most frequently used base architecture and was employed by 64% of the teams. The remaining teams utilized V-Net [Milletari et al., 2016], ResNet [He et al., 2016], QuickNAT [Roy et al., 2019], and DeepMedic [Kamnitsas et al., 2016] basic architectures or a combination of them. Adaptive Moment Estimation (Adam) was the most commonly applied optimizer (61%), followed by Stochastic Gradient Descent (SGD) (33%). Most of the teams used a DSC (29%) or Cross Entropy (CE) (21%) loss function.

The rankings for both the development and mystery phases are provided in Table 5.34. The median and IQR of DSC values are provided to give further insights. In addition, we show the performance of the raw metric values for all data sets and ROIs in Figures 5.29 (development phase) and 5.30 (mystery phase). For the development phase, the lowest median of the mean DSC performance across algorithms was achieved for the pancreatic tumor mass region (0.21). On the other hand, liver segmentation yielded the highest scores (0.94). For the mystery phase, the colon cancer ROI was the most difficult task with a median of mean DSC of only 0.16. The spleen was the data set with the highest median of mean DSC (0.94).

Table 5.34: Test-based rankings for the development and mystery phases of the Medical Segmentation Decathlon (MSD) challenge. For every algorithm, the median and Interquartile Range (IQR) of the Dice Similarity Coefficient (DSC) values of all 19 participating teams are provided.

The development phase				The mystery phase			
Rank	**Team identifier**	**Median DSC**	**IQR DSC**	**Rank**	**Team identifier**	**Median DSC**	**IQR DSC**
1	*nnU-Net*	0.79	(0.61,0.88)	1	*nnU-Net*	0.71	(0.58,0.82)
2	*K.A.V.athlon*	0.77	(0.58,0.87)	2	*NVDLMED*	0.69	(0.55,0.79)
3	*NVDLMED*	0.78	(0.57,0.87)	3	*K.A.V.athlon*	0.67	(0.49,0.80)
4	*Lupin*	0.75	(0.52,0.86)	4	*LS Wang's Group*	0.64	(0.46,0.78)
5	*CerebriuDIKU*	0.76	(0.51,0.88)	5	*MIMI*	0.65	(0.45,0.75)
6	*LS Wang's Group*	0.75	(0.51,0.88)	6	*CerebriuDIKU*	0.56	(0.15,0.71)
7	*MIMI*	0.73	(0.51,0.86)	7	*Whale*	0.55	(0.20,0.68)
8	*Whale*	0.65	(0.28,0.83)	8	*UBIlearn*	0.55	(0.05,0.69)
9	*VST*	0.69	(0.39,0.84)	9	*Jiafucang*	0.48	(0.04,0.67)
10	*UBIlearn*	0.72	(0.40,0.85)	10	*Lupin*	0.57	(0.19,0.69)
11	*A-REUMI01*	0.70	(0.42,0.85)	11	*LfB*	0.49	(0.16,0.64)
12	*BCVuniandes*	0.70	(0.42,0.86)	12	*A-REUMI01*	0.51	(0.14,0.65)
13	*BUT*	0.72	(0.40,0.84)	13	*VST*	0.41	(0.00,0.64)
14	*LfB*	0.68	(0.43,0.82)	14	*AI-Med*	0.33	(0.01,0.52)
15	*Jiafucang*	0.49	(0.11,0.81)	15.5	*Lesswire1*	0.40	(0.08,0.52)
16	*AI-Med*	0.63	(0.30,0.79)	15.5	*BUT*	0.38	(0.01,0.60)
17	*Lesswire1*	0.65	(0.33,0.79)	17	*RegionTec*	0.29	(0.00,0.50)
18	*EdwardMa12593*	0.31	(0.01,0.69)	18	*BCVuniandes*	0.10	(0.01,0.38)
19	*RegionTec*	0.57	(0.19,0.73)	19	*EdwardMa12593*	0.08	(0.01,0.17)

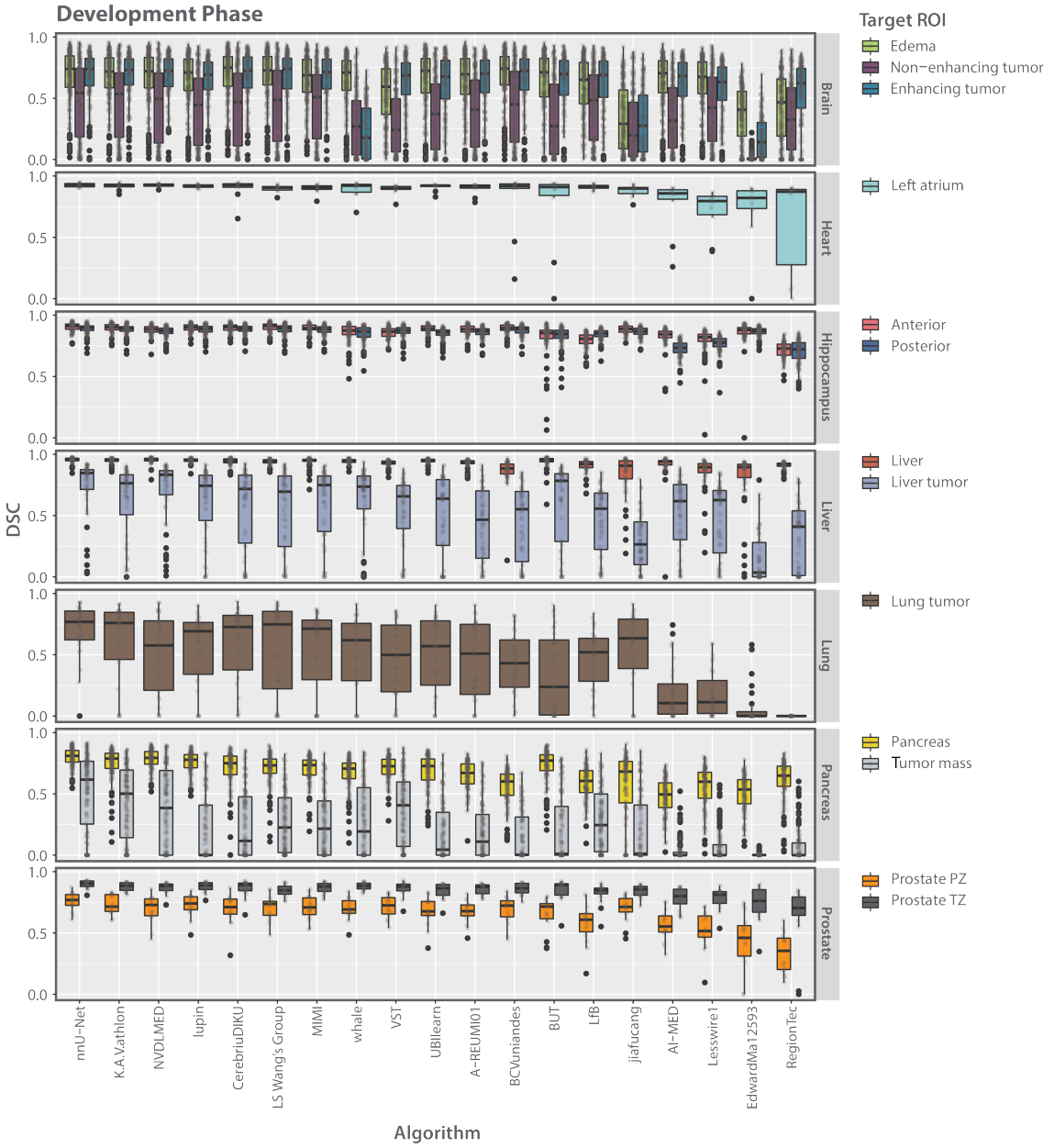


Figure 5.29: Dots- and boxplots illustrating the individual algorithm performances for the development phase of the Medical Segmentation Decathlon (MSD) challenge. Metric values are shown for the Dice Similarity Coefficient (DSC). Results are grouped by data set and target Region of Interest (ROI). Figure adapted from [Antonelli/Reinke et al., 2022].

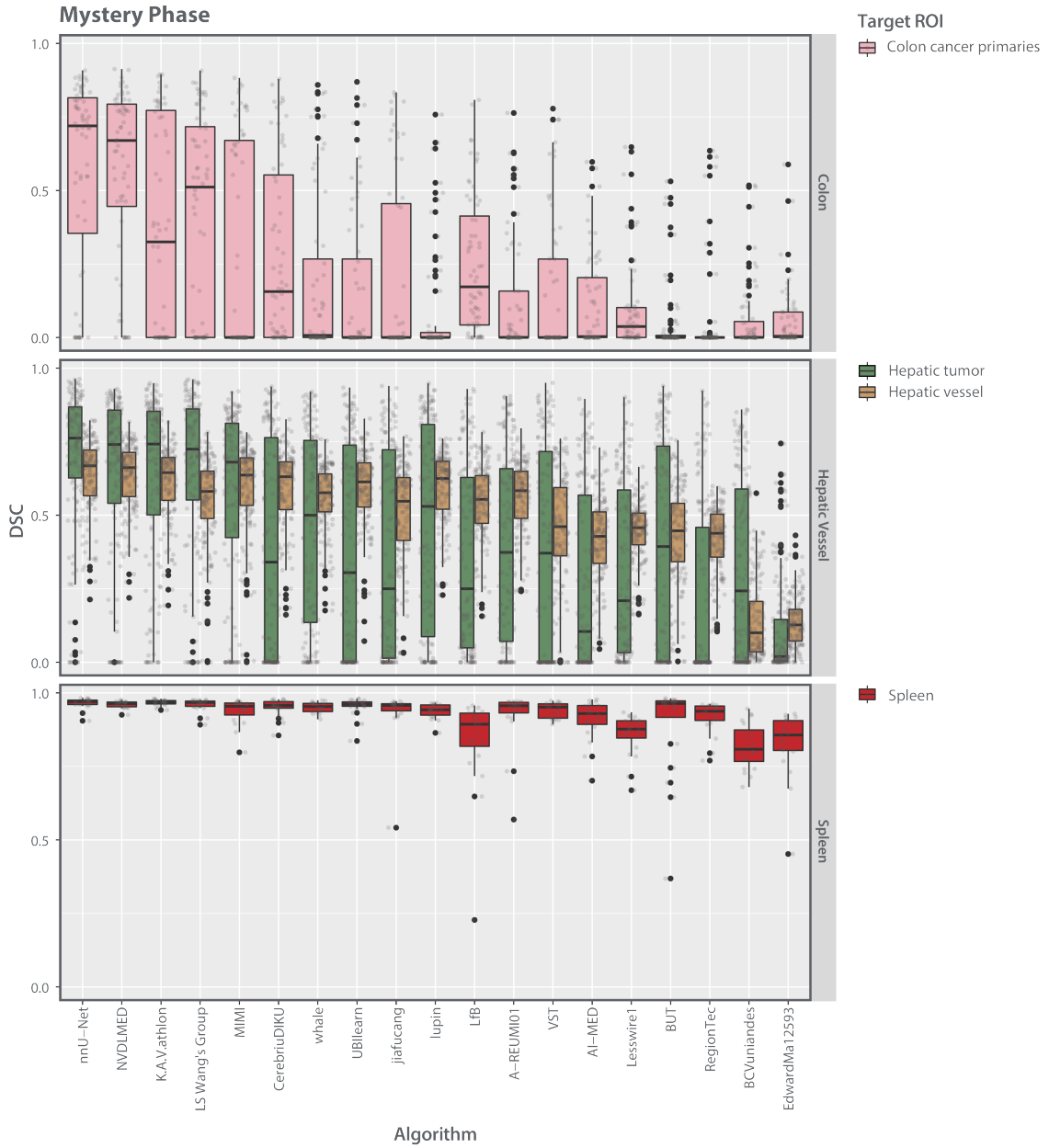


Figure 5.30: Dots- and boxplots illustrating the individual algorithm performances for the mystery phase of the Medical Segmentation Decathlon (MSD) challenge. Metric values are shown for the Dice Similarity Coefficient (DSC). Results are grouped by data set and target Region of Interest (ROI). Figure adapted from [Antonelli/Reinke et al., 2022].

We analyzed the ranking uncertainty by utilizing bootstrapping techniques. Kendall's τ was computed between the original and all bootstrapped rankings, resulting in a median (IQR) Kendall's τ was 0.94 (0.91,0.95) for the colon data set, 0.99 (0.98, 0.99) for the hepatic vessel data set and 0.92 (0.89, 0.94) for the spleen data set. This means that the mystery phase ranks remained relatively constant even when subjected to minor changes. Kendall's τ distribution was comparable to the development phase rankings. The results of the ranking uncertainty analysis are summarized in Figure 5.32 in form of a stacked frequency plot. It can directly be seen that the winning team *nnU-Net* was extremely stable against small perturbations and was the winner for nearly all of the seventeen ROIs, while the remaining teams exhibited higher variability of ranks. This was similar when applying different ranking schemes, as exemplarily shown in Figure 5.31 for the hepatic vessel task. The winning team was superior to the others in most of the cases, the algorithms in the middle of the leaderboard interchanged a lot. This trend could be observed for the remaining ROIs as well.

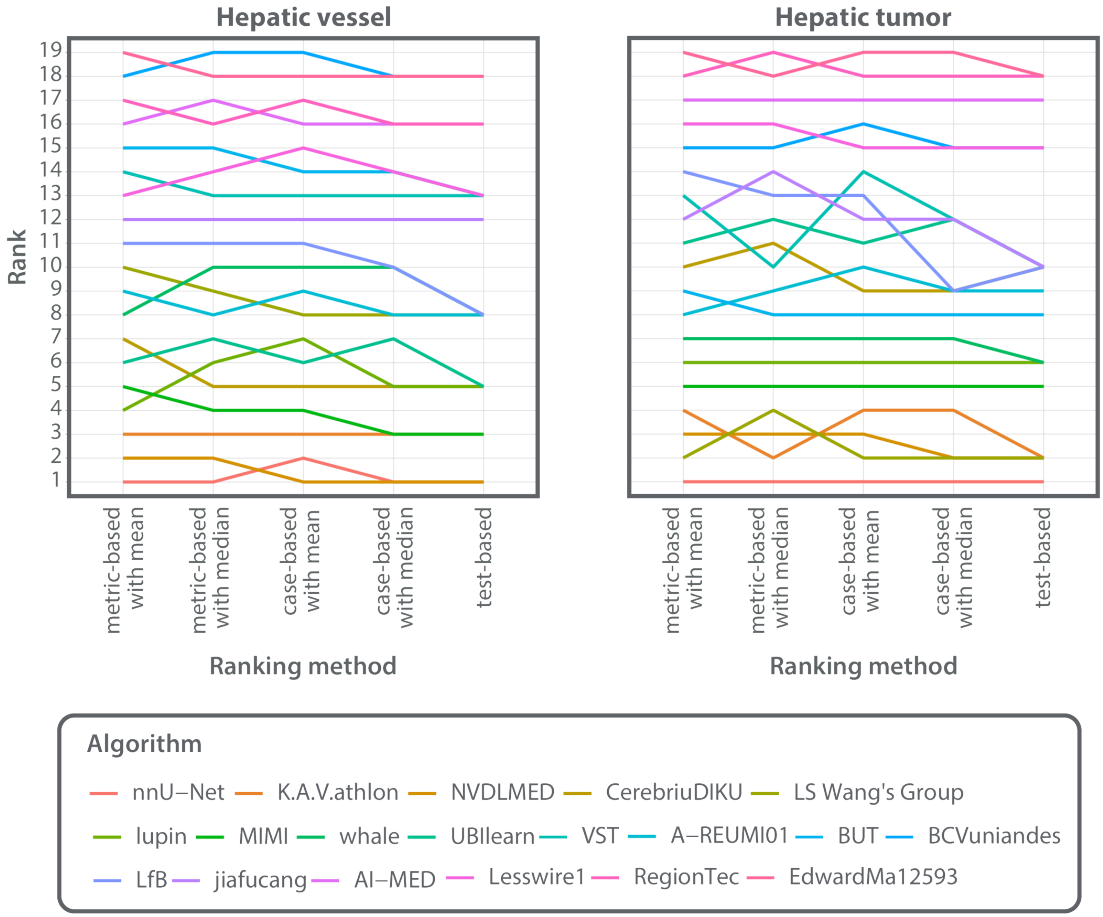


Figure 5.31: Lineplots illustrating the ranking robustness across four different ranking methods for the hepatic vessel data set for the Medical Segmentation Decathlon (MSD) challenge. The height of the line represents the associated rank for each ranking scheme encoded on the x-axis. Each of the 19 algorithms is represented by a differently colored line. Figure adapted from [Antonelli/Reinke et al., 2022].

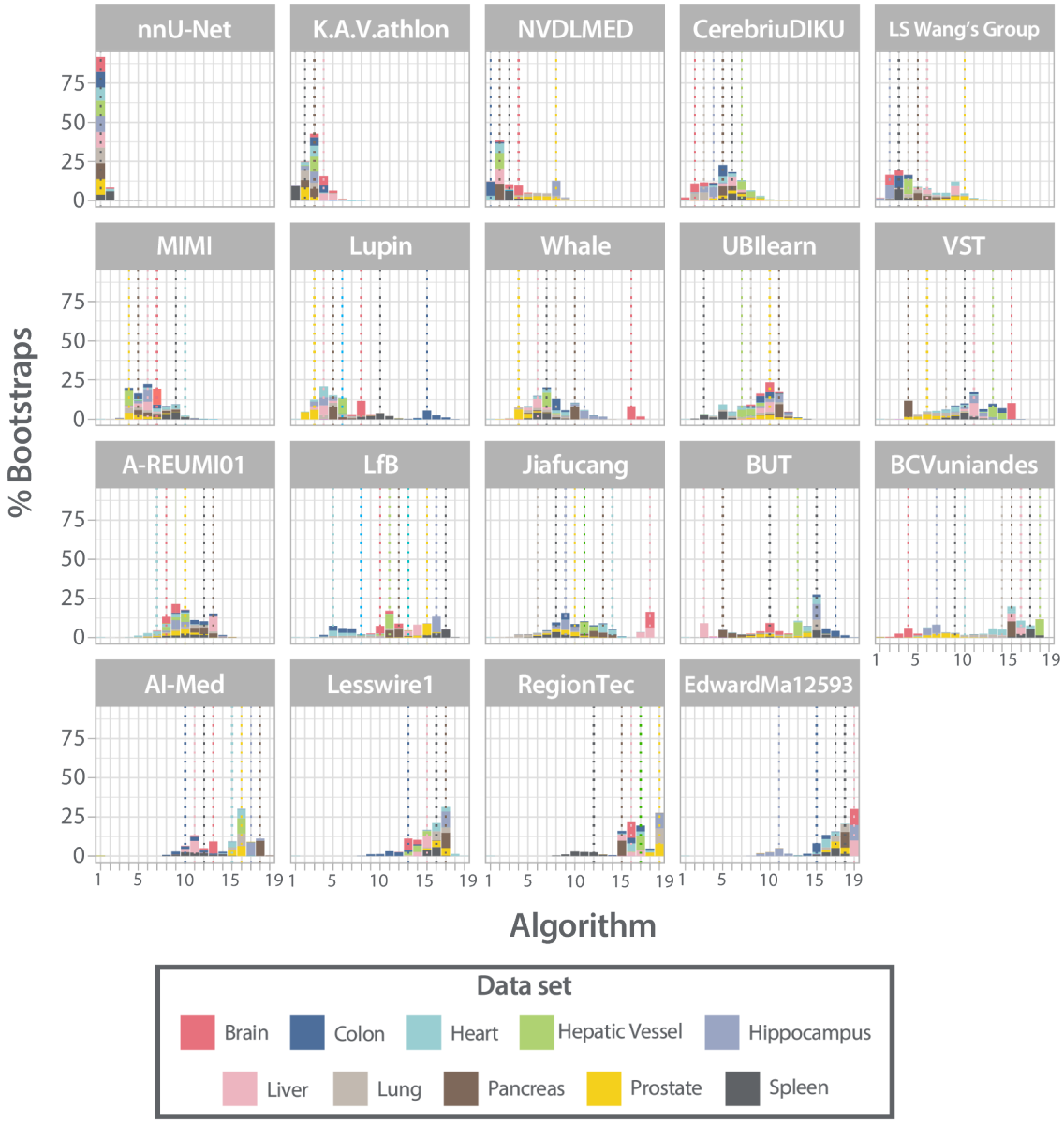


Figure 5.32: Stacked frequency plot illustrating the ranking uncertainty for all data sets of the Medical Segmentation Decathlon (MSD) challenge. Rankings were based on the Dice Similarity Coefficient (DSC). The proportion of achieved ranks (x-axis) for all bootstrap samples is provided on the y-axis. The plot is color-coded by the task and one plot is shown for every participating algorithm. In the case of algorithms that achieved the same ranking in multiple tasks for the whole assessment set, vertical lines appear on top of each other. Figure adapted from [Antonelli/Reinke et al., 2022].

To investigate similarities between the ten data sets based on the performances and ranks, a cluster dendrogram was generated as shown in Figure 5.33. Based on this hierarchical clustering, similarities between several data sets were found. For instance, the lung and hepatic vessel data set were found to be very similar. Both of them share a relatively low median of mean DSC performances (0.51 for the lung tumor, 0.55 for the hepatic vessel, and 0.38 for the hepatic tumor ROI). The liver and pancreas data sets were also found to be very similar. However, they may be similar in terms of rankings, but the performances were much lower for the pancreas task (median of mean DSC of 0.69 for the pancreas and 0.21 for the pancreatic tumor mass) compared to the liver (median of mean DSC of 0.94 for liver and 0.54 for the liver tumor). Although the colon and spleen data sets are very similar, their median of mean performances comes with a substantial difference from 0.16 for the colon data set and 0.94 for the spleen data set.

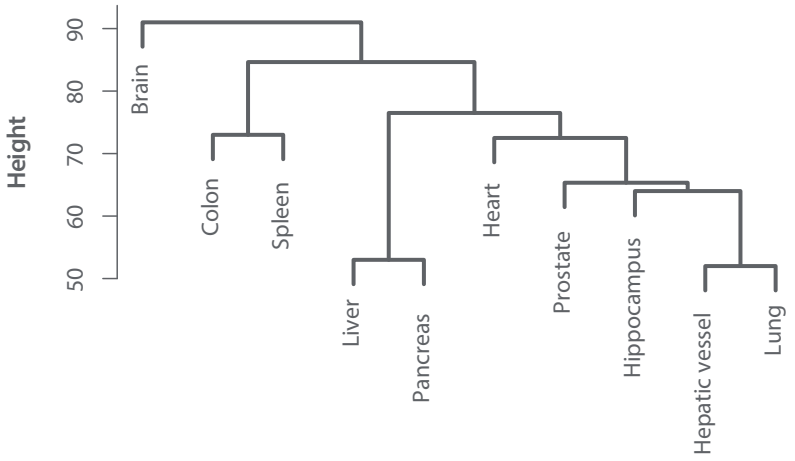


Figure 5.33: Dendrogram illustrating the similarity of data sets for the Medical Segmentation Decathlon (MSD) challenge. It is based on hierarchical clustering and depicts clusters according to Spearman's footrule for the average agglomeration.

5.3.4 Discussion

Challenge rankings are very sensitive to several design choices, as we showed in Section 4.3. This results in a high ranking uncertainty in many cases. In this section, we presented several analysis and visualization techniques that can help to overcome the problem of ranking sensitivity. After revisiting our challenge database, we found that only a limited amount of challenges used advanced visualization techniques beyond simple ranking tables and boxplots. Although boxplots are very suitable for visualizing the raw metric values, they do not help in quantifying ranking uncertainty. We applied the visualization techniques presented in Section 5.3.2 to one simulated and two biomedical image analysis challenges: the RobustMIS and the MSD challenges. The purpose of the simulated challenge was to show the advantages of advanced visualization techniques. They were designed to highlight different and extreme use cases, in which rankings were completely stable and distinguishable (best-case simulation), due to chance (fully random simulation), or yielded the same score and rank for every algorithm (worst-case simulation). As we showed at the beginning of the results of this section, nearly a third of previous challenges presented their results only in the form of a ranking table. The rankings for the best-case and

the fully random simulation would have been exactly the same. Without further visualization and analysis, one would not have observed the differences between the tasks or algorithms. While the dots- and boxplots already reveal the differences in the tasks, the ranking robustness, and uncertainty analysis plots are highly important to show that the rankings of the fully random simulation are completely unstable compared to the best-case simulation.

Both biomedical challenges focus on segmentation problems and are composed of multiple tasks. While the RobustMIS challenge is composed of three tasks for three stages and two metrics, the rankings for the MSD were computed for 17 target ROIs and two distinct metrics. For the MSD, we decided to compute one aggregated ranking over all ROIs and data sets for the development and the mystery phase. However, the ranking uncertainty analysis was still performed on ROI and data set level to reveal insights from the diverse data sets. For this challenge, the winning team *nnU-Net* was extremely distinct from all other teams and significantly superior for almost all tasks. Rankings varied substantially, especially in the middle of the leaderboard. However, the winner and its two successors were stable, although the latter two often interchanged their (second and third) ranks. Given a large number of tasks, the across-task analysis was especially interesting. It not only revealed the superiority of the winning methods for most data sets, but also showed task similarities based on the rankings. These, however, should be interpreted with care, as they do not reflect the raw performance scores. For example, although the colon task was the most difficult and the spleen task was the easiest, the clustering analysis showed them to be very similar.

On the other hand, two ranking schemes were defined for the RobustMIS challenge, for which the results were presented separately. The accuracy ranking was very stable in contrast to the robustness ranking, which showed a high level of variability. The original robustness rankings should therefore be interpreted with care, keeping in mind that the winner was not stable. Especially for the binary segmentation task, teams *haoyun*, *CASIA SRL*, and *www* all were on ranks 1 to 3 at similar frequencies throughout the bootstrap rankings. For the instrument instance segmentation task, one team yielded a rank variation of 5 ranks for the robustness ranking. In the accuracy rankings, most algorithms were significantly superior to their successors. This was not the case for the robustness rankings. This indicates that many algorithms showed poor results in terms of robustness with many teams yielding a 5% percentile of 0.00.

Not all of the presented analysis techniques should be applied to all challenges. The podium plots, for instance, were very complex to read for both RobustMIS and MSD challenges. This is because they either have too many participants, as for the MSD, or too many individual data points, as for the RobustMIS challenge. However, they can be very insightful for small numbers of cases and algorithms, as for example shown in Schellenberg et al. [2022].

The two example challenges chosen for this section are very large in terms of validation: a large number of tasks, multiple metrics, and large data sets. This results in very long reports produced by the *challengeR* toolkit. In such cases, researchers need to carefully review all visualizations so as not to miss important information when selecting figures for their publications. The plots provided for the characterization of algorithms and tasks are especially helpful for large challenges (see Figures 5.28 and 5.32). It should further be noted that the presented visualization techniques are mainly designed for validation per image, not per data set. Thus, in many cases, they can not directly be applied to image-level classification or object detection tasks. However,

the bootstrapping analysis, for example, is still applicable to examine variations in the rankings when the data set is slightly modified. Furthermore, segmentation remains the most frequently applied problem in challenges (see Section 3.1.3). The analysis can therefore still be applied to numerous competitions.

5.3.5 Conclusion

All in all, the insights provided by the advanced visualization techniques, which would not have been obtainable with pure ranking tables, substantially increased the interpretability of challenge results. They show details of algorithm performances that are very important for the practitioner when choosing a specific algorithm for a new problem. For example, based on the analysis, a researcher could be confident that the winning algorithm of the MSD, team *nnU-Net*, would solve a range of different semantic segmentation approaches. However, the same researcher would know to be careful when choosing an algorithm from the *RobustMIS* challenge if the focus is algorithm robustness, since many teams yielded similarly (low) results.

It should be noted that the usage of the suggested visualization techniques is far from being restricted to challenge results. Indeed, they can be applied to most of the currently published papers. Many journals nowadays require the comparison of the new method or approach to baseline or state-of-the-art methods. Those benchmarking experiments can therefore be interpreted as "mini-challenges", for which all of the presented analysis methods can easily be applied. Similarly, they can be used for hyperparameter tuning or model selection, and beyond biomedical applications. The insights provided by uncertainty analyses are very helpful for future algorithm development and add tremendous value over simple ranking tables. Finally, they help to avoid inadequate algorithms that might only have been superior to others by chance being spread across the community or even translated into clinical practice.

5.4 Improving common practice of reporting and analyses

Disclosure

This work has been supervised by Lena Maier-Hein. It has been conducted together with several co-authors, especially Matthias Eisenmann, Tobias Roß, and Annette Kopp-Schneider. The work has been published in *Medical Image Analysis* [Maier-Hein et al., 2020] and submitted to *Medical Image Analysis* [Roß et al., 2021]. Please refer to Chapter A.1 for full disclosure.

5.4.1 Introduction

Biomedical challenges are typically organized to gain insights or to find algorithms that best solve a particular research question [Mendrik and Aylward, 2019]. With this reasoning in mind, challenges, their winners, and their results have a substantial impact on future scientific work. Thus, a high-quality challenge design is key for ensuring that we can trust challenge results. Eddy et al. [2012] argue that transparency and validation are the most important aspects to consider when establishing confidence and trust in biomedical models. While we have presented strategies for improving the validation of challenges (Sections 5.2 and 5.3), here, we focus on further advancing the transparency of challenge results.

> *'The one 'practice' that can be universally commended is the transparent and complete reporting of all facets of a study, allowing a critical reader to evaluate the work and fully understand its strengths and limitations.'*
>
> — [Nichols et al., 2017]

The structured challenge submission system was the first step toward increased challenge transparency. However, we demonstrate in this section that this step alone was insufficient. Although organizers could easily use the information provided in the challenge applications, most of them did not make use of it in challenge reports. Since transparency and reproducibility of results are crucial for ensuring high-quality research [Baker, 2016; McNutt, 2014; Schulz et al., 2010], we went a step further by introducing a reporting guideline for challenges, similar to popular guidelines and checklists for reporting clinical trials, as presented in Section 2.5. The reporting guideline for biomedical challenges included a refined version of the parameter list that was initially used for the structured challenge submission system in Section 5.1. We registered the guideline with the Enhancing the QUAlity and Transparency Of health Research (EQUATOR) network [Altman et al., 2008] and present the individual parameters in this section.

Challenge reports should ideally not only present the raw results of a challenge. One parameter of the reporting guideline is *further analyses* of challenge results (see Section 5.4.3). This includes all kinds of analyses helpful to draw meaningful conclusions from the results besides the ranking analysis. As challenges play a major role in driving scientific research, we should ask ourselves how we can leverage more from challenge results. For example, one intriguing subject is what distinguishes difficult-to-process challenge images for the algorithms. Identifying characteristics

of the worst-case images can be extremely useful for future algorithm development [Roß et al., 2021]. While this question was sometimes investigated manually [Roß/Reinke et al., 2020; Allan et al., 2021, 2020], this process is time-consuming and may be biased and subjective. Therefore, we propose a statistical approach using Linear Mixed Models (LMMs) to identify image characteristics that positively or negatively influenced the performances of challenge participants. We demonstrate the value of such an analysis for the Robust Medical Instrument Segmentation (RobustMIS) challenge.

To summarize, the contribution of this section is twofold. First, we present the newly designed challenge reporting guideline and show how we can further apply other strategies from clinical trials to challenges. Second, we show how we can better learn from challenges.

Specific findings in Part 4

Challenges are not reported comprehensively.

Challenge algorithms are not reproducible.

Research questions investigated in this chapter

Can we apply clinical trial concepts to challenge reporting?

How can we gain more insights from challenge results?

5.4.2 Methods

In this section, we review the influence of our structured challenge submission system on challenge reporting. As an additional step, we introduce the concept of challenge registration. To formalize our challenge parameter list, we converted a refined version into a challenge reporting guideline. We conclude with a mixed model-based strength- and weakness analysis of challenge results. All methodological details are provided in the following paragraphs.

Structured challenge submission system

In Section 5.1, we introduced a structured challenge submission system for biomedical challenges. In this section, we show the improved reporting practice based on the parameters captured within the Object Database (ODB) of the submission system. For the Medical Image Computing and Computer Assisted Interventions (MICCAI) 2018 conference, we further compared the parameters stored in the system – as accepted by the MICCAI 2018 challenge chairs – with the information actually reported on the websites for discrepancies.

Challenge registration

We leveraged best practices from clinical trials by reviewing the concept of clinical trial registration. The International Committee of Medical Journal Editors (ICMJE) "requires, and recommends that all medical journal editors require registration of clinical trials in a public trials registry at or before the time of first patient enrollment as a condition of consideration for publication" [ICMJE, 2013]. The purposes and benefits of clinical trial registration are described in Zarin and Keselman

[2007] and summarized in Table 5.35. The concept was translated to biomedical challenges and is presented in Section 5.4.3.

Table 5.35: Clinical trial registry purposes and benefits for various groups. Table courtesy of Zarin and Keselman [2007].

Registry purpose	**Group that benefits**
Fulfill ethical obligations to participants and the research community	Patients, the general public, the research community
Provide information to potential participants and referring clinicians	Patients, clinicians
Reduce publication bias	Users of the medical literature
Help editors and others understand the context of study results	Journal editors, users of the medical literature
Promote more efficient allocation of research funds	Granting agencies, the research community
Help institutional review boards (IRB) determine the appropriateness of a research study	IRB, ethicists

Reporting guideline

Based on the feedback after the first year of challenge submissions using the structured challenge submission system, the MICCAI Board challenge working group refined the challenge parameter list. We especially clarified the parameter descriptions that were misunderstood by the users, namely the challenge organizers, and removed redundant information by merging the parameters that were individually listed for the training and test data sets. We then provided the refined list of 48 parameters to all members of the Biomedical Image Analysis ChallengeS (BIAS) initiative in form of a survey asking for agreement with the name and explanation of each parameter as well as constructive feedback in case of disagreement, similar to the first survey mentioned in the previous sections. Furthermore, we asked the survey respondents to rate every parameter by its importance (absolutely essential for challenge result interpretation and/or challenge participation, should be included, may be omitted) and whether it is essential for challenge review. We, as the MICCAI Board challenge working group, addressed all the comments and organized a final meeting with all members to address conflicts. In the end, we created a final list of 42 main parameters with 79 sub-parameters. For example, the main parameter *challenge name* comes with two sub-parameters 'provide a representative name of the challenge' and 'provide the acronym of the challenge (if any)'.

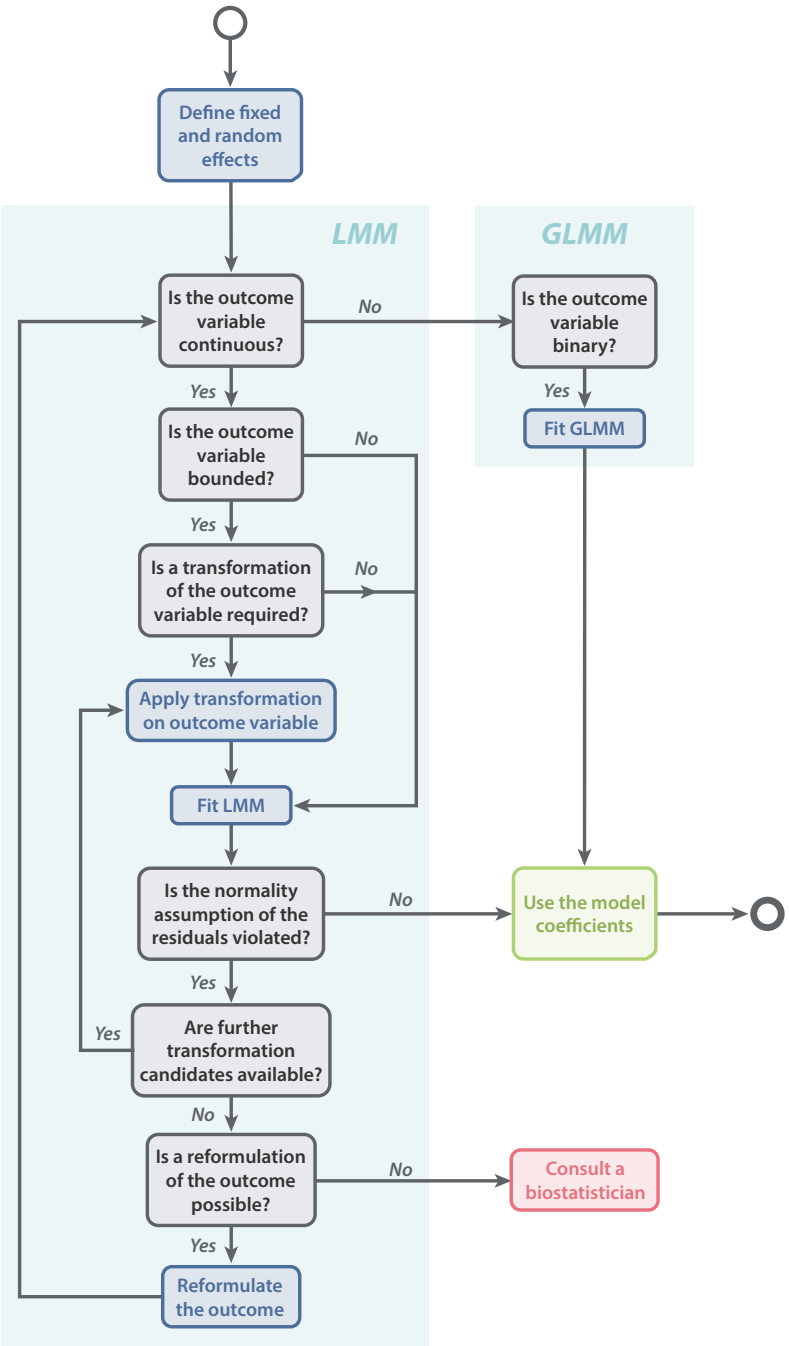


Figure 5.34: Workflow diagram for applying the suggested mixed model analysis to challenge results. Based on the outcome variable, a Linear Mixed Model (LMM) or Generalized Linear Mixed Model (GLMM) can be applied. In the case of a complex situation, we recommend consulting a biostatistician.

Mixed model-based strength-weakness analysis

Biomedical data is often subject to hierarchy and correlation between subjects, e.g. patients from the same hospital or scanner. Thus, results should be analyzed by a mixed model rather than regular linear models. To determine how widespread this analysis technique is, we investigated how many challenges used mixed models. We reviewed all 5,390 papers that were published in the proceedings of the MICCAI conferences in the period of 2004 to 2021. All papers were searched for 44 terms related to mixed model analysis (see Appendix A.9) to find out how many papers are actually utilizing those models.

To investigate the question of how to leverage more from challenge results, we suggest studying which image characteristics made an image particularly hard to process for challenge participants. We propose the following procedure using mixed models. We instantiated this procedure for the instrument instance segmentation task of the RobustMIS challenge (see Sections 4.4 and 5.3 as well as Appendix A.6 for a detailed overview of the challenge). An overview of the mixed model-based challenge result analysis is provided in Figure 5.34.

1. **Identification of potential failure sources:** First, *potential failure sources* of algorithms should be identified. Based on the challenge and data set, failure sources may, for instance, relate to the imaging protocol (e.g. yielding underexposure), the imaging devices (e.g. yielding dirty lenses), or the actual target structure (e.g. yielding overlapping objects). The identification of failure cases and image characteristics should rely on extensive literature research, expert knowledge, and personal experience.

2. **Metadata annotation:** The identified error sources should be semantically annotated among the challenge test cases (or parts of the test data).

3. **Mixed model analysis:** LMMs (or GLMMs) can be used to quantify the impact of the identified image characteristics on algorithm performance. For this purpose, the performance of challenge algorithms for an image is modeled as a function of the annotated image characteristics that were identified as potential failure sources.

5.4.3 Results

Recent reporting practice and challenge registration

To overcome the lack of reporting, we implemented the structured challenge submission system, which was used for multiple conferences, as detailed in Section 5.1.3. For the 138 challenges and 255 corresponding tasks that were accepted through the system, a median percentage of 100% (Interquartile Range (IQR): (100%, 100%); minimum: 93%) was reported, given the requirement to fill at least 90% of parameters. A median percentage of 100% was achieved for all conferences. The lowest numbers of parameters were reported at the MICCAI 2020 conference with a minimum of 93% parameters reported. For the IEEE International Symposium on Biomedical Imaging (ISBI) 2021 conference, all challenges reported all parameters.

Exemplarily for MICCAI 2018, we revisited the challenge websites five months after their acceptance and compared the reported information to the parameters given at the point of acceptance (see Figure 5.35 for an overview). A median of 40% (IQR: (31%, 57%); minimum: 9%; maximum: 89%) of parameters was reported as stated in the respective challenge designs submitted to the

structured submission system and accepted by the reviewers. However, a median of 49% (IQR: (29%, 57%); minimum: 9%; maximum: 83%) of parameters were not provided on the websites although all information was available in the challenge designs they submitted earlier. Finally, a median of 9% (IQR: (6%, 17%); minimum: 3%; maximum: 29%) of parameters on the websites differed from the accepted challenge designs.

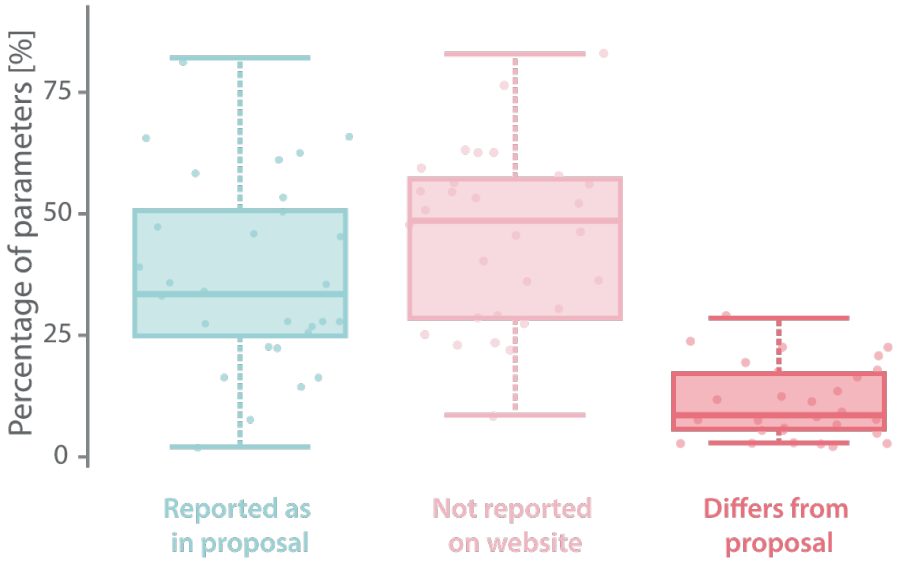


Figure 5.35: Percentage of parameters reported on the challenge websites as stated in the accepted challenge proposal (left), not reported on the challenge websites (middle), and reported differently from the proposal (right). One data point corresponds to one of the accepted MICCAI 2018 challenges.

Challenge registration

Given these findings, which were contrary to our recommendations that challenge organizers should publish the complete challenge design before the challenge takes place, the BIAS initiative examined best practices in the field of clinical trials. We introduced the concept of challenge registration for MICCAI challenges. After acceptance, the complete challenge design documents that were produced by the structured challenge submission system were uploaded to Zenodo [Peters et al., 2017], a platform for open science, that additionally assigns a Digital Object Identifier (DOI) to every challenge design document. The DOIs were subsequently added to the website of the MICCAI Board challenge working group (which is now the MICCAI Special Interest Group (SIG) for Challenges) [MICCAI_SIG_for_Challenges, 2018]. Thus, even if the information was missing from the challenge websites, details were still publicly available and visible to all potential challenge participants. Challenge organizers further agreed that all changes to the challenge design should be reported back to the respective MICCAI challenges team. This process was handled via the structured challenge submission system and yielded a new challenge design document that highlighted all changes. If accepted by the MICCAI challenges team, the updated document was added as a new version to Zenodo. As a consequence, 88 MICCAI challenges were registered in the period from 2020 to 2022, from which 19 reported changes to their challenge design documents.

Reporting guideline

To enhance the transparency of reporting, we designed a reporting guideline for biomedical image analysis challenges by the BIAS initiative based on the challenge parameter list (cf. Table 4.5). The guideline was not intended to force authors to report results in the presented order, but rather proposed to write a challenge report in the typical way research papers are written. Authors could choose a different ordering or renaming of sections as long as the relevant items were reported. The proposed structure intersects the methods section into challenge organization, mission of the challenge, challenge data sets, and assessment methods. The concrete checklist is given in Appendix A.8. It contains a column, in which challenge organizers or reviewers could insert the page number at which the respective information is provided to allow for easy and quick information transfer. Before presenting the structure and checklist items, we specify the changes with respect to the previous challenge parameter list.

The reporting guideline described the items that should be addressed in a full challenge report. Therefore, checklist items for the title, abstract, keywords, introduction, discussion, and conclusion were added. The method sections 'challenge organization', 'mission of the challenge', 'challenge data sets', and 'assessment method' were kept in the BIAS reporting guideline. 'Participation conditions' were merged with the 'challenge organization', similar to the 'study conditions', from which parameters were moved either to the mission of the challenge (*study cohort, context information*, and *imaging modality(ies)*) or challenge data sets (*center(s), acquisition device(s)* and *protocol(s)*, and *operator(s)*).

The BIAS reporting guideline was redesigned to group several checklist items into one parameter. For example, the new parameter *data source(s)* contained the parameters *acquisition device(s)*, *details on the data acquisition*, *center(s)/institute(s)*, and *characteristics of the subjects involved in the data acquisition* (former: *operator(s)*). In some cases, parameters of the previous challenge parameter list were merged into one checklist item, such as all parameters separately listed for training and test data, which were combined into one item, as information was often found to be redundant. In the case of different specifications for training and test sets, organizers were asked to highlight them. Finally, several parameters were renamed or provided with a more detailed description to ensure clarity. For example, the parameter *study cohort* was misunderstood by many users of the challenge submission system and was therefore renamed to *challenge cohort*.

In the following, all parameters and checklist items listed in the *BIAS* reporting guideline are presented:

Title, abstract, and keywords The *title* of the challenge report emphasizes the challenge mission. It should name the imaging modality, problem category, and state that the report is about a biomedical image analysis challenge organized in a specific year. In the report's *abstract*, a summary of the challenge goal, the results, and conclusions should be given. The main challenge characteristics (e.g. problem category and imaging modalities) should be provided as *keywords*.

Introduction The challenge report introduction motivates the topic of the challenge from a *technical* as well as a *biomedical* point of view and highlights the envisioned biomedical/technical impact. The *primary challenge objective* should be provided based on the motivation including concrete *statements of the challenge's task(s)*.

Methods – Challenge organization A substantial portion of a challenge is about challenge planning. Many necessary parameters belong to the overall challenge organization and may be of less interest to the report reader. We therefore suggest moving this information to the appendix. Regardless of where the information is given, the challenge organization section should comprise the *representative challenge name* and *acronym*, if any. Names and affiliations of all members of the *organizing team* should be provided. The intended challenge *submission cycle* (or: life cycle type) provides the reader with information on whether the challenge is a one-time or repeated event with a fixed submission deadline or whether it is open for submissions after the conference. The *event*, for example, the conference, associated with the challenge should be named, as well as the platform, such as grand-challenge.org or Kaggle, which was used for the challenge should be reported along with its *URL*.

Several participation policies should be defined, starting with the *allowed user interaction* of the submitted methods. A challenge may for example only accept automatic methods, or also allow for semi-automated algorithms. If so, a strategy to handle different method types in the challenge ranking(s) (if any) should be provided, for example providing separate leaderboards. A *training data policy* defines whether challenge participants are allowed to utilize other data than the challenge training data for optimizing their methods, such as private data or publicly available data sets. In addition, a particular *participation policy for members of the organizer's institutes* implies whether or not members of the organizer's institutions are listed in a ranking in case of participation or whether they are not allowed to participate at all. The *award policy* and *results announcement policy* define the challenge prizes and how results are revealed, for example, only the top three methods could be announced or participants may choose whether or not to be listed online. Finally, the *publication policy* provides details on which members of a participating team are listed as co-authors of a publication, or whether they are allowed to publish their methods based on the challenge results before the challenge publication is accepted by a journal.

Challenge organizers should describe in detail how challenge participants should *submit* their results, for example by providing Docker containers or uploading the outputs of their methods to an evaluation platform. In addition, *submission instructions* should be provided, indicating the concrete output format and instructions to upload/generate the desired submission outputs. Furthermore, participants need to know whether a *pre-evaluation* is possible before they submit their final results, for example, if participants are allowed to submit their results several times to check their current scores.

The *challenge schedule* should be provided as a timetable, including the most important dates for challenge participants, i.e. release date(s) of training, validation, and test data, the registration period, submission deadlines, workshop day (if any), and the release date(s) of the final results. Furthermore, organizers should indicate whether *ethics approval* was needed (e.g. not necessary for fully anonymized data), and if so, link to the approval's document or provide the ID (if available). In addition, the *data usage agreement* including the concrete *license* clearly states how the challenge data can be used and distributed by challenge participants, and general researchers. Authors should include the *code availability* of the organizer's validation software and participating methods' code. Finally, *conflicts of interest* should be provided, especially giving information to organizers who had access to the test data annotations, along with *author contributions*.

Methods – Mission of the challenge This section should provide deeper insights into the concrete challenge mission. First, the *field(s) of application* that the challenge is targeting should be mentioned. This may be the diagnosis of patients, intervention planning, or similar. The *problem category* indicates whether algorithms target semantic segmentation, object detection, or other image processing tasks.

The *challenge* and *target cohort* should be described. The cohorts may differ from each other, since the target cohort is focusing on patients (more general: subjects/objects) who should be targeted in the final biomedical application (e.g. patients with specific criteria undergoing a specific surgery), whereas the challenge cohort means the subjects/objects from whom/which the challenge data was acquired (e.g. healthy volunteers). It is important to differentiate both cohorts since there may be a domain gap between the challenge results and the real application given model training on the challenge cohort data.

Authors should describe the *imaging modality/-(ies)* used for the challenge data and provide *context information along with the images* corresponding to the image data (e.g. tumor volume), the patients (subjects/objects) (e.g. gender, age) and the acquisition process (e.g. calibration for an image modality). This information may differ for the challenge and target cohort.

Similarly to the differentiation between challenge and target cohort, one should differentiate between the *data origin*, i.e. the region(s) of the subjects/objects in the final biomedical application (for example a liver tumor in a Computed Tomography (CT) image), and the *algorithm target*, i.e. the region(s) of the subjects/objects the participating methods were designed to focus on.

Finally, the *assessment aim* should be mentioned, which refers to the properties that should be optimized by the participating methods. This may be the accurate segmentation of a tumor or the short computing time of the algorithms. The assessment aim should be reflected in the performance metrics.

Methods – Challenge data sets This section describes the challenge data in detail. It should start with a concrete explanation of the *data sources*, involving a description of the *device(s)* used for data acquisition as well as relevant details on the *data acquisition process*, including an acquisition protocol, if any. The *center(s)* in which the data was acquired or the platform from which the data was taken should be described. Finally, relevant *characteristics of the subjects* involved in the data acquisition process should be given, such as the level of expertise of the surgeon.

The *training and test case characteristics* first define what the challenge defines as one *case* (see Section 2.1.2), followed by the number of training, validation, and test cases as well as the *total number of cases*. A *justification of why these numbers were chosen* and *how they were split* into the training, validation, and test sets should be included in the section. This should be followed by *further important characteristics* of the training, validation, and test sets, such as the class distribution in comparison to the real-world distribution.

The *annotation characteristics* describe the *method used to determine the reference annotation*, for example, manual annotation or annotation done by automatic methods. The information should be complemented by providing the *instructions given to the annotators* (if any). This should include the complete annotation protocol, explaining how to annotate common and corner cases of the data. *Details on the subjects/objects/algorithms* generating the annotations should be given, such as the level of expertise in the domain and annotation for every annotator. In the case of multiple annotators, the *method used to merge multiple annotations* for a single case should be specified, such as a majority vote or consulting another observer in case of ambiguities.

Data pre-processing methods should be described as well as *potential error sources* related to the image annotation, for example by providing inter- and intra-rater variability, and other relevant sources of error.

Methods – Assessment method This section focuses on the validation of participating methods. First, the *metric(s) used to assess a property of an algorithm* should be described in detail and should be chosen according to the assessment aim described above. The metric choice should be *justified*. The metrics should be followed by *methods used to compute a performance rank* for the participating teams. For example, the ranking scheme may be metric-, case-, or test-based (cf. Section 2.1.2). Specifically, authors should indicate how to deal with *submissions with empty results* and different user interaction methods. The ranking method should be well *justified*. Finally, the *statistical analyses* should be described in detail. It should include details on missing data handling, ranking variability, and the used software products. Similar to the above, the statistical methods need to be *justified*.

Results – Challenge outcome The results section focuses on the concrete challenge outcome. Specifically, this includes information on the participating methods and their results. The section should start with providing information on the *challenge submissions*, including the *number of registrations*, *number of participating teams* with valid submissions, and the *number of participating teams mentioned in the paper*. The last two numbers may be similar but a challenge with many submissions may choose to only report the top-ranked methods, such as the top ten. For those teams, a team identifier and a method description, including the choice of hyperparameters and concrete architecture should be given. If possible, a link to the team's code is preferable.

Aggregated metric values for every participating team including a measure of variability, such as the standard deviation or IQR should be reported. In the case of per-image validation, ideally, the distribution of metric values should be visualized, for example in the form of dots- and boxplots. Metric values should be followed or reported along with the *rankings* including the number of submissions per participating team. The *statistical analysis* of the rankings should be provided. Finally, potential *further analyses* can be reported, such as inter-algorithm variability, ranking uncertainty analyses, common problems of the submitted algorithms, or the performance of a combination of algorithms via ensembling.

Discussion The discussion section *summarizes* the results of the challenge. The section should describe the *biomedical and technical impact* of the challenge according to the challenge motivation described in the introduction. The concrete challenge results should be discussed, indicating whether the *problem is solved* with the challenge results and *providing advantages and disadvantages of the submitted algorithms*. An *analysis of individual cases* would be beneficial

for readers, stating for which cases the algorithms performed poorly, arguing that future work should focus on those cases. *Limitations of the challenge* related to the challenge design and execution should be described as well as *recommendations for future work*. The challenge report should end with a *concise conclusion* based on the challenge results.

Mixed model-based strength-weakness analysis

As introduced in the previous section, a challenge report could encompass further analyses, going beyond ranking tables and analyses. We present an analysis of image characteristics that were identified as particularly challenging for the challenge participants of the RobustMIS challenge. Before the results are presented, we give an overview of the usage of mixed model analyses in current publications. We found that only 1.7% of the MICCAI 2004 to 2021 conferences papers utilized mixed models although non-independent data is very widespread in biomedical use cases. Most of those publications use hierarchical or mixed models to explore the variability of groups or differences between them (e.g. [Swee and Grbić, 2014; Kim et al., 2015; Kutra et al., 2012]).

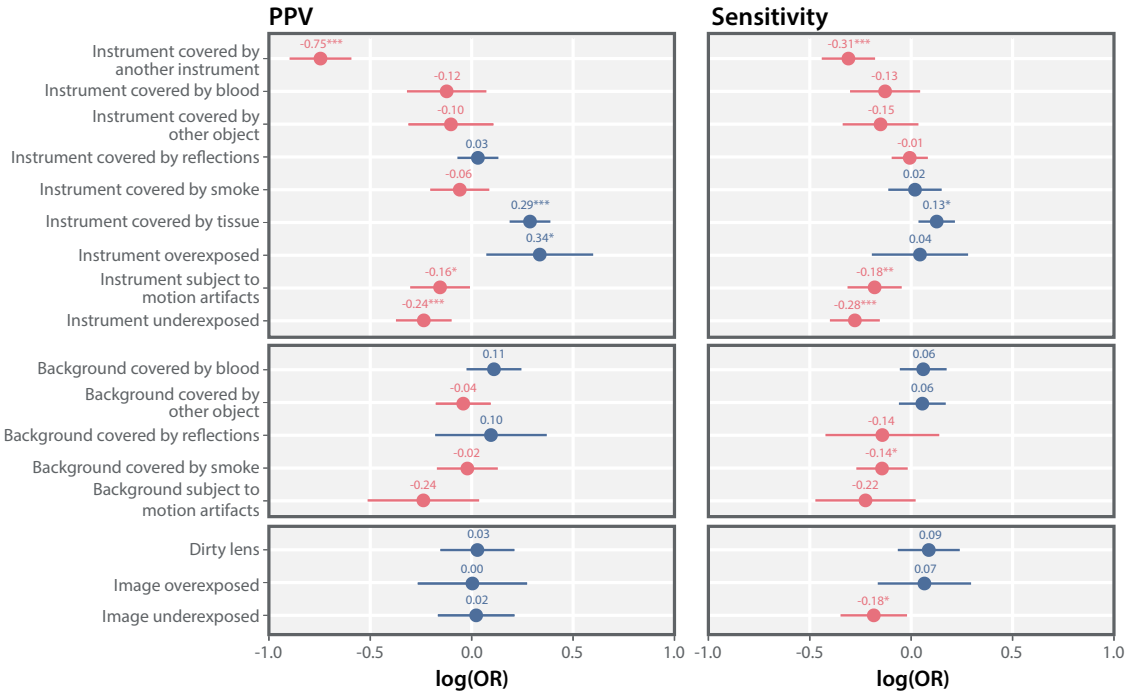
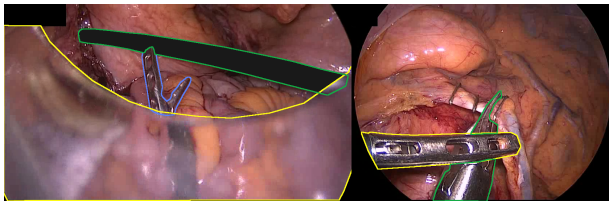
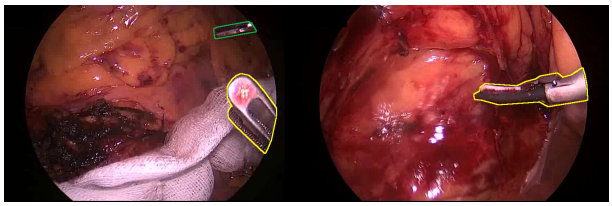
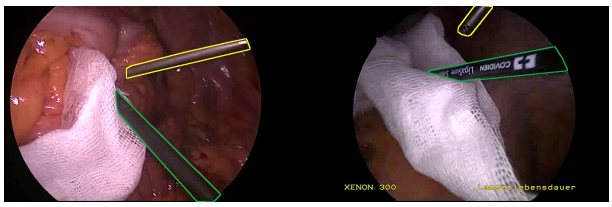
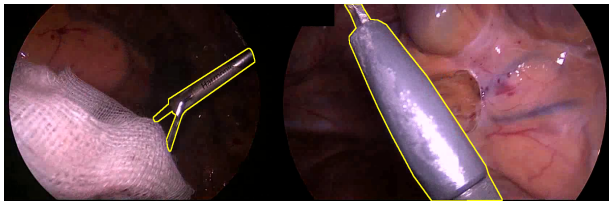


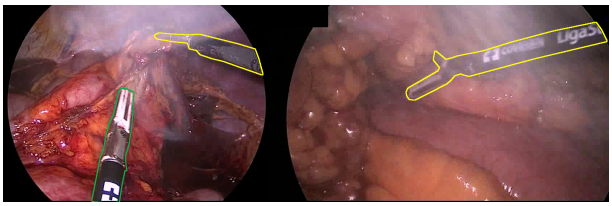
Figure 5.36: Overview of significant positive and negative effects of image characteristics for the Positive Predictive Value (PPV) and Sensitivity for the top five algorithms of the Robust Medical Instrument Segmentation (RobustMIS) challenge simultaneously. The effects are shown in the form of the logarithm of the Odds Ratio (OR), log(OR). Positive effects are shown in blue, negative effects in red with asterisks indicating significant effects (*: $p \leq 0.05$, **: $p \leq 0.01$, ***: $p \leq 0.001$). The confidence intervals are shown by horizontal lines.

Identification of failure cases In the case of the RobustMIS challenge, we combined our personal knowledge gathered during the data set annotation and generation [Roß/Reinke et al., 2020; Maier-Hein et al., 2021] with the results from analyzing the literature related to artifacts in endoscopic imaging [Ali et al., 2020, 2021a; Funke et al., 2018; Soberanis-Mukul et al., 2020], endoscopy applications in image analysis [Maier-Hein et al., 2014; Ali et al., 2021b], and challenges in the context of endoscopy (e.g. [Bodenstedt et al., 2018; Allan et al., 2020]). Based on this analysis, we identified potential sources of errors, from which three characterize the whole image, and nine characterize the background (five image characteristics) and/or the surgical instrument(s) (nine image characteristics) visible in the image. Thus, a total of 17 image characteristics were considered potential error sources. They are presented in Tables 5.36 - 5.38.

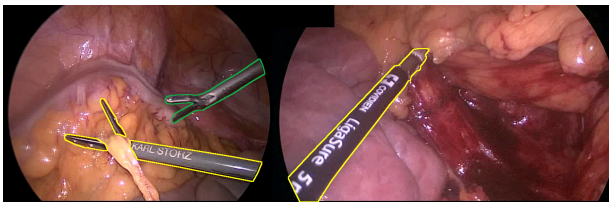
Table 5.36: List of **instrument-related image characteristics** that were identified as failure sources in the Robust Medical Instrument Segmentation (RobustMIS) challenge.

Image characteristic	***Example images***
Instrument covered by another instrument	
Instrument covered by blood	
Instrument covered by other object	
Instrument covered by reflections	

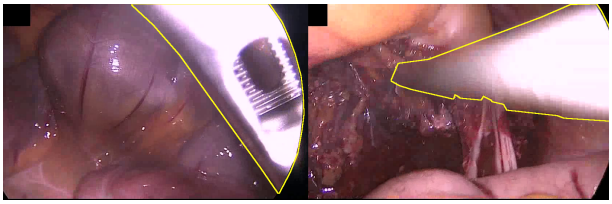
Instrument covered by smoke



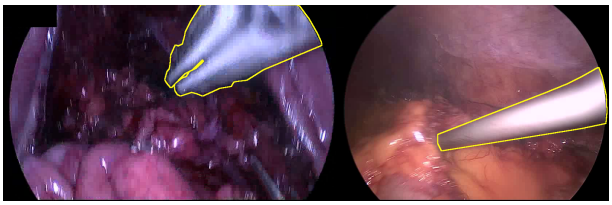
Instrument covered by tissue



Instrument overexposed



Instrument subject to motion artifacts



Instrument underexposed

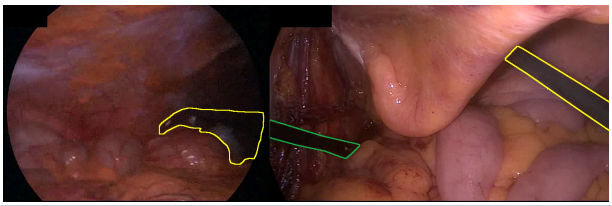


Table 5.37: List of **background-related image characteristics** that were identified as failure sources in the Robust Medical Instrument Segmentation (RobustMIS) challenge.

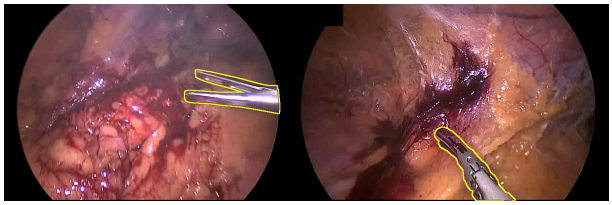
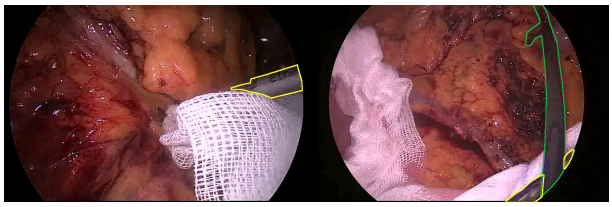
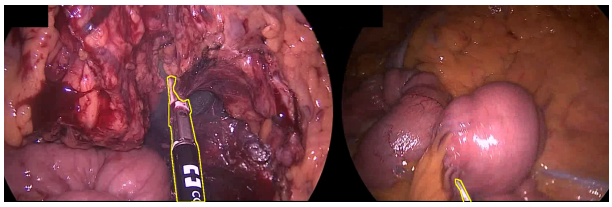
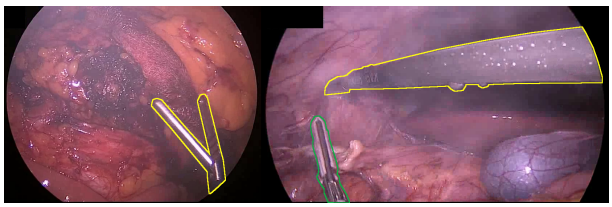
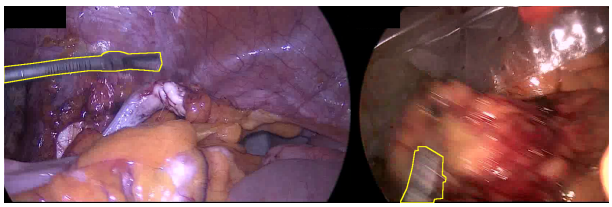
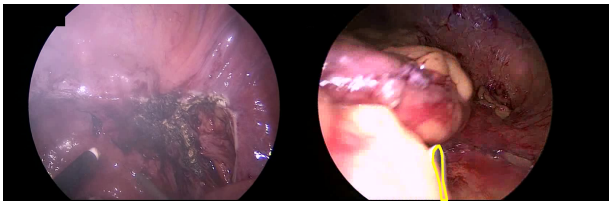
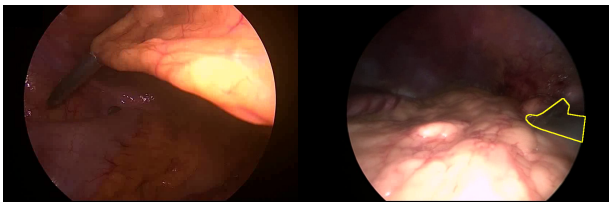
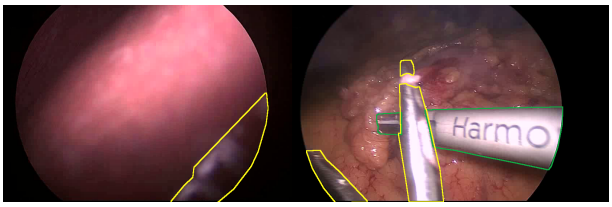
Image characteristic	***Example images***
Background covered by blood	
Background covered by other object	
Background covered by reflections	
Background covered by smoke	
Background subject to motion artifacts	

Table 5.38: List of **global image characteristics** that were identified as failure sources in the Robust Medical Instrument Segmentation (RobustMIS) challenge.

Image characteristic	***Example images***
Image overexposed	
Image underexposed	
Dirty lens	

Metadata annotation A trained engineer annotated the image characteristics given in Tables 5.36 - 5.38 for the test cases of the RobustMIS challenge. The annotation was performed per instrument instance, resulting in a total of 29,718 annotations for all instrument instances of the data set.

Mixed model analysis Given the hierarchical data structure of the RobustMIS challenge, which is illustrated in Figure 5.14(b), we explored a LMM to represent the performance of challenge algorithms for an image as a function of the annotated image characteristics that were identified as potential failure sources. The model then aimed to predict the effect of the image characteristics on the algorithm performance. For the RobustMIS challenge, we chose to model the Multi-Instance Dice Similarity Coefficient (MI DSC) performance of algorithms, which was one of the primary challenge metrics. However, the value range of the MI DSC, $[0, 1]$, violated the necessary assumption that the outcome variable (here: the aggregated MI DSC scores) should follow a normal distribution. Any transformation, such as the *logit* function, did not yield the desired approximate normal distribution of the residuals. We thus reformulated the problem to a binary pixel-level classification problem, for which we could use a GLMM instead.

The MI DSC metric is equal to the per-instance DSC, which can be reformulated in terms of PPV and Sensitivity (see Equation 2.48). We performed our analysis per instrument instance with the new target metrics PPV and Sensitivity. For every pixel, we checked whether it was correctly classified as an instrument based on two perspectives: The True Positive (TP) instances were either related to the reference mask (Sensitivity) or the prediction mask (PPV). More precisely, PPV and Sensitivity were computed for every image $i \in I$, every instrument instance $j \in J$, and every challenge participant $p \in P$, with I, J, and P referring to the total number of images, instrument instances, and challenge participants.

- → The **Sensitivity** per image, instrument, and participant would then determine the probability that a *pixel of the reference mask is present in the prediction mask*.
- → The **PPV** per image, instrument, and participant corresponds to the probability of a *pixel of the predicted mask being present in the reference mask*.

This binary pixel-level classification per instrument instance can be viewed as a Bernoulli experiment, being feasible for GLMMs (see Section 2.2.3 for a detailed description of GLMMs).

Strength-weakness analysis For this model, we used the 17 identified image characteristics as fixed effects. The patient, image, instrument instance, and challenge participant were modeled as random effects. With this setup, we assessed the effect of image characteristics on the global algorithm performance. However, the analysis could similarly be performed for every algorithm individually to analyze the effects of image characteristics on the performance of the single participants. In this case, the MI DSC scores would only be used from the respective participant and the algorithm would be removed from the random effects. Please note that we implemented this model for the top five challenge participants only (teams *fisensee*, *Uniandes*, *caresyntax*, *SQUASH* and *www*). With this subset of algorithms, we aimed to only focus on the most successful algorithms and their failure cases.

To interpret the model outcome, the regression coefficients for the fixed effects, β, can be regarded as the log OR [McCulloch and Searle, 2004], which measures the logarithm of the ratio of odds of the occurrence of the outcome when the image characteristics are present and absent (OR). A positive log OR refers to an increase in the chance that a good performance score is achieved for the respective image characteristics, which can be interpreted as a positive effect on the algorithm performance. On the other hand, a negative log OR refers to a negative effect on the metric value, i.e. the chance is decreased.
This analysis was performed from two perspectives: First, we analyzed the effects of image characteristics globally over the top five challenge participants. In the second step, we repeated the experiment for the individual algorithms to check whether the same trends were observed for individual participants. Figure 5.36 shows the effect of image characteristics on the MI DSC scores in form of the log OR globally for the top five challenge participants. According to the results of the analysis, the following image characteristics yielded a **significant negative effect** on the algorithm performances (sorted by decreasing order of magnitude of negative effects):

1. Instrument covered by another instrument
2. Instrument subject to motion artifacts
3. Underexposed instruments

4. Backgrounds covered by smoke
5. Underexposed images

Significantly positive effects were observed in the case of overexposed instruments or instruments covered by tissue. We found similar effects for the analysis of individual algorithms, as shown in Table 5.39 and described in the following paragraph.

Team *fisensee* was the only challenge participant for which no significant negative effect was found for instruments covered by another instrument and underexposed instruments, but additionally suffered from images in which the background was covered by smoke. On the other hand, this team profited from a background covered with blood. While the background did not yield any significant effects on the global performance, team *Uniandes* was negatively affected if the background was covered by reflections, smoke, or if it was subject to motion artifacts. It was the only team for which a background covered by other objects had a positive effect on the Sensitivity. Images covered by blood, other objects, and smoke in addition to a background that was subject to motion artifacts negatively influenced the results of team *caresyntax*. But similarly to team *fisensee*, this team was positively affected by a background covered by blood. Team *SQUASH* was the only team showing a positive effect from instruments covered by smoke, while team *www* suffered from a background in which smoke was visible.

5.4.4 Discussion

In previous sections, we showed that the lack of reporting is a huge issue for challenges. While the introduction of the challenge submission system and review process (see Section 5.1) was already an important step, we showed in this section that the challenge organizers in practice did not make use of this information source. With the introduction of challenge registration, we followed quality measures from clinical trials to make sure the relevant information on challenge designs was publicly available. Even if challenge organizers decide to publish only a fraction of the parameters on their websites and in the challenge reports, the full information is still accessible to the public. In addition, it is easy to track the changes of a challenge design, similar to a versioning control in software engineering. Changes that were not reasonable will be discussed with the challenge organizers to ensure the high quality of the challenge. Although this process is currently only implemented for MICCAI challenges, we hope to introduce this concept to other conferences and challenge hosts, as well as to enhance the quality of challenges in general. As MICCAI is currently the largest conference in the field of medical image analysis, most of the biomedical image analysis challenges since 2020 have been registered, thus increasing transparency and mitigating the potential danger of challenge manipulation (see Section 4.1).

Challenges are often published in high-ranked journals, thus, it is required that they provide a transparent and complete overview of all facets of their design and results. Reporting guidelines have been used in the context of clinical trials for many years and substantially increased the quality and interpretability of reports [Schulz et al., 2010]. Creating such a guideline for challenges comes with several advantages: First, for the authors of a challenge report, the writing process is facilitated by the guidance of the checklist. In addition, the guideline helps verify whether the included information is complete. This step is useful for both authors and reviewers of the report. Finally, a completed checklist including information on where to find respective details facilitates the retrieval of relevant information and improves the reproducibility of the challenge results.

Table 5.39: Overview of significant positive and negative effects of image characteristics for the Positive Predictive Value (PPV) and Sensitivity. Values are shown for the *global* mixed model analysis and the analysis for the top five algorithms of the Robust Medical Instrument Segmentation (RobustMIS) challenge (*fisensee*, *Uniandes*, *caresyntax*, *SQUASH* and *www*). Positive effects are indicated by a blue cell and a + symbol (+: $p \leq 0.05$, ++: $p \leq 0.01$, +++: $p \leq 0.001$). Negative effects are analogously shown by red cells and a - symbol.

Instrument-related image characteristics

Image characteristic	***PPV***						***Sensitivity***					
	Global	*fisensee*	*Uniandes*	*caresyntax*	*SQUASH*	*www*	*Global*	*fisensee*	*Uniandes*	*caresyntax*	*SQUASH*	*www*
Instrument covered by another instrument	- - -		- - -	- - -	- - -	- - -	- - -		- -	- - -	- -	- - -
Instrument covered by blood										-		
Instrument covered by other object									-	-		
Instrument covered by reflections												
Instrument covered by smoke				-							+	
Instrument covered by tissue	+++	++	++	+++	++	+++	+				+	+
Instrument overexposed	+	+++	+++	++	+++	+						
Instrument subject to motion artifacts	-		- -	-		-	- -	- -	- - -	- -		- -
Instrument underexposed	- - -		- -	- - -	- -	- - -	- - -	- - -	- - -	- -		- - -

Background-related image characteristics

Image characteristic	***PPV***						***Sensitivity***					
	Global	*fisensee*	*Uniandes*	*caresyntax*	*SQUASH*	*www*	*Global*	*fisensee*	*Uniandes*	*caresyntax*	*SQUASH*	*www*
Background covered by blood		+		+						+		
Background covered by other object									+			
Background covered by reflections									- - -			
Background covered by smoke							-	- -	-			-
Background subject to motion artifacts				-					-			

Global image-related image characteristics

Image characteristic	***PPV***						***Sensitivity***					
	Global	*fisensee*	*Uniandes*	*caresyntax*	*SQUASH*	*www*	*Global*	*fisensee*	*Uniandes*	*caresyntax*	*SQUASH*	*www*
Image overexposed												
Image underexposed							-		-			- -
Dirty lens												

The guideline has been used by several challenge organizers so far (e.g. [Oreiller et al., 2022; Timmins et al., 2021; Antonelli/Reinke et al., 2022; Roß/Reinke et al., 2020]). Furthermore, the journal *Medical Image Analysis* requires challenge reports to complete the BIAS checklist along with the submission of the publication. This journal is therefore following the same best practices as done for clinical trials.

Moreover, the guideline encourages challenge organizers to perform an in-depth analysis of the challenge results, which may even go beyond ranking uncertainty analysis (see Section 5.3). Challenges are typically organized for a specific purpose and to solve a research question. An analysis of failure cases informs the research community which problems are already (nearly) solved and which types of problems would need further investigation. We performed such a statistical analysis for the RobustMIS challenge in the form of GLMMs, addressing the hierarchical data structure of the challenge. Although this technique is rather new to many researchers from the Machine Learning (ML) community – less than 2% of MICCAI publications utilized mixed models in the past 18 years – it comes with several advantages over traditional analysis methods. A random forest, for example, would give a ranking of the failure cases without quantifying whether they had a positive or negative effect on the algorithm performance.

A limitation of the failure analysis study is the additional manual annotation effort. Creating high-quality annotations is typically cost- and time-consuming. However, we argue that such an effort substantially helps in driving the research on tackling common failure cases of challenges. By using quality-controlled crowdsourcing [Maier-Hein et al., 2014] or automatic techniques of image characteristic annotation, the annotation effort could be further compensated.

Previous literature in the field of minimally invasive surgery has investigated efforts to manually review failure cases [Maier-Hein et al., 2014; Ali et al., 2021b, 2020, 2021a; Funke et al., 2018; Soberanis-Mukul et al., 2020; Bodenstedt et al., 2018; Allan et al., 2020; Roß/Reinke et al., 2020]. However, our statistical analysis revealed that some of the suggested failure cases actually do not have a statistically significant negative effect on the challenge participants. For example, the mentioned studies disclose the presence of blood and reflections as harmful to the algorithm performance. In contrast, our study showed that blood present in the background of images can even increase the algorithm performance for some of the challenge participants. Our analysis further reveals strong positive effects of instruments that are overexposed or covered by tissue. This is probably due to the fact that instruments better stand out against the background and are more easily recognizable for a ML algorithm.

The strongest negative effect was found for instruments covered by other instruments. Indeed, in those cases, many algorithms interpreted the disconnected instrument as two individual instruments, decreasing the performance. This might be a result of the algorithms' widespread use of the Mask R-CNN architecture (see Section 4.4). As the backbone of the network suggests bounding boxes for relevant regions in an image, overlapping instrument instances may yield multiple suggested bounding boxes. This could further explain why team *fisensee* was the only participant not suffering from this image characteristic, as this team used a different network architecture. To sum up, such a failure analysis could be used for further algorithm development, specifically tailored to the image characteristics that were hard to interpret, as done in Roß et al. [2021].

5.4.5 Conclusion

In this section, we presented the concept of challenge registration in addition to a challenge reporting guideline. Both approaches successfully compensate for problems related to the lack of transparency and manipulation of challenges. In addition, the reporting guideline includes items reflecting our solutions proposed in this part of the thesis. For example, we advice challenge organizers to justify the choice of metrics and rankings, and to apply techniques for ranking uncertainty analyses. Moreover, we introduce a further analysis technique of challenge results by making use of mixed models for failure analysis of challenge algorithms. This kind of investigation can be extremely useful for future algorithm development and is not restricted to challenges only. As we showed, the technique can also be applied to the results of a single algorithm and can thus be used for individual strength-weakness analysis.

6 | Discussion

The three main parts of this thesis contain a detailed discussion of their respective results in each section. The discussions can be found in Section 3.1.4 for Chapter 3 for a review of the status quo. Sections 4.1.4, 4.2.4, 4.3.4, and 4.4.4 discuss the results for Chapter 4 related to flaws of biomedical image analysis validation. Sections 5.1.4, 5.2.4, 5.3.4, and 5.4.4 discuss the findings of Chapter 5 related to improving common practice. In this section, we provide an overarching discussion of the work presented in this thesis and and put the results into a broader context.

Challenge design
Biomedical image analysis challenges have led to a tremendous increase in the comparability of algorithms. However, challenges are very heterogeneous in their design, raising the question of whether challenge results are actually meaningful for practical applications. Moreover, the community agrees that challenges should undergo more quality control. We demonstrated that challenges can easily be manipulated if the design has severe issues. Since their implementation, our structured challenge submission system and suggested peer-review process have therefore ensured that only high-quality challenges were accepted for prominent conferences. Of note, although researchers openly welcomed our findings, as indicated by the invitations to multiple talks and respective feedback, the research practice did not directly improve substantially. As shown in Section 5.4, simply raising awareness of the problems and even implementing a structured challenge submission system were not enough to break standard practices, suggesting that even more quality-control measures were required. Indeed, only through combining all the contributions presented in this thesis could improvements be enforced.

Metrics
In Section 4.2, we provided evidence that validation metrics come with severe limitations. We showed that the meaningfulness of challenge results and validation in general suffers from the common practice of not taking these limitations into account, and, indeed, selecting metrics based on their popularity rather than their suitability to the underlying clinical or biological problem. In a recent study [Tran et al., 2022], we further demonstrated that even conventional hyperparameter settings, such as the selection of localization thresholds for the computation of the Average Precision (AP) metric, should not indiscriminately be transferred to any problem since they may not always result in clinically useful outcomes. Obtaining a comprehensive overview of metric-related pitfalls is very difficult, which may be one of the reasons why many researchers pick metrics based on related literature rather than considering their own specific problems and

data sets. The complexity of metric selection and its dependence on many variables may be another reason for the current poor selection practice. To empower researchers to select adequate metrics, we first presented an extensive overview of metric-related pitfalls, which served as the basis for the subsequent problem-aware metric recommendation framework. With our framework, we hope to steer the research community toward a more problem-aware validation practice.

It must be noted that the vast amount of knowledge incorporated in the framework resulted in a complex structure of the recommendations that may not be easy to follow. Before finalizing the framework, we gathered feedback from the general public, and its complexity garnered the most criticism. However, given the complexity of the problem, we could not reduce it substantially, since this would have implied an incomplete view to the problem. To reduce the complexity for the end users and ensure that it can easily be adopted by the research community, we are currently working on a web-based toolkit. The toolkit could also overcome another potential limitation of the framework: The individual metric mappings directly show all metric candidates at a glance. This may potentially bias researchers towards answering the questions of the problem fingerprint such that a certain, desired metric is selected, rather than answering them impartially. The toolkit will present questions in the context of the biomedical research question, without explicitly showing which choices lead to which metrics at the first glance.

Rankings

We demonstrated how the choice of metrics can substantially impact challenge rankings. Rankings are also influenced by the strategy of aggregation, aggregation operator as well as the underlying data and annotators. Their high sensitivity to the chosen ranking scheme makes them particularly susceptible to manipulation. A sophisticated ranking scheme and analysis are extremely important to obtain meaningful results and thus ensure that the winning algorithm is truly superior. These conclusions are also important for general algorithm validation and benchmarking. Researchers are typically expected to compare the results of their methods against state-of-the-art algorithms. If rankings can easily be tuned and no further analysis is provided, both meaning and fairness of the results are endangered. We therefore recommended using advanced visualization and analysis tools to better understand the distribution of algorithm performance and the stability of rankings, and created an open-source toolkit to this end. Future work may include considerations of how to define a ranking scheme for a given problem to reflect the domain interest., for example by searching for a pareto-optimal ranking scheme.

A limitation of the presented toolkit lies in the fact that it was designed for a per-case- or per-patient validation, meaning that at least one metric value is expected for every case or patient. This means that per-data set validation, such as typically done in image-level classification or object detection, is not directly supported. However, for such validation schemes, a ranking per class is directly implied from the metric calculation since it yields a single score per data set per metric. For multiple classes, the tool could be adjusted to generate a ranking over classes. Furthermore, the bootstrapping and uncertainty techniques implemented in the toolkit can be used also for per-data set validation problems. It should be noted that the toolkit is currently implemented in the `R` language, which is more prominent in the statistical research community, but rather unpopular among biomedical image analysis data scientists. We are therefore planning to also extend it to the `Python` programming language in the future.

Reporting and analyses
A transparent reporting of the challenge design and results is one of the most important steps for good scientific practice [Altman et al., 2008]. However, neither the challenge design nor the participating algorithms are typically reported in sufficient detail, with parameters that are important for meaningful interpretation of the challenge often missing. If important parameters, such as the inter-rater variability, are not provided, it is difficult to assess whether the chosen metrics or ranking schemes are appropriate. Moreover, we demonstrated that it is often not possible to reimplement the challenge algorithms, although the purpose of challenges is to find a methodology that succeeds in solving a specific research question. The scientific community is naturally interested in adopting the new state of the art for future research, and understandably frustrated if models cannot be reproduced. To overcome this problem, we proposed a reporting guideline and the concept of challenge registration, following best practice principles from clinical trials. This step facilitates the information exchange between the challenge organizers and other researchers and also helps in reporting general algorithm validation. In addition, we showed that we can learn even more from challenges by performing a statistics-based strength-weakness analysis of algorithms. We think that this step further increases the impact of challenges as it becomes immediately evident which aspects of a research problem are sufficiently solved and which aspects should be further analyzed in future algorithm development.

Of note, while the proposed solutions substantially improved the overall quality of challenges, we still cannot fully enforce their usage. The concept of challenge registration, for example, is currently only enforced for Medical Image Computing and Computer Assisted Interventions (MICCAI) challenges. While MICCAI typically comprises the largest amount of challenges in the biomedical image analysis domain, other communities did not adopt this concept yet. Although we are actively working towards broadening the scope of this concept to include other well-known stakeholders like IEEE International Symposium on Biomedical Imaging (ISBI) and Dialogue on Reverse Engineering Assessment and Methods (DREAM), we have not yet reached all potential challenge hosts. This is especially difficult for large platforms like Kaggle, which do yet not have a comprehensive quality control process for challenges. Additionally, although it was made clear to challenge organizers that modifications to their challenge design must be submitted again and discussed with the respective challenge chairs so that the changes are reflected in the versioning of the uploaded design papers, very few researchers actually adhered to this restriction. We are currently lacking staff resources to enforce the new practices in a strict manner. However, since researchers are becoming more and more aware of the aforementioned issues, the MICCAI Special Interest Group for Challenges is currently discussing possibilities to address this issue, for example by gathering funding for specifically working towards a stricter and monitored quality control mechanism.

Data and annotations
One of the key conclusions of the presented work is the formulation of a meaningful, precise, and feasible challenge objective. Challenge results are only interpretable and useful for further research if the challenge goal and design are clear and sound. Our work focused on the design, validation, reporting, and analysis of challenge results. However, we would also like to draw attention to the importance of challenge data. As mentioned before, the data impacts the metrics and challenge rankings. Therefore, data sets need to be created carefully. For example, a data set should be realistic and ideally drawn from several institutions and different scanners to reflect a broad range of imaging situations. However, determining the optimal sample size

for a data set [Jain and Chandrasekaran, 1982; Raudys et al., 1991; Kalayeh and Landgrebe, 1983; Bonett, 2002; Shoukri et al., 2004; Sim and Wright, 2005] or avoiding biases [Torralba and Efros, 2011; Zendel et al., 2017; Deng et al., 2009; Everingham et al., 2015] is often not straightforward.

Related to data is the quality of the reference annotations. The algorithms can only be as good as the data they are learning from. We should therefore make sure that the annotations are generated from multiple annotators to decrease the effects of inter-rater variability [Joskowicz et al., 2019]. Furthermore, in Rädsch et al. [2022], we showed that the annotation protocol should be thoroughly designed, should explain edge cases, and contain several illustrative examples. We notably showed that annotation protocols only containing little information yield low-quality annotations since ambiguities are not resolved for annotators. Annotations should be generated by domain experts or professional annotators [Rädsch et al., 2022; Wang et al., 2015; Bernard et al., 2015]. However, the creation of high-quality labels is costly, both in terms of money and resources. Offering incentives for high-quality data creation and acquisition could be one possible approach [Von Ahn and Dabbish, 2004]. Another approach to enhance the data set quality was chosen by the MICCAI Kidney Tumor Segmentation (KiTS) challenge [Heller et al., 2019]. After the organizers released the training data set, they specifically asked participants and the community for feedback on the data. The raised issues were then resolved in a second iteration of the training data and incorporated into the test set as well. This practice could even be used for other parts of the challenge design, such as metrics or ranking schemes. Furthermore, researchers are becoming more and more aware of the importance of high quality data. Journals have followed this trend by specifically accepting the submission of data set publications. For example, we published the data set related to the Robust Medical Instrument Segmentation (RobustMIS) challenge in the Nature Scientific Data journal [Maier-Hein et al., 2021], which specifically asked for concrete quality measures such as a technical validation of the data set. With this approach, published data sets can be assured to be of higher quality.

Collaborative challenges

The reasoning behind organizing challenges is to examine a specific research problem. This raises the question of whether we should continue to put focus on winning challenges and presenting rankings. Collaborative challenges pose an alternative, in which teams work together by combining their methods and expertise to find a solution. This practice has been successfully used in the mathematical community [Išgum et al., 2015; Barnes et al., 2017]. At MICCAI, the Cardiac Resynchronization Therapy Electrophysiological (CRT-EPiggy) challenge [Camara, 2019] began as a competitive challenge, but the organizers soon realized that collaboration between teams yielded better results. However, the incentives for participating in a challenge may decrease if no winner is announced. A future research direction may be searching for incentives to participate in collaborative challenges [Barnes et al., 2017; Peng et al., 2015]. In their analysis of Kaggle competitions, Tauchert et al. [2020] have shown that the award money from featured challenges is seven times higher than award money from research. Increasing funding and sponsoring possibilities for challenges could thus be a viable future direction to enhance scientific progress.

Future work

Although we presented several methods that already heavily improved the impact and quality of biomedical image analysis challenges, there is still the potential for more improvements and intriguing research questions remain unanswered. One such question of importance for future research is that of why a challenge participant won a challenge. Isensee et al. [2021], for instance,

reviewed the roughly one hundred algorithms that participated in the KiTS challenge. The U-Net architecture was used among the top teams but was also spread across the entire leaderboard. The authors could not determine whether a specific modification of the network architecture succeeded over others. This raises the question of whether the network architecture is the most important factor for winning a challenge or whether other aspects such as the data augmentation strategy or ensembling are more relevant. In a recent study[1], we initiated an international survey that was distributed to all MICCAI and ISBI challenges, asking participants about their methods, incentives, key design choices, and general challenge strategies. In this way, we hope to analyse the differences between winning and non-winning challenge participants. So far, for example, we found that challenge winners rated analyzing failure cases and knowing the state-of-the-art methods in the field as crucially important to win a challenge.

Another future direction lies in the adaptation of the concept of clinical trial certification to challenges. For example, accepted randomized and double-blinded prospective trials promise a high quality due to the study design. Translated to challenge design, a certified challenge would imply a specific degree of quality, such as complete reporting, metrics chosen according to our recommendation framework, or data sets reviewed for their quality.

Finally, in this thesis, we adopted many concepts from clinical trials, but we could also analyze practices from other communities. For example, in the Natural Language Processing (NLP) domain, a recently published dynamic benchmarking website [Kiela et al., 2021] offers a dynamic collection of data applied to the models. The aim is working towards the question of "how well (...) Artificial Intelligence (AI) systems perform when interacting with humans" [Kiela et al., 2021]. This research question would also be relevant for biomedical challenges to assess the performance of algorithms in a simulation of real-world usage in clinical practice.

[1] work in progress, thus not yet published

7 Summary of Contributions and Conclusion

In this thesis, several aspects of biomedical image analysis challenges and the validation of Artificial Intelligence (AI) algorithms were analyzed. Overall, this thesis resulted in

- **10 first-author publications** (2 in peer-reviewed journals; 1 full paper at a peer-reviewed conference; 3 abstracts/short papers at conference proceedings or conference workshops; 2 preprints submitted to peer-reviewed journals; 2 invited commentaries/mini reviews),
- **5 second-author publications** (3 in peer-reviewed journals; 2 submitted to peer-reviewed conference proceedings),
- **15 invited talks at major international events**,
- **5 awards**,
- **3 memberships of recognized international consortia**, and
- **3 invitations for being a challenges (co-) chair at internationally recognized conferences**,

as detailed in the following Section 7.1. I subsequently give an overview of my publications in Section 7.2 and end with a conclusion to the thesis in Section 7.3.

7.1 Summary of Contributions

Contribution 1: Comprehensive and structured analysis of challenges
In this thesis, the first comprehensive and structured analysis of more than 500 challenge tasks was performed. During our analysis, it was found that challenges are the state-of-the-art technique for the assessment of AI algorithms in the biomedical image analysis community and that challenges are available for a variety of problem categories and applications. However, it was also discovered that the design of challenges was heterogeneous and not standardized. In addition, the general community was concerned about challenge quality and asked for more quality control and recommendations.

Contribution 1 has led to three publications [Maier-Hein et al., 2018; Reinke et al., 2021c; Bron et al., 2021]. Furthermore, two colleagues and I were invited to give talks at the Endoscopic Vision

Challenge and the Tutorial on Designing Benchmarks and Challenges for Measuring Algorithm Performance in Biomedical Image Analysis at the Medical Image Computing and Computer Assisted Interventions (MICCAI) 2017 conference.

> **HYPOTHESIS 1: Current biomedical challenge design is heavily flawed**
>
> **SPECIFIC FINDINGS**
>
> **Specific Finding 1.1:** Security holes in challenge design can easily be exploited.
>
> **PROPOSED SOLUTIONS**
>
> **Solution 1.1:** Structured challenge submission system and introduction of challenge peer review process.
>
> **Solution 1.2:** New challenge design and reporting guideline.

Contribution 2: Revelation of flaws in challenge design
It was shown that weaknesses in the challenge design may, in theory, easily be exploited by researchers. Due to ranking sensitivity and poor reporting practices, rankings could be manipulated by challenge organizers to change the challenge winners. On the other hand, security flaws could also be exploited by challenge participants via selective test case submission.

Contribution 3: Structured challenge submission system
A structured challenge submission system was proposed, forcing challenge organizers to enter complete information about their challenge designs prior to submitting their challenges to a conference. In addition, a standard peer review process for challenge design was introduced to ensure that only high-quality challenges are accepted to large conferences, such as MICCAI.

Contributions 2 and 3 have led to a publication that I presented as an oral at the MICCAI conference in 2018 [Reinke et al., 2018a], which was highlighted in the *Computer Vision News* for MICCAI 2018[1], and several publications of challenge reports based on this structured submission system [Antonelli/Reinke et al., 2022; Roß/Reinke et al., 2020; Zimmerer et al., 2022; Pati et al., 2021; Wagner et al., 2021]. Furthermore, I was invited to write a review of an epilepsy challenge [Reinke, 2021], give a keynote at the MICCAI 2019 Workshop on Large-Scale Annotation of Biomedical Data and Expert Label Synthesis, and give a talk at the seminar on challenges of the African Network for Artificial Intelligence in Biomedical Imaging (AFRICAI) in 2022. Moreover, I won the third place in the MICCAI 2020 Educational Challenge together with colleagues on how to organize a challenge. Since 2019, I am a member of the organizing committee of the MICCAI Endoscopic Vision Challenge. In addition, at MICCAI 2020, I was one of the conference challenge co-chairs and the challenges chair for MICCAI 2021 and Medical Imaging with Deep Learning conference (MIDL) 2020.

[1] https://www.rsipvision.com/ComputerVisionNews-2018October/45/

Hypothesis 2: Common image analysis metrics do not reflect the biomedical domain interest

Specific findings

Specific Finding 2.1: Information on metric pitfalls is currently inaccessible.

Specific Finding 2.2: Metrics are widely chosen based on popularity rather than their clinical suitability, ignoring their pitfalls.

Proposed solutions

Solution 2.1: Metric recommendation framework.

Contribution 4: Revelation of metric-related pitfalls in biomedical image analysis
We were the first to comprehensively present metric-related pitfalls in the context of image-level classification, semantic segmentation, object detection, and instance segmentation problems. Knowing about metric limitations is crucial for adequate metric selection and meaningful validation of algorithm performance. However, many researchers do not select suitable metrics for their specific research problems.

Contribution 5: Problem-aware metric recommendation framework
We therefore proposed our metric recommendation framework that suggests appropriate metric candidates for the specific research problem based on a problem fingerprint that abstracts from the underlying domain. Problem-aware decision guides assist researchers in the process. We believe that this framework will improve the overall quality of overall image analysis in the community by ensuring that only appropriate metrics are applied.

Contributions 4 and 5 resulted in the acceptance of an abstract submitted to the Medical Imaging Meets Neural Information Processing Systems (NeurIPS) workshop 2022 and two short papers at the MIDL conference [Reinke et al., 2021b, 2022a], one of which was highlighted in *MIDL Vision 2022*[2]. For the 2021 short paper, I won the *best short oral presentation* award and was asked to moderate the MIDL 2021 workshop on "how to write an award-winning paper". In addition, we submitted two papers to Nature Methods [Reinke et al., 2021a; Maier-Hein et al., 2022]. Reinke et al. [2021a] was presented as a picture story and already received 60 citations in roughly 1.5 years. We have written a blogpost on metric-related pitfalls[3] for the MICCAI 2021 Educational Challenge, for which I won the first prize. Furthermore, based on this work, I was invited to give several talks, namely for the Medical Open Network for Artificial Intelligence (MONAI) 2021 bootcamp, the 2022 International Conference on Medical Imaging and Case Reports (MICR), the 2022 annual meeting of the Society for Imaging Informatics in Medicine (SIIM), the 2022 workshop on Responsible Machine Learning in Healthcare, the 2022 Deutsches Krebsforschungszentrum (German Cancer Research Center) (DKFZ) data science seminar. Finally, I was invited to give keynotes and talks for the Informatics for Life event 2022, an AFRICAI event, and the Workshop for Effectively Communicating Bioimage Analysis 2024.

[2] https://www.rsipvision.com/MIDL2022/16/

[3] https://medium.com/miccai-educational-initiative/a-discovery-dive-into-the-world-of-evaluation-dos-don-ts-and-other-considerations-4189ab46fe06

Hypothesis 3: Challenge rankings are highly unstable

Specific findings

Specific Finding 3.1: Rankings are highly sensitive to various design parameters.

Proposed solutions

Solution 3.1: Methods and toolkit for analyzing and visualizing benchmarking results.

Contribution 6: Revelation of ranking uncertainty in challenges
It was demonstrated that a variety of distinct design choices, such as the selected metric, the aggregation method, or the annotator, had a substantial impact on challenge ranks. Thus, rankings are often very unstable, yielding inconsistent winners.

Contribution 7: Open-source toolkit for advanced ranking analysis and visualization
To compensate for this issue, several advanced visualization techniques and methods for analyzing the ranking uncertainty were presented, which were implemented in an open-source toolkit for analyzing challenges and benchmarking results. We further showed how these techniques enhance the interpretability and the value of challenges.

Contribution 6 finding was part of the previously mentioned study [Maier-Hein et al., 2018]. In addition, we published Contribution 7 [Wiesenfarth et al., 2021]. Based on our findings, I was invited to give two talks, at the 2020 European Lab for Learning & Intelligent Systems (ELLIS) Health Workshop and at the weiss AI in Surgery mini symposium. I furthermore was awarded the best Ph.D. talk presentation at the Ph.D. retreat 2021 of the DKFZ International Ph.D. Program in Heidelberg (Helmholtz International Graduate School for Cancer Research).

Hypothesis 4: Challenge results are not reproducible

Specific findings

Specific Finding 4.1: Challenges are not reported comprehensively and are not reproducible.

Specific Finding 4.2: Challenge algorithms are not reproducible.

Proposed solutions

Solution 4.1: New challenge reporting guideline.

Solution 4.2: Challenge registration.

Solution 4.3: Challenge algorithm strength-weakness analysis.

Contribution 8: Revelation of non-reproducibility of challenge results
Finally, we uncovered that both challenge design and participating algorithms are often not reproducible. Challenge design is often not reported comprehensively. In addition, method descriptions provided by challenge participants are frequently insufficient to reproduce the results without the actual source code being publicly available. This constitutes extremely poor research practice and may lead to severe consequences because it facilitates cheating.

Contribution 9: Challenge reporting guideline and registration
We thus implemented a reporting guideline for biomedical image analysis challenges to enhance the transparency of challenge reports. It was complemented by introducing the concept of challenge registration, for which the complete challenge design that was inserted in the structured challenge submission system was made publicly available. This step ensures that the respective information of the challenge design is already disposable upon and prior to challenge execution.

Contribution 10: Algorithm strength-weakness analysis
Finally, we showed that challenge result analysis can be enhanced by strength-weakness analyses of the participating algorithms. Such an analysis could subsequently be used for future algorithm development to specifically address problematic image characteristics in new or adapted models.

Contribution 8 was part of the previously mentioned study [Maier-Hein et al., 2018]. Our reporting guideline from Contribution 9 was published [Maier-Hein et al., 2020] and registered with the Enhancing the QUAlity and Transparency Of health Research (EQUATOR) network[4] [Altman et al., 2008]. We further submitted our statistical strength-weakness analysis from Contribution 10 to Medical Image Analysis [Roß et al., 2021]. Based on our work, a colleague and I were invited to give a talk at the 2020 Workshop on Data Curation, Standardisation, and Algorithm Benchmarking. In addition, I was one of the founding members of the Biomedical Image Analysis ChallengeS (BIAS) initiative and currently serve as an active member of the MONAI Working Group Evaluation, Reproducibility, Benchmarks, and its benchmarking task force lead. Furthermore, I was an active member of the MICCAI Board Challenge Working Group and was elected as the secretary of the MICCAI Special Interest Group (SIG) for Challenges. Finally, together with a colleague, I won the 3rd price of the 2022 Women in MICCAI Inspirational Leadership Legacy interview award.

7.2 Own publications

This section lists all papers I (co-)authored. The first part of the list contains first authorships, including peer-reviewed journal publications and conference proceedings (including short papers and abstracts, as indicated in the respective publication), and other types of publication (preprints and invited mini reviews). In addition, co-authorships are listed, containing peer-reviewed journal publications, and other types of publication (preprints). The first author(s) for every paper is/are underlined, while my name is emphasized by using bold face font. In some of the publications, the first authorship was shared, indicated by two authors being underlined and being marked with an asterisk. In the case of preprints, the journal to which it was submitted to is mentioned.

[4] https://www.equator-network.org/reporting-guidelines/bias-transparent-reporting-of-biomedical-image-analysis-challenges/

First authorships – peer-reviewed journal publications and conference proceedings

Annika Reinke[*] /Michela Antonelli[*] et al. *The Medical Segmentation Decathlon*. ***Nature Communications*** 13(1): 1–13, 2022.

Annika Reinke[*] /Tobias Ross[*] et al. *Comparative validation of multi-instance instrument segmentation in endoscopy: results of the ROBUST-MIS 2019 challenge*. ***Medical Image Analysis***, 101920, 2020.

Annika Reinke[*] /Matthias Eisenmann[*] et al. *How to exploit weaknesses in biomedical challenge design and organization*. In ***International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)***, 388–395. Springer, 2018.

Annika Reinke[*] /Lena Maier-Hein[*] et al. *Metrics Reloaded*. In ***Medical Imaging meets NeurIPS***, 2022.

Annika Reinke[*] /Lena Maier-Hein[*] et al. *Metrics Reloaded – A new recommendation framework for biomedical image analysis validation*. In ***Medical Imaging with Deep Learning (MIDL, Short paper)***, 2022.

Annika Reinke[*] et al. *Common limitations of performance metrics in biomedical image analysis*. In ***Medical Imaging with Deep Learning (MIDL, Short paper)***, 2021.

First authorships – other

Annika Reinke[*] /Lena Maier-Hein[*] et al. *Metrics Reloaded: Pitfalls and recommendations for image analysis validation*. arXiv preprint arXiv:2104.05642, 2022. ***Submitted to Nature Methods with a positive Pre-submission Request.***

Annika Reinke et al. *Common limitations of image processing metrics: A picture story*. arXiv preprint arXiv:2104.05642, 2021. ***Submitted to Nature Methods.***

Annika Reinke. *Bring the model to the data: The Deep Learning Epilepsy Detection Challenge*. ***EBioMedicine***, 66, 2021.

Annika Reinke et al. *Common Pitfalls and Recommendations for Grand Challenges in Medical Artificial Intelligence*. ***European Urology Focus***, 2021.

Co-Authorships – peer-reviewed journal publications

Esther E Bron, Stefan Klein, **Annika Reinke** et al. *Ten years of image analysis and machine learning competitions in dementia*. ***NeuroImage***, 119083, 2022.

David Zimmerer, Peter M Full, Fabian Isensee, Paul Jäger, Tim Adler, Jens Petersen, Gregor Köhler, Tobias Ross, **Annika Reinke** et al. *MOOD 2020: A public Benchmark for Out-of-Distribution Detection and Localization on medical Images.* ***IEEE Transactions on Medical Imaging***, 2022.

Melanie Schellenberg, Kris K Dreher, Niklas Holzwarth, Fabian Isensee, **Annika Reinke** et al. *Semantic segmentation of multispectral photoacoustic images using deep learning*. ***Photoacoustics***, 100341, 2022.

[*]Shared first authors

Manuel Wiesenfarth, **Annika Reinke** et al. *Methods and open-source toolkit for analyzing and visualizing challenge results*. ***Scientific Reports***, 11(1):1—15, 2021.

Lena Maier-Hein, **Annika Reinke** et al. *BIAS: Transparent reporting of biomedical image analysis challenges*. Medical Image Analysis, 66:101796, 2020.

Lena Maier-Hein*, Martin Wagner*, Tobias Ross, **Annika Reinke** et al. *Heidelberg Colorectal Data Set for Surgical Data Science in the Sensor Operating Room*. ***Nature Scientific Data*** 8(1):1–11, 2021.

Lena Maier-Hein*, Matthias Eisenmann*, **Annika Reinke** et al. *Why rankings of biomedical image analysis competitions should be interpreted with care*. ***Nature Communications***, 9(1):1–13, 2018.

Co-Authorships – other

Tim Rädsch, **Annika Reinke** et al. *Labeling instructions matter in biomedical image analysis*. arXiv preprint arXiv:2207.09899, 2022. ***Submitted major revision to Nature Machine Intelligence.***

Thuy Nuong Tran, Tim Adler, Amine Yamlahi, Evangelia Christodoulou, Patrick Godau, **Annika Reinke** et al. *Sources of performance variability in deep learning-based polyp detection*. arXiv preprint arXiv:2211.09708, 2022. ***Submitted to the Information Processing in Computer-Assisted Interventions (IPCAI) 2023 conference.***

Tobias Roß*, Pierangela Bruno*, **Annika Reinke** et al. *How can we learn (more) from challenges? A statistical approach to driving future algorithm development*. arXiv preprint arXiv:2106.09302, 2021. ***Submitted major revision to Medical Image Analysis.***

Martin Wagner, Beat-Peter Müller-Stich, Anna Kisilenko, Duc Tran, Patrick Heger, Lars Mündermann, David M Lubotsky, Benjamin Müller, Tornike Davitashvili, Manuela Capek, **Annika Reinke** et al. *Comparative Validation of Machine Learning Algorithms for Surgical Workflow and Skill Analysis with the HeiChole Benchmark*. arXiv preprint arXiv:2109.14956, 2021. ***Submitted major revision to Medical Image Analysis.***

Sarthak Pati, Ujjwal Baid, Maximilian Zenk, Brandon Edwards, Micah Sheller, G Anthony Reina, Patrick Foley, Alexey Gruzdev, Jason Martin, Shadi Albarqouni, Yong Chen, Russell, Taki Shinohara, **Annika Reinke** et al. *The Federated Tumor Segmentation (FeTS) Challenge*. ***arXiv preprint*** arXiv:2105.05874, 2021.

*Shared first authors

7.3 Conclusion

By systematically assessing the suitability of AI algorithms in solving a particular research problem in a comparative manner as well as providing new data sets for benchmarking and biomedical image analysis challenges can offer a powerful tool that drives future research. However, as we uncovered in this thesis, their potential is substantially diminished by numerous fundamental flaws in challenge design, metrics, rankings, and reporting. Importantly, these flaws also crucially impact benchmarking and validation in biomedical image analysis and even in computer vision in general. To address these issues, we introduced a range of solutions that have already begun to improve the interpretability and quality of challenges conducted within the scope of major conferences in the field. These include a structured challenge submission system, which was used for MICCAI, IEEE International Symposium on Biomedical Imaging (ISBI), and MIDL challenges since 2018, a challenge reporting guideline, which the Medical Image Analysis journal requires to be followed for the publication of challenge reports, and the concept of challenge registration, applied to MICCAI challenges since 2020. Furthermore, our uncertainty-aware ranking analysis toolkit was already used by several challenges and benchmarking studies and substantially increased interpretability of results. Finally, we introduced a problem-aware metrics recommendation framework, which is currently transferred into an online toolkit to ensure wide applicability of the framework towards high-quality and problem-driven metric selection.

Armato et al. [2020] argue that a biomedical challenge "is an academic exercise that, although undeniably important, does not directly advance the field." This assessment may be true in the case of low-quality challenges with severe flaws in the design. However, if organized with a clear, transparent, and meaningful objective based on which the entire design, validation, and reporting pipeline is built, challenges do have the potential to generate substantial scientific progress in the research field and lead to tangible practical advancements. They present unique opportunities for us to gauge the abilities, strengths, and weaknesses of the current state of the art, and use these findings to tailor future algorithm development. In uncovering and solving critical issues regarding their transparency, reliability, and robustness, the work presented in this thesis ultimately enables a higher level of trust to be placed into challenges, and thus advances their significance in the comprehensive validation of AI algorithms to be used for clinical practice.

Bibliography

Sharib Ali, Felix Zhou, Barbara Braden, Adam Bailey, Suhui Yang, Guanju Cheng, Pengyi Zhang, Xiaoqiong Li, Maxime Kayser, Roger D Soberanis-Mukul, et al. **An objective comparison of detection and segmentation algorithms for artefacts in clinical endoscopy**. *Scientific reports*, 10(1):1–15, 2020.

Sharib Ali, Mariia Dmitrieva, Noha Ghatwary, Sophia Bano, Gorkem Polat, Alptekin Temizel, Adrian Krenzer, Amar Hekalo, Yun Bo Guo, Bogdan Matuszewski, et al. **Deep learning for detection and segmentation of artefact and disease instances in gastrointestinal endoscopy**. *Medical Image Analysis*, page 102002, 2021a.

Sharib Ali, Felix Zhou, Adam Bailey, Barbara Braden, James E East, Xin Lu, and Jens Rittscher. **A deep learning framework for quality assessment and restoration in video endoscopy**. *Medical Image Analysis*, 68: 101900, 2021b.

Max Allan, Satoshi Kondo, Sebastian Bodenstedt, Stefan Leger, Rahim Kadkhodamohammadi, Imanol Luengo, Felix Fuentes, Evangello Flouty, Ahmed Mohammed, Marius Pedersen, et al. **2018 Robotic Scene Segmentation Challenge**. *arXiv preprint arXiv:2001.11190*, 2020.

Max Allan, Jonathan Mcleod, Cong Cong Wang, Jean Claude Rosenthal, Ke Xue Fu, Trevor Zeffiro, Wenyao Xia, Zhu Zhanshi, Huoling Luo, Xiran Zhang, et al. **Stereo Correspondence and Reconstruction of Endoscopic Data Challenge**. *arXiv preprint arXiv:2101.01133*, 2021.

Douglas G Altman, Iveta Simera, John Hoey, David Moher, and Ken Schulz. **EQUATOR: reporting guidelines for health research**. *Open Medicine*, 2(2):e49, 2008.

Shadi AlZu'bi, Mohammed Shehab, Mahmoud Al-Ayyoub, Yaser Jararweh, and Brij Gupta. **Parallel implementation for 3d medical volume fuzzy segmentation**. *Pattern Recognition Letters*, 130:312–318, 2020.

Vincent Andrearczyk, Valentin Oreiller, Sarah Boughdad, Catherine Cheze Le Rest, Hesham Elhalawani, Mario Jreige, John O Prior, Martin Vallières, Dimitris Visvikis, Mathieu Hatt, et al. **Overview of the HECKTOR challenge at MICCAI 2021: automatic head and neck tumor segmentation and outcome prediction in PET/CT images**. In *3D Head and Neck Tumor Segmentation in PET/CT Challenge*, pages 1–37. Springer, 2021.

Michela/Annika Antonelli/Reinke, Spyridon Bakas, Keyvan Farahani, Annette Kopp-Schneider, Bennett A Landman, Geert Litjens, Bjoern Menze, Olaf Ronneberger, Ronald M Summers, et al. **The medical segmentation decathlon**. *Nature Communications*, 13(1):1–13, 2022. Shared first authors: Michela Antonelli, Annika Reinke.

Guilherme Aresta, Teresa Araújo, Scotty Kwok, Sai Saketh Chennamsetty, Mohammed Safwan, Varghese Alex, Bahram Marami, Marcel Prastawa, Monica Chan, Michael Donovan, et al. **Bach: Grand challenge on breast cancer histology images**. *Medical image analysis*, 56:122–139, 2019.

Ignacio Arganda-Carreras, Srinivas C Turaga, Daniel R Berger, Dan Cireşan, Alessandro Giusti, Luca M Gambardella, Jürgen Schmidhuber, Dmitry Laptev, Sarvesh Dwivedi, Joachim M Buhmann, et al. **Crowdsourcing the creation of image segmentation algorithms for connectomics**. *Frontiers in neuroanatomy*, page 142, 2015.

Samuel G Armato, Keyvan Farahani, and Habib Zaidi. **Biomedical image analysis challenges should be considered as an academic exercise, not an instrument that will move the field forward in a real, practical way**. *Medical Physics*, 47(6):2325–2328, 2020.

Saeid Asgari Taghanaki, Kumar Abhishek, Joseph Paul Cohen, Julien Cohen-Adad, and Ghassan Hamarneh. **Deep semantic segmentation of natural and medical images: a review**. *Artificial Intelligence Review*, 54 (1):137–178, 2021.

John Attia. **Moving beyond sensitivity and specificity: using likelihood ratios to help interpret diagnostic tests**. *Australian prescriber*, 26(5):111–113, 2003.

Marc Aubreville, Nikolas Stathonikos, Christof A Bertram, Robert Klopleisch, Natalie ter Hoeve, Francesco Ciompi, Frauke Wilm, Christian Marzahl, Taryn A Donovan, Andreas Maier, et al. **Mitosis domain generalization in histopathology images–The MIDOG challenge**. *arXiv preprint arXiv:2204.03742*, 2022.

Leonardo Ayala, Fabian Isensee, Sebastian J Wirkert, Anant S Vemuri, Klaus H Maier-Hein, Baowei Fei, and Lena Maier-Hein. **Band selection for oxygenation estimation with multispectral/hyperspectral imaging**. *Biomedical Optics Express*, 13(3):1224–1242, 2022.

Min Bai and Raquel Urtasun. **Deep watershed transform for instance segmentation**. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5221–5229, 2017.

Spyridon Bakas, Mauricio Reyes, Andras Jakab, Stefan Bauer, Markus Rempfler, Alessandro Crimi, Russell Takeshi Shinohara, Christoph Berger, Sung Min Ha, Martin Rozycki, et al. **Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the BRATS challenge**. *arXiv preprint arXiv:1811.02629*, 2018.

Monya Baker. **Reproducibility crisis**. *Nature*, 533(26):353–66, 2016.

D Bamira and MH Picard. **Imaging: Echocardiology—Assessment of Cardiac Structure and Function**. *Elsevier*, 2018.

Andriy I Bandos, Howard E Rockette, Tao Song, and David Gur. **Area under the free-response ROC curve (FROC) and a related summary index**. *Biometrics*, 65(1):247–256, 2009.

David Barnes, Trena Wilkerson, and Michelle Stephan. **Contributing to the development of grand challenges in maths education**. In *Proceedings of the 13th International Congress on Mathematical Education*, pages 703–704. Springer, Cham, 2017.

Christoph Baur, Stefan Denner, Benedikt Wiestler, Nassir Navab, and Shadi Albarqouni. **Autoencoders for unsupervised anomaly segmentation in brain MR images: a comparative study**. *Medical Image Analysis*, 69:101952, 2021.

Colin Begg, Mildred Cho, Susan Eastwood, Richard Horton, David Moher, Ingram Olkin, Roy Pitkin, Drummond Rennie, Kenneth F Schulz, David Simel, et al. **Improving the quality of reporting of randomized controlled trials: the CONSORT statement**. *Jama*, 276(8):637–639, 1996.

Babak Ehteshami Bejnordi, Mitko Veta, Paul Johannes Van Diest, Bram Van Ginneken, Nico Karssemeijer, Geert Litjens, Jeroen AWM Van Der Laak, Meyke Hermsen, Quirine F Manson, Maschenka Balkenhol, et al. **Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer**. *Jama*, 318(22):2199–2210, 2017.

Miroslav Beneš and Barbara Zitová. **Performance evaluation of image segmentation algorithms on microscopic image data**. *Journal of microscopy*, 257(1):65–85, 2015.

Jorge Bernal, Aymeric Histace, Marc Masana, Quentin Angermann, Cristina Sánchez-Montes, Cristina Rodríguez de Miguel, Maroua Hammami, Ana García-Rodríguez, Henry Córdova, Olivier Romain, et al. **GTCreator: a flexible annotation tool for image-based datasets**. *International journal of computer assisted radiology and surgery*, 14(2):191–201, 2019.

Olivier Bernard, Johan G Bosch, Brecht Heyde, Martino Alessandrini, Daniel Barbosa, Sorina Camarasu-Pop, Frederic Cervenansky, Sébastien Valette, Oana Mirea, Michel Bernier, et al. **Standardized evaluation system for left ventricular segmentation algorithms in 3D echocardiography**. *IEEE transactions on medical imaging*, 35(4):967–977, 2015.

Christopher M Bishop and Nasser M Nasrabadi. **Pattern recognition and machine learning**, volume 4. Springer, 2006.

Sebastian Bodenstedt, Max Allan, Anthony Agustinos, Xiaofei Du, Luis Garcia-Peraza-Herrera, Hannes Kenngott, Thomas Kurmann, Beat Müller-Stich, Sebastien Ourselin, Daniil Pakhomov, et al. **Comparative evaluation of instrument segmentation and tracking methods in minimally invasive surgery**. *arXiv preprint arXiv:1805.02475*, 2018.

Douglas G Bonett. **Sample size requirements for estimating intraclass correlations with desired precision**. *Statistics in medicine*, 21(9):1331–1335, 2002.

Patrick M Bossuyt and Johannes B Reitsma. **The STARD initiative**. *The Lancet*, 361(9351):71, 2003.

Patrick M Bossuyt, Johannes B Reitsma, David E Bruns, Constantine A Gatsonis, Paul P Glasziou, Les M Irwig, Jeroen G Lijmer, David Moher, Drummond Rennie, and Henrica CW de Vet. **Special Reports-Towards Complete and Accurate Reporting of Studies of Diagnostic Accuracy: The STARD Initiative**. *Radiology-Radiological Society of North America*, 226(1):24–28, 2003.

PM Bossuyt, JB Reitsma, DE Bruns, CA Gatsonis, et al. **STARD 2015: an updated list of essential items for reporting diagnostic accuracy studies.[Online]. BMJ 351, h5527**, 2015.

G. Bradski. **The OpenCV Library**. *Dr. Dobb's Journal of Software Tools*, 2000.

BRATS2015. **BRATS2015**. `https://www.smir.ch/BRATS/Start2015`, 2015. (Accessed on 06/22/2022).

Robert L Brennan and Dale J Prediger. **Coefficient kappa: Some uses, misuses, and alternatives**. *Educational and psychological measurement*, 41(3):687–699, 1981.

Glenn W Brier et al. **Verification of forecasts expressed in terms of probability**. *Monthly weather review*, 78 (1):1–3, 1950.

Esther E Bron, Stefan Klein, Annika Reinke, Janne M Papma, Lena Maier-Hein, Daniel C Alexander, and Neil P Oxtoby. **Ten years of image analysis and machine learning competitions in dementia**. *arXiv preprint arXiv:2112.07922*, 2021.

Bernice B Brown. **Delphi process: a methodology used for the elicitation of opinions of experts**. Technical report, Rand Corp Santa Monica CA, 1968.

Jason Brownlee. **ROC Curves and Precision-Recall Curves for Imbalanced Classification**. `https://machinelearningmastery.com/roc-curves-and-precision-recall-curves-for-imbalanced-classification/`, 2020. (Accessed on 06/28/2022).

Nithin Buduma, Nikhil Buduma, and Joe Papa. **Fundamentals of deep learning**. " O'Reilly Media, Inc.", 2022.

Neil G Burnet, Simon J Thomas, Kate E Burton, and Sarah J Jefferies. **Defining the tumour and target volumes for radiotherapy**. *Cancer Imaging*, 4(2):153, 2004.

Oscar Camara. **Best (and worst) practices for organizing a challenge on cardiac biophysical models during ai summer: the crt-epiggy19 challenge**. In *International Workshop on Statistical Atlases and Computational Models of the Heart*, pages 329–341. Springer, 2019.

Andrea Caponnetto, Charles A Micchelli, Massimiliano Pontil, and Yiming Ying. **Universal multi-task kernels**. *The Journal of Machine Learning Research*, 9:1615–1646, 2008.

Aaron Carass, Snehashis Roy, Adrian Gherman, Jacob C Reinhold, Andrew Jesson, Tal Arbel, Oskar Maier, Heinz Handels, Mohsen Ghafoorian, Bram Platel, et al. **Evaluating white matter lesion segmentations with refined Sørensen-Dice analysis**. *Scientific reports*, 10(1):1–19, 2020.

Rubén Cárdenes, Rodrigo de Luis-Garcia, and Meritxell Bach-Cuadra. **A multidimensional segmentation evaluation for medical image data**. *Computer methods and programs in biomedicine*, 96(2):108–124, 2009.

Carlos Castilla, Martin Maška, Dmitry V Sorokin, Erik Meijering, and Carlos Ortiz-de Solórzano. **3-D quantification of filopodia in motile cancer cells**. *IEEE transactions on medical imaging*, 38(3):862–872, 2018.

Weijie Chen, Maryellen L Giger, and Ulrich Bick. **A fuzzy c-means (fcm)-based approach for computerized segmentation of breast lesions in dynamic contrast-enhanced mr images1**. *Academic radiology*, 13(1): 63–72, 2006.

Bowen Cheng, Ross Girshick, Piotr Dollár, Alexander C Berg, and Alexander Kirillov. **Boundary IoU: Improving Object-Centric Image Segmentation Evaluation**. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15334–15342, 2021.

Davide Chicco and Giuseppe Jurman. **The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation**. *BMC genomics*, 21(1):1–13, 2020.

Davide Chicco, Niklas Tötsch, and Giuseppe Jurman. **The Matthews correlation coefficient (MCC) is more reliable than balanced accuracy, bookmaker informedness, and markedness in two-class confusion matrix evaluation**. *BioData mining*, 14(1):1–22, 2021.

Nancy Chinchor. **MUC-4 Evaluation Metrics**. In *Proceedings of the 4th Conference on Message Understanding*, MUC4 '92, page 22–29, USA, 1992. Association for Computational Linguistics. ISBN 1558602739. doi: 10.3115/1072064.1072067. URL https://doi.org/10.3115/1072064.1072067.

Francois Chollet et al. **Keras**. https://github.com/fchollet/keras, 2015. (Accessed on 05/19/2022).

Celia Chui, Maryam Kouchaki, and Francesca Gino. **"Many others are doing it, so why shouldn't I?": How being in larger competitions leads to more cheating**. *Organizational Behavior and Human Decision Processes*, 164:102–115, 2021.

William S Cleveland. **Visualizing data**. Hobart press, 1993.

Paul Clough, Henning Müller, and Mark Sanderson. **The CLEF 2004 cross-language image retrieval track**. In *Workshop of the Cross-Language Evaluation Forum for European Languages*, pages 597–613. Springer, 2004.

Noel Codella, Veronica Rotemberg, Philipp Tschandl, M Emre Celebi, Stephen Dusza, David Gutman, Brian Helba, Aadi Kalloo, Konstantinos Liopyris, Michael Marchetti, et al. **Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic)**. *arXiv preprint arXiv:1902.03368*, 2019.

Noel CF Codella, David Gutman, M Emre Celebi, Brian Helba, Michael A Marchetti, Stephen W Dusza, Aadi Kalloo, Konstantinos Liopyris, Nabin Mishra, Harald Kittler, et al. **Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic)**. In *2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018)*, pages 168–172. IEEE, 2018.

Jacob Cohen. **A coefficient of agreement for nominal scales**. *Educational and psychological measurement*, 20(1):37–46, 1960.

Jérémie F Cohen, Daniël A Korevaar, Douglas G Altman, David E Bruns, Constantine A Gatsonis, Lotty Hooft, Les Irwig, Deborah Levine, Johannes B Reitsma, Henrica CW De Vet, et al. **STARD 2015 guidelines for reporting diagnostic accuracy studies: explanation and elaboration**. *BMJ open*, 6(11):e012799, 2016.

Olivier Commowick, Audrey Istace, Michael Kain, Baptiste Laurent, Florent Leray, Mathieu Simon, Sorina Camarasu Pop, Pascal Girard, Roxana Ameli, Jean-Christophe Ferré, et al. **Objective evaluation of multiple sclerosis lesion segmentation using a data management and processing infrastructure**. *Scientific reports*, 8(1):1–17, 2018.

Nancy R Cook. **Use and misuse of the receiver operating characteristic curve in risk prediction**. *Circulation*, 115(7):928–935, 2007.

Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Scharwächter, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. **The Cityscapes Dataset**. In *CVPR Workshop on The Future of Datasets in Vision*, 2015.

Lester A Critchley, Anna Lee, and Anthony M-H Ho. **A critical review of the ability of continuous cardiac output monitors to measure trends in cardiac output**. *Anesthesia & Analgesia*, 111(5):1180–1192, 2010.

George Cybenko, Dianne P O'Leary, and Jorma Rissanen. **The Mathematics of Information Coding, Extraction and Distribution**, volume 107. Springer Science & Business Media, 1998.

Astrid Dannenberg and Elina Khachatryan. **A comparison of individual and group behavior in a competition with cheating opportunities**. *Journal of Economic Behavior & Organization*, 177:533–547, 2020.

Brian A Davey and Hilary A Priestley. **Introduction to lattices and order**. Cambridge university press, 2002.

Jesse Davis and Mark Goadrich. **The relationship between Precision-Recall and ROC curves**. In *Proceedings of the 23rd international conference on Machine learning*, pages 233–240, 2006.

Benoit M Dawant, Rui Li, Brian Lennon, and Senhu Li. **Semi-automatic segmentation of the liver and its evaluation on the MICCAI 2007 grand challenge data set**. *3D Segmentation in The Clinic: A Grand Challenge*, pages 215–221, 2007.

Laura Daza, Angela Castillo, María Escobar, Sergio Valencia, Bibiana Pinzón, and Pablo Arbeláez. **Lucas: Lung cancer screening with multimodal biomarkers**. In *Multimodal Learning for Clinical Decision Support and Clinical Image-Based Procedures*, pages 115–124. Springer, 2020.

Elizabeth R DeLong, David M DeLong, and Daniel L Clarke-Pearson. **Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach**. *Biometrics*, pages 837–845, 1988.

Janez Demšar. **Statistical comparisons of classifiers over multiple data sets**. *The Journal of Machine Learning Research*, 7:1–30, 2006.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. **Imagenet: A large-scale hierarchical image database**. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

Lee R Dice. **Measures of the amount of ecologic association between species**. *Ecology*, 26(3):297–302, 1945.

Docker. **Docker**. `https://www.docker.com/`, 2022. (Accessed on 06/17/2022).

David M Eddy, William Hollingworth, J Jaime Caro, Joel Tsevat, Kathryn M McDonald, and John B Wong. **Model transparency and validation: a report of the ISPOR-SMDM Modeling Good Research Practices Task Force–7**. *Medical Decision Making*, 32(5):733–743, 2012.

Bradley Efron and Robert J Tibshirani. **An introduction to the bootstrap**. CRC press, 1994.

Manuel JA Eugster, Torsten Hothorn, and Friedrich Leisch. **Exploratory and inferential analysis of benchmark experiments**. Technical Report 030, Department of Statistics, University of Munich, 2008.

Manuel JA Eugster, Torsten Hothorn, and Friedrich Leisch. **Domain-based benchmark experiments: Exploratory and inferential analysis**. *Austrian Journal of Statistics*, 41(1):5–26, 2012.

Mark Everingham, SM Ali Eslami, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. **The pascal visual object classes challenge: A retrospective**. *International journal of computer vision*, 111 (1):98–136, 2015.

FacebookResearch. **Detectron2**. `https://github.com/facebookresearch/detectron2`, 2019. (Accessed on 05/18/2022).

Jingtao Fan, Lu Fang, Jiamin Wu, Yuchen Guo, and Qionghai Dai. **From brain science to artificial intelligence**. *Engineering*, 6(3):248–252, 2020.

Gunnar Farnebäck. **Two-frame motion estimation based on polynomial expansion**. In *Scandinavian conference on Image analysis*, pages 363–370. Springer, 2003.

Luciana Ferrer. **Analysis and Comparison of Classification Metrics**. *arXiv preprint arXiv:2209.05355*, 2022.

Martin Feuerman and Allen R Miller. **Relationships between statistical measures of agreement: sensitivity, specificity and kappa**. *Journal of evaluation in clinical practice*, 14(5):930–933, 2008.

James Fishbaugh, Marcel Prastawa, Bo Wang, Patrick Reynolds, Stephen Aylward, and Guido Gerig. **Data-driven rank aggregation with application to grand challenges**. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 754–762. Springer, 2017.

Isabel Funke, Sebastian Bodenstedt, Carina Riediger, Jürgen Weitz, and Stefanie Speidel. **Generative adversarial networks for specular highlight removal in endoscopic images**. In *Medical Imaging 2018: Image-Guided Procedures, Robotic Interventions, and Modeling*, volume 10576, page 1057604. International Society for Optics and Photonics, 2018.

Joel J Gagnier, Gunver Kienle, Douglas G Altman, David Moher, Harold Sox, and David Riley. **The CARE guidelines: consensus-based clinical case reporting guideline development**. *Journal of medical case reports*, 7(1):1–6, 2013.

James Gareth, Witten Daniela, Hastie Trevor, and Tibshirani Robert. **An introduction to statistical learning: with applications in R**. Spinger, 2013.

Guido Gerig, Matthieu Jomier, and Miranda Chakos. **Valmet: A new validation tool for assessing and improving 3D object segmentation**. In *International conference on medical image computing and computer-assisted intervention*, pages 516–523. Springer, 2001.

Ross Girshick. **Fast r-cnn**. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015.

Xavier Glorot and Yoshua Bengio. **Understanding the difficulty of training deep feedforward neural networks**. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 249–256. JMLR Workshop and Conference Proceedings, 2010.

Tilmann Gneiting and Adrian E Raftery. **Strictly proper scoring rules, prediction, and estimation**. *Journal of the American statistical Association*, 102(477):359–378, 2007.

Orcun Goksel, Antonio Foncubierta-Rodríguez, Oscar Alfonso Jiménez del Toro, Henning Müller, Georg Langs, Marc-André Weber, Bjoern H Menze, Ivan Eggel, Katharina Gruenberg, Marianne Winterstein, et al. **Overview of the VISCERAL Challenge at ISBI 2015.** In *VISCERAL Challenge@ ISBI*, pages 6–11, 2015.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. **Deep Learning**. MIT Press, 2016. `http://www.deeplearningbook.org`.

Mark J Gooding, Annamarie J Smith, Maira Tariq, Paul Aljabar, Devis Peressutti, Judith van der Stoep, Bart Reymen, Daisy Emans, Djoya Hattu, Judith van Loon, et al. **Comparative evaluation of autocontouring in clinical practice: a practical method using the Turing test**. *Medical physics*, 45(11):5105–5115, 2018.

Jan Gorodkin. **Comparing two K-category assignments by a K-category correlation coefficient**. *Computational biology and chemistry*, 28(5-6):367–374, 2004.

Samuel Gratzl, Alexander Lex, Nils Gehlenborg, Hanspeter Pfister, and Marc Streit. **Lineup: Visual analysis of multi-attribute rankings**. *IEEE transactions on visualization and computer graphics*, 19(12):2277–2286, 2013.

Sebastian Gruber and Florian Buettner. **Trustworthy Deep Learning via Proper Calibration Errors: A Unifying Approach for Quantifying the Reliability of Predictive Uncertainty**. *arXiv preprint arXiv:2203.07835*, 2022.

Zaiwang Gu, Jun Cheng, Huazhu Fu, Kang Zhou, Huaying Hao, Yitian Zhao, Tianyang Zhang, Shenghua Gao, and Jiang Liu. **Ce-net: Context encoder network for 2d medical image segmentation**. *IEEE transactions on medical imaging*, 38(10):2281–2292, 2019.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. **On calibration of modern neural networks**. In *International conference on machine learning*, pages 1321–1330. PMLR, 2017.

Yanming Guo, Yu Liu, Theodoros Georgiou, and Michael S Lew. **A review of semantic segmentation using deep neural networks**. *International journal of multimedia information retrieval*, 7(2):87–93, 2018.

Agrim Gupta, Piotr Dollar, and Ross Girshick. **LVIS: A dataset for large vocabulary instance segmentation**. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5356–5364, 2019.

Metin N Gurcan, Anant Madabhushi, and Nasir Rajpoot. **Pattern recognition in histopathological images: An ICPR 2010 contest**. In *International Conference on Pattern Recognition*, pages 226–234. Springer, 2010.

James A Hanley and Barbara J McNeil. **The meaning and use of the area under a receiver operating characteristic (ROC) curve.** *Radiology*, 143(1):29–36, 1982.

Trevor Hastie, Robert Tibshirani, Jerome H Friedman, and Jerome H Friedman. **The elements of statistical learning: data mining, inference, and prediction**, volume 2. Springer, 2009.

Trine B Haugen, Steven A Hicks, Jorunn M Andersen, Oliwia Witczak, Hugo L Hammer, Rune Borgli, Pål Halvorsen, and Michael Riegler. **Visem: A multimodal video dataset of human spermatozoa**. In *Proceedings of the 10th ACM Multimedia Systems Conference*, pages 261–266, 2019.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. **Deep residual learning for image recognition**. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. **Mask r-cnn**. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.

Tobias Heimann, Bram Van Ginneken, Martin A Styner, Yulia Arzhaeva, Volker Aurich, Christian Bauer, Andreas Beck, Christoph Becker, Reinhard Beichel, György Bekes, et al. **Comparison and evaluation of methods for liver segmentation from CT datasets**. *IEEE transactions on medical imaging*, 28(8):1251–1265, 2009.

Nicholas Heller, Niranjan Sathianathen, Arveen Kalapara, Edward Walczak, Keenan Moore, Heather Kaluzniak, Joel Rosenberg, Paul Blake, Zachary Rengel, Makinna Oestreich, et al. **The kits19 challenge data: 300 kidney tumor cases with clinical context, ct semantic segmentations, and surgical outcomes**. *arXiv preprint arXiv:1904.00445*, 2019.

Steven A Hicks, Inga Strümke, Vajira Thambawita, Malek Hammou, Michael A Riegler, Pål Halvorsen, and Sravanthi Parasa. **On evaluation metrics for medical applications of artificial intelligence**. *Scientific Reports*, 12(1):1–9, 2022.

Peter Hirsch, Lisa Mais, and Dagmar Kainmueller. **PatchPerPix for instance segmentation**. *arXiv preprint arXiv:2001.07626*, 2020.

Sture Holm. **A simple sequentially rejective multiple test procedure**. *Scandinavian journal of statistics*, pages 65–70, 1979.

Kurt Hornik and David Meyer. **Deriving consensus rankings from benchmarking experiments**. In *Advances in data analysis*, pages 163–170. Springer, 2007.

Zilong Hu, Jinshan Tang, Ziming Wang, Kai Zhang, Ling Zhang, and Qingling Sun. **Deep learning for image-based cancer detection and diagnosis- A survey**. *Pattern Recognition*, 83:134–149, 2018.

Arnaud Huaulmé, Duygu Sarikaya, Kévin Le Mut, Fabien Despinoy, Yonghao Long, Qi Dou, Chin-Boon Chng, Wenjun Lin, Satoshi Kondo, Laura Bravo-Sánchez, et al. **MIcro-surgical anastomose workflow recognition challenge report**. *Computer Methods and Programs in Biomedicine*, 212:106452, 2021.

Peter J Huber. **Robust estimation of a location parameter**. In *Breakthroughs in statistics*, pages 492–518. Springer, 1992.

Daniel P Huttenlocher, Gregory A. Klanderman, and William J Rucklidge. **Comparing images using the Hausdorff distance**. *IEEE Transactions on pattern analysis and machine intelligence*, 15(9):850–863, 1993.

Gareth Iacobucci. **Covid-19: Prevalence has quadrupled in England since start of September, study shows**, 2020.

ICMJE. **Clinical Trials**. `https://www.icmje.org/recommendations/browse/publishing-and-editorial-issues/clinical-trial-registration.html`, 2013. (Accessed on 07/18/2022).

Vladimir Iglovikov and Alexey Shvets. **Ternausnet: U-net with vgg11 encoder pre-trained on imagenet for image segmentation**. *arXiv preprint arXiv:1801.05746*, 2018.

John PA Ioannidis. **Why most published research findings are false**. *PLoS medicine*, 2(8):e124, 2005.

Fabian Isensee, Paul Jäger, Jakob Wasserthal, David Zimmerer, Jens Petersen, Simon Kohl, Justus Schock, Andre Klein, Tobias Roß, Sebastian Wirkert, Peter Neher, Stefan Dinkelacker, Gregor Köhler, and Klaus Maier-Hein. **Batchgenerators - a python framework for data augmentation**. doi:10.5281/zenodo.3632567, 2020. (Accessed on 06/03/2022).

Fabian Isensee, Paul F Jaeger, Simon AA Kohl, Jens Petersen, and Klaus H Maier-Hein. **nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation**. *Nature methods*, 18(2): 203–211, 2021.

Ivana Išgum, Manon JNL Benders, Brian Avants, M Jorge Cardoso, Serena J Counsell, Elda Fischi Gomez, Laura Gui, Petra S Hűppi, Karina J Kersbergen, Antonios Makropoulos, et al. **Evaluation of automatic neonatal brain segmentation algorithms: the NeoBrainS12 challenge**. *Medical image analysis*, 20(1):135–151, 2015.

Paul Jaccard. **The distribution of the flora in the alpine zone. 1**. *New phytologist*, 11(2):37–50, 1912.

Paul Ferdinand Jäger. **Challenges and Opportunities of End-to-End Learning in Medical Image Classification**. PhD thesis, Karlsruher Institut für Technologie (KIT), 2020.

Anil K Jain and Balakrishnan Chandrasekaran. **39 Dimensionality and sample size considerations in pattern recognition practice**. *Handbook of statistics*, 2:835–855, 1982.

Pierre Jannin, Christophe Grova, and Calvin R Maurer. **Model for defining and reporting reference-based validation protocols in medical image processing**. *International Journal of Computer Assisted Radiology and Surgery*, 1(2):63–73, 2006.

Leo Joskowicz, D Cohen, N Caplan, and J Sosna. **Inter-observer variability of manual contour delineation of structures in CT**. *European radiology*, 29(3):1391–1399, 2019.

Kaggle. **Data Science Bowl 2017**. `https://www.kaggle.com/competitions/data-science-bowl-2017`, 2016. (Accessed on 07/26/2022).

HM Kalayeh and David A Landgrebe. **Predicting the required number of training samples**. *IEEE transactions on pattern analysis and machine intelligence*, 6:664–667, 1983.

Konstantinos Kamnitsas, Enzo Ferrante, Sarah Parisot, Christian Ledig, Aditya V Nori, Antonio Criminisi, Daniel Rueckert, and Ben Glocker. **DeepMedic for brain tumor segmentation**. In *International workshop on Brainlesion: Glioma, multiple sclerosis, stroke and traumatic brain injuries*, pages 138–149. Springer, 2016.

Maurice G Kendall. **A new measure of rank correlation**. *Biometrika*, 30(1/2):81–93, 1938.

Justin Ker, Lipo Wang, Jai Rao, and Tchoyoson Lim. **Deep learning applications in medical image analysis**. *Ieee Access*, 6:9375–9389, 2017.

Muhammad Attique Khan, Imran Ashraf, Majed Alhaisoni, Robertas Damaševičius, Rafal Scherer, Amjad Rehman, and Syed Ahmad Chan Bukhari. **Multimodal brain tumor classification using deep learning and robust feature selection: A machine learning application for radiologists**. *Diagnostics*, 10(8):565, 2020.

Jack Kiefer and Jacob Wolfowitz. **Stochastic estimation of the maximum of a regression function**. *The Annals of Mathematical Statistics*, pages 462–466, 1952.

Douwe Kiela, Max Bartolo, Yixin Nie, Divyansh Kaushik, Atticus Geiger, Zhengxuan Wu, Bertie Vidgen, Grusha Prasad, Amanpreet Singh, Pratik Ringshia, et al. **Dynabench: Rethinking benchmarking in NLP**. *arXiv preprint arXiv:2104.14337*, 2021.

Hosung Kim, Claude Lepage, Alan C Evans, A James Barkovich, and Duan Xu. **NEOCIVET: Extraction of cortical surface and analysis of neonatal gyrification using a modified CIVET pipeline**. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 571–579. Springer, 2015.

Diederik P Kingma and Jimmy Ba. **Adam: A method for stochastic optimization**. *arXiv preprint arXiv:1412.6980*, 2014.

Alexander Kirillov, Kaiming He, Ross Girshick, Carsten Rother, and Piotr Dollár. **Panoptic segmentation**. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9404–9413, 2019.

Hoon Ko, Heewon Chung, Wu Seong Kang, Kyung Won Kim, Youngbin Shin, Seung Ji Kang, Jae Hoon Lee, Young Jun Kim, Nan Yeol Kim, Hyunseok Jung, et al. **COVID-19 pneumonia diagnosis using a simple 2D deep learning framework with a single chest CT image: model development and validation**. *Journal of medical Internet research*, 22(6):e19569, 2020.

Florian Kofler, Ivan Ezhov, Fabian Isensee, Fabian Balsiger, Christoph Berger, Maximilian Koerner, Johannes Paetzold, Hongwei Li, Suprosanna Shit, Richard McKinley, et al. **Are we using appropriate segmentation metrics? Identifying correlates of human expert perception for CNN training beyond rolling the DICE coefficient**. *arXiv preprint arXiv:2103.06205*, 2021.

Florian Kofler, Suprosanna Shit, Ivan Ezhov, Lucas Fidon, Rami Al-Maskari, Hongwei Li, Harsharan Bhatia, Timo Loehr, Marie Piraud, Ali Erturk, et al. **blob loss: instance imbalance aware loss functions for semantic segmentation**. *arXiv preprint arXiv:2205.08209*, 2022.

Ron Kohavi et al. **A study of cross-validation and bootstrap for accuracy estimation and model selection**. *Ijcai*, 14(2):1137–1145, 1995.

Thijs Kooi. **Evaluation curves for object detection algorithms in medical images**. `https://medium.com/lunit/evaluation-curves-for-object-detection-algorithms-in-medical-images-4b083fddce6e`, 2021a. (Accessed on 06/28/2022).

Thijs Kooi. **Evaluation curves for object detection algorithms in medical images**. `https://medium.com/lunit/evaluation-curves-for-object-detection-algorithms-in-medical-images-4b083fddce6e`, 2021b. (Accessed on 07/18/2022).

Michal Kozubek. **Challenges and benchmarks in bioimage analysis**. *Focus on Bio-Image Informatics*, pages 231–262, 2016.

Harold W Kuhn. **The Hungarian method for the assignment problem**. *Naval research logistics quarterly*, 2 (1-2):83–97, 1955.

Ananya Kumar, Percy S Liang, and Tengyu Ma. **Verified uncertainty calibration**. *Advances in Neural Information Processing Systems*, 32, 2019.

Dominik Kutra, Axel Saalbach, Helko Lehmann, Alexandra Groth, Sebastian PM Dries, Martin W Krueger, Olaf Dössel, and Jürgen Weese. **Automatic multi-model-based segmentation of the left atrium in cardiac MRI scans**. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 1–8. Springer, 2012.

Sabine Landau, Morven Leese, Daniel Stahl, and Brian S Everitt. **Cluster analysis**. John Wiley & Sons, 2011.

Jochen K Lennerz, Ursula Green, Drew FK Williamson, and Faisal Mahmood. **A unifying force for the realization of medical AI**. *npj Digital Medicine*, 5(1):1–3, 2022.

Qing Li, Weidong Cai, Xiaogang Wang, Yun Zhou, David Dagan Feng, and Mei Chen. **Medical image classification with convolutional neural network**. In *2014 13th international conference on control automation robotics & vision (ICARCV)*, pages 844–848. IEEE, 2014.

H.-T. I. P. S. Lin, M.-F. Balcan, R. Hadsell, and M. Ranzato. **What we learned from NeurIPS 2020 reviewing process**. `https://medium.com/@NeurIPSConf/what-we-learned-from-neurips-2020-reviewing-process-e24549eea38f`, 2020. (Accessed on 08/02/2022).

Shili Lin. **Rank aggregation methods**. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2(5):555–570, 2010.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. **Microsoft coco: Common objects in context**. In *European conference on computer vision*, pages 740–755. Springer, 2014.

Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. **Feature pyramid networks for object detection**. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017a.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. **Focal loss for dense object detection**. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017b.

Tsung-Yu Lin, Aruni RoyChowdhury, and Subhransu Maji. **Bilinear cnn models for fine-grained visual recognition**. In *Proceedings of the IEEE international conference on computer vision*, pages 1449–1457, 2015.

Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen Awm Van Der Laak, Bram Van Ginneken, and Clara I Sánchez. **A survey on deep learning in medical image analysis**. *Medical image analysis*, 42:60–88, 2017.

Wei Liu, Andrew Rabinovich, and Alexander C Berg. **Parsenet: Looking wider to see better**. *arXiv preprint arXiv:1506.04579*, 2015.

Xiaoxuan Liu, Samantha Cruz Rivera, David Moher, Melanie J Calvert, and Alastair K Denniston. **Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension**. *bmj*, 370, 2020.

George F Luger. **Artificial intelligence: structures and strategies for complex problem solving**. Pearson education, 2005.

Kaveh Madani, Laura Read, and Laleh Shalikarian. **Voting under uncertainty: a stochastic framework for analyzing group decision making problems**. *Water resources management*, 28(7):1839–1856, 2014.

Aouatef Mahani and Ahmed Riad Baba Ali. **Classification problem in imbalanced datasets**. *Recent Trends in Computational Intelligence*, pages 1–23, 2019.

Oskar Maier. **Metric measures (medpy.metric)**. `https://loli.github.io/medpy/metric.html`, 2013. (Accessed on 05/02/2022).

Oskar Maier, Bjoern Menze, and Mauricio Reyes. **ISLES: Ischemic Stroke Lesion Segmentation Challenge 2015**. `http://www.isles-challenge.org/ISLES2015/`, 2015. (Accessed on 06/22/2022).

Lena Maier-Hein, Anja Groch, Adrien Bartoli, Sebastian Bodenstedt, G Boissonnat, P-L Chang, NT Clancy, Daniel S Elson, Sven Haase, Eric Heim, et al. **Comparative validation of single-shot optical techniques for laparoscopic 3-D surface reconstruction**. *IEEE transactions on medical imaging*, 33(10):1913–1930, 2014.

Lena Maier-Hein, Matthias Eisenmann, Annika Reinke, Sinan Onogur, Marko Stankovic, Patrick Scholz, Tal Arbel, Hrvoje Bogunovic, Andrew P Bradley, Aaron Carass, Carolin Feldmann, Alejandro F Frangi, Peter M Full, Bram van Ginneken, Allan Hanbury, Katrin Honauer, Michal Kozubek, Bennett A Landman, Keno März, Oskar Maier, Klaus Maier-Hein, Bjoern H Menze, Henning Müller, Peter F Neher, Wiro Niessen, Nasir Rajpoot, Gregory C Sharp, Korsuk Sirinukunwattana, Stefanie Speidel, Christian Stock, Danail Stoyanov, Abdel A Taha, Fons Van der Sommen, Ching-Wei Wang, Marc-André Weber, Guoyan Zheng, Pierre Jannin, and Annette Kopp-Schneider. **Why rankings of biomedical image analysis competitions should be interpreted with care**. *Nature communications*, 9(1):1–13, 2018. Shared first authors: Lena Maier-Hein, Matthias Eisenmann.

Lena Maier-Hein, Annika Reinke, Michal Kozubek, Anne L Martel, Tal Arbel, Matthias Eisenmann, Allan Hanbury, Pierre Jannin, Henning Müller, Sinan Onogur, Julio Saez-Rodriguez, Bram van Ginneken, Annette Kopp-Schneider, and Bennett A Landman. **BIAS: Transparent reporting of biomedical image analysis challenges**. *Medical image analysis*, 66:101796, 2020.

Lena Maier-Hein, Martin Wagner, Tobias Ross, Annika Reinke, Sebastian Bodenstedt, Peter M Full, Hellena Hempe, Diana Mindroc-Filimon, Patrick Scholz, Thuy Nuong Tran, et al. **Heidelberg colorectal data set for surgical data science in the sensor operating room**. *Scientific data*, 8(1):1–11, 2021.

Lena Maier-Hein, Annika Reinke, Evangelia Christodoulou, Ben Glocker, Patrick Godau, Fabian Isensee, Jens Kleesiek, Michal Kozubek, Mauricio Reyes, Michael A Riegler, et al. **Metrics reloaded: Pitfalls and recommendations for image analysis validation**. *arXiv preprint arXiv:2206.01653*, 2022.

Lisa Mais, Peter Hirsch, and Dagmar Kainmueller. **Patchperpix for instance segmentation**. In *European Conference on Computer Vision*, pages 288–304. Springer, 2020.

Martin Maška, Vladimír Ulman, David Svoboda, Pavel Matula, Petr Matula, Cristina Ederra, Ainhoa Urbiola, Tomás España, Subramanian Venkatesan, Deepak MW Balak, et al. **A benchmark for comparison of cell tracking algorithms**. *Bioinformatics*, 30(11):1609–1617, 2014.

Matterport. **Matterport Mask R-CNN**. `https://github.com/matterport/Mask_RCNN`, 2017. (Accessed on 05/19/2022).

Brian W Matthews. **Comparison of the predicted and observed secondary structure of T4 phage lysozyme**. *Biochimica et Biophysica Acta (BBA)-Protein Structure*, 405(2):442–451, 1975.

Pavel Matula, Martin Maška, Dmitry V Sorokin, Petr Matula, Carlos Ortiz-de Solórzano, and Michal Kozubek. **Cell tracking accuracy measurement based on comparison of acyclic oriented graphs**. *PloS one*, 10(12): e0144959, 2015.

Charles E McCulloch and Shayle R Searle. **Generalized, linear, and mixed models**. John Wiley & Sons, 2004.

Warren S McCulloch and Walter Pitts. **A logical calculus of the ideas immanent in nervous activity**. *The bulletin of mathematical biophysics*, 5(4):115–133, 1943.

Marcia McNutt. **Reproducibility**, 2014.

G. Meissner, A. Nern, Z. Dorman, DePasquale G.M., K. Forster, T. Gibney, Hausenfluck J.H., Y. He, N. Iyer, J. Jeter, et al. **A searchable image resource of Drosophila GAL4-driver expression patterns with single neuron resolution**. *BioRxiv*, page 2020.05.29.080473, 2022.

Adriënne M Mendrik and Stephen R Aylward. **A framework for challenge design: Insight and deployment challenges to address medical image analysis problems**. *arXiv preprint arXiv:1911.08531*, 2019.

Bjoern H Menze, Andras Jakab, Stefan Bauer, Jayashree Kalpathy-Cramer, Keyvan Farahani, Justin Kirby, Yuliya Burren, Nicole Porz, Johannes Slotboom, Roland Wiest, et al. **The multimodal brain tumor image segmentation benchmark (BRATS)**. *IEEE transactions on medical imaging*, 34(10):1993–2024, 2014.

MICCAI_SIG_for_Challenges. **MICCAI registered challenges**. `http://www.miccai.org/special-interest-groups/challenges/miccai-registered-challenges/`, 2018. (Accessed on 07/18/2022).

Charles A Micchelli and Massimiliano Pontil. **On learning vector-valued functions**. *Neural computation*, 17 (1):177–204, 2005.

Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. **V-net: Fully convolutional neural networks for volumetric medical image segmentation**. In *2016 fourth international conference on 3D vision (3DV)*, pages 565–571. IEEE, 2016.

Sudhanshu K Mishra. **The most representative composite rank ordering of multi-attribute objects by the particle swarm optimization**. *Available at SSRN 1326386*, 2009.

Tom M Mitchell and Tom M Mitchell. **Machine learning**, volume 1. McGraw-hill New York, 1997.

Pawel Mlynarski, Hervé Delingette, Hamza Alghamdi, Pierre-Yves Bondiau, and Nicholas Ayache. **Anatomically consistent CNN-based segmentation of organs-at-risk in cranial radiotherapy**. *Journal of Medical Imaging*, 7(1):014502, 2020.

MONAI-developers. **MONAI**. `https://github.com/Project-MONAI/MONAI/tree/dev/monai/metrics`, 2019. (Accessed on 05/02/2022).

John Mongan, Linda Moy, and Charles E Kahn Jr. **Checklist for artificial intelligence in medical imaging (CLAIM): a guide for authors and reviewers**. *Radiology. Artificial Intelligence*, 2(2), 2020.

Keelin Murphy, Bram Van Ginneken, Joseph M Reinhardt, Sven Kabus, Kai Ding, Xiang Deng, Kunlin Cao, Kaifang Du, Gary E Christensen, Vincent Garcia, et al. **Evaluation of registration methods on thoracic CT: the EMPIRE10 challenge**. *IEEE transactions on medical imaging*, 30(11):1901–1920, 2011.

Mahdi Pakdaman Naeini, Gregory Cooper, and Milos Hauskrecht. **Obtaining well calibrated probabilities using bayesian binning**. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.

Yukiko Nagao, Mika Sakamoto, Takumi Chinen, Yasushi Okada, and Daisuke Takao. **Robust classification of cell cycle phase and biological feature extraction by image-based deep learning**. *Molecular biology of the cell*, 31(13):1346–1354, 2020.

Ying-Hwey Nai, Bernice W Teo, Nadya L Tan, Sophie O'Doherty, Mary C Stephenson, Yee Liang Thian, Edmund Chiong, and Anthonin Reilhac. **Comparison of metrics for the evaluation of medical segmentations using prostate MRI dataset**. *Computers in Biology and Medicine*, 134:104497, 2021.

Tanya Nair. **Exploring uncertainty measures in deep networks for multiple sclerosis lesion detection and segmentation**. PhD thesis, McGill University, 2018.

RB Nelsen. **Kendall tau metric**. *Encyclopaedia of mathematics*, 3:226–227, 2001.

Zhen-Liang Ni, Gui-Bin Bian, Xiao-Hu Zhou, Zeng-Guang Hou, Xiao-Liang Xie, Chen Wang, Yan-Jie Zhou, Rui-Qi Li, and Zhen Li. **Raunet: Residual attention u-net for semantic segmentation of cataract surgical instruments**. In *International Conference on Neural Information Processing*, pages 139–149. Springer, 2019.

Zhen-Liang Ni, Gui-Bin Bian, Guan-An Wang, Xiao-Hu Zhou, Zeng-Guang Hou, Xiao-Liang Xie, Zhen Li, and Yu-Han Wang. **BARNet: bilinear attention network with adaptive receptive fields for surgical instrument segmentation**. *arXiv preprint arXiv:2001.07093*, 2020.

Thomas E Nichols, Samir Das, Simon B Eickhoff, Alan C Evans, Tristan Glatard, Michael Hanke, Nikolaus Kriegeskorte, Michael P Milham, Russell A Poldrack, Jean-Baptiste Poline, et al. **Best practices in data analysis and sharing in neuroimaging using MRI**. *Nature neuroscience*, 20(3):299–303, 2017.

Nudrat Nida, Aun Irtaza, Ali Javed, Muhammad Haroon Yousaf, and Muhammad Tariq Mahmood. **Melanoma lesion detection and segmentation using deep region based convolutional neural network and fuzzy C-means clustering**. *International journal of medical informatics*, 124:37–48, 2019.

Stanislav Nikolov, Sam Blackwell, Alexei Zverovitch, Ruheena Mendes, Michelle Livne, Jeffrey De Fauw, Yojan Patel, Clemens Meyer, Harry Askham, Bernadino Romera-Paredes, et al. **Clinically applicable segmentation of head and neck anatomy for radiotherapy: deep learning algorithm development and validation study**. *Journal of Medical Internet Research*, 23(7):e26151, 2021.

Jeremy Nixon, Michael W Dusenberry, Linchuan Zhang, Ghassen Jerfel, and Dustin Tran. **Measuring Calibration in Deep Learning.** *CVPR Workshops*, 2(7), 2019.

OpenCV. **Tutorial Optical Flow**. `https://docs.opencv.org/3.4/d4/dee/tutorial_optical_flow.html`, 2000. (Accessed on 05/24/2022).

Valentin Oreiller, Vincent Andrearczyk, Mario Jreige, Sarah Boughdad, Hesham Elhalawani, Joel Castelli, Martin Vallières, Simeng Zhu, Juanying Xie, Ying Peng, et al. **Head and neck tumor segmentation in PET/CT: the HECKTOR challenge**. *Medical image analysis*, 77:102336, 2022.

Trishan Panch, Heather Mattie, and Leo Anthony Celi. **The "inconvenient truth" about AI in healthcare**. *NPJ digital medicine*, 2(1):1–3, 2019.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. **Pytorch: An imperative style, high-performance deep learning library**. *Advances in neural information processing systems*, 32, 2019.

Sarthak Pati, Ujjwal Baid, Maximilian Zenk, Brandon Edwards, Micah Sheller, G Anthony Reina, Patrick Foley, Alexey Gruzdev, Jason Martin, Shadi Albarqouni, Russell Yong Chen, Taki Shinohara, Annika Reinke, David Zimmerer, John B Freymann, Justin S Kirby, Christos Davatzikos, Rivka R Colen, Aikaterini Kotrotsou, Daniel Marcus, Mikhail Milchenko, Arash Nazer, Hassan Fathallah-Shaykh, Roland Wiest, Andras Jakab, Marc-Andre Weber, Abhishek Mahajan, Lena Maier-Hein, Jens Kleesiek, Bjoern Menze, Klaus Maier-Hein, and Spyridon Bakas. **The Federated Tumor Segmentation (FeTS) Challenge**. *arXiv preprint arXiv:2105.05874*, 2021.

Hanchuan Peng, Michael Hawrylycz, Jane Roskams, Sean Hill, Nelson Spruston, Erik Meijering, and Giorgio A Ascoli. **BigNeuron: large-scale 3D neuron reconstruction from optical microscopy images**. *Neuron*, 87(2): 252–256, 2015.

Isabella Peters, Peter Kraker, Elisabeth Lex, Christian Gumpenberger, and Juan Ignacio Gorraiz. **Zenodo in the spotlight of traditional and new metrics**. *Frontiers in Research Metrics and Analytics*, 2:13, 2017.

Hung Viet Pham, Shangshu Qian, Jiannan Wang, Thibaud Lutellier, Jonathan Rosenthal, Lin Tan, Yaoliang Yu, and Nachiappan Nagappan. **Problems and opportunities in training deep learning software systems: An analysis of variance**. In *Proceedings of the 35th IEEE/ACM international conference on automated software engineering*, pages 771–783, 2020.

Paolo Piccinini, Andrea Prati, and Rita Cucchiara. **Real-time object detection and localization with SIFT-based clustering**. *Image and Vision Computing*, 30(8):573–587, 2012.

David L Poole and Alan K Mackworth. **Artificial Intelligence: foundations of computational agents**. Cambridge University Press, 2010.

Teodora Popordanoska, Raphael Sayer, and Matthew B Blaschko. **A Consistent and Differentiable Lp Canonical Calibration Error Estimator**. *arXiv preprint arXiv:2210.07810*, 2022.

Nicolas Posocco and Antoine Bonnefoy. **Estimating Expected Calibration Errors**. In *International Conference on Artificial Neural Networks*, pages 139–150. Springer, 2021.

Charles Poynton. **Digital video and HD: Algorithms and Interfaces**. Elsevier, 2012.

PyTorchLightning. **All TorchMetrics**. `https://torchmetrics.readthedocs.io/en/stable/metrics.html`, 2020. (Accessed on 05/02/2022).

Joaquin Quinonero-Candela, Carl Edward Rasmussen, Fabian Sinz, Olivier Bousquet, and Bernhard Schölkopf. **Evaluating predictive uncertainty challenge**. In *Machine Learning Challenges Workshop*, pages 1–27. Springer, 2005.

Tim Rädsch, Annika Reinke, Vivienn Weru, Minu D Tizabi, Nicholas Schreck, A Emre Kavur, Bünyamin Pekdemir, Tobias Roß, Annette Kopp-Schneider, and Lena Maier-Hein. **Labeling instructions matter in biomedical image analysis**. *arXiv preprint arXiv:2207.09899*, 2022.

Sarunas J Raudys, Anil K Jain, et al. **Small sample size effects in statistical pattern recognition: Recommendations for practitioners**. *IEEE Transactions on pattern analysis and machine intelligence*, 13(3):252–264, 1991.

A Ramaswamy Reddy, EV Prasad, and LSS Reddy. **Abnormality detection of brain mr image segmentation using iterative conditional mode algorithm**. *Int J Appl Inform Syst*, 5(2):56–66, 2013.

Annika Reinke. **Bring the model to the data: The Deep Learning Epilepsy Detection Challenge**. *EBioMedicine*, 66, 2021.

Annika Reinke, Matthias Eisenmann, Sinan Onogur, Marko Stankovic, Patrick Scholz, Peter M Full, Hrvoje Bogunovic, Bennett A Landman, Oskar Maier, Bjoern Menze, Gregory C Sharp, Korsuk Sirinukunwattana, Stefanie Speidel, Fons Van der Sommen, Guoyan Zheng, Henning Müller, Michal Kozubek, Tal Arbel, Andrew P Bradley, Pierre Jannin, Annette Kopp-Schneider, and Lena Maier-Hein. **How to exploit weaknesses in biomedical challenge design and organization**. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 388–395. Springer, 2018a. Shared first authors: Annika Reinke, Matthias Eisenmann.

Annika Reinke, Sinan Onogur, Matthias Eisenmann, and Lena Maier-Hein. **Structured Challenge Submission System**. `https://www.biomedical-challenges.org/`, 2018b. (Accessed on 11/18/2022).

Annika Reinke, Matthias Eisenmann, Minu D Tizabi, Carole H Sudre, Tim Rädsch, Michela Antonelli, Tal Arbel, Spyridon Bakas, M Jorge Cardoso, Veronika Cheplygina, Keyvan Farahani, Ben Glocker, Doreen Heckmann-Nötzel, Fabian Isensee, Pierre Jannin, Charles Kahn, Jens Kleesiek, Tahsin Kurc, Michal Kozubek, Bennett A Landman, Geert Litjens, Klaus Maier-Hein, Anne L Martel, Henning Müller, Jens Petersen, Mauricio Reyes, Nicola Rieke, Bram Stieltjes, Ronald M Summers, Sotirios A Tsaftaris, Bram van Ginneken, Annette Kopp-Schneider, Paul Jäger, and Lena Maier-Hein. **Common limitations of image processing metrics: A picture story**. *arXiv preprint arXiv:2104.05642*, 2021a.

Annika Reinke, Matthias Eisenmann, Minu Dietlinde Tizabi, Carole H Sudre, Tim Rädsch, Michela Antonelli, Tal Arbel, Spyridon Bakas, M Jorge Cardoso, Veronika Cheplygina, et al. **Common limitations of performance metrics in biomedical image analysis**. In *Medical Imaging with Deep Learning*, 2021b.

Annika Reinke, Minu D Tizabi, Matthias Eisenmann, and Lena Maier-Hein. **Common Pitfalls and Recommendations for Grand Challenges in Medical Artificial Intelligence**. *European Urology Focus*, 2021c.

Annika Reinke, Lena Maier-Hein, Evangelia Christodoulou, Ben Glocker, Patrick Scholz, Fabian Isensee, Jens Kleesiek, Michal Kozubek, Mauricio Reyes, Michael Alexander Riegler, et al. **Metrics Reloaded-A new recommendation framework for biomedical image analysis validation**. In *Medical Imaging with Deep Learning*, 2022a. Shared first authors: Annika Reinke and Lena Maier-Hein.

Annika Reinke, Lena Maier-Hein, Patrick Godau, Evangelia Christodoulou, Ben Glocker, Patrick Scholz, Fabian Isensee, Jens Kleesiek, Michal Kozubek, Mauricio Reyes, Michael Alexander Riegler, et al. **Metrics Reloaded**. In *Medical Imaging meets NeurIPS*, 2022b. Shared first authors: Annika Reinke and Lena Maier-Hein.

Torsten Rohlfing, Natalie M Zahr, Edith V Sullivan, and Adolf Pfefferbaum. **The SRI24 multichannel atlas of normal adult human brain structure**. *Human brain mapping*, 31(5):798–819, 2010.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. **U-net: Convolutional networks for biomedical image segmentation**. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

Frank Rosenblatt. **The perceptron: a probabilistic model for information storage and organization in the brain.** *Psychological review*, 65(6):386, 1958.

Azriel Rosenfeld and John L Pfaltz. **Sequential operations in digital picture processing**. *Journal of the ACM (JACM)*, 13(4):471–494, 1966.

Tobias Roß, Pierangela Bruno, Annika Reinke, Manuel Wiesenfarth, Lisa Koeppel, Peter M Full, Bünyamin Pekdemir, Patrick Godau, Darya Trofimova, Fabian Isensee, Sara Moccia, Francesco Calimeri, Beat P Müller-Stich, Annette Kopp-Schneider, and Lena Maier-Hein. **How can we learn (more) from challenges? A statistical approach to driving future algorithm development**. *arXiv preprint arXiv:2106.09302*, 2021. Shared first authors: Tobias Roß and Pierangela Bruno.

Antoine Rosset, Luca Spadola, and Osman Ratib. **OsiriX: an open-source software for navigating in multidimensional DICOM images**. *Journal of digital imaging*, 17(3):205–216, 2004.

Holger R Roth, Ziyue Xu, Carlos Tor Diez, Ramon Sanchez Jacob, Jonathan Zember, Jose Molto, Wenqi Li, Sheng Xu, Baris Turkbey, Evrim Turkbey, et al. **Rapid artificial intelligence solutions in a pandemic—The COVID-19-20 Lung CT Lesion Segmentation Challenge**. *Medical Image Analysis*, page 102605, 2022.

Kenneth J Rothman. **Epidemiology: an introduction**. Oxford university press, 2012.

Abhijit Guha Roy, Sailesh Conjeti, Nassir Navab, Christian Wachinger, Alzheimer's Disease Neuroimaging Initiative, et al. **QuickNAT: A fully convolutional network for quick and accurate segmentation of neuroanatomy**. *NeuroImage*, 186:713–727, 2019.

Tobias/Annika Roß/Reinke, Peter M. Full, Martin Wagner, Hannes Kenngott, Martin Apitz, Hellena Hempe, Diana Mindroc Filimon, Patrick Scholz, Thuy Nuong Tran, Pierangela Bruno, Pablo Arbeláez, Gui-Bin Bian, Sebastian Bodenstedt, Jon Lindström Bolmgren, Laura Bravo-Sánchez, Hua-Bin Chen, Cristina González, Dong Guo, Pål Halvorsen, Pheng-Ann Heng, Enes Hosgor, Zeng-Guang Hou, Fabian Isensee, Debesh Jha, Tingting Jiang, Yueming Jin, Kadir Kirtac, Sabrina Kletz, Stefan Leger, Zhixuan Li, Klaus H. Maier-Hein, Zhen-Liang Ni, Michael A. Riegler, Klaus Schoeffmann, Ruohua Shi, Stefanie Speidel, Michael Stenzel, Isabell Twick, Gutai Wang, Jiacheng Wang, Liansheng Wang, Lu Wang, Yujie Zhang, Yan-Jie Zhou, Lei Zhu, Manuel Wiesenfarth, Annette Kopp-Schneider, Beat P. Müller-Stich, and Lena Maier-Hein. **Comparative validation of multi-instance instrument segmentation in endoscopy: results of the ROBUST-MIS 2019 challenge**. *Medical Image Analysis*, page 101920, 2020. ISSN 1361-8415. doi: https://doi.org/10.1016/j.media.2020.101920. Shared first authors: Tobias Roß, Annika Reinke.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. **Imagenet large scale visual recognition challenge**. *International journal of computer vision*, 115(3):211–252, 2015.

Stuart J Russell and Peter Norvig. **Artificial Intelligence: a modern approach**. Pearson, 4 edition, 2020.

Anindo Saha, Joeran Bosma, Jasper Linmans, Matin Hosseinzadeh, and Henkjan Huisman. **Anatomical and Diagnostic Bayesian Segmentation in Prostate MRI – Should Different Clinical Objectives Mandate Different Loss Functions?** *arXiv preprint arXiv:2110.12889*, 2021.

Cristina Sánchez-Montes, Francisco Javier Sánchez, Jorge Bernal, Henry Córdova, María López-Cerón, Miriam Cuatrecasas, Cristina Rodríguez De Miguel, Ana García-Rodríguez, Rodrigo Garcés-Durán, María Pellisé, et al. **Computer-aided prediction of polyp histology on white light colonoscopy using surface pattern analysis**. *Endoscopy*, 51(03):261–265, 2019.

Melanie Schellenberg, Kris K Dreher, Niklas Holzwarth, Fabian Isensee, Annika Reinke, Nicholas Schreck, Alexander Seitel, Minu D Tizabi, Lena Maier-Hein, and Janek Gröhl. **Semantic segmentation of multispectral photoacoustic images using deep learning**. *Photoacoustics*, 26:100341, 2022.

Kenneth F Schulz, Douglas G Altman, and David Moher. **CONSORT 2010 statement: updated guidelines for reporting parallel group randomised trials**. *Trials*, 11(1):1–8, 2010.

Christof Schuster. **A note on the interpretation of weighted kappa and its relations to other rater agreement statistics for metric scales**. *Educational and Psychological Measurement*, 64(2):243–253, 2004.

Scikitlearn. **Sklearn Metrics**. `https://scikit-learn.org/stable/modules/classes.html#module-sklearn.metrics`, 2007. (Accessed on 05/02/2022).

Anjany Sekuboyina, Malek E Husseini, Amirhossein Bayat, Maximilian Löffler, Hans Liebl, Hongwei Li, Giles Tetteh, Jan Kukačka, Christian Payer, Darko Štern, et al. **VerSe: a vertebrae labelling and segmentation benchmark for multi-detector CT images**. *Medical image analysis*, 73:102166, 2021.

Arnaud Arindra Adiyoso Setio, Alberto Traverso, Thomas De Bel, Moira SN Berens, Cas Van Den Bogaard, Piergiorgio Cerello, Hao Chen, Qi Dou, Maria Evelina Fantacci, Bram Geurts, et al. **Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: the LUNA16 challenge**. *Medical image analysis*, 42:1–13, 2017.

Hoo-Chang Shin, Holger R Roth, Mingchen Gao, Le Lu, Ziyue Xu, Isabella Nogues, Jianhua Yao, Daniel Mollura, and Ronald M Summers. **Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning**. *IEEE transactions on medical imaging*, 35(5):1285–1298, 2016.

Suprosanna Shit, Johannes C Paetzold, Anjany Sekuboyina, Ivan Ezhov, Alexander Unger, Andrey Zhylka, Josien PW Pluim, Ulrich Bauer, and Bjoern H Menze. **clDice-a novel topology-preserving loss function for tubular structure segmentation**. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16560–16569, 2021.

C Shorten and TM Khoshgoftaar. **A survey on Image Data Augmentation for Deep Learning**. *Journal of Big Data*, 6:60, 2019.

Mohamed M Shoukri, MH Asyali, and A Donner. **Sample size requirements for the design of reliability study: review and new results**. *Statistical methods in medical research*, 13(4):251–271, 2004.

Julius Sim and Chris C Wright. **The kappa statistic in reliability studies: use, interpretation, and sample size requirements**. *Physical therapy*, 85(3):257–268, 2005.

Karen Simonyan and Andrew Zisserman. **Very deep convolutional networks for large-scale image recognition**. *arXiv preprint arXiv:1409.1556*, 2014.

Amber L Simpson, Michela Antonelli, Spyridon Bakas, Michel Bilello, Keyvan Farahani, Bram Van Ginneken, Annette Kopp-Schneider, Bennett A Landman, Geert Litjens, Bjoern Menze, et al. **A large annotated medical image dataset for the development and evaluation of segmentation algorithms**. *arXiv preprint arXiv:1902.09063*, 2019.

Ana-Maria Šimundić. **Measures of diagnostic accuracy: basic definitions**. *ejifcc*, 19(4):203, 2009.

Korsuk Sirinukunwattana, Josien PW Pluim, Hao Chen, Xiaojuan Qi, Pheng-Ann Heng, Yun Bo Guo, Li Yang Wang, Bogdan J Matuszewski, Elia Bruni, Urko Sanchez, et al. **Gland segmentation in colon histology images: The glas challenge contest**. *Medical image analysis*, 35:489–502, 2017.

Roger D Soberanis-Mukul, Shadi Albarqouni, and Nassir Navab. **Polyp-artifact relationship analysis using graph inductive learned representations**. *arXiv preprint arXiv:2009.07109*, 2020.

Charles Spearman. **The proof and measurement of association between two things**. *The American journal of psychology*, 100(3/4):441–471, 1987.

Stephen V Stehman. **Selecting and interpreting measures of thematic classification accuracy**. *Remote sensing of Environment*, 62(1):77–89, 1997.

Robert Stine. **An introduction to bootstrap methods: Examples and ideas**. *Sociological Methods & Research*, 18(2-3):243–291, 1989.

Mervyn Stone. **Cross-validatory choice and assessment of statistical predictions**. *Journal of the royal statistical society: Series B (Methodological)*, 36(2):111–133, 1974.

Cecilia Summers and Michael J Dinneen. **Nondeterminism and instability in neural network optimization**. In *International Conference on Machine Learning*, pages 9913–9922. PMLR, 2021.

Shiliang Sun, Zehui Cao, Han Zhu, and Jing Zhao. **A survey of optimization methods from a machine learning perspective**. *IEEE transactions on cybernetics*, 50(8):3668–3681, 2019.

Joshua KY Swee and Saša Grbić. **Advanced transcatheter aortic valve implantation (TAVI) planning from CT with ShapeForest**. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 17–24. Springer, 2014.

Eulalia Szmidt and Janusz Kacprzyk. **The Spearman and Kendall rank correlation coefficients between intuitionistic fuzzy sets.** In *EUSFLAT Conf.*, pages 521–528, 2011.

Abdel Aziz Taha and Allan Hanbury. **Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool**. *BMC medical imaging*, 15(1):1–28, 2015.

Christoph Tauchert, Peter Buxmann, and Jannis Lambinus. **Crowdsourcing Data Science: A Qualitative Analysis of Organizations' Usage of Kaggle Competitions**. In *Proceedings of the 53rd Hawaii International Conference on System Sciences*, 2020.

Alaa Tharwat. **Classification assessment methods**. *Applied Computing and Informatics*, 2020.

Henry C Thode. **Testing for normality**. CRC press, 2002.

Kimberley M Timmins, Irene C van der Schaaf, Edwin Bennink, Ynte M Ruigrok, Xingle An, Michael Baumgartner, Pascal Bourdon, Riccardo De Feo, Tommaso Di Noto, Florian Dubost, et al. **Comparing methods of detecting and segmenting unruptured intracranial aneurysms on TOF-MRAS: The ADAM challenge**. *Neuroimage*, 238:118216, 2021.

Laszlo Tirian and Barry J Dickson. **The VT GAL4, LexA, and split-GAL4 driver line collections for targeted expression in the Drosophila nervous system**. *BioRxiv*, page 198648, 2017.

Torch. **PyTorch: Models and pre-trained weights**. `https://pytorch.org/vision/stable/models.html`, 2017. (Accessed on 05/24/2022).

Antonio Torralba and Alexei A Efros. **Unbiased look at dataset bias**. In *CVPR 2011*, pages 1521–1528. IEEE, 2011.

Thuy N Tran, Tim Adler, Amine Yamlahi, Evangelia Christodoulou, Patrick Godau, Annika Reinke, Minu D Tizabi, Peter Sauer, Tillmann Persicke, Jörg G. Albert, and Lena Maier-Hein. **Sources of performance variability in deep learning-based polyp detection**. *arXiv preprint arXiv:2211.09708*, 2022.

Vladimír Ulman, Martin Maška, Klas EG Magnusson, Olaf Ronneberger, Carsten Haubold, Nathalie Harder, Pavel Matula, Petr Matula, David Svoboda, Miroslav Radojevic, et al. **An objective comparison of cell-tracking algorithms**. *Nature methods*, 14(12):1141–1152, 2017.

Uchila N Umesh, Robert A Peterson, and Matthew H Sauber. **Interjudge agreement and the maximum value of kappa**. *Educational and Psychological Measurement*, 49(4):835–850, 1989.

Femke Vaassen, Colien Hazelaar, Ana Vaniqui, Mark Gooding, Brent van der Heyden, Richard Canters, and Wouter van Elmpt. **Evaluation of measures for assessing time-saving of automatic organ-at-risk segmentation in radiotherapy**. *Physics and Imaging in Radiation Oncology*, 13:1–6, 2020.

Juozas Vaicenavicius, David Widmann, Carl Andersson, Fredrik Lindsten, Jacob Roll, and Thomas Schön. **Evaluating model calibration in classification**. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 3459–3467. PMLR, 2019.

Bram Van Ginneken, Tobias Heimann, and Martin Styner. **3D segmentation in the clinic: A grand challenge**. In *MICCAI workshop on 3D segmentation in the clinic: a grand challenge*, volume 1, pages 7–15, 2007.

Bram Van Ginneken, Samuel G Armato III, Bartjan de Hoop, Saskia van Amelsvoort-van de Vorst, Thomas Duindam, Meindert Niemeijer, Keelin Murphy, Arnold Schilham, Alessandra Retico, Maria Evelina Fantacci, et al. **Comparing and combining algorithms for computer-aided detection of pulmonary nodules in computed tomography scans: the ANODE09 study**. *Medical image analysis*, 14(6):707–722, 2010.

Jan P Vandenbroucke, Erik von Elm, Douglas G Altman, Peter C Gøtzsche, Cynthia D Mulrow, Stuart J Pocock, Charles Poole, James J Schlesselman, Matthias Egger, and Strobe Initiative. **Strengthening the Reporting of Observational Studies in Epidemiology (STROBE): explanation and elaboration**. *Annals of internal medicine*, 147(8):W–163, 2007.

Gaël Varoquaux and Veronika Cheplygina. **Machine learning for medical imaging: methodological failures and recommendations for the future**. *NPJ digital medicine*, 5(1):1–8, 2022.

Gaël Varoquaux and Olivier Colliot. **Evaluating machine learning models and their diagnostic value**, 2022.

Baptiste Vasey, Myura Nagendran, Bruce Campbell, David A Clifton, Gary S Collins, Spiros Denaxas, Alastair K Denniston, Livia Faes, Bart Geerts, Mudathir Ibrahim, et al. **Reporting guideline for the early stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI**. *bmj*, 377, 2022.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. **Attention is all you need**. *Advances in neural information processing systems*, 30, 2017.

Andrew J Vickers and Elena B Elkin. **Decision curve analysis: a novel method for evaluating prediction models**. *Medical Decision Making*, 26(6):565–574, 2006.

Andrew J Vickers, Ben Van Calster, and Ewout W Steyerberg. **Net benefit approaches to the evaluation of prediction models, molecular markers, and diagnostic tests**. *bmj*, 352, 2016.

Luis Von Ahn and Laura Dabbish. **Labeling images with a computer game**. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 319–326, 2004.

Martin Wagner, Beat-Peter Müller-Stich, Anna Kisilenko, Duc Tran, Patrick Heger, Lars Mündermann, David M Lubotsky, Benjamin Müller, Tornike Davitashvili, Manuela Capek, Annika Reinke, Armine Vardazaryan Tong Yu, Chinedu Innocent Nwoye, Nicolas Padoy, Xinyang Liu, Eung-Joo Lee, Constantin Disch, Hans Meine, Tong Xia, Fucang Jia, Satoshi Kondo, Wolfgang Reiter, Yueming Jin, Yonghao Long, Meirui Jiang, Qi Dou, Pheng Ann Heng, Isabell Twick, Kadir Kirtac, Enes Hosgor, Jon Lindström Bolmgren, Michael Stenzel, Björn von Siemens, Hannes G Kenngott, Felix Nickel, Moritz von Frankenberg, Franziska Mathis-Ullrich, Lena Maier-Hein, Stefanie Speidel, and Sebastian Bodenstedt. **Comparative Validation of Machine Learning Algorithms for Surgical Workflow and Skill Analysis with the HeiChole Benchmark**. *arXiv preprint arXiv:2109.14956*, 2021.

Ching-Wei Wang, Cheng-Ta Huang, Meng-Che Hsieh, Chung-Hsing Li, Sheng-Wei Chang, Wei-Cheng Li, Rémy Vandaele, Raphaël Marée, Sébastien Jodogne, Pierre Geurts, et al. **Evaluation and comparison of anatomical landmark detection methods for cephalometric x-ray images: a grand challenge**. *IEEE transactions on medical imaging*, 34(9):1890–1900, 2015.

Ziheng Wang and Ann Majewicz Fey. **Deep learning with convolutional neural network for objective skill evaluation in robot-assisted surgery**. *International journal of computer assisted radiology and surgery*, 13(12):1959–1970, 2018.

Matthijs J Warrens. **Some paradoxical results for the quadratically weighted kappa**. *Psychometrika*, 77(2): 315–323, 2012.

Michael Weller, Martin Van Den Bent, Kirsten Hopkins, Jörg C Tonn, Roger Stupp, Andrea Falini, Elizabeth Cohen-Jonathan-Moyal, Didier Frappaz, Roger Henriksson, Carmen Balana, et al. **EANO guideline for the diagnosis and treatment of anaplastic gliomas and glioblastoma**. *The lancet oncology*, 15(9):e395–e403, 2014.

Brady T West, Kathleen B Welch, and Andrzej T Galecki. **Linear mixed models: a practical guide using statistical software**. Chapman and Hall/CRC, 2006.

Jay West, J Michael Fitzpatrick, Matthew Y Wang, Benoit M Dawant, Calvin R Maurer Jr, Robert M Kessler, Robert J Maciunas, Christian Barillot, Didier Lemoine, Andre Collignon, et al. **Comparison and evaluation of retrospective intermodality brain image registration techniques**. *Journal of computer assisted tomography*, 21(4):554–568, 1997.

David Widmann, Fredrik Lindsten, and Dave Zachariah. **Calibration tests in multi-class classification: A unifying framework**. *Advances in Neural Information Processing Systems*, 32, 2019.

Maarit Widmann. **Cohen's Kappa: what it is, when to use it, how to avoid pitfalls**. `https://www.knime.com/blog/cohens-kappa-an-overview`, 2020. (Accessed on 06/28/2022).

Manuel Wiesenfarth, Annika Reinke, Bennett A Landman, Matthias Eisenmann, Laura Aguilera Saiz, M Jorge Cardoso, Lena Maier-Hein, and Annette Kopp-Schneider. **challengeR - Methods and open-source toolkit for analyzing and visualizing challenge results**. `https://github.com/wiesenfa/challengeR`, 2019. (Accessed on 06/14/2022).

Manuel Wiesenfarth, Annika Reinke, Bennett A Landman, Matthias Eisenmann, Laura Aguilera Saiz, M Jorge Cardoso, Lena Maier-Hein, and Annette Kopp-Schneider. **Methods and open-source toolkit for analyzing and visualizing challenge results**. *Scientific Reports*, 11(1):1–15, 2021.

Varduhi Yeghiazaryan and Irina Voiculescu. **An overview of current evaluation methods used in medical image segmentation**. *Department of Computer Science, University of Oxford*, 2015.

Varduhi Yeghiazaryan and Irina D Voiculescu. **Family of boundary overlap metrics for the evaluation of medical image segmentation**. *Journal of Medical Imaging*, 5(1):015006, 2018.

William J Youden. **Index for rating diagnostic tests**. *Cancer*, 3(1):32–35, 1950.

Paul A Yushkevich, Yang Gao, and Guido Gerig. **ITK-SNAP: An interactive tool for semi-automatic segmentation of multi-modality biomedical images**. In *2016 38th annual international conference of the IEEE engineering in medicine and biology society (EMBC)*, pages 3342–3345. IEEE, 2016.

Sonia Zachariah, William Wykes, and David Yorston. **Grading diabetic retinopathy (DR) using the Scottish grading protocol**. *Community eye health*, 28(92):72, 2015.

Deborah A Zarin and Alla Keselman. **Registering a clinical trial in ClinicalTrials. gov**. *Chest*, 131(3):909–912, 2007.

Oliver Zendel, Markus Murschitz, Martin Humenberger, and Wolfgang Herzner. **How good is my test data? Introducing safety analysis for computer vision**. *International Journal of Computer Vision*, 125(1):95–109, 2017.

Ying Zhang, Yubin Xie, Wenzhong Liu, Wankun Deng, Di Peng, Chenwei Wang, Haodong Xu, Chen Ruan, Yongjie Deng, Yaping Guo, et al. **DeepPhagy: a deep learning framework for quantitatively measuring autophagy activity in Saccharomyces cerevisiae**. *Autophagy*, 16(4):626–640, 2020.

David Zimmerer, Peter M Full, Fabian Isensee, Paul Jäger, Tim Adler, Jens Petersen, Gregor Köhler, Tobias Ross, Annika Reinke, Antanas Kascenas, et al. **MOOD 2020: A public Benchmark for Out-of-Distribution Detection and Localization on medical Images**. *IEEE Transactions on Medical Imaging*, 2022.

Kelly H Zou, William M Wells III, Ron Kikinis, and Simon K Warfield. **Three validation metrics for automated probabilistic image segmentation of brain tumours**. *Statistics in medicine*, 23(8):1259–1282, 2004.

List of Acronyms

Adam Adaptive Moment Estimation
AFRICAI African Network for Artificial Intelligence in Biomedical Imaging
AI Artificial Intelligence
AP Average Precision
ASSD Average Symmetric Surface Distance
AUC Area under the Curve
AUROC Area under the Receiver Operating Characteristic Curve
BA Balanced Accuracy
BARNet Bilinear Attention Network
BCE Binary Cross Entropy
BIAS Biomedical Image Analysis ChallengeS
BM Bookmaker Informedness
BMI Body Mass Index
Boundary IoU Boundary Intersection over Union
Box/Approx IoU Box/Approximation Intersection over Union
Box IoU Box Intersection over Union
BRATS Brain Tumor Image Segmentation
BS Brier Score
BSS Brier Skill Score
clDice Centerline Dice Similarity Coefficient
CAMELYON Cancer Metastases in Lymph Nodes Challenge
CARE Case Reporting Guidelines
CE Cross Entropy
CI Confidence Interval
CK Cohen's Kappa
CLAIM Checklist for Artificial Intelligence in Medical Imaging
COCO Common Objects in COntext
CONSORT Consolidated Standards of Reporting Trials
CNN Convolutional Neural Network
CRT-EPiggy Cardiac Resynchronization Therapy Electrophysiological
CT Computed Tomography
CWCE Class-wise Calibration Error
DAC Dense Atrous Convolution

DECIDE-AI Developmental and Exploratory Clinical Investigations of DEcision support systems driven by Artificial Intelligence
DKFZ Deutsches Krebsforschungszentrum (German Cancer Research Center)
DL Deep Learning
DOI Digital Object Identifier
DPANet Dense Pyramid Attention Network
DREAM Dialogue on Reverse Engineering Assessment and Methods
DSC Dice Similarity Coefficient
EC Expected Cost
ECE Expected Calibration Error
ECE^{KDE} Expected Calibration Error Kernel Density Estimate
ECG Electrocardiogram
ELLIS European Lab for Learning & Intelligent Systems
EQUATOR Enhancing the QUAlity and Transparency Of health Research
FLAIR Fluid-attenuated Inversion Recovery
FN False Negative
FP False Positive
FPN Feature Pyramid Network
FPPI False Positives per Image
FROC Free-Response Receiver Operating Characteristic
GlaS Gland Segmentation
GLMM Generalized Linear Mixed Model
GPU Graphics Processing Unit
GUI Graphical User Interface
HD Hausdorff Distance
HD95 Hausdorff Distance 95 Percentile
HeiCo Heidelberg Colorectal
ImLC Image-level Classification
InS Instance Segmentation
ICC Interclass Correlation
ICMJE International Committee of Medical Journal Editors
ImageCLEF Image Cross Language Evaluation Forum
IoR Intersection over Reference
IoU Intersection over Union
IQR Interquartile Range
IRB institutional review boards
IPCAI Information Processing in Computer-Assisted Interventions
ISBI IEEE International Symposium on Biomedical Imaging
ISIC International Skin Imaging Collaboration
ISLES Ischemic Stroke Lesion Segmentation
J Youden's Index
JVM Java Virtual Machine
KCE Kernel Calibration Error
KI Künstliche Intelligenz

KiTS Kidney Tumor Segmentation
LMM Linear Mixed Model
LR+ Positive Likelihood Ratio
MONAI Medical Open Network for Artificial Intelligence
NaN Not a Number
NeurIPS Neural Information Processing Systems
NLP Natural Language Processing
ML Machine Learning
MASD Mean Absolute Surface Distance
Mask IoU Mask Intersection over Union
mAP mean Average Precision
MCC Matthews Correlation Coefficient
MCE Maximum Calibration Error
MICCAI Medical Image Computing and Computer Assisted Interventions
MICR Medical Imaging and Case Reports
MIDL Medical Imaging with Deep Learning conference
ML Machine Learning
MRI Magnetic Resonance Imaging
mp-MRI Multiparametric Magnetic Resonance Imaging
MPRAGE Magnetization Prepared Rapid Acquisition Gradient Echo
MS Multiple Sclerosis
TOF-MRA Time of Flight Magnetic Resonance Angiographs
MI Mutual Information
MI DSC Multi-Instance Dice Similarity Coefficient
MI NSD Multi-Instance Normalized Surface Distance
MSD Medical Segmentation Decathlon
NB Net Benefit
NLL Negative Log Likelihood
NN Neural Network
NPV Negative Predictive Value
NSD Normalized Surface Distance
ObD Object Detection
ODB Object Database
OR Odds Ratio
PPV Positive Predictive Value
PQ Panoptic Quality
PR Precision-Recall
PSR Proper Scoring Rules
PZ Peripheral Zone
Q-Q plot Quantile-Quantile plot
RI Rand Index
RAUNet Residual attention U-Net
RBC Red Blood Cell
RBS Root Brier Score

ResNet Residual Neural Network
REST Representational State Transfer
RobustMIS Robust Medical Instrument Segmentation
ROC Receiver Operating Characteristic
ROI Region of Interest
RPN Region Proposal Network
SemS Semantic Segmentation
SGD Stochastic Gradient Descent
SIG Special Interest Group
SIIM Society for Imaging Informatics in Medicine
STARD Standards for Reporting of Diagnostic Accuracy Studies
STROBE Strengthening the Reporting of Observational studies in Epidemiology
TN True Negative
TNR True Negative Rate
TP True Positive
TCE Top-label Calibration Error
TE Echo Time
TI Inversion Time
TR Repetition Time
TPR True Positive Rate
TZ Transition Zone
US Ultrasound
VS Volumetric Similarity
WBC White Blood Cell
WCK Weighted Cohen's Kappa
WSI Whole Slide Imaging
Xth Percentile HD Xth Percentile Hausdorff Distance

List of Figures

1.1 Top: Goals and advantages of organizing challenges. Bottom: Common steps in organizing a biomedical image analysis challenge. . . . 3
1.2 Overview of contributions. . . . 6

2.1 Problem categories covered in this thesis that are illustrated for three different application domains: radiology, cell biology, and surgery . . . 9
2.2 Illustration of bootstrapping in the case of ranking variability. . . . 16
2.3 Residual block as proposed by He et al. [2016]. . . . 23
2.4 U-Net architecture as proposed by Ronneberger et al. [2015]. . . . 24
2.5 Mask R-CNN architecture as proposed by He et al. [2017]. . . . 25
2.6 Visualization of a binary confusion matrix. . . . 27
2.7 Illustration of thresholding of predicted class scores. . . . 28
2.8 Image-level counting metrics as functions of the prevalence for fixed Sensitivity and Specificity. 29
2.9 Overview of the most frequent per-class counting classification metrics. . . . 32
2.10 Overview of the most frequent multi-class counting classification metrics. . . . 36
2.11 Visualization of a confusion matrix with C classes and N samples. . . . 37
2.12 Multi-threshold metrics and per-class counting metrics with application-driven thresholds. 39
2.13 Most frequently used overlap-based metrics. . . . 42
2.14 Illustration of the Centerline Dice Similarity Coefficient (clDice), measuring the connectivity of structures. . . . 43
2.15 Boundary-based metrics that aggregate shortest distances of boundaries A and B. . . . 45
2.16 Boundary-based metrics based on hyperparameters. . . . 47
2.17 Overview of (a) center-/point-based and (b) overlap-based localization criteria. . . . 49
2.18 Localization criteria may discard spatial information and should be well motivated by the underlying task. . . . 50
2.19 Example of an assignment strategy in the case of two predictions that have been assigned to the same reference object. . . . 51
2.20 Principle of the Free-Response Receiver Operating Characteristic (FROC) Score. . . . 53
2.21 Validation at object level can be done on a per-data-set or per-image basis. . . . 54
2.22 Principle of the Panoptic Quality (PQ). . . . 55
2.23 Use cases of calibration assessment . . . 57
2.24 Illustration of different definitions of calibration for the exemplary computation of the calibration error. . . . 57

3.1 Overview and statistics of biomedical image analysis challenges. . . . 71
3.2 Overview of the results of the community survey with $n = 295$ respondents. . . . 73

4.1 Effect of exchanging parameters of the ranking designs for one example MICCAI 2015 segmentation task. . . . 81
4.2 Metric pitfalls taxonomy. . . . 87
4.3 Pitfalls related to a mismatch between the problem category and the metric. . . . 89
4.4 Pitfalls related to disregard of the domain interest: Importance of structure boundaries, center, and volume. . . . 91
4.5 Pitfalls related to disregard of the domain interest: Unequal severity of class confusions. . . . 93
4.6 Pitfalls related to disregard of the domain interest: Importance of comparability across data sets. . . . 95
4.7 Pitfalls related to disregard of the domain interest: Importance of confidence values and benefit-cost analysis. . . . 97
4.8 Pitfalls related to disregard of properties of the target structure: Small structure sizes and high variability of structure sizes. . . . 99
4.9 Pitfalls related to disregard of properties of the target structure: Complex shapes, overlapping and disconnected structures. . . . 101
4.10 Pitfalls related to disregard of properties of the data set: Class imbalance, small sample sizes. 103
4.11 Pitfalls related to disregard of properties of the data set: High inter-rater variability, spatial outliers in the reference and empty reference or prediction. . . . 105
4.12 Pitfalls related to disregard of properties of the algorithm output: Possibility of overlapping predictions and unavailability of predicted class scores. . . . 107
4.13 Pitfalls related to inadequate metric aggregation: Hierarchical structure of classes, unequal handling of classes, and per-class aggregation. . . . 109
4.14 Pitfalls related to inadequate metric aggregation: Ignoring hierarchical data structure and missing value handling. . . . 111
4.15 Pitfalls related to inadequate metric aggregation: Lack of stratification, interdependencies, and disregarding number of images. . . . 113
4.16 Pitfalls related to inadequate ranking scheme: Ignoring metric relationships and ranking variability. . . . 115
4.17 Pitfalls related to inadequate metric implementation: Non-standardized metric definition and sensitivity to hyperparameters. . . . 117
4.18 Pitfalls related to inadequate metric implementation: cutoff value, discretization issues, definition of class labels, and assessing multiple properties. . . . 119
4.19 Pitfalls related to inadequate metric implementation: Pitfalls regarding multi-threshold metrics. 121
4.20 Pitfalls related to inadequate metric implementation: Center- or point-based localization criteria. . . . 123
4.21 Pitfalls related to inadequate metric reporting. . . . 125
4.22 Pitfalls related to inadequate interpretation of metric values. . . . 127
4.23 Ranking robustness for several ranking design changes for the MICCAI 2015 segmentation task (n = 56). . . . 136
4.24 Ranking robustness shown for the MICCAI 2015 segmentation task (n = 56). . . . 137
4.25 Qualitative analysis of deficiencies in the seven method descriptions across several different aspects of implementation. . . . 149
4.26 Dots- and boxplots showing the individual algorithm performance for Stage 3 of the Robust Medical Instrument Segmentation (RobustMIS) challenge for the Multi-Instance Dice Similarity Coefficient (MI DSC). . . . 150

5.1 Overview and statistics of biomedical image analysis challenges from **2004–2016** and **2018–2022**. 158
5.2 Overview of the metric recommendation framework. . . . 165
5.3 Mapping M1: Problem category mapping to identify the problem fingerprint F1.1. . . . 167
5.4 Mapping M2: Multi-class counting metrics. . . . 178
5.5 Mapping M3 (Part 1): Per-class counting metrics. . . . 179
5.6 Mapping M3 (Part 2): Per-class counting metrics. . . . 180

5.7 Mapping M4: Multi-threshold metrics. 181
5.8 Mapping M5: Calibration metrics. 182
5.9 Mapping M6: Overlap-based metrics. 183
5.10 Mapping M7: Boundary-based metrics. 184
5.11 Mapping M8: Localization criteria. 185
5.12 Mapping M9: Assignment strategies. 186
5.13 Overview of the employed visualization techniques. 211
5.14 Overview of the Robust Medical Instrument Segmentation (RobustMIS) challenge. 215
5.15 Overview of the Medical Segmentation Decathlon (MSD) challenge. 218
5.16 Dots- and boxplots illustrating the individual algorithm performances for the simulated challenge for the best-case, fully random, and worst-case simulations. 220
5.17 Podium plots illustrating the individual algorithm performances for the simulated challenge for the best-case, fully random, and worst-case simulations. 221
5.18 Ranking heatmaps for the simulated challenge for the best-case, fully random, and worst-case simulations. 222
5.19 Blob plots visualizing ranking uncertainty over 1,000 bootstrap samples for the simulated challenge for the best-case, fully random, and worst-case simulations. 223
5.20 Violin plots visualizing Kendall's τ values over 1,000 bootstrap samples for the simulated challenge for the best-case, fully random, and worst-case simulations. 224
5.21 Significance maps visualizing pairwise significant test results for a Wilcoxon signed rank test for the simulated challenge for the best-case, fully random, and worst-case simulations. . . 224
5.22 Lineplots illustrating the ranking robustness across four different ranking methods for the simulated challenge for the best-case, fully random, and worst-case simulations. 225
5.23 Dots- and boxplots illustrating the individual algorithm performances for Stage 3 of the Robust Medical Instrument Segmentation (RobustMIS) challenge for the binary and instrument instance segmentation tasks. 228
5.24 Ranking heatmaps for Stage 3 of the Robust Medical Instrument Segmentation (RobustMIS) challenge for the binary and instrument instance segmentation tasks. 229
5.25 Blob plots visualizing ranking uncertainty over 1,000 bootstrap samples for Stage 3 of the Robust Medical Instrument Segmentation (RobustMIS) challenge for the binary and instrument instance segmentation tasks and the accuracy and robustness rankings. 231
5.26 Significance maps visualizing pairwise significant test results for a Wilcoxon signed rank test for Stage 3 of the Robust Medical Instrument Segmentation (RobustMIS) challenge for the binary and instrument instance segmentation tasks. 232
5.27 Dots- and boxplots for all stages of the mean performance of every participating team of the Robust Medical Instrument Segmentation (RobustMIS) challenge for the binary and instrument instance segmentation tasks. 233
5.28 Blob plots visualizing ranking uncertainty over 1,000 bootstrap samples for all stages separately for every participating team of the Robust Medical Instrument Segmentation (RobustMIS) challenge for the instrument instance segmentation task. 233
5.29 Dots- and boxplots illustrating the individual algorithm performances for the development phase of the Medical Segmentation Decathlon (MSD) challenge. 235
5.30 Dots- and boxplots illustrating the individual algorithm performances for the mystery phase of the Medical Segmentation Decathlon (MSD) challenge. 236
5.31 Lineplots illustrating the ranking robustness across four different ranking methods for the hepatic vessel data set for the Medical Segmentation Decathlon (MSD) challenge 237
5.32 Stacked frequency plot illustrating the ranking uncertainty for all data sets of the Medical Segmentation Decathlon (MSD) challenge. Rankings were based on the Dice Similarity Coefficient (DSC). 238
5.33 Dendrogram illustrating the similarity of data sets for the Medical Segmentation Decathlon (MSD) challenge. 239
5.34 Workflow diagram for applying the suggested mixed model analysis to challenge results. . . 245

5.35 Percentage of parameters reported on the challenge websites as stated in the accepted challenge proposal, not reported on the challenge websites, and reported differently from the proposal. . . . 247
5.36 Overview of significant positive and negative effects of image characteristics for the Positive Predictive Value (PPV) and Sensitivity for the top five algorithms of the Robust Medical Instrument Segmentation (RobustMIS) challenge simultaneously. . . . 252

A.1 Profile of the Sensitivity. . . . 314
A.2 Profile of the Specificity. . . . 315
A.3 Profile of the Positive Predictive Value (PPV). . . . 316
A.4 Profile of the Negative Predictive Value (NPV). . . . 317
A.5 Profile of the Positive Likelihood Ratio (LR+). . . . 318
A.6 Profile of the F_β Score. . . . 319
A.7 Profile of the Net Benefit (NB). . . . 320
A.8 Profile of the Panoptic Quality (PQ). . . . 321
A.9 Profile of the False Positives per Image (FPPI). . . . 322
A.10 Profile of the Accuracy. . . . 323
A.11 Profile of the Balanced Accuracy (BA). . . . 324
A.12 Profile of the Matthews Correlation Coefficient (MCC). . . . 325
A.13 Profile of Weighted Cohen's Kappa (WCK). . . . 326
A.14 Profile of the Expected Cost (EC). . . . 327
A.15 Profile of Area under the Receiver Operating Characteristic Curve (AUROC). . . . 328
A.16 Profile of Average Precision (AP). . . . 329
A.17 Profile of FROC Score. . . . 330
A.18 Profile of the Dice Similarity Coefficient (DSC). . . . 331
A.19 Profile of the Intersection over Union (IoU). . . . 332
A.20 Profile of the Centerline Dice Similarity Coefficient (clDice). . . . 333
A.21 Profile of the Intersection over Reference (IoR). . . . 334
A.22 Profile of the Hausdorff Distance (HD). . . . 335
A.23 Profile of the Hausdorff Distance 95 Percentile (HD95). . . . 336
A.24 Profile of the Average Symmetric Surface Distance (ASSD). . . . 337
A.25 Profile of the Mean Absolute Surface Distance (MASD). . . . 338
A.26 Profile of the Normalized Surface Distance (NSD). . . . 339
A.27 Profile of the Boundary Intersection over Union (Boundary IoU). . . . 340
A.28 Profile of the Center Distance. . . . 341
A.29 Profile of the Point inside Mask/ Box/ Approximation criterion. . . . 342
A.30 Profile of the Brier Score (BS). . . . 343
A.31 Profile of the Negative Log Likelihood (NLL). . . . 344
A.32 Profile of the Root Brier Score (RBS). . . . 345
A.33 Profile of the Expected Calibration Error (ECE). . . . 346
A.34 Profile of the Class-wise Calibration Error (CWCE). . . . 347
A.35 Profile of the Kernel Calibration Error (KCE). . . . 348
A.36 Profile of the Expected Calibration Error Kernel Density Estimate (ECE^{KDE}). . . . 349
A.37 Instantiation of the metric recommendation framework for common biomedical image classification use cases. . . . 351
A.38 Instantiation of the metric recommendation framework for common biomedical semantic segmentation use cases. . . . 352
A.39 Instantiation of the metric recommendation framework for common biomedical object detection use cases. . . . 353
A.40 Instantiation of the metric recommendation framework for common biomedical instance segmentation use cases. . . . 354

List of Tables

4.1 Overview of pitfalls related to image-level classification and calibration metrics. 128
4.2 Overview of pitfalls related to segmentation metrics. 129
4.3 Overview of pitfalls related to object detection metrics. 130
4.4 Summary of the most important design choices that could be retrieved from the method descriptions for the Robust Medical Instrument Segmentation (RobustMIS) challenge participants. Used abbreviations: Adaptive Moment Estimation (Adam), Binary Cross Entropy (BCE), Cross Entropy (CE), Dice Similarity Coefficient (DSC), Residual Neural Network (ResNet), Region Proposal Network (RPN), Stochastic Gradient Descent (SGD). A detailed description of all methods and assumptions drawn from the descriptions is provided in Appendix A.5. Table adapted from Roß/Reinke et al. [2020]. 144
4.5 List of parameters important for reporting a biomedical image analysis challenge, including the coverage of 549 tasks reporting each parameter in percent. 146
4.6 Overview of shortcomings of method descriptions from the participating teams of the Robust Medical Instrument Segmentation (RobustMIS) challenge. 147
4.7 Accuracy rankings for Stage 3 of the Robust Medical Instrument Segmentation (RobustMIS) challenge of the seven competing teams based on the Multi-Instance Dice Similarity Coefficient (MI DSC). 150
4.8 Robustness rankings for Stage 3 of the Robust Medical Instrument Segmentation (RobustMIS) challenge of the seven competing teams based on the 5% quantile (Q5) of the Multi-Instance Dice Similarity Coefficient (MI DSC). 151

5.1 Problem fingerprint for the general problem category (Image-level Classification (ImLC), Semantic Segmentation (SemS), Object Detection (ObD), Instance Segmentation (InS)). Each problem fingerprint comes with an identifier (ID), an illustration, and a description. 166
5.2 Problem fingerprint for domain interest problem characteristics. 168
5.3 Problem fingerprint for target structure problem characteristics. 173
5.4 Problem fingerprint for data set problem characteristics. 174
5.5 Problem fingerprint for algorithm output problem characteristics. 175
5.6 Decision guide 2.1 for choosing between Weighted Cohen's Kappa (WCK) and the Expected Cost (EC) as a multi-class counting metric. 187
5.7 Decision guide 2.2 for choosing between the Balanced Accuracy (BA) and the Expected Cost (EC) as a per-class counting metric. 187
5.8 Decision guide 2.3 for choosing between Balanced Accuracy (BA), Matthews Correlation Coefficient (MCC), and the normalized variant of the Expected Cost (EC) as a multi-class counting metric. 188
5.9 Decision guide 3.2 for choosing between the Net Benefit (NB) and the Expected Cost (EC) as a per-class counting metric. 190
5.10 Decision guide 3.3 for choosing between the LR+ and Sensitivity as a per-class counting metric. 191

5.11 Decision guide 3.4 for choosing between the LR+, Sensitivity, and the F_β Score and the Panoptic Quality (PQ) as a per-class counting metric. . . . 192
5.12 Decision guide 3.5 for determining the hyperparameter β of the F_β Score as a per-class counting metric. . . . 192
5.13 Decision guide 3.6 for choosing between the F_β Score and the Panoptic Quality (PQ) as a per-class counting metric. . . . 193
5.14 Decision guide 4.1 for choosing between Average Precision (AP) and the Area under the Receiver Operating Characteristic Curve (AUROC) as a multi-threshold metric. . . . 193
5.15 Decision guide 4.2 for choosing between Average Precision (AP) and the Free-Response Receiver Operating Characteristic (FROC) Score as a multi-threshold metric. . . . 194
5.16 Decision guide 5.1 for choosing between Brier Score (BS) and Negative Log Likelihood (NLL) as a calibration metric. . . . 195
5.17 Decision guide 5.2 for choosing between top-label assessment (Expected Calibration Error (ECE)) or focus on all predicted class scores (Class-wise Calibration Error (CWCE) and Expected Calibration Error Kernel Density Estimate (ECE^{KDE})) for calibration assessment. . . . 196
5.18 Decision guide 6.1 for choosing between the Dice Similarity Coefficient (DSC) and Intersection over Union (IoU) as an overlap-based metric. . . . 196
5.19 Decision guide 6.2 for determining the hyperparameter β of the F_β Score as an overlap-based metric. . . . 197
5.20 Decision guide 7.1 for choosing between the Boundary Intersection over Union (Boundary IoU) and Normalized Surface Distance (NSD) as a boundary-based metric. . . . 197
5.21 Decision guide 7.2 for choosing between the Average Symmetric Surface Distance (ASSD) and Mean Absolute Surface Distance (MASD) as a boundary-based metric. . . . 198
5.22 Decision guide 7.3 for choosing between the Hausdorff Distance (HD) and the X^{th} Percentile Hausdorff Distance (X^{th} Percentile HD) as a boundary-based metric. . . . 198
5.23 Decision guide 8.1 for choosing between the Boundary IoU, Mask Intersection over Union (Mask IoU), and IoR as a localization criterion. . . . 199
5.24 Decision guide 8.2 for choosing between the Center Distance, the Point inside Mask/Box/Approximation (Approx), and the Mask Intersection over Union (Mask IoU) > 0.5. as a localization criterion. . . . 200
5.25 Decision guide 8.3 for determining the localization threshold. . . . 201
5.26 Decision guide 9.1 for choosing between the Greedy (by "localization criterion" Matching, the Optimal (Hungarian) Matching and the Matching via "localization criterion" > 0.5 as an assignment strategy 201
5.27 Overview of how pitfalls were addressed in the metrics recommendation framework. . . . 202
5.28 Overview of metric application recommendations. . . . 204
5.29 Rankings for the three tasks of the simulated challenge. Rankings are provided for a generic metric bounded between 0 and 1 for a metric-based aggregation with the mean. . . . 220
5.30 Rankings for Stage 3 of the binary segmentation task of the Robust Medical Instrument Segmentation (RobustMIS) challenge. Rankings are provided for the Dice Similarity Coefficient (DSC) (top) and the Normalized Surface Distance (NSD) (bottom) metrics. Accuracy rankings are shown on the left and are based on the proportion of significant tests divided by the number of algorithms (Prop. Sign). The robustness rankings are shown on the right and are based on the 5% quantile (Q5) of the DSC/NSD. . . . 226
5.31 Ranking for Stage 3 of the instrument detection task of the Robust Medical Instrument Segmentation (RobustMIS) challenge for the F_1 Score. . . . 227

5.32 Rankings for Stage 3 of the instrument instance segmentation task of the Robust Medical Instrument Segmentation (RobustMIS) challenge. Rankings are provided for the Multi-Instance Dice Similarity Coefficient (MI DSC) (top) and the Multi-Instance Normalized Surface Distance (MI NSD) (bottom) metrics. Accuracy rankings are shown on the left and are based on the proportion of significant tests divided by the number of algorithms (Prop. Sign). The robustness rankings are shown on the right and are based on the 5% quantile (Q5) of the MI DSC/MI NSD. 227
5.33 Descriptive statistics of Kendall's τ for comparing the original ranking with 1,000 bootstrapped rankings for Stage 3 of the Robust Medical Instrument Segmentation (RobustMIS) challenge. Mean, Median, 25% Quartile (Q25) and 75% Quartile (Q75) of Kendall's τ are provided for different ranking schemes for the Dice Similarity Coefficient (DSC), Multi-Instance Dice Similarity Coefficient (MI DSC), Normalized Surface Distance (NSD) and Multi-Instance Dice Similarity Coefficient (MI DSC) metrics. 230
5.34 Test-based rankings for the development and mystery phases of the Medical Segmentation Decathlon (MSD) challenge. For every algorithm, the median and Interquartile Range (IQR) of the Dice Similarity Coefficient (DSC) values of all 19 participating teams are provided. 234
5.35 Clinical trial registry purposes and benefits for various groups. Table courtesy of Zarin and Keselman [2007]. 244
5.36 List of **instrument-related image characteristics** that were identified as failure sources in the Robust Medical Instrument Segmentation (RobustMIS) challenge. 253
5.37 List of **background-related image characteristics** that were identified as failure sources in the Robust Medical Instrument Segmentation (RobustMIS) challenge. 255
5.38 List of **global image characteristics** that were identified as failure sources in the Robust Medical Instrument Segmentation (RobustMIS) challenge. 256
5.39 Overview of significant positive and negative effects of image characteristics for the Positive Predictive Value (PPV) and Sensitivity. 259

A.1 List of mixed model analysis-related search phrases, together with how often they appeared in the 5,390 papers that were submitted to the Medical Image Computing and Computer Assisted Interventions (MICCAI) 2004—2021 conferences. 375

A | Appendix

A.1 Contributions

The goal of this section is to emphasize my contributions and set them apart from those of the broader team or consortia. For every section of Chapters 3 - 5, this section provides a disclosure with the acknowledgements of the main team members working on the respective results.

Most of my work was carried out together with large international consortia to reflect the general opinion of the research community. Large parts of the presented work would not have been possible without the consensus agreement from the broad community. Those consortia were primarily led by myself, Prof. Dr. Lena Maier-Hein, and other core team members of the respective research. Furthermore, I was assisted by several team membres of the Division of Intelligent Medical Systems (IMSY) headed by Prof. Dr. Lena Maier-Hein. Below, I list my concrete contributions of the results presented in Chapter 3 and Sections 4.1 - 5.4.

A.1.1 Descriptive analysis of common practice

A large group of researchers from more than 30 institutions throughout the world collaborated on this project, which finally resulted in the publication [Maier-Hein et al., 2018]. The organization of the work and the conception of the challenge parameter list was done by Prof. Dr. Lena Maier-Hein, Matthias Eisenmann, myself, Sinan Onogur, and Patrick Scholz. Matthias Eisenmann, myself, and Sinan Onogur implemented the `Eclipse Meta Model` for the structured challenge capturing and analysis, which was mainly performed by the three of us, given the large amount of challenges analyzed and the multi-observer nature of the experiments. I organized, prepared, and evaluated the community survey described in this section.

A.1.2 Revealing flaws of common practice

Challenge design (Section 4.1) This part of the work was initiated and organized by myself and Prof. Dr. Lena Maier-Hein, with support from Matthias Eisenmann, and Prof. Dr. Annette Kopp-Schneider, and was published in [Reinke et al., 2018a]. The data collection was based on the MICCAI 2015 segmentation challenges and mainly carried out by Matthias Eisenmann. The descriptive analysis of all MICCAI challenges from the data described in the above paragraph was implemented by myself. I, Prof. Annette Kopp-Schneider, and Prof. Lena Maier-Hein worked on the conception of the statistical analysis as well as its design and implementation. The conception of the experiments was done by myself and Prof. Dr. Lena Maier-Hein, while I focused on analysing the rankings.

Metrics (Section 4.2) The work was initiated, coordinated, and designed by myself, Prof. Dr. Lena Maier-Hein, and Dr. Paul F. Jäger. The three of us formed the core team of the international Delphi expert consortium comprising more than 40 international researchers. The work was submitted to Nature Methods and is available as a dynamically updated preprint [Reinke et al., 2021a], submitted to Nature Methods as well, and a MIDL short paper [Reinke et al., 2021b]. The primary literature research was designed by myself, Dr. Minu D. Tizabi, and Prof. Dr. Lena Maier-Hein and was mainly carried out by me and Dr. Minu D. Tizabi. The Delphi process was led by myself, Dr. Paul F. Jäger, and Prof. Dr. Lena Maier-Hein, while the analysis and evaluation of the results was done by myself. Furthermore, I created the examples from the pitfall suggestions, computed all metric values, and designed all figures by myself. The metric values in the figures were reviewed by Tim Rädsch and Dr. Carole H. Sudre.

Rankings (Section 4.3) Similarly, a large group of researchers from more than 30 institutions throughout the world collaborated on this project, which finally resulted in the publication [Maier-Hein et al., 2018].Similar to the paragraphs above, the data collection was based on the MICCAI 2015 segmentation challenges and mainly carried out by Matthias Eisenmann, assisted by mayself. The conception of the statistical analysis was designed and implemented by Prof. Dr. Annette Kopp-Schneider, myself, Matthias Eisenmann, and Prof. Dr. Lena Maier-Hein.

Reporting and analyses (Section 4.4) Similarly, a large group of researchers from more than 30 institutions throughout the world collaborated on the first part of the section, which finally resulted in the publication [Maier-Hein et al., 2018]. All captured challenges served as the basis for the descriptive analysis of challenge design reporting. The results were analyzed by myself, Matthias Eisenmann, and Sinan Onogur, as described above. The second part of the section (reporting of algorithms) was based on the internship results of Georg Grab, whom I supervised. The project was designed by myself and we worked together on the methodology and the conception of the experiments. The implementation itself was done by Georg Grab, given the fact I was one of the primary challenge organizers. I was therefore biased in favor of the implementations because I was familiar with some of the specifics of the participating methods. The final decision on how to deal with ambiguities in the method description was done by myself.

A.1.3 Improving common practice

Challenge design (Section 5.1) The structured challenge submission system was mainly implemented by myself as lead developer, together with Sinan Onogur and Matthias Eisenmann. The experiments based on the challenges captured within the submission system were planned and conducted by myself. The challenge review process was introduced in the same step by the MICCAI Board Challenge Working Group, of which I am a working member. The Working Group is now known as the MICCAI SIG for Challenges and I was elected as its secretary. I managed the challenge review process for MICCAI 2018 and 2020 as challenges co-chair, and for MICCAI 2021 and MIDL 2020 as challenges chair, with help from the respective challenges teams.

Metrics (Section 5.2) The work was initiated, coordinated and designed by myself, Dr. Paul F. Jäger, and Prof. Dr. Lena Maier-Hein as the core team of the international Delphi expert consortium comprising more than 70 international researchers. The work was submitted to Nature Methods and is available as a preprint [Maier-Hein et al., 2022]. It was further accepted as a MIDL short paper [Reinke et al., 2022a] and as an abstract for the Medical Imaging Meets NeurIPS workshop [Reinke et al., 2022b]. We organized the workshops and the surveys, while I led the results evaluation of the questionnaires. We coordinated the work of the expert groups and designed and refined the metric mappings, problem fingerprints, and decision guides based on suggestions from the expert teams and feedback from the surveys. The visual conception of the mappings was done by myself, with support from Matthias Eisenmann. Lead by myself, we instantiated several biomedical use cases. This analysis was mainly done together with Dr. Paul F. Jäger,

Patrick Scholz, and Michael Baumgartner. I analyzed the metric selection of four biomedical image analysis challenges by myself.

Rankings (Section 5.3) The work was initiated, coordinated and designed by Dr. Manuel Wiesenfarth, myself, Matthias Eisenmann,Prof. Dr. Lena Maier-Hein, and Prof. Dr. Annette Kopp-Schneider and was published in [Wiesenfarth et al., 2021]. I designed and implemented the analysis of common practice in challenge result visualization. The open-source toolkit, including the visualization techniques, was implemented by Dr. Manuel Wiesenfarth as lead developer, Matthias Eisenmann, myself, Emre Kavur, and Laura Aguilera Saiz. The RobustMIS challenge was mainly organized by myself, Dr. Tobias Roß, and Prof. Dr. Lena Maier-Hein. I performed the ranking uncertainty analysis of the challenge. The work was published in [Roß/Reinke et al., 2020]. The MSD challenge was mainly organized by myself, Michela Antonelli, Prof. Dr. Lena Maier-Hein, and Prof. Dr. M. Jorge Cardoso. Again, I performed the ranking uncertainty analysis, supported by Dr. Manuel Wiesenfarth. The work was published in [Antonelli/Reinke et al., 2022].

Reporting and analyses (Section 5.4) As described above, the structured challenge submission system was mainly implemented by myself as lead developer, together with Sinan Onogur and Matthias Eisenmann. The descriptive analysis of the captured challenges was done by myself. The MICCAI Board Challenge Working Group introduced the process of challenge registration. The concrete process of putting the challenge design documents online and reviewing changes in challenge design was done by myself since 2020*. The BIAS reporting guideline (published in [Maier-Hein et al., 2020]) was designed by the BIAS initiative, founded by the MICCAI Board challenge working group. The guideline was based on the challenge parameter list, which was mainly designed by Prof. Dr. Lena Maier-Hein, myself, Matthias Eisenmann, and Sinan Onogur. The refinement of the list was achieved via a survey, implemented and evaluated by myself. Dr. Tobias Roß, myself, Prof. Dr. Lena Maier-Hein, and Prof. Dr. Annette Kopp-Schneider initiated, designed and implemented the mixed model analysis of challenge results. I reviewed all MICCAI 2004-2021 papers for their usage of mixed models.

*http://www.miccai.org/special-interest-groups/challenges/miccai-registered-challenges/

A.2 Ranking schemes used in this thesis

In this section, we formally present the three ranking schemes investigated in this thesis. Algorithm 1 presents the **metric-based ranking**, Algorithm the **case-based ranking**, and Algorithm the **test-based ranking**.

Algorithm 1 Metric-based ranking scheme [Maier-Hein et al., 2018]

for each participating team $p_i, i = 1, \ldots, N_p$ **do**
    Calculate the performance $m_k(p_i, c_j)$ for every case $c_j, j = 1, \ldots, N_c$ and every metric, $m_k, k = 1, \ldots, N_m$.
    In case of a missing value, i.e. $m_k(p_i, c_j) = \texttt{NaN}$, decide on a missing value strategy, for example by penalizing this case with the worst possible metric value.

    Option 1 (aggregate over cases first) :
        For each metric m_k, determine a metric-specific score $s_k(p_i)$ by aggregating $m_k(p_i, c_j)$ over all cases $c_j, j = 1, \ldots, N_c$, for example using the mean, median or a specific percentile.
        if $N_m > 1$ (multiple metrics) **then**
            Aggregate $s_k(p_i)$ over all metrics $m_k, k = 1, \ldots, N_m$ to achieve one score $s(p_i)$ per participating team p_i.
        else if $N_m = 1$ (single metric) **then**
            The score $s(p_i)$ is equal to $s_k(p_i)$ for every participating team p_i.
        end if

    Option 2 (aggregate over metric values first) :
        For each case c_j, determine a case-specific score $s_j(p_i)$ by aggregating $m_k(p_i, c_j)$ over all metrics $m_k, k = 1, \ldots, N_m$, for example using the mean, median or a specific percentile.
        if $N_c > 1$ (multiple cases) **then**
            Aggregate $s_j(p_i)$ over all cases $c_j, j = 1, \ldots, N_c$ to achieve one score $s(p_i)$ per participating team p_i.
        else if $N_c = 1$ (single case) **then**
            The score $s(p_i)$ is equal to $m_k(p_i, c_j)$ for every participating team p_i.
        end if

end for
Compute final the rank $r(p_i)$ for each participating team based on the scores $s(p_i), i = 1, \ldots, N_p$.

Algorithm 2 Case-based ranking scheme [Maier-Hein et al., 2018]

for each case $c_j, j = 1, \ldots, N_c$ **do**
  Calculate the performance $m_k(p_i, c_j)$ for every participating team $p_i, i = 1, \ldots, N_p$ and every metric, $m_k, k = 1, \ldots, N_m$.
  Determine a metric- and case-specific rank $r_{j,k}(p_i)$ for every participating team $p_i, i = 1, \ldots, N_p$ based on the performance $m_k(p_i, c_j)$.
  In case of a missing value, i.e. $m_k(p_i, c_j) = \texttt{NaN}$, assign the worst possible rank to $r_{j,k}(p_i)$.

  Option 1 (aggregate over cases first) :
    Determine an overall rank $r_k(p_i)$ over all cases $c_j, j = 1, \ldots, N_c$ for each metric m_k by aggregating over $r_{j,k}(p_i)$, for example using the mean, median or a specific percentile.
    Compute the final rank $r(p_i)$ for each participating team by aggregating the overall ranks $r_k(p_i)$ over all metrics $m_k, k = 1, \ldots, N_m$, for example using the mean, median or a specific percentile.

  Option 2 (aggregate over metric values first) :
    Determine an overall rank $r_j(p_i)$ over all metrics $m_k, k = 1, \ldots, N_m$ for each case c_j by aggregating over $r_{j,k}(p_i)$, for example using the mean, median or a specific percentile.
    Compute the final rank $r(p_i)$ for each participating team by aggregating the overall ranks $r_j(p_i)$ over all cases $c_j, j = 1, \ldots, N_c$, for example using the mean, median or a specific percentile.

end for

Algorithm 3 Test-based ranking scheme [Maier-Hein et al., 2018]

Select the significance level α (e.g. 5%).
for each metric $m_k, k = 1, \ldots, N_k$ **do**
  Calculate the performance $m_k(p_i, c_j)$ for every participating team $p_i, i = 1, \ldots, N_p$ and every case, $c_j, j = 1, \ldots, N_c$.
  Use a Wilcoxon signed rank test with level α to perform all pairwise comparisons between the participating teams $(p_i, p_{i'})$ based on the performances $m_k(p_i, c_j)$ and $m_k(p_{i'}, c_j)$.
  Determine a significance score $s_k(p_i), i = 1, \ldots, N_p$ which is equal to the number of significant test results, i.e. the number of participating teams that perform significantly worse compared to team p_i.
  Compute the ranking $r_k(p_i)$ based on the significance scores $s_k(p_i)$, ordering by the highest significant score (descending).
end for
Compute final the ranking $r(p_i)$ for each participating team by aggregating the metric-specific ranks $r_k(p_i)$ for every metric $m_k, k = 1, \ldots, N_k$.

A.3 Profile of common validation metrics

In this section, we present validation metrics for image-level classification (Chapter 2.4.1), semantic segmentation (Chapter 2.4.2), object detection (Chapter 2.4.3), and instance segmentation (Chapter 2.4.4). For every metric, we provide a concrete profile including relevant information, limitations, and recommendations. The following metrics are presented:

Per-class counting metrics

Sensitivity [Tharwat, 2020]: Figure A.1
Specificity [Tharwat, 2020]: Figure A.2
Positive Predictive Value (PPV) [Tharwat, 2020]: Figure A.3
Negative Predictive Value (NPV) [Tharwat, 2020]: Figure A.4
Positive Likelihood Ratio (LR+) [Attia, 2003]: Figure A.5
F_β Score [Chinchor, 1992]: Figure A.6
Net Benefit (NB) [Vickers and Elkin, 2006]: Figure A.7
Panoptic Quality (PQ) [Kirillov et al., 2019]: Figure A.8
False Positives per Image (FPPI) [Van Ginneken et al., 2010]: Figure A.9

Multi-class counting metrics

Accuracy [Tharwat, 2020]: Figure A.10
Balanced Accuracy (BA) [Tharwat, 2020]: Figure A.11
Matthews Correlation Coefficient (MCC) [Matthews, 1975]: Figure A.12
Weighted Cohen's Kappa (WCK) [Cohen, 1960]: Figure A.13
Expected Cost (EC) [Bishop and Nasrabadi, 2006; Hastie et al., 2009; Ferrer, 2022]: Figure A.14

Multi-threshold metrics

Area under the Receiver Operating Characteristic Curve (AUROC) [Hanley and McNeil, 1982]: Figure A.15
Average Precision (AP) [Lin et al., 2014; Everingham et al., 2015]: Figure A.16
Free-Response Receiver Operating Characteristic (FROC) Score [Van Ginneken et al., 2010; Bandos et al., 2009]: Figure A.17

Overlap-based metrics

Dice Similarity Coefficient (DSC) [Dice, 1945]: Figure A.18
Intersection over Union (IoU) [Jaccard, 1912]: Figure A.19
Centerline Dice Similarity Coefficient (clDice) [Shit et al., 2021]: Figure A.20
Intersection over Reference (IoR) [Maška et al., 2014]: Figure A.21

Distance-based metrics

Hausdorff Distance (HD) [Huttenlocher et al., 1993]: Figure A.22
Hausdorff Distance 95 Percentile (HD95) [Huttenlocher et al., 1993]: Figure A.23
Average Symmetric Surface Distance (ASSD) [Yeghiazaryan and Voiculescu, 2015]: Figure A.24
Mean Absolute Surface Distance (MASD) [Beneš and Zitová, 2015]: Figure A.25
Normalized Surface Distance (NSD) [Nikolov et al., 2021]: Figure A.26
Boundary Intersection over Union (Boundary IoU) [Cheng et al., 2021]: Figure A.27
Center Distance [Gurcan et al., 2010]: Figure A.28

Point-based metric: **Point inside Mask/ Box/ Approximation criterion** (Figure A.29).

Calibration metrics

Brier Score (BS) [Yeghiazaryan and Voiculescu, 2015]: Figure A.30
Negative Log Likelihood (NLL) [Cybenko et al., 1998]: Figure A.31
Root Brier Score (RBS) [Gruber and Buettner, 2022]: Figure A.32
Expected Calibration Error (ECE) [Naeini et al., 2015]: Figure A.33
Class-wise Calibration Error (CWCE) [Kumar et al., 2019]: Figure A.34
Kernel Calibration Error (KCE) [Widmann, 2020; Gruber and Buettner, 2022]: Figure A.35
Expected Calibration Error Kernel Density Estimate (ECE^{KDE}) [Popordanoska et al., 2022]: Figure A.36

A.3.1 Per-class counting metrics

SENSITIVITY
Recall, True positive rate (TPR), Hit rate

$$\text{Sensitivity} = \frac{TP}{TP + FN} = \frac{\blacksquare}{\blacksquare\ \blacksquare}$$

VALUE RANGE: [0, 1] ↑

DESCRIPTION
Sensitivity measures how good a method is in classifying truly positive samples as positive.

DEFINITION
[Tharwat, 2020]

RELATIONS
Sensitivity = 1 - False Negative Rate (FNR)

CARDINALITIES

TP	FP	FN	TN
●	○	●	○

PREVALENCE DEPENDENCY? ○

RECOMMENDED FOR

ImLC	SemS	ObD	InS
●	○	●	●

METRIC FAMILY

Counting metric	Multi-threshold metric	Distance-based metric
●	○	○

RELEVANT LIMITATIONS
- Sensitivity does not penalize FP predictions (Fig. 5.4 (a)).
- Sensitivity is undefined if the reference is empty (Fig. 5.10 (c)), and ad hoc rules need to be established depending on the use case.
- Sensitivity depends on the definition of the positive class (Fig. 5.17 (c)).
- A trivial classifier always predicting the positive class yields a Sensitivity of 1, but a Specificity of 0.

RECOMMENDATIONS
- Sensitivity should be preferred for comparisons across data sets given its prevalence independence.
- Sensitivity should be preferred in the presence of class imbalances.
- Sensitivity is easy to understand and interpret and helps interpreting test results for clinical decision making.
- Sensitivity should be reported together with Specificity.

Figure A.1: Profile of the Sensitivity [Tharwat, 2020]. This per-class counting metric relies on the True Positive (TP) and False Negative (FN) cardinalities of the confusion matrix (see Figure 2.6) and ranges between 0 and 1. It is recommended to be used for Image-level Classification (ImLC), Object Detection (ObD), and Instance Segmentation (InS) problems. Further abbreviations: False Positive (FP), True Negative (TN), and Semantic Segmentation (SemS).

SPECIFICITY
Selectivity, True negative rate (TNR)

$$\text{Specificity} = \frac{TN}{TN + FP} =$$

VALUE RANGE: [0, 1] ↑

CARDINALITIES

TP	FP	FN	TN
○	●	○	●

PREVALENCE DEPENDENCY? ○

DESCRIPTION
Specificity measures how good a method is in classifying truly negative samples as negative.

DEFINITION
[Tharwat, 2020]

RELATIONS
Specificity = 1 - False Positive Rate (FPR)

RECOMMENDED FOR

ImLC	SemS	ObD	InS
●	○	○	○

METRIC FAMILY

Counting metric	Multi-threshold metric	Distance-based metric
●	○	○

RELEVANT LIMITATIONS
- Specificity does not penalize FN predictions (Fig. 5.4 (a)).
- Specificity is undefined if both TN and FP are zero (Fig. 5.10 (c)), and ad hoc rules need to be established depending on the use case.
- Specificity depends on the definition of the positive class (Fig. 5.17 (c)).
- Specificity depends on the definition of TN (undefined for ObD and InS).
- A trivial classifier always predicting the positive class yields a Specificity of 1, but a Sensitivity of 0.

RECOMMENDATIONS
- Specificity should be preferred for comparisons across data sets given its prevalence independence.
- Specificity should be preferred in the presence of class imbalances.
- Specificity is easy to understand and interpret and helps interpreting test results forclinical decision making.
- Specificity should be reported together with Sensitivity.

Figure A.2: Profile of the Specificity [Tharwat, 2020]. This per-class counting metric relies on the True Negative (TN) and False Positive (FP) cardinalities of the confusion matrix (see Figure 2.6) and ranges between 0 and 1. It is recommended to be used for Image-level Classification (ImLC) problems only as it relies on the definition of TN. Further abbreviations: True Positive (TP), False Positive (FP), Semantic Segmentation (SemS), Object Detection (ObD), and Instance Segmentation (InS).

POSITIVE PREDICTIVE VALUE (PPV)
Precision

$$PPV = \frac{TP}{TP + FP} = \frac{\blacksquare}{\blacksquare\blacksquare} \quad \text{if Prevalence} = 0.5$$

$$PPV_{corrected} = \frac{Sensitivity \bullet Prevalence}{Sensitivity \bullet Prevalence + (1 - Specificity) \bullet (1 - Prevalence)}$$

VALUE RANGE: [0, 1] ↑

DESCRIPTION
PPV represents the probability of a positive prediction corresponding to an actual positive sample.

DEFINITION
[Tharwat, 2020]

RELATIONS
PPV is the complement of False discovery rate (FDR): PPV = 1 - FDR

PREVALENCE DEPENDENCY? ●

CARDINALITIES

TP	FP	FN	TN
●	●	○	○

RECOMMENDED FOR

ImLC	SemS	ObD	InS
●	○	●	●

METRIC FAMILY

Counting metric	Multi-threshold metric	Distance-based metric
●	○	○

RELEVANT LIMITATIONS
- PPV does not penalize FN predictions (Fig. 5.4 (a)).
- PPV is prevalence-dependent, thus not comparable across data sets if not corrected (Fig. 5.5) with different prevalences if not corrected.
- The expected value for an uninformed classifier depends on class imbalance.
- The higher the prevalence, the higher the expected value (Fig. 5.9 (a))
- PPV is undefined if the prediction is empty (Fig. 5.10 (c)), and ad hoc rules need to be established depending on the use case.
- PPV depends on the definition of the positive class (Fig. 5.17 (c)).

RECOMMENDATIONS
- The prevalence-corrected PPV should be preferred in the presence of class imbalances.
- PPV should be preferred if a standardized metric definition is desired.
- For comparison across data sets with different prevalences, the prevalence-corrected version or alternatively the positive Likelihood ratio (LR+) are more appropriate, since they are prevalence independent

Figure A.3: Profile of the Positive Predictive Value (PPV) [Tharwat, 2020]. This per-class counting metric relies on the True Positive (TP) and False Positive (FP) cardinalities of the confusion matrix (see Figure 2.6) and ranges between 0 and 1. It is recommended to be used for Image-level Classification (ImLC), Object Detection (ObD) and Instance Segmentation (InS) problems. Further abbreviations: False Negative (FN), True Negative (TN), and Semantic Segmentation (SemS).

NEGATIVE PREDICTIVE VALUE (NPV)

$$NPV = \frac{TN}{TN + FN} =$$  if Prevalence $= 0.5$

$$NPV_{corrected} = \frac{\text{Specificity} \cdot (1 - \text{Prevalence})}{\text{Specificity} \cdot (1 - \text{Prevalence}) + (1 - \text{Sensitivity}) \cdot \text{Prevalence}}$$

VALUE RANGE: [0, 1] ↑

DESCRIPTION
NPV represents the probability of a negative predicti-on corresponding to an actual negative sample.

DEFINITION
[Tharwat, 2020]

RELATIONS
NPV is the complement of False omission rate (FOR):
NPV = 1 - FOR

PREVALENCE DEPENDENCY? ●

CARDINALITIES

TP	FP	FN	TN
○	○	●	●

RECOMMENDED FOR

ImLC	SemS	ObD	InS
●	○	○	○

METRIC FAMILY

Counting metric	Multi-threshold metric	Distance-based metric
●	○	○

RELEVANT LIMITATIONS
- NPV does not penalize FP predictions (Fig. 5.4 (b)).
- NPV is prevalence-dependent, thus not comparable across data sets across data sets with different prevalences if not corrected (Fig. 5.5).
- The expected value for an uninformed classifier depends on class imbalance.
- The lower the prevalence, the higher the expected value. (Fig. 5.9 (a)).
- NPV is undefined if both, TN and FN, are zero (Fig. 5.10 (c)), and ad hoc rules need to be established depending on the use case.
- NPV value depends on the definition of the positive class (Fig. 5.17 (c)).
- NPV depends on the definition of TN (undefined for ObD and InS).

RECOMMENDATIONS
- For comparison across data sets with different prevalences, the prevalence-corrected version or alternatively the negative Likelihood ratio (LR-) are more appropriate, since they are prevalence-independent.
- In general NPV has a very narrow use case, since it requires the definition of a positive class which is then usually the more important one. Furthermore in these cases a typical class imbalance towards the negative class renders (the not corrected version of) NPV less meaningful.

Figure A.4: Profile of the Negative Predictive Value (NPV) [Tharwat, 2020]. This per-class counting metric relies on the True Negative (TN) and False Negative (FN) cardinalities of the confusion matrix (see Figure 2.6) and ranges between 0 and 1. It is recommended to be used for Image-level Classification (ImLC) problems only as it relies on the definition of TN. Further abbreviation: True Positive (TP), False Positive (FP), Semantic Segmentation (SemS), Object Detection (ObD), and InS.

POSITIVE LIKELIHOOD RATIO (LR+)

Likelihood ratio positive, Likelihood ratio for positive results

$$LR+ = \frac{\text{Sensitivity}}{1 - \text{Specificity}} = \frac{\square}{\square} \Big/ \left(1 - \frac{\square}{\square}\right)$$

VALUE RANGE: $[0, \infty)$ ↑

DESCRIPTION
LR+ indicates the factor by which a positive prediction occurs more frequently among actual positive samples than among actual negative samples.

DEFINITION
[Attia 2003]

CARDINALITIES

TP	FP	FN	TN
●	●	●	●

RECOMMENDED FOR

ImLC	SemS	ObD	InS
●	○	○	○

METRIC FAMILY

Counting metric	Multi-threshold metric	Distance-based metric
●	○	○

PREVALENCE DEPENDENCY? ○

RELEVANT LIMITATIONS
- LR+ is undefined if the Specificity is 1 (Fig. 5.10 (c)), and ad hoc rules need to be established depending on the use case.
- LR+ depends on the definition of the positive class (Fig. 5.17 (c)).
- LR+ depends on the definition of TN (undefined for ObD and InS).
- LR+ may yield the same value for different Sensitivity and Specificity scores.

RECOMMENDATIONS
- LR+ should be preferred for a comparison across data sets given its prevalence independence.
- LR+ should be preferred in the presence of class imbalances.

Figure A.5: Profile of the Positive Likelihood Ratio (LR+) [Attia, 2003]. This per-class counting metric relies on the Sensitivity and Specificity, therefore on all entries of the confusion matrix (True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN); see Figure 2.6) and ranges between 0 and ∞. It is recommended to be used for Image-level Classification (ImLC) problems only as it relies on the definition of TN. Further abbreviations: Semantic Segmentation (SemS), Object Detection (ObD), Instance Segmentation (InS).

F_β SCORE

$$F_\beta \text{ Score} = (1+\beta^2)\frac{\text{PPV} \bullet \text{Sensitivity}}{\beta^2 \bullet \text{PPV} + \text{Sensitivity}}$$

$$= \frac{(1+\beta^2) \bullet \text{TP}}{(1+\beta^2) \bullet \text{TP} + \beta^2 \bullet \text{FN} + \text{FP}} = \frac{(1+\beta^2) \bullet \blacksquare}{(1+\beta^2) \bullet \blacksquare + \beta^2 \bullet \subset + \blacksquare}$$

VALUE RANGE: [0, 1] ↑

DESCRIPTION
The F_β Score weights PPV (FP) and Sensitivity (FN) with the parameter β.

The special case of $\beta = 1$ is the harmonic mean of PPV and Sensitivity and is a common metric in segmentation problems (here usually referred to as DSC).

DEFINITION
[Chinchor 1992]

PREVALENCE DEPENDENCY? ●

CARDINALITIES

TP	FP	FN	TN
●	●	●	○

RECOMMENDED FOR

ImLC	SemS	ObD	InS
●	●	●	●

METRIC FAMILY

Counting metric	Multi-threshold metric	Distance-based metric
●	○	○

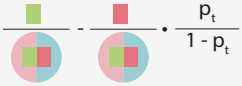
RELEVANT LIMITATIONS
- F_β Score is unaware of the structure shape and center (SemS, InS; Figs. 5.3 (a) and (b)).
- F_β Score is prevalence-dependent, thus not comparable across data sets (Fig. 5.5 (b)) with different prevalences.
- F_β Score treats tiny structures differently than larger ones (SemS, InS; Fig. 5.7 (a)).
- F_β Score is unable to compensate labeling imprecisions (SemS, InS; Fig. 5.10 (a)).
- F_β Score is undefined if both, reference and prediction, are empty (Fig. 5.10 (c)), and ad hoc rules need to be established depending on the use case.
- F_β Score depends on the definition of the positive class (Fig. 5.17 (c)).

RECOMMENDATIONS
- F_β Score should be preferred if class confusions are of unequal severity, e.g. to differently penalize over- and undersegmentation.
- In InS problems, F_β Score should be preferred if the detection quality should be assessed independently from the segmentation quality.

Figure A.6: Profile of the F_β Score [Chinchor, 1992]. This per-class counting metric relies on the True Positive (TP), False Positive (FP) and False Negative (FN) cardinalities of the confusion matrix (see Figure 2.6) and ranges between 0 and 1. It is recommended to be used for Image-level Classification (ImLC), Semantic Segmentation (SemS), Object Detection (ObD) and InS problems. Further abbreviation: True Negative (TN).

NET BENEFIT (NB)

$$NB = \frac{TP}{TP + TN + FP + FN} - \frac{FP}{TP + TN + FP + FN} \cdot \frac{p_t}{1 - p_t}$$



VALUE RANGE: [-1, 1] ↑

DESCRIPTION

NB evaluates the quality of a model intended to support a specific clinical decision. Net benefit gives the 'net' proportion of TPs that results from a prediction. This is equivalent to the proportion of TPs in the absence of FPs.
To derive it, the total count of FPs is weighted by the task-related risk threshold (= exchange rate between the benefit of TPs and harm of FPs), then subtracted from the total count of TPs, and the result is divided by data set size. When varying the risk threshold over a 'reasonable range' of possible thresholds, plotting NB by risk threshold yields a decision curve. It is a strictly proper performance measure.

DEFINITION

[Vickers and Elkin, 2006]

RECOMMENDED FOR

ImLC	SemS	ObD	InS
●	○	○	○

PREVALENCE DEPENDENCY? ●

METRIC FAMILY

Counting metric	Multi-threshold metric	Distance-based metric
●	○	○

CARDINALITIES

TP	FP	FN	TN
●	●	●	●

RELEVANT LIMITATIONS

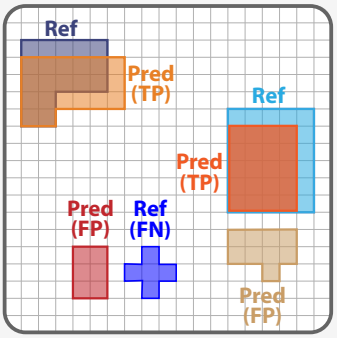
- NB depends on the definition of TN (undefined for ObD and InS).
- NB requires the availability of predicted class scores.
- NB can only be applied if a range of relevant decision thresholds can be defined. If problem-specific penalties can be provided, these can be used to calculate the decision threshold. If no task-related thresholds or penalties can be provided, the prediction model has little clinical use.

RECOMMENDATIONS

- NB is preferred if (1) the predicted class scores indicate the risk associated with a case belonging to a particular class and (2) there is a (range of reasonable) task-related risk threshold(s) or problem-specific penalties.
- The clinical decision that the model is intended to support should be clearly defined.
- NB is strictly proper, hence poor discrimination and miscalibration lead to lower values for NB.

Figure A.7: Profile of the Net Benefit (NB) [Vickers and Elkin, 2006]. This per-class counting metric relies on all cardinalities of the confusion matrix (True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN); see Figure 2.6) and ranges between -1 and 1. It is recommended to be used for Image-level Classification (ImLC) problems only.

PANOPTIC QUALITY (PQ)



Reference (Ref) instances

Predicted (Pred) instances

$$PQ = \frac{\sum_{(Ref,\,Pred)\in TP} IoU(Ref, Pred)}{|TP| + 0.5\,|FP| + 0.5\,|FN|}$$

$$= \underbrace{\frac{\sum_{(Ref,\,Pred)\in TP} IoU(Ref, Pred)}{|TP|}}_{\text{Segmentation quality}} \cdot \underbrace{\frac{|TP|}{|TP| + 0.5\,|FP| + 0.5\,|FN|}}_{\text{Detection quality}}$$

$$= \frac{IoU(\,,\,) + IoU(\,,\,)}{|\{\,,\,\}|} \cdot \frac{|\{\,,\,\}|}{|\{\,,\,\}| + 0.5\,|\{\,,\,\}| + 0.5\,|\{\,\}|}$$

VALUE RANGE: [0, 1] ↑

DESCRIPTION

PQ assesses segmentation and detection quality in one metric. The segmentation quality is measured by averaging the IoU scores of all TP instances. The detection quality is measured by the F_1 Score. While in the F_1 Score, each TP counts as "1", PQ replaces this "1" score in the nominator with the IoU score of each TP. The F_1 Score as a detection metric implies two cutoffs:
- A prior cutoff on a localisation criterion for matching and
- a prior cutoff on object class scores to generate a confusion matrix.

In theory, F_1 Scores would have to be reported as F_1 Score@class_score_cutoff_X@IoU_cutoff_Y. While multi-threshold metrics like AP or FROC Score ease these assumptions by averaging performance over class score thresholds (we still report AP@IoU_cutoff_Y), PQ eases the assumptions by directly incorporating the localisation performance.

DEFINITION

[Kirillov et al., 2019]

CARDINALITIES

TP	FP	FN	TN
●	●	●	○

RECOMMENDED FOR

ImLC	SemS	ObD	InS
○	○	○	●

METRIC FAMILY

Counting metric	Multi-threshold metric	Distance-based metric
●	○	○

RELEVANT LIMITATIONS

Only inspecting the PQ value does not show the tradeoff between segmentation and detection quality. One of the terms may overrule the other without being noticed (Fig. 5.17 (d)).

RECOMMENDATIONS

PQ should be preferred if the detection and segmentation quality should be assessed in a single score, e.g. when the overall best model needs to be selected, which can naturally not be done naturally based on two separate metrics. For diagnostics, individual detection and segmentation performance should also be reported.

Figure A.8: Profile of the Panoptic Quality (PQ) [Kirillov et al., 2019]. This per-class counting metric measures the segmentation and detection quality simultaneously based on the True Positive (TP), False Positive (FP) and False Negative (FN) cardinalities of the confusion matrix (see Figure 2.6) and ranges between 0 and 1. It is recommended to be used for Instance Segmentation (InS) problems. Further abbreviations: Image-level Classification (ImLC), Object Detection (ObD), Semantic Segmentation (SemS), and True Negative (TN).

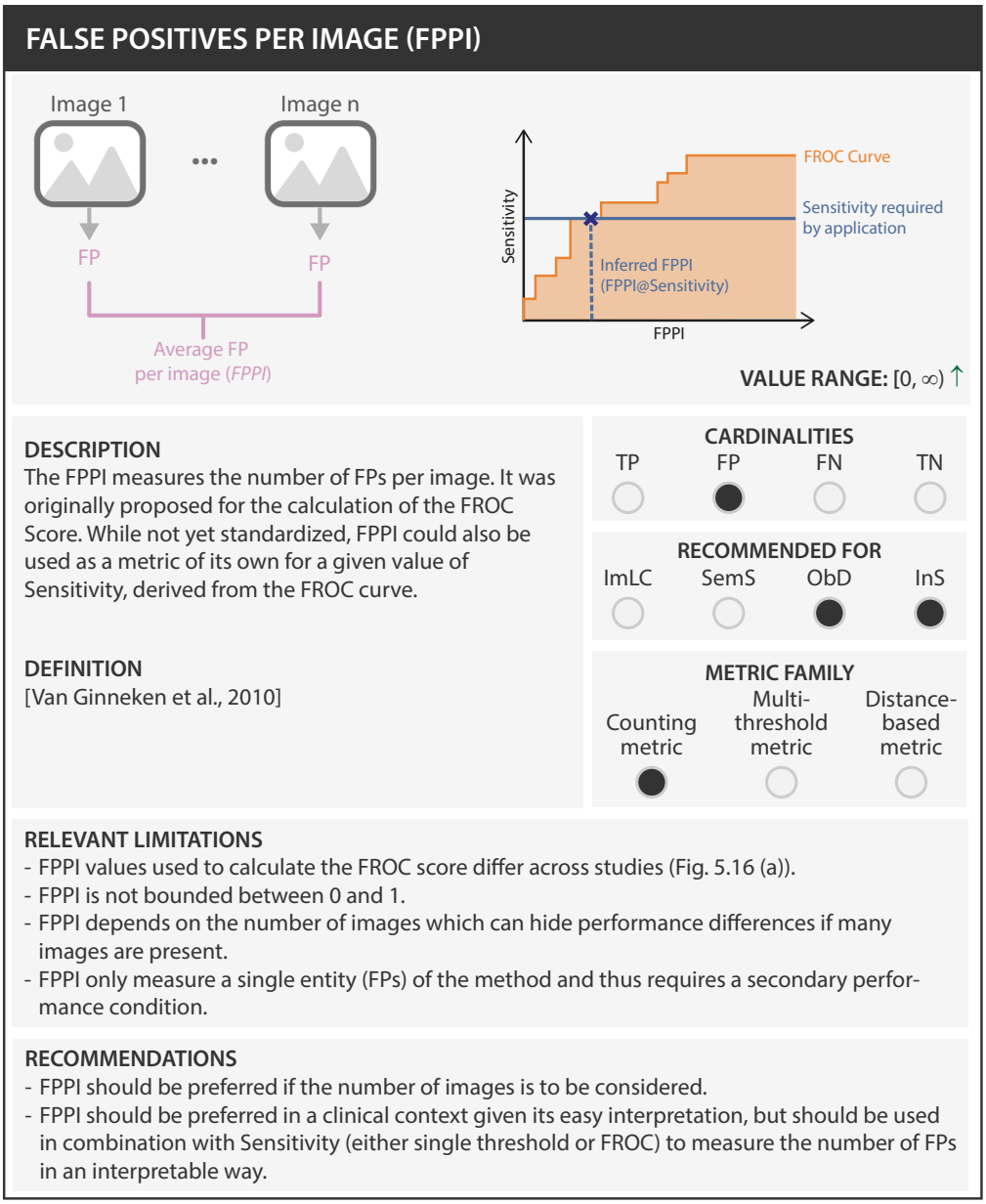


Figure A.9: Profile of the False Positives per Image (FPPI). This per-class counting metric relies on the False Positive (FP) cardinality of the confusion matrix (see Figure 2.6) and ranges between 0 and ∞. It is recommended to be used for Object Detection (ObD) and Instance Segmentation (InS) problems. Further abbreviations: Image-level Classification (ImLC), Semantic Segmentation (SemS), True Positive (TP), True Negative (TN), and False Negative (FN).

A.3.2 Multi-class counting metrics

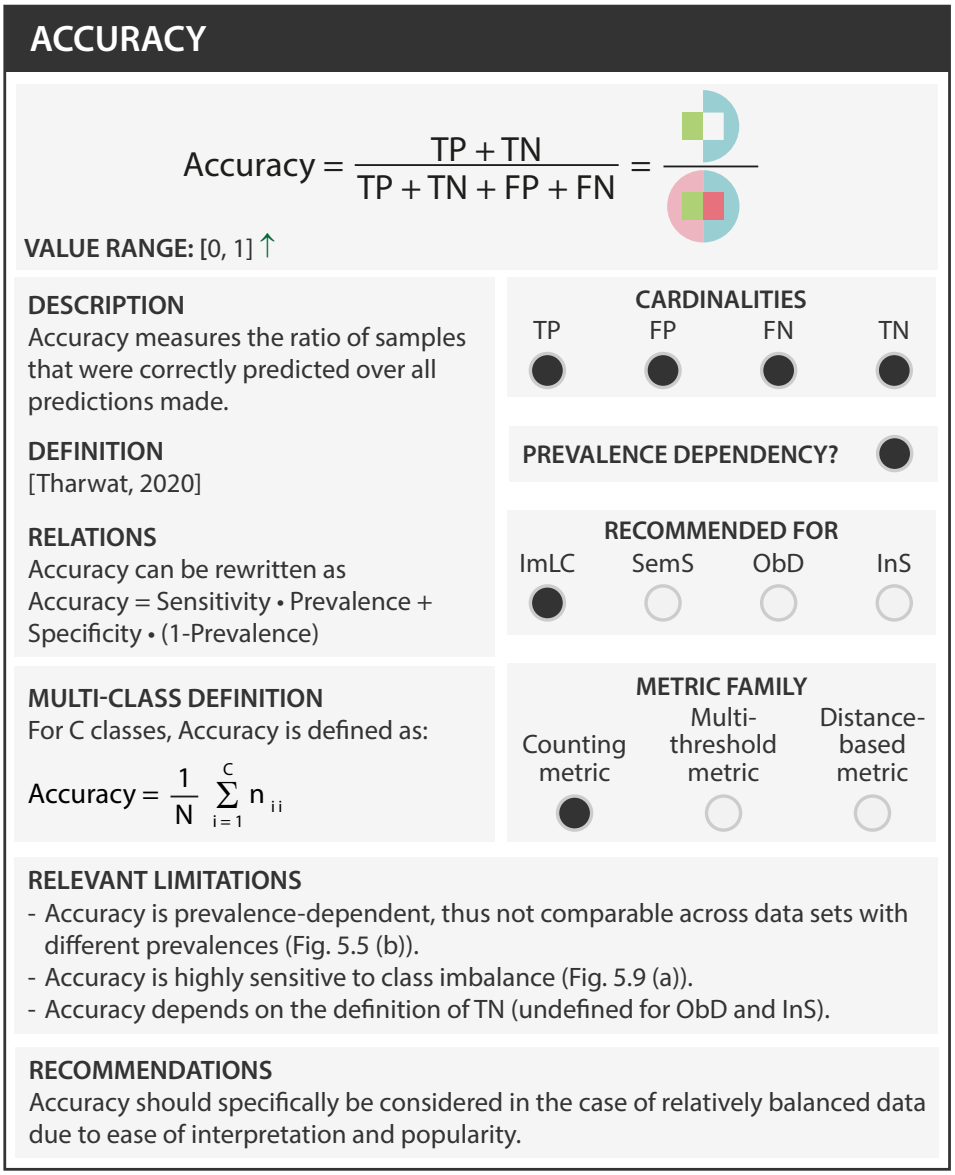


Figure A.10: Profile of the Accuracy [Tharwat, 2020]. This multi-class counting metric relies on all cardinalities of the confusion matrix (True Positive (TP), False Positive (FP), False Negative (FN), and True Negative (TN); see Figure 2.6) and ranges between 0 and 1. For the multi-class definition, please refer to Figure 2.11 for the illustration of the $C \times C$ confusion matrix. n_{ii} refers to its diagonal entries and N to the total number of samples. It is recommended to be used for Image-level Classification (ImLC) problems only as it relies on the definition of TN. Further abbreviations: Semantic Segmentation (SemS), Object Detection (ObD), Instance Segmentation (InS).

BALANCED ACCURACY (BA)

$$BA = \frac{1}{2}(\text{Sensitivity} + \text{Specificity})$$

$$= \frac{1}{2}\left(\frac{TP}{TP + FN} + \frac{TN}{TN + FP}\right) = \frac{1}{2}\left(\frac{\square}{\square} + \frac{\square}{\square}\right)$$

VALUE RANGE: [0, 1] ↑

DESCRIPTION
BA measures the arithmetic mean of Sensitivities for each class, i.e. for each class it measures the fraction of actual positive samples that were predicted as such.

RELATIONS
- BA is closely related to J: $BA = \frac{J + 1}{2}$
- For a prevalence of 50%, BA = Accuracy and Cohen's Kappa = 2BA - 1

DEFINITION
[Tharwat, 2020]

CARDINALITIES

TP	FP	FN	TN
●	●	●	●

PREVALENCE DEPENDENCY? ○

RECOMMENDED FOR

ImLC	SemS	ObD	InS
●	○	○	○

MULTI-CLASS DEFINITION
For C classes, BA is defined as the arithmetic mean of Sensitivities per class:

$$BA = \frac{1}{N}\sum_{i=1}^{C}\frac{n_{ii}}{n_{ii} + n_{i\cdot}}$$

METRIC FAMILY

Counting metric	Multi-threshold metric	Distance-based metric
●	○	○

RELEVANT LIMITATIONS
- BA can be misleading in the case of ordinal scales (Fig. 5.4 (c)).
- BA can be overly optimistic in the case of class imbalances (Fig. 5.9 (b)).
- BA may yield the same value for different Sensitivity and Specificity scores.
- BA depends on the definition of TN (undefined for ObD and InS).

RECOMMENDATIONS
BA should be specifically considered …
- … in the presence of high class imbalance.
- … if a comparison of performance across data sets with different prevalences is desired.

Figure A.11: Profile of the Balanced Accuracy (BA) [Tharwat, 2020]. This multi-class counting metric relies on all cardinalities of the confusion matrix (True Positive (TP), False Positive (FP), False Negative (FN), and True Negative (TN); see Figure 2.6) and ranges between 0 and 1. For the multi-class definition, please refer to Figure 2.11 for the illustration of the multi-class $C \times C$ confusion matrix. n_{ii} refers to its diagonal entries, $n_{i\cdot}$ to the sum of entries per row i, and N to the total number of samples. It is recommended to be used for Image-level Classification (ImLC) problems only as it relies on the definition of TN. Further abbreviations: Semantic Segmentation (SemS), Object Detection (ObD), Instance Segmentation (InS).

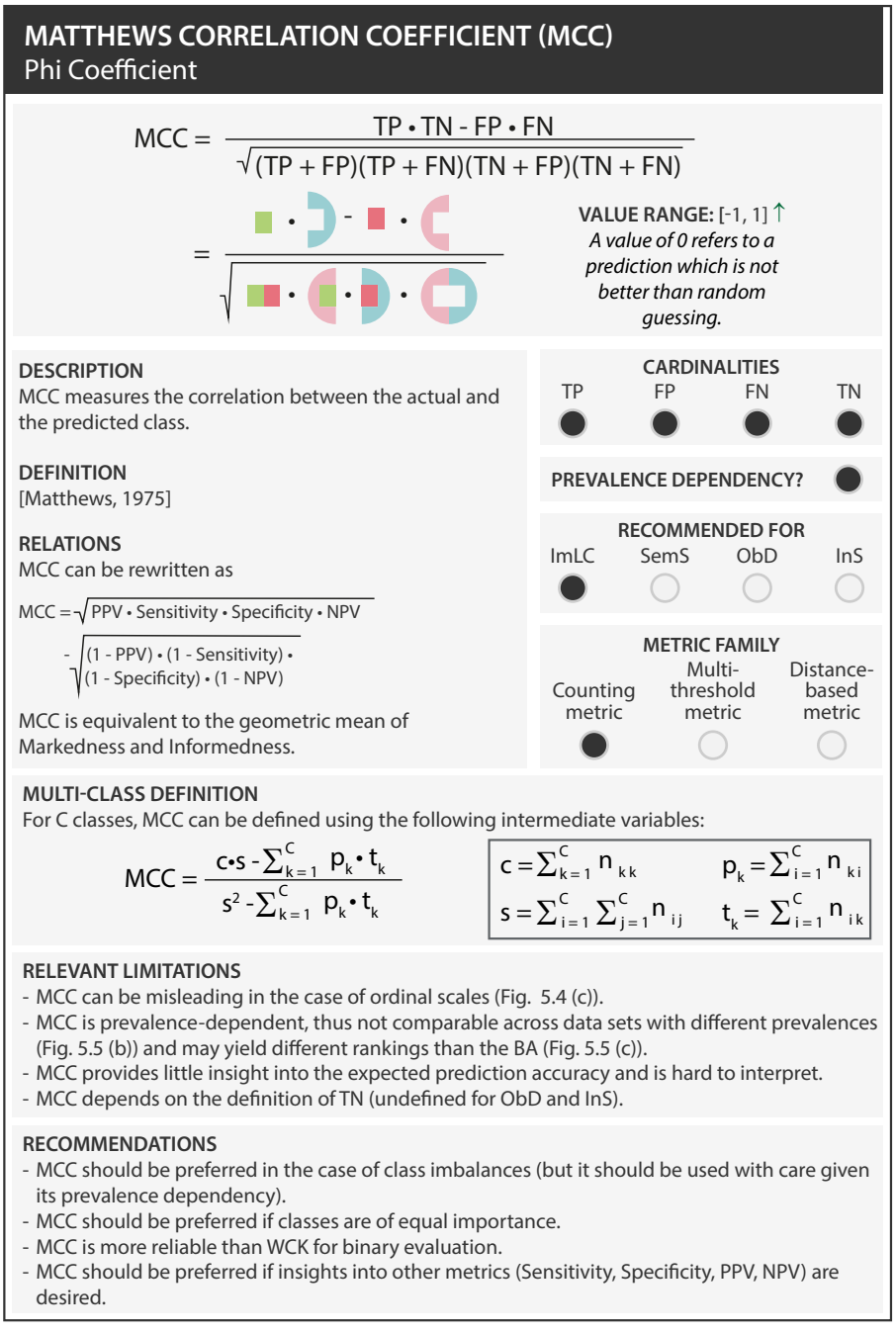


Figure A.12: Profile of the Matthews Correlation Coefficient (MCC) [Matthews, 1975]. This multi-class counting metric relies on all cardinalities of the confusion matrix (True Positive (TP), False Positive (FP), False Negative (FN), and True Negative (TN); see Figure 2.6) and ranges between -1 and 1. For the multi-class definition, please refer to Figure 2.11 for the illustration of the multi-class $C \times C$ confusion matrix. n_{ij} refers to its entries for row i and column j. It is recommended to be used for Image-level Classification (ImLC) problems only as it relies on the definition of TN. Further abbreviation: Semantic Segmentation (SemS), Object Detection (ObD), InS. Further abbreviations: Semantic Segmentation (SemS), Object Detection (ObD), Instance Segmentation (InS).

WEIGHTED COHEN'S KAPPA (WCK)

Weighted Cohen's Kappa Coefficient, Weighted Kappa Statistic, Weighted Kappa Score

$$WCK = \frac{p_0^w - p_e^w}{1 - p_e^w}$$

$$p_0^w = \frac{w_{TP}TP + w_{TN}TN + w_{FP}FP + w_{FN}FN}{TP + TN + FP + FN}$$

$$p_e^w = w_{TP}\frac{(TP + FP)(TP + FN)}{TP + TN + FP + FN} + w_{TN}\frac{(TN + FP)(TN + FN)}{TP + TN + FP + FN} + w_{FN}\frac{(FN + FP)(FN + TN)}{TP + TN + FP + FN} + w_{FP}\frac{(FP + TP)(FP + TN)}{TP + TN + FP + FN}$$

VALUE RANGE: $(-\infty, \infty)$ ↑

Typically, w_{TP} and w_{TN} are chosen equal to zero, while w_{FP} and $w_{FN} > 0$, yielding a value range of [-1, 1]. A value of 0 refers to a prediction which is not better than random guessing.

DESCRIPTION

WCK calculates the degree of agreement between the reference and prediction while incorporating the agreement resulting from chance.

In contrast to CK, it takes into account the degree of disagreement between reference and prediction.

RECOMMENDED FOR

ImLC	SemS	ObD	InS
●	○	○	○

CARDINALITIES

TP	FP	FN	TN
●	●	●	●

PREVALENCE DEPENDENCY? ●

DEFINITION

[Cohen, 1960]

METRIC FAMILY

Counting metric	Multi-threshold metric	Distance-based metric
●	○	○

MULTI-CLASS DEFINITION

For C classes, WCK can be defined as: $WCK = \left(\sum_{i=1}^{C}\sum_{j=1}^{C} w_{ij} \cdot n_{ij}\right) / \left(\sum_{i=1}^{C}\sum_{j=1}^{C} w_{ij} \cdot \frac{n_{i\cdot} \cdot n_{\cdot j}}{N^2}\right)$

RELEVANT LIMITATIONS

- WCK is prevalence-dependent, thus not comparable across data sets (Fig. 5.5 (b)) and may yield different rankings than the BA (Fig. 5.5 (c)).
- WCK is more likely to be higher when the actual class distribution is balanced (Fig. 5.21 (c)).
- WCK can be misleading for label shifts in ordinal classes.
- WCK provides little insight into the expected prediction accuracy and is hard to interpret.
- The selection of weights to different types of mistakes is problem-dependent and requires domain knowledge.
- WCK depends on the definition of TN (undefined for ObD and InS).

RECOMMENDATIONS

- WCK should be preferred in the case of class imbalances (but should be used with care given its prevalence dependency).
- WCK should be preferred if problem-specific penalties are desired.
- Quadratically WCK is a popular choice for ordinal classification problems such as disease grading.
- WCK is commonly used to measure inter-rater agreement.

Figure A.13: Profile of Weighted Cohen's Kappa (WCK) [Cohen, 1960]. This multi-class counting metric relies on all cardinalities of the confusion matrix (True Positive (TP), False Positive (FP), False Negative (FN), and True Negative (TN); see Figure 2.6) and ranges between -1 and 1. For the multi-class definition, please refer to Figure 2.11 for the illustration of the multi-class $C \times C$ confusion matrix. n_{ij} refers to its entries for row i and column j, $n_{i\cdot}$ to the sum of entries per row i, $n_{\cdot j}$ to the sum of entries per column j, and N to the total number of samples. It is recommended to be used for Image-level Classification (ImLC) problems only as it relies on the definition of TN. Further abbreviations: Semantic Segmentation (SemS), Object Detection (ObD), Instance Segmentation (InS).

EXPECTED COST (EC)
Expected prediction error, expected loss

$$EC = C_{miss} \cdot P_{miss} \cdot P_{tar} + C_{FA} \cdot P_{FA} \cdot (1 - P_{tar}) = C_{miss} \cdot \frac{FN}{TP + FN} \cdot \frac{TP + FN}{TP + TN + FP + FN} + C_{FA} \cdot \frac{FP}{TN + FP} \cdot \left(1 - \frac{TP + FN}{TP + TN + FP + FN}\right)$$

VALUE RANGE: $[0, \infty)$ ↓
EC can be assumed to be positive if costs are non-negative, which can be done without loss of generality.

DESCRIPTION
EC is a generalization of the probability of error (which is, in turn, 1 - Accuracy) for cases in which errors cannot all be considered to have equally severe consequences. It is defined as the expectation of the cost, where the cost incurred on a certain sample depends on the sample's class and the decision made for that sample. In practice, the expectation can be estimated as a simple average of the costs over the evaluation samples.
EC describes the weighted sum of error rates with P_{miss} being the FN rate and P_{FA} being the FP rate. P_{tar} refers to the prior probability (prevalence). C_{miss} and C_{FA} denote estimated costs of the error rates and can be adjusted as a weighting of them. It can be used to measure discrimination and calibration in one score.

VARIANT
Normalized EC (EC_{norm}): normalizes EC by the EC of a naive system.

DEFINITION
[Bishop and Nasrabadi, 2006; Hastie et al., 2009; Ferrer, 2022]

CARDINALITIES

TP	FP	FN	TN
●	●	●	●

PREVALENCE DEPENDENCY? ◐
Both options are possible depending on how the priors are set in the definition of the metric.

RECOMMENDED FOR

ImLC	SemS	ObD	InS
●	○	●	●

MULTI-CLASS DEFINITION
For C classes, EC is defined as:

$$EC = \sum_{i=1}^{C} \sum_{j=1}^{C} P_i \cdot w_{ij} \cdot \frac{n_{ij}}{n_{i\cdot}}$$

METRIC FAMILY

Counting metric	Multi-threshold metric	Distance-based metric
●	○	○

RELEVANT LIMITATIONS
- EC depends on the definition of TN (undefined for ObD and InS).
- EC is rather uncommon and can therefore not be used for comparison with other publications.

RECOMMENDATIONS
- EC should be preferred if problem-specific penalties are available.
- EC can be used as a prevalence-dependent and prevalence-independent metric.
- EC can also be used for decomposing the system performance into discrimination and calibration.

Figure A.14: Profile of the Expected Cost (EC) [Bishop and Nasrabadi, 2006; Hastie et al., 2009; Ferrer, 2022]. This multi-class counting metric relies on all cardinalities of the confusion matrix (True Positive (TP), False Positive (FP), False Negative (FN), and True Negative (TN); see Figure 2.6) and ranges between 0 and ∞, given their problem-specific weights. For the multi-class definition, please refer to Figure 2.11 for the illustration of the multi-class $C \times C$ confusion matrix. n_{ij} refers to its entries for row i and column j, $n_{i\cdot}$ to the sum of entries per row i, and P_i to the priors/prevalences per class i. It is recommended to be used for Image-level Classification (ImLC) problems only as it relies on the definition of TN. Further abbreviations: Semantic Segmentation (SemS), Object Detection (ObD), Instance Segmentation (InS).

A.3.3 Multi-threshold metrics

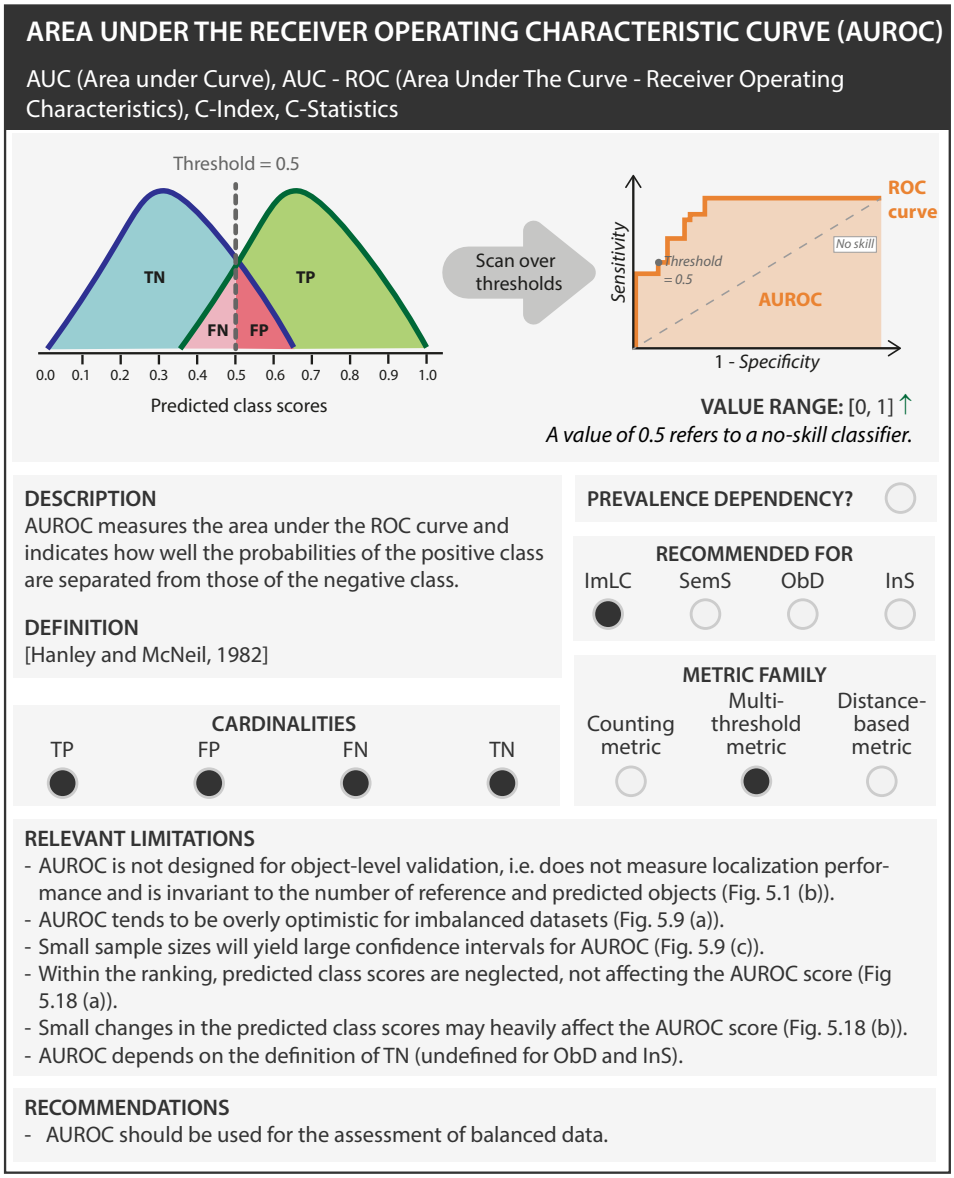
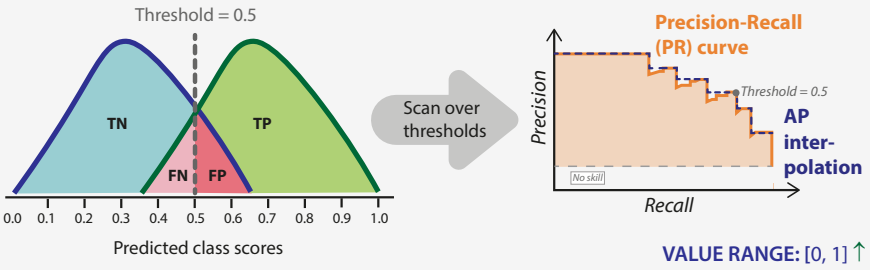


Figure A.15: Profile of AUROC. This multi-threshold counting metric relies on all cardinalities of the dynamic confusion matrix based on multiple thresholds (True Positive (TP), False Positive (FP), False Negative (FN), and True Negative (TN); see Figures 2.6 and 2.7) and ranges between 0 and 1. It is recommended to be used for Image-level Classification (ImLC) problems only as it relies on the definition of TN. Further abbreviations: Semantic Segmentation (SemS), Object Detection (ObD), Instance Segmentation (InS).

AVERAGE PRECISION (AP)



DESCRIPTION

AP measures the interpolated area under the PR curve. Imagine sorting all predictions by descending predicted probabilities (with scores of equal likelihood put side by side). From top to bottom now iteratively declare each sample as positive and thus continuously increase the positive predicted samples. At each point calculate Precision (= PPV) and Recall (= Sensitivity). While Sensitivity must increase while going through the list, PPV might either increase (if the newly positive declared sample is a true positive) or decrease (if the newly positive declared sample is a false positive). After iterating through the whole list, average the PPV scores, weighted by the increase in Sensitivity of each step.

In terms of the PR curve, a monotonic curve is created from the original curve and subsequently (typically done in OD and IS problems), PPV values are selected at certain Sensitivity points on the curve. Thus, AP is determined as the weighted mean of PPVs at each threshold where the weight represents the change in Sensitivity from the previous threshold.

The **mean Average Precision (mAP)** is a commonly used extension, which measures the average of the AP over multiple classes.

DEFINITION

[Lin et al., 2014, Everingham et al., 2015]

PREVALENCE DEPENDENCY? ●

METRIC FAMILY

Counting metric	Multi-threshold metric	Distance-based metric
○	●	○

CARDINALITIES

TP	FP	FN	TN
●	●	●	○

RECOMMENDED FOR

ImLC	SemS	ObD	InS
●	○	●	●

RELEVANT LIMITATIONS

- Within the ranking, predicted class scores are neglected, not affecting the AP score (Fig. 5.18 (a)).
- Small changes in the predicted class scores may heavily affect the AP score (Fig. 5.18 (b)).
- AP does not take into account the total number of images (Fig. 5.14 (c).
- AP depends on PPV, which is prevalence-dependent, thus it is not comparable across data sets with different prevalences (Fig. 5.5).
- [OD, IS] The number of predictions per image is usually limited to a maximum number which can influence the AP score.

RECOMMENDATIONS

- AP should be preferred in the case of class imbalances (but it should be used with care given its prevalence dependency).
- AP should be preferred if a standardized definition is needed.

Figure A.16: Profile of AP [Lin et al., 2014; Everingham et al., 2015]. This multi-threshold counting metric relies on the True Positive (TP), False Positive (FP) and False Negative (FN) cardinalities of the dynamic confusion matrix based on multiple thresholds (see Figures 2.6 and 2.7) and ranges between 0 and 1. It is recommended to be used for Image-level Classification (ImLC), Object Detection (ObD) and InS problems. Further abbreviations: Positive Predictive Value (PPV), Semantic Segmentation (SemS), True Negative (TN).

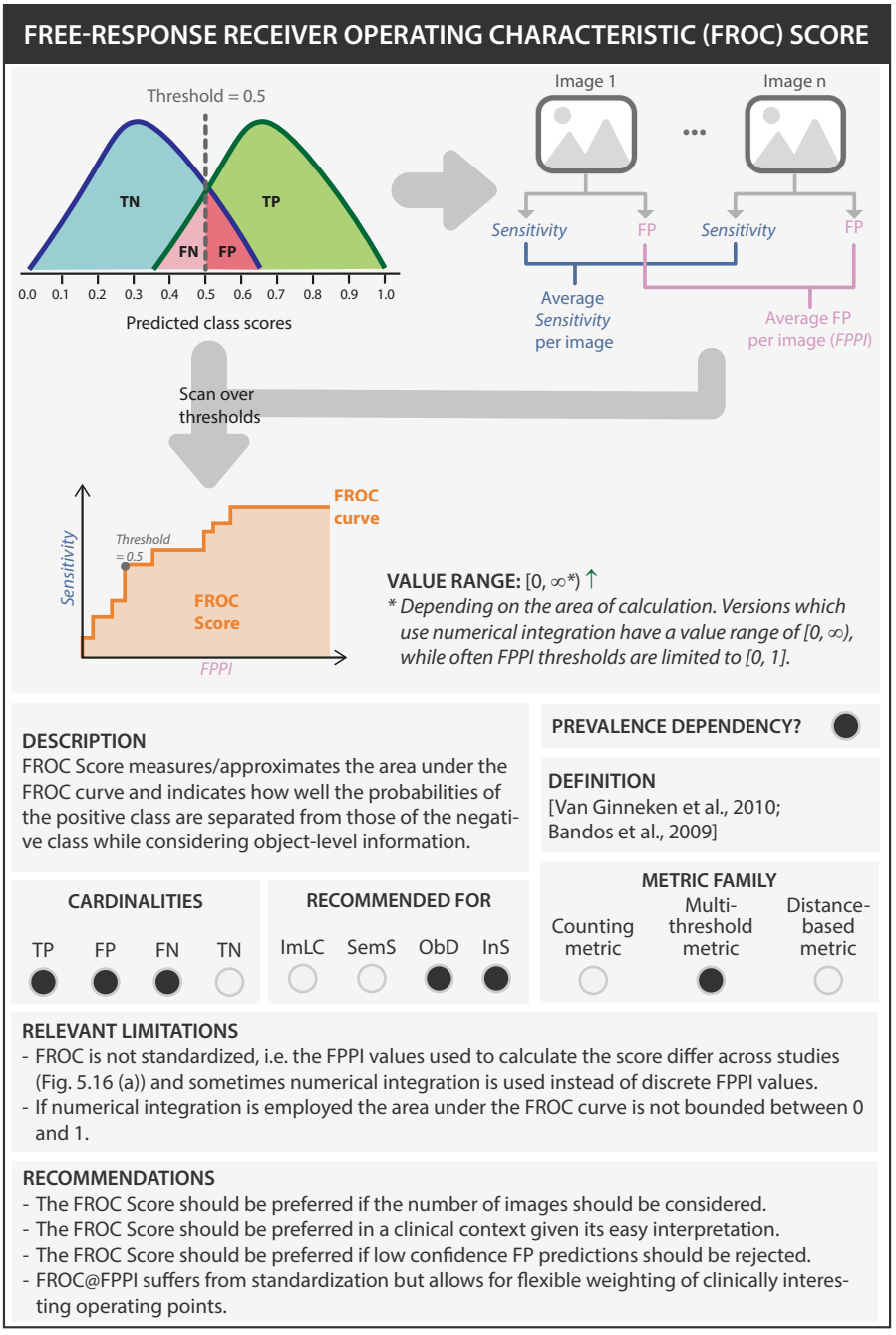
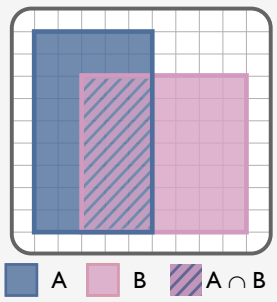


Figure A.17: Profile of FROC Score [Van Ginneken et al., 2010; Bandos et al., 2009]. This multi-threshold counting metric relies on the True Positive (TP), False Positive (FP) and False Negative (FN) cardinalities of the dynamic confusion matrix based on multiple thresholds (see Figures 2.6 and 2.7) and ranges between 0 and 1. In contrast to AUROC, it operates on object-level. It is recommended to be used for Image-level Classification (ImLC), Object Detection (ObD) and InS problems. Further abbreviation: Semantic Segmentation (SemS).

A.3.4 Overlap-based metrics

DICE SIMILARITY COEFFICIENT (DSC)
Sørensen–Dice Coefficient, F_1 Score, Balanced F Score



$$DSC(A,B) = \frac{2\,|A \cap B|}{|A| + |B|} = \frac{2\,PPV \cdot Sensitivity}{PPV + Sensitivity}$$

VALUE RANGE: [0, 1] ↑

DESCRIPTION
DSC measures the overlap between two structures.

DEFINITION
[Dice, 1945]

RELATIONS
DSC is closely related to the IoU: $DSC = \frac{2\,IoU}{1 + IoU}$

DSC is equal to the F_1 Score ($\beta = 1$ in F_β Score).

CARDINALITIES

TP	FP	FN	TN
●	●	●	○

RECOMMENDED FOR

ImLC	SemS	ObD	InS
●	●	●	●

METRIC FAMILY

Counting metric	Multi-threshold metric	Distance-based metric
●	○	○

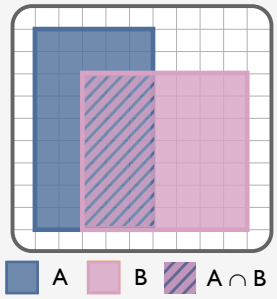
RELEVANT LIMITATIONS
- DSC is unaware of shapes, distances and centers (Figs. 5.3 (a) and (b), Figs. 5.8 (a) and (b)).
- DSC treats oversegmentation and undersegmentation differently (Fig. 5.4 (b)).
- DSC penalizes missed pixels more in small objects (Figs. 5.7 (a)).
- DSC does not compensate for inter-rater variability (Fig. 5.10 (a)).
- DSC is undefined if both the reference and prediction are empty (Fig. 5.10 (c)), and ad hoc rules need to be established depending on the use case.

RECOMMENDATIONS
- An overlap-based metric, such as DSC, should be used in most cases for segmentation assessment. An exception is the case of consistently tiny structures along with a noisy reference.
- DSC should generally be used in combination with a boundary-based metric due to its shape unawareness.

Figure A.18: Profile of the Dice Similarity Coefficient (DSC) [Dice, 1945]. This overlap-based or per-class counting metric measures the overlap between structures based on the True Positive (TP), False Positive (FP) and False Negative (FN) cardinalities of the confusion matrix (see Figure 2.6) and ranges between 0 and 1. It is recommended to be used for Semantic Segmentation (SemS), Object Detection (ObD) and InS problems. Further abbreviations: Image-level Classification (ImLC), True Negative (TN).

INTERSECTION OVER UNION (IoU)

Jaccard Index, Tanimoto Coefficient



$$IoU(A,B) = \frac{\blacksquare}{\blacksquare + \blacksquare - \blacksquare} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|} = \frac{|A \cap B|}{|A \cup B|} = \frac{PPV \cdot Sensitivity}{PPV + Sensitvity - PPV \cdot Sensitivity}$$

VALUE RANGE: [0, 1] ↑

DESCRIPTION

IoU measures the overlap between two structures. It is often referred to as **Box IoU** when comparing bounding boxes, **Mask IoU** when comparing segmentation masks or **Approx IoU** when comparing approximations of objects beyond bounding boxes.

DEFINITION

[Jaccard, 1912]

RELATIONS

IoU is closely related to the DSC: $IoU = \frac{DSC}{2 - DSC}$

CARDINALITIES

TP	FP	FN	TN
●	●	●	○

RECOMMENDED FOR

ImLC	SemS	ObD	InS
○	●	●	●

METRIC FAMILY

Counting metric	Multi-threshold metric	Distance-based metric
●	○	○

RELEVANT LIMITATIONS

- IoU is unaware of shapes, distances and centers (Figs. 5.3 (a) and (b), Figs. 5.8 (a) and (b)).
- IoU treats oversegmentation and undersegmentation differently (Fig. 5.4 (b)).
- IoU penalizes missed pixels more in small objects (Figs. 5.7 (a)).
- Box IoU is not a good representation of complex or disconnected structures (Fig. 5.8 (d)).
- IoU does not compensate for inter-rater variability (Fig. 5.10 (a)).
- IoU is undefined if both the reference and prediction are empty (Fig. 5.10 (c)), and ad hoc rules need to be established depending on the use case.
- A loose IoU-based localization criterion can be deceived by large predictions (Fig. 5.16 (c)).
- IoU behaves differently in 2D and 3D settings (Fig. 5.21 (b)).

RECOMMENDATIONS

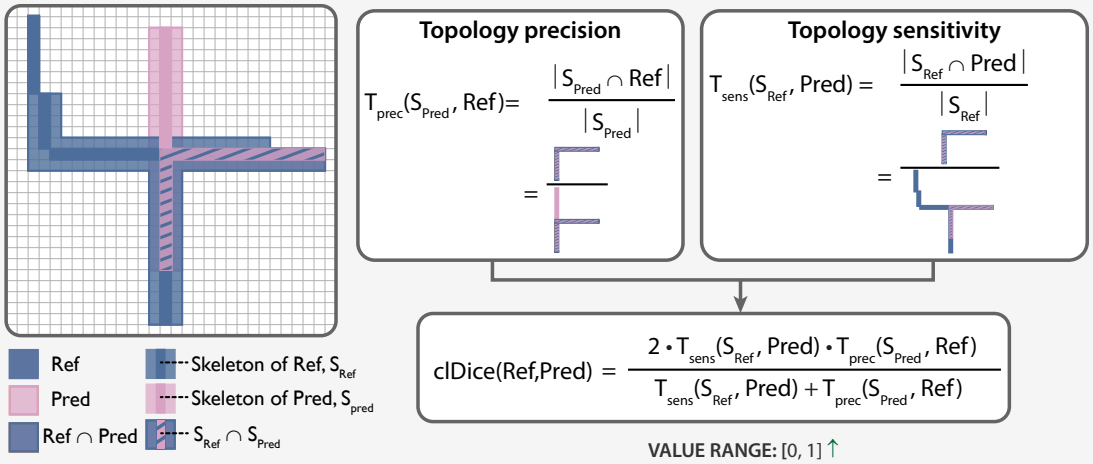
IoU for assessment of segmentation quality:

- An overlap-based metric, such as IoU, should be used in most cases for segmentation assessment. An exception is the case of consistently tiny structures along with a noisy reference.
- IoU should generally be used in combination with a boundary-based metric due to its shape unawareness.avoided, such as the distance parameter in Boundary IoU.

IoU should not be used localization criterion if contour agreement is important for deciding on a match between predicted and reference object.

Figure A.19: Profile of the Intersection over Union (IoU) [Jaccard, 1912]. This overlap-based or per-class counting metric measures the overlap between structures based on the True Positive (TP), False Positive (FP) and False Negative (FN) cardinalities of the confusion matrix (see Figure 2.6) and ranges between 0 and 1. It is recommended to be used for Semantic Segmentation (SemS), Object Detection (ObD) and InS problems. Further abbreviations: Image-level Classification (ImLC), True Negative (TN).

CENTER LINE DICE SIMILARITY COEFFICIENT (clDICE)



VALUE RANGE: [0, 1] ↑

DESCRIPTION
clDice measures the overlap between two tubular shaped structures. The formula is similar to the usual DSC, but relies on the topology precision and sensitivity which are based on the skeletons of the structures.

DEFINITION
[Shit et al., 2021]

CARDINALITIES

TP	FP	FN	TN
●	●	●	○

RECOMMENDED FOR

ImLC	SemS	ObD	InS
○	●	○	●

METRIC FAMILY

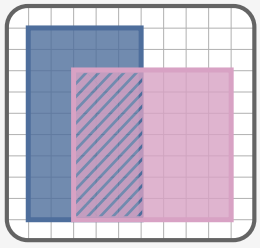
Counting metric	Multi-threshold metric	Distance-based metric
●	○	○

RELEVANT LIMITATIONS
- clDice is unaware of distances and center points (Figs. 5.3 (a) and (b), Fig. 5.8 (a)).
- clDice penalizes missed pixels more in small objects (Figs. 5.7 (a)).
- clDice is undefined if the reference and prediction are empty (Fig. 5.10 (c)), and ad hoc rules need to be established depending on the use case.

RECOMMENDATIONS
- clDice should be preferred over the standard DSC in the presence of tubular structures.
- clDice should be preferred over the standard DSC if structure center lines are of particular interest.

Figure A.20: Profile of the Centerline Dice Similarity Coefficient (clDice) [Shit et al., 2021]. This overlap-based or per-class counting metric measures the overlap between tubular structures based on the True Positive (TP), False Positive (FP) and False Negative (FN) cardinalities of the confusion matrix (see Figure 2.6) and ranges between 0 and 1. It is recommended to be used for Semantic Segmentation (SemS) and InS problems. Further abbreviations: Image-level Classification (ImLC), Object Detection (ObD), and True Negative (TN).

INTERSECTION OVER REFERENCE (IoR)
Pixel-level Sensitivity



$$IoR(A,B) = \frac{\square}{\square} = \frac{|A \cap B|}{|A|}$$

VALUE RANGE: [0, 1] ↑

DESCRIPTION
The IoR measures the overlap between two structures. It is defined as the pixel- level Sensitivity and only f ocuses on FN pixels.

DEFINITION
[Maška et al., 2014]

RECOMMENDED FOR

ImLC	SemS	ObD	InS
○	○	●	●

CARDINALITIES

TP	FP	FN	TN
●	○	●	○

METRIC FAMILY

Counting metric	Multi-threshold metric	Distance-based metric
●	○	○

RELEVANT LIMITATIONS
- IoR penalizes missed pixels more in small objects (Fig. 5.7 (a)).
- IoR treats oversegmentation and undersegmentation differently (Fig. 5.4 (b)).
- IoR is unaware of shapes, distances and centers (Figs. 5.3 (a) and (b), Figs. 5.8 (a) and (b)).
- IoR is undefined if the reference and prediction are empty (Fig. 5.10 (c)), and ad hoc rules need to be established depending on the use case.
- A loose IoR-based localization criterion can be deceived by large predictions (Fig. 5.16 (c)).
- IoR is rather uncommon and can therefore not be used for comparison with other publications.
- IoR is typically used only as a localization criterion.

RECOMMENDATIONS
- IoR should be used if there is no interest in predicting object boundaries.
- IoR should be preferred to IoU in the case of touching structures, in which one prediction may overlap multiple reference objects. In this case, assignment of one prediction to multiple reference objects must be enabled and a penalty for this type of algorithm error ("non-split error") introduced.

Figure A.21: Profile of the Intersection over Reference (IoR) [Maška et al., 2014]. This overlap-based or per-class counting metric measures the overlap between structures based on the True Positive (TP) and False Negative (FN) cardinalities of the confusion matrix (see Figure 2.6) and ranges between 0 and 1. It is recommended to be used as a localization criterion for Object Detection (ObD) and InS problems. Further abbreviations: Image-level Classification (ImLC), Semantic Segmentation (SemS), and True Negative (TN).

A.3.5 Distance-based metrics

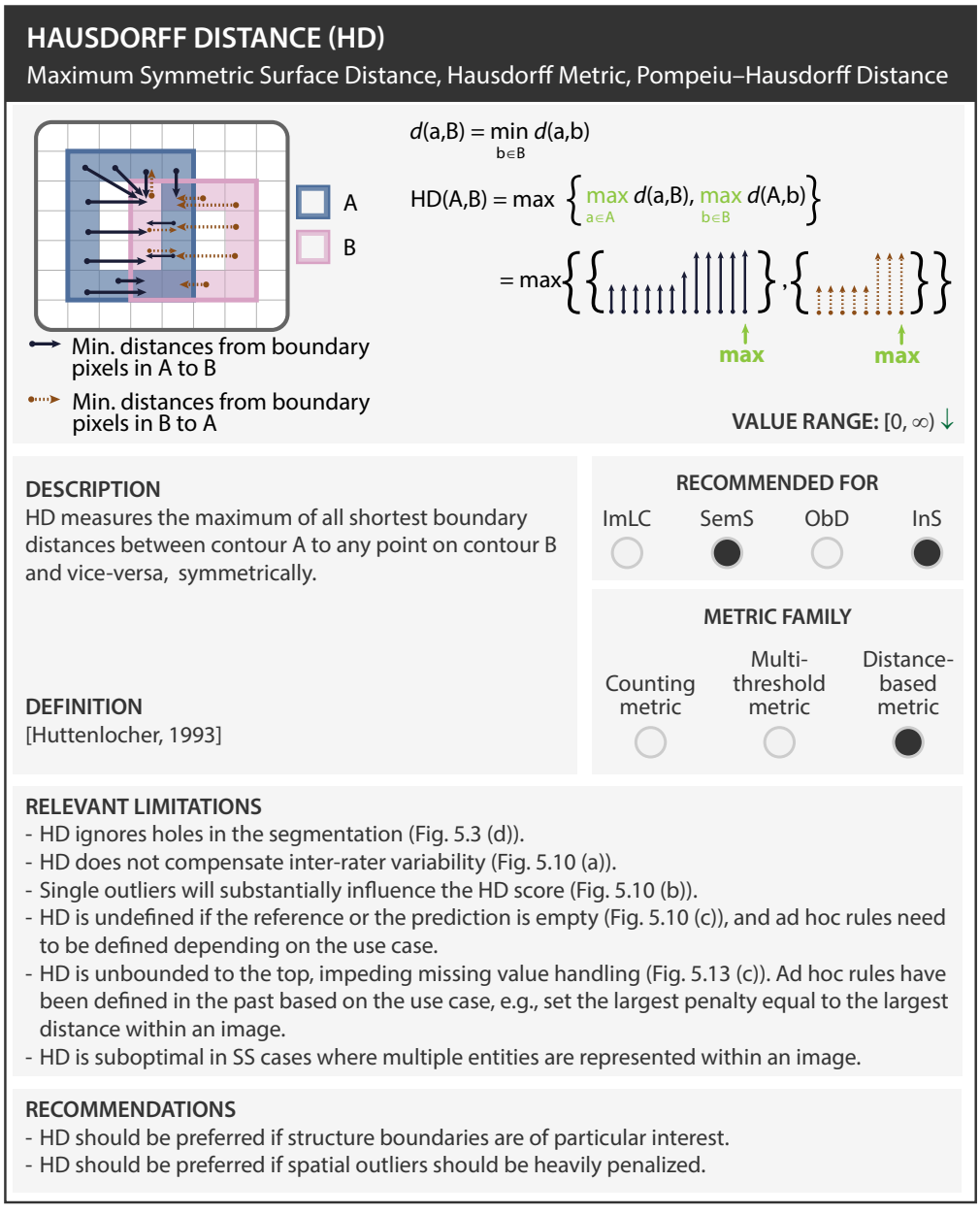
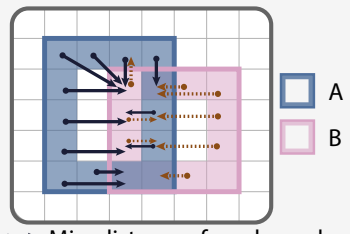


Figure A.22: Profile of the Hausdorff Distance (HD) [Huttenlocher et al., 1993]. This distance-based metric and ranges between 0 and ∞, where a value of 0 corresponds to a perfect prediction. It is recommended to be used for Semantic Segmentation (SemS) and Instance Segmentation (InS) problems. Further abbreviations: Image-level Classification (ImLC), Object Detection (ObD).

HAUSDORFF DISTANCE 95 PERCENTILE (HD95)



$$d_{95}(A,B) = \underset{a \in A}{x_{95}} \left\{ \min_{b \in B} d(a,b) \right\}$$

$$HD95(A,B) = \max \left\{ d_{95}(A,B), d_{95}(B,A) \right\}$$

$$= \max \left\{ \{ \}, \{ \} \right\}$$

x_{95} x_{95}

VALUE RANGE: $[0, \infty)$ ↓

DESCRIPTION
HD95 measures the maximum of all the 95 percentile of all shortest boundary distances from each structure to the other. The 95 percentile can be replaced by any other percentile.

DEFINITION
[Huttenlocher, 1993]

RECOMMENDED FOR

ImLC	SemS	ObD	InS
○	●	○	●

METRIC FAMILY

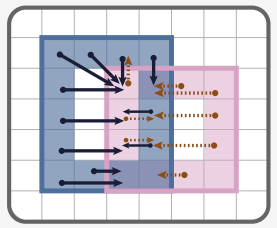
Counting metric	Multi-threshold metric	Distance-based metric
○	○	●

RELEVANT LIMITATIONS
- HD95 ignores holes in the segmentation (Fig. 5.3 (d)).
- HD95 does not compensate inter-rater variability (Fig. 5.10 (a)).
- HD95 is undefined if the reference or the prediction is empty (Fig. 5.10 (c)), and ad hoc rules need to be defined depending on the use case.
- HD95 is unbounded to the top, impeding missing value handling (Fig. 5.13 (c)). Ad hoc rules have been defined in the past based on the use case, e.g., set the largest penalty equal to the largest distance within an image.
- HD95 is suboptimal in SS cases where multiple entities are represented within an image.

RECOMMENDATIONS
- HD95 should be preferred if structure boundaries are of particular interest.
- HD95 should be preferred if (extreme) spatial outliers should be compensated.

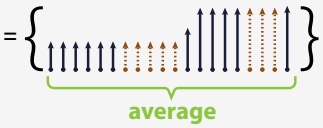
Figure A.23: Profile of the Hausdorff Distance 95 Percentile (HD95) [Huttenlocher et al., 1993]. This distance-based metric ranges between 0 and ∞, where a value of 0 corresponds to a perfect prediction. It can be generalized to all percentiles by replacing the 95th percentile with any other. It is recommended to be used for Semantic Segmentation (SemS) and Instance Segmentation (InS) problems. Further abbreviations: Image-level Classification (ImLC), Object Detection (ObD).

AVERAGE SYMMETRIC SURFACE DISTANCE (ASSD)



$$d(a,B) = \min_{b \in B} d(a,b)$$

$$ASSD(A,B) = \frac{\sum_{a \in A} d(a,B) + \sum_{b \in B} d(A,b)}{|A| + |B|}$$



VALUE RANGE: $[0, \infty)$ ↓

DESCRIPTION
ASSD measures the average of all shortest boundary distances between contour A to any point on contour B and vice versa, symmetrically.

DEFINITION
[Yeghiazaryan, Varduhi and Voiculescu, 2015]

RECOMMENDED FOR

ImLC	SemS	ObD	InS
○	●	○	●

METRIC FAMILY

Counting metric	Multi-threshold metric	Distance-based metric
○	○	●

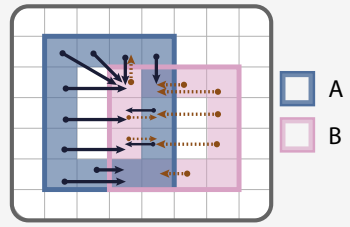
RELEVANT LIMITATIONS
- ASSD ignores holes in the segmentation (Fig. 5.3 (d)).
- ASSD does not compensate for inter-rater variability (Fig. 5.10 (a)).
- ASSD is undefined if the reference or the prediction is empty (Fig. 5.10 (c)), and ad hoc rules need to be established depending on the use case.
- ASSD is unbounded to the top, impeding missing value handling (Fig. 5.13 (c)). Ad hoc rules have been defined in the past based on the use case, e.g., set the largest penalty equal to the largest distance within an image.

RECOMMENDATIONS
- ASSD should be preferred if structure boundaries are of particular interest.
- ASSD should be preferred if the structure sizes in the data are of low variability.
- ASSD is recommended to compare boundaries of single individual objects and areas. It may be less suitable to compare relationships between surfaces of many dense objects.

Figure A.24: Profile of the Average Symmetric Surface Distance (ASSD) [Yeghiazaryan and Voiculescu, 2015]. This distance-based metric ranges between 0 and ∞, where a value of 0 corresponds to a perfect prediction. It is recommended to be used for Semantic Segmentation (SemS) and Instance Segmentation (InS) problems. Further abbreviations: Image-level Classification (ImLC), Object Detection (ObD).

MEAN AVERAGE SURFACE DISTANCE (MASD)

MEAN SURFACE DISTANCE



$$d(a,B) = \min_{b \in B} d(a,b)$$

$$MASD(A,B) = \frac{1}{2}\left(\frac{\sum_{a \in A} d(a,B)}{|A|} + \frac{\sum_{b \in B} d(A,b)}{|B|}\right)$$

$$= \left\{\left\{\right\}_{average}, \left\{\right\}_{average}\right\}_{average}$$

VALUE RANGE: $[0, \infty)$ ↓

DESCRIPTION
MASD measures the average of averages of all shortest distances from all sampled points on one boundary to any other point on another boundary.

DEFINITION
[Beneš and Zitová, 2015]

RECOMMENDED FOR

ImLC	SemS	ObD	InS
○	●	○	●

METRIC FAMILY

Counting metric	Multi-threshold metric	Distance-based metric
○	○	●

RELEVANT LIMITATIONS
- MASD ignores holes in the segmentation (Fig. 5.3 (d)).
- MASD only depends on one boundary, if the prediction is very small and next to the reference-boundary (Fig. 5.7 (c)).
- MASD does not compensate for inter-rater variability (Fig. 5.10 (a)).
- MASD is undefined if the reference or the prediction is empty (Fig. 5.10 (c)), and ad hoc rules need to be established depending on the use case.
- MASD is unbounded to the top, impeding missing value handling (Fig. 5.13 (c)). Ad hoc rules have been defined in the past based on the use case, e.g., set the largest penalty equal to the largest distance within an image.

RECOMMENDATIONS
- MASD should be preferred if structure boundaries are of particular interest.
- MASD should be preferred if the structure sizes in the data are of high variability.
- MASD makes most sense to measure the distance of a test segmentation to a reference segmentation, not in the other direction.

Figure A.25: Profile of the Mean Absolute Surface Distance (MASD) [Beneš and Zitová, 2015]. This distance-based metric ranges between 0 and ∞, where a value of 0 corresponds to a perfect prediction. It is recommended to be used for Semantic Segmentation (SemS) and Instance Segmentation (InS) problems. Further abbreviations: Image-level Classification (ImLC), Object Detection (ObD).

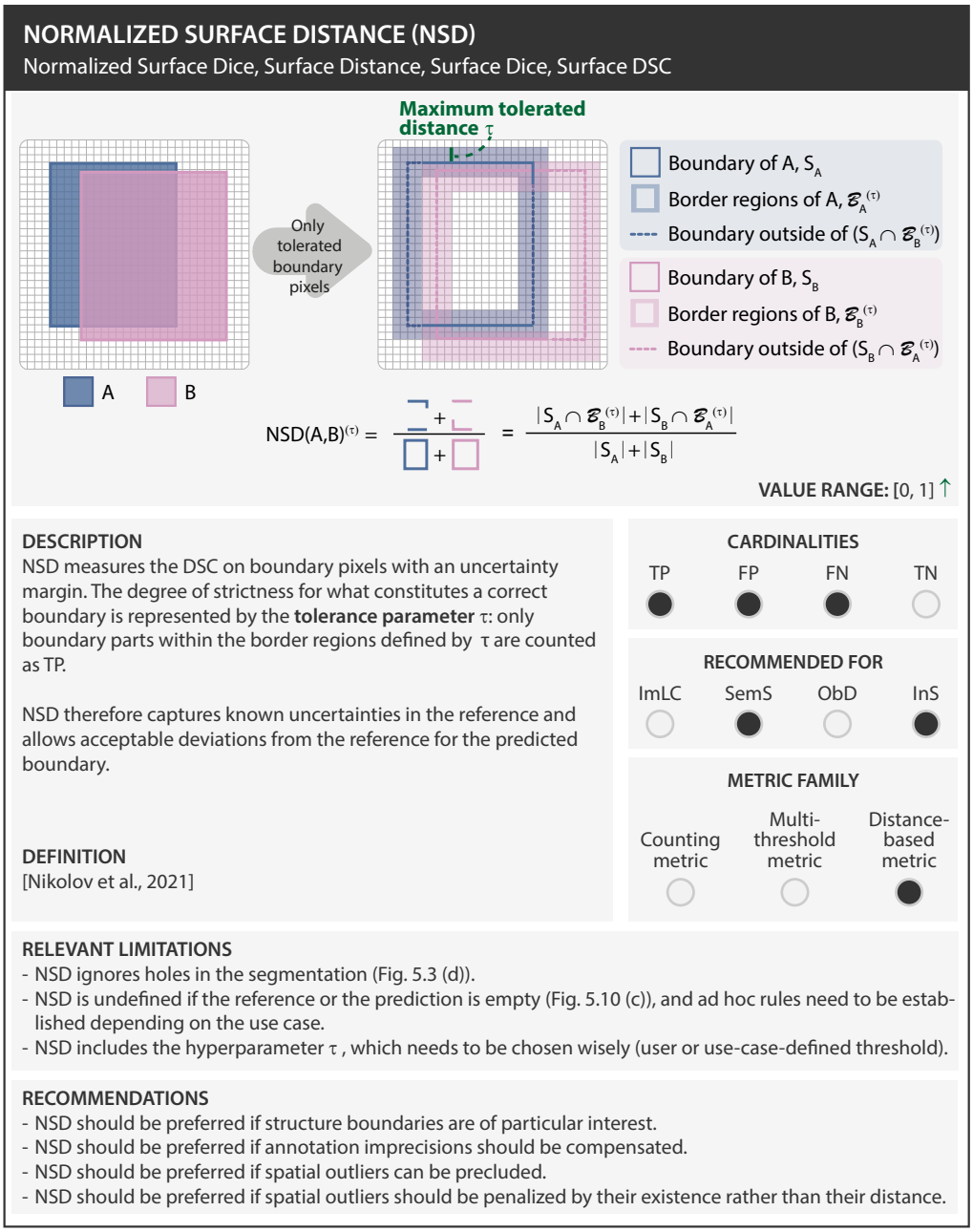
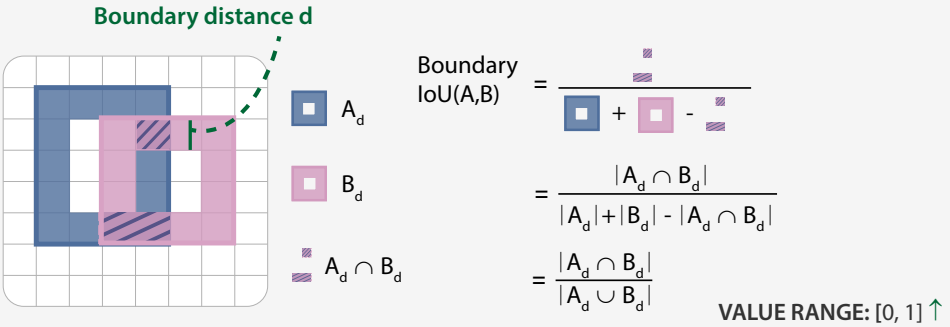


Figure A.26: Profile of the Normalized Surface Distance (NSD) [Tharwat, 2020]. This distance-based metric relies on the True Positive (TP), False Positive (FP), False Negative (FN) cardinalities of the confusion matrix (see Figure 2.6) and ranges between 0 and 1. It features a hyperparameter, specifying the maximum tolerated distance. It is recommended to be used for Semantic Segmentation (SemS) and Instance Segmentation (InS) problems. Further abbreviations: True Negative (TN), Image-level Classification (ImLC), Object Detection (ObD).

BOUNDARY INTERSECTION OVER UNION (BOUNDARY IoU)



VALUE RANGE: [0, 1] ↑

DESCRIPTION
Boundary IoU measures the overlap of two structures, without imposing a tolerance, by computing the IoU of the boundaries up to a pre-defined width.
The **width** is controlled by the parameter d.

DEFINITION
[Cheng et al., 2021]

RECOMMENDED FOR

ImLC	SemS	ObD	InS
○	●	●	●

CARDINALITIES

TP	FP	FN	TN
●	●	●	○

METRIC FAMILY

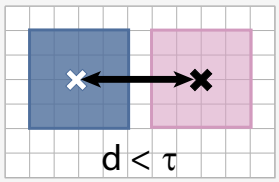
Counting metric	Multi-threshold metric	Distance-based metric
○	○	●

RELEVANT LIMITATIONS
- Boundary IoU may yield a perfect value for only predicting a boundary (Fig. 5.3 (c)), and ad hoc rules need to be established depending on the use case.
- Boundary IoU is undefined if the reference and prediction are empty (Fig. 5.10 (c)).

RECOMMENDATIONS
- Boundary IoU should be preferred if structure boundaries are of particular interest.
- Boundary IoU should be preferred if boundary errors should be penalized as severe inconsistencies.
- Boundary IoU should be preferred if spatial outliers can be precluded.
- Boundary IoU should be preferred if spatial outliers should be penalized by their existence rather than their distance.

Figure A.27: Profile of the Boundary Intersection over Union (Boundary IoU) [Cheng et al., 2021]. This distance-based metric relies on the True Positive (TP), False Positive (FP), False Negative (FN) cardinalities of the confusion matrix (see Figure 2.6) and ranges between 0 and 1. It features a hyperparameter, specifying the width to the boundaries. It is recommended to be used for Semantic Segmentation (SemS), Object Detection (ObD) and Instance Segmentation (InS) problems. Further abbreviations: True Negative (TN), Image-level Classification (ImLC).

CENTER DISTANCE
Centroid distance, center distance-based criterion



Reference
Prediction
Reference center
Prediction center

VALUE RANGE: [0, ∞) ↓

DESCRIPTION
The Center Distance measures the (typically Euclidean) distance between the reference and predicted center point of an object.
The prediction is considered as a hit, if the distance is smaller than a predefined threshold τ. Center Distance can also be used as a performance metric.

DEFINITION
[Gurcan et al., 2010]

RECOMMENDED FOR

ImLC	SemS	ObD	InS
○	○	●	●

METRIC FAMILY

Counting metric	Multi-threshold metric	Distance-based metric
○	○	●

RELEVANT LIMITATIONS
- The Center Distance ignores the overlap between objects (Fig. 5.19 (b)).
- The Center Distance is a bad approximation of tubular structures (Fig. 5.19 (d)).
- The Center Distance is often only a rough approximation of the real object center distance because the object center is often computed imprecisely (e.g., as the bounding box center, intensity maximum position, etc.).
- The Center Distance is typically used only as a localization criterion.

RECOMMENDATIONS
The Center Distance should be preferred if only a rough estimate of the object location is available (only position, imprecise bounding box, complex shape, etc.).

Figure A.28: Profile of the Center Distance [Gurcan et al., 2010]. This distance-based metric ranges between 0 and ∞. It is recommended to be used as a localization criterion for Object Detection (ObD) and Instance Segmentation (InS) problems. Further abbreviations: Image-level Classification (ImLC), Semantic Segmentation (SemS).

A.3.6 Point-based metrics

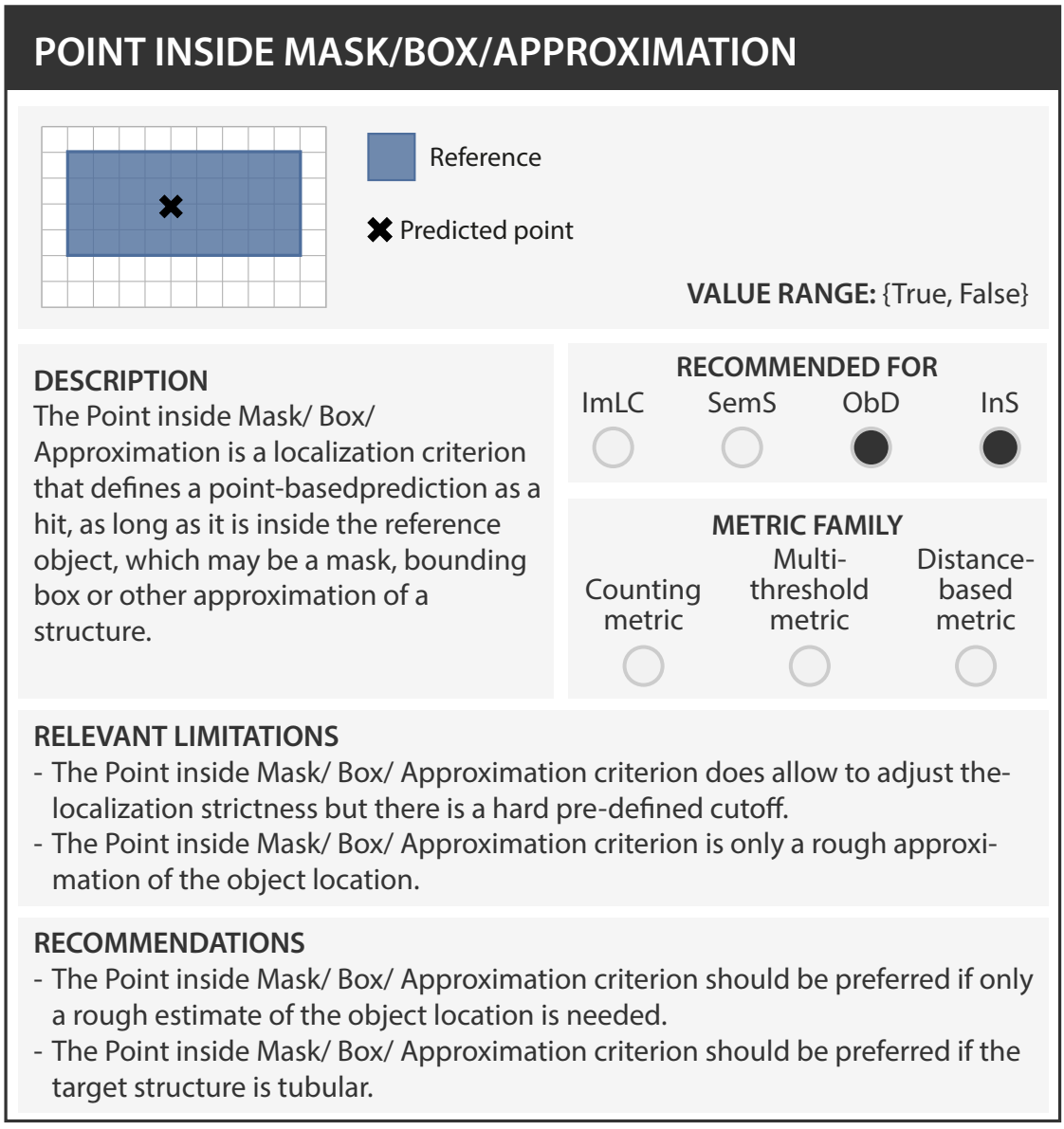


Figure A.29: Profile of the Point inside Mask/ Box/ Approximation criterion. This point-based metric does not rely on the cardinalities of the confusion matrix and is a binary decision. It is recommended to be used as a localization criterion for Object Detection (ObD) and Instance Segmentation (InS) problems. Further abbreviations: True Positive (TP), False Positive (FP), False Negative (FN), True Negative (TN), Image-level Classification (ImLC), Semantic Segmentation (SemS).

A.3.7 Calibration metrics

BRIER SCORE (BS)

$$BS = \frac{1}{n} \sum_{t=1}^{n} (p_t - y_t)^2$$

VALUE RANGE: [0, 1] ↓

DESCRIPTION
The BS measures the calibration quality of a prediction compared and is a proper scoring rule. It is the mean squared error of a predicted class score p_t and the actual outcome y_t, typically defined as 1 or 0.

VARIANT
Brier Skill Score (BSS): normalizes BS by the BS of a naive system.

DEFINITION
[Gneiting and Raftery, 2007]

RECOMMENDED FOR

ImLC	SemS	ObD	InS
●	○	●	●

METRIC FAMILY

Counting metric	Multi-threshold metric	Distance-based metric	Calibration metric
○	○	○	●

TYPE OF CALIBRATION

Top-label	Marginal	Canonical
○	○	●

RELEVANT LIMITATIONS
- BS does not take into account ordinal grading (Fig. 5.4 (d)).
- BS simultaneously assesss the discrimination and calibration performance in one score and can thus only be used for relative assessment of calibration.
- BS is highly prevalence-dependent, implying that scores may drastically change when the prevalence changes.
- Predicted class scores linked to sporadic events have little effect on the score.

RECOMMENDATIONS
BS should be preferred if the calibration and discrimination should be assessed in a single score for comparative calibration asssessment.

Figure A.30: Profile of the Brier Score (BS) [Yeghiazaryan and Voiculescu, 2015]. It is recommended to be used for Image-level Classification (ImLC), Object Detection (ObD) and Instance Segmentation (InS) problems. Further abbreviations: Semantic Segmentation (SemS).

NEGATIVE LOG LIKELIHOOD (NLL)

$$NLL = -\frac{1}{n}\sum_{t=1}^{n} y_t \bullet \log(p_t)$$

n = 2:

$$NLL = -(y_t \bullet \log(p_t) + (1 - y_t) \bullet \log(1 - p_t))$$

VALUE RANGE: $[0, \infty)$ ↓

DESCRIPTION
The NLL measures the calibration quality of a prediction. It is the negative logarithm of a predicted class score p_t and the actual outcome y_t, typically defined as 1 or 0.

DEFINITION
[Cybenko et al., 1998]

RECOMMENDED FOR

ImLC	SemS	ObD	InS
●	○	○	○

METRIC FAMILY

Counting metric	Multi-threshold metric	Distance-based metric	Calibration metric
○	○	○	●

TYPE OF CALIBRATION

Top-label	Marginal	Canonical
○	○	●

RELEVANT LIMITATIONS
- NLL does not take into account ordinal grading (Fig. 5.4 (d)).
- NLL simultaneously assesss the discrimination and calibration performance in one score and can thus only be used for relative assessment of calibration
- NLL introduces a stronger penalization of tail probabilities, i.e. overconfident predictions lead to higher losses and conservative models are favored.

RECOMMENDATIONS
NLL should be preferred if the calibration and discrimination should be assessed in a single score for comparative calibration asssessment.

Figure A.31: Profile of the Negative Log Likelihood (NLL) [Cybenko et al., 1998]. It is recommended to be used for Image-level Classification (ImLC) problems. Further abbreviations: Instance Segmentation (InS), Object Detection (ObD), Semantic Segmentation (SemS).

ROOT BRIER SCORE (RBS)

$$RBS = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(p_t - y_t)^2}$$

VALUE RANGE: [0, 1] ↓

DESCRIPTION
The RBS measures the calibration quality of a prediction and is a proper scoring rule. It is the square root of the mean squared error of a predicted class score p_t and the actual outcome y_t, typically defined as 1 or 0.

RBS calculates an asymptotically tight and unbiased upper bound of the canonical calibration error.

DEFINITION
[Gruber and Buettner, 2022]

RECOMMENDED FOR

ImLC	SemS	ObD	InS
●	○	●	●

METRIC FAMILY

Counting metric	Multi-threshold metric	Distance-based metric	Calibration metric
○	○	○	●

TYPE OF CALIBRATION

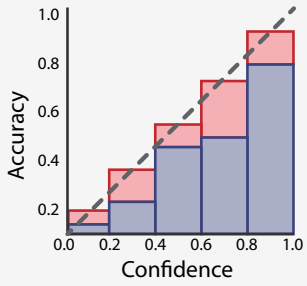
Top-label	Marginal	Canonical
○	○	●

RELEVANT LIMITATIONS
- RBS does not take into account ordinal grading (Fig. 5.4 (d)).
- For small sample sizes, the upper bound's tightness is unknown, i.e. it is unclear to what extent RBS over-estimates the canonical calibration error.

RECOMMENDATIONS
- RBS provides an interpretable measure for calibration.
- RBS should be used as an unbiased upper bound of the canonical calibration error.

Figure A.32: Profile of the Root Brier Score (RBS) [Gruber and Buettner, 2022]. This calibration metric ranges between 0 and 1, where a value of 0 corresponds to a perfectly calibrated prediction. It is recommended to be used for Image-level Classification (ImLC), Object Detection (ObD), and Instance Segmentation (InS) problems. Further abbreviations: Semantic Segmentation (SemS).

EXPECTED CALIBRATION ERROR (ECE)



BINNED DENSITY ESTIMATE

$$ECE = \sum_{m=1}^{n} \frac{|B_m|}{n} |Accuracy(B_m) - Confidence(B_m)|$$

$$= \text{Weighted Average} \{\ ,\ ,\ ,\ ,\ \}$$

VALUE RANGE: [0, 1] ↑

GENERAL DEFINITION

$$ECE = \mathbb{E}\left(\left|P\left(Y = \underset{j \in |K|}{\text{argmax}}\ g(X)_j \mid \underset{j \in |K|}{\max}\ g(X)_j\right) - \underset{j \in |K|}{\max}\ g(X)_j\right|\right)$$

DESCRIPTION

The ECE measures the calibration quality of a prediction and is the (top-label) ℓ_1 calibration error.

For a binned strategy of estimation of the conditional expectation of the calibration error, ECE is the weighted average between the amount of correct predictions (Accuracy) and the average predicted class scores (confidence) per bin B_m.

DEFINITION

[Naeini et al., 2015]

RECOMMENDED FOR

ImLC	SemS	ObD	InS
●	○	●	●

METRIC FAMILY

Counting metric	Multi-threshold metric	Distance-based metric	Calibration metric
○	○	○	●

TYPE OF CALIBRATION

Top-label	Marginal	Canonical
●	○	○

RELEVANT LIMITATIONS

- The number and size of the bins in the ECE is not standardized (Fig. 5.17 (b)).
- ECE is dependent on the sample size (Fig. 5.6 (c)).
- ECE may indicate a perfect calibration, although the predicted class scores are not fully calibrated.
- ECE only measures the calibration quality, not the discrimination.
- ECE only provides a discretization of the conditional expectation of the calibration error, which is not differentiable.

RECOMMENDATIONS

- ECE should be preferred if an interpretable estimate of the calibration error is requested.
- ECE should be preferred if class confusions are of equal importance.
- ECE should be preferred if only calibration quality should be assessed (without assessing the discrimination).
- ECE should be preferred if the empirical success rate should be visualized.

Figure A.33: Profile of the Expected Calibration Error (ECE) [Naeini et al., 2015]. It is recommended to be used for Image-level Classification (ImLC), Object Detection (ObD), and Instance Segmentation (InS) problems. Further abbreviations: Semantic Segmentation (SemS).

CLASS-WISE CALIBRATION ERROR (CWCE)

$$\text{CWCE} = \left(\sum_{k=1}^{K} w_k \, \mathbb{E}\left(\left(g(X)_k - P(Y = k \mid g(X)_k) \right)^2 \right) \right)^{1/2}$$

DESCRIPTION

CWCE measures the class-wise calibration performance as a specific case of the calibration error. The classes are weighted with class-specific weights w_k. CWCE can also be reported per class.

DEFINITION

[Kumar et al., 2019]

RECOMMENDED FOR

ImLC	SemS	ObD	InS
●	○	●	●

METRIC FAMILY

Counting metric	Multi-threshold metric	Distance-based metric	Calibration metric
○	○	○	●

TYPE OF CALIBRATION

Top-label	Marginal	Canonical
○	●	○

RELEVANT LIMITATIONS

- If a binning strategy is applied for density estimation, the number and size of the bins is not standardized (Fig. 5.17 (b)).
- CWCE may indicate a perfect calibration, although the scores are not calibrated w.r.t the canoncal calibration condition.
- CWCE only measures the calibration quality, not the discrimination.

RECOMMENDATIONS

- CWCE should be used for a class-wise calibration assessment.
- CWCE should be used in the case of a mismatch between class prevalences and class importance.
- CWCE should be used if only calibration quality should be assessed (without assessing the discrimination).
- CWCE should also be used to report per-class performance.

Figure A.34: Profile of the Class-wise Calibration Error (CWCE) [Kumar et al., 2019]. It is recommended to be used for Image-level Classification (ImLC), Object Detection (ObD), and Instance Segmentation (InS) problems. Further abbreviations: Semantic Segmentation (SemS).

KERNEL CALIBRATION ERROR (KCE)

$$\mathrm{KCE} = \left\| \mathbb{E}\left((g(X) - P(Y \mid g(X))) \bullet k(g(X), \cdot) \right) \right\|_{\mathcal{H}}$$

DESCRIPTION
KCE measures the canonical calibration performance as a specific case of the calibration error. It is based on a matrix-valued kernel k from a reproducing kernel Hilbert space $\mathcal{H}$.

KCE is an unbiased estimator of the calibration error.

DEFINITION
[Widmann et al., 2019; Gruber and Buettner, 2022]

RECOMMENDED FOR

ImLC	SemS	ObD	InS
●	○	●	●

METRIC FAMILY

Counting metric	Multi-threshold metric	Distance-based metric	Calibration metric
○	○	○	●

TYPE OF CALIBRATION

Top-label	Marginal	Canonical
○	○	●

RELEVANT LIMITATIONS
KCE cannot be used as an interpretable estimate of the calibration error and should only be used for comparative calibration assessment.

RECOMMENDATIONS
- KCE should be preferred for a canonical calibration assessment.
- KCE should be preferred in the case of no mismatch between class prevalences and class importance.
- KCE should be used for comparative calibration assessment as an unbiased estimator of the calibration error.

Figure A.35: Profile of the Kernel Calibration Error (KCE) [Widmann, 2020; Gruber and Buettner, 2022]. It is recommended to be used for Image-level Classification (ImLC), Object Detection (ObD), and Instance Segmentation (InS) problems. Further abbreviations: Semantic Segmentation (SemS).

EXPECTED CALIBRATION ERROR KERNEL DENSITY ESTIMATE (ECE^{KDE})

$$ECE^{KDE} = \frac{1}{n} \sum_{j=1}^{n} \left| \widehat{\mathbb{E}\left(y \mid g(X)\right)}\big|_{g(x_j)} - g(x_j) \right|$$

$$\mathbb{E}\left(y \mid g(X)\right) \approx \frac{\sum_{j=1}^{n} k(g(X), g(x_j)) y_j}{\sum_{j=1}^{n} k(g(X), g(x_j))} =: \widehat{\mathbb{E}\left(y \mid g(X)\right)}$$

$\widehat{\mathbb{E}\left(y \mid g(X)\right)}\big|_{g(x_j)}$:

$\widehat{\mathbb{E}\left(y \mid g(X)\right)}$ at $g(X) = g(x_j)$

DESCRIPTION

ECE^{KDE} uses a kernel density estimate for estimating the conditional expectation in contrast to the binning strategy applied by the standard ECE.

DEFINITION

[Popordanoska et al., 2022].

RECOMMENDED FOR

ImLC	SemS	ObD	InS
●	○	●	●

METRIC FAMILY

Counting metric	Multi-threshold metric	Distance-based metric	Calibration metric
○	○	○	●

TYPE OF CALIBRATION

Top-label	Marginal	Canonical
○	○	●

RELEVANT LIMITATIONS

- ECE^{KDE} does not scale to large number of classes.
- ECE^{KDE} is a biased estimator, particularly unreliable for small sample sizes.

RECOMMENDATIONS

- ECE^{KDE} should be preferred for a canonical calibration assessment.
- ECE^{KDE} should be preferred in the case of no mismatch between class prevalences and class importance and if class confusions are of unequal severity.
- ECE^{KDE} should be preferred if only calibration quality should be assessed (without assessing the discrimination).

Figure A.36: Profile of the Expected Calibration Error Kernel Density Estimate (ECE^{KDE}) [Popordanoska et al., 2022]. This calibration metric ranges between 0 and 1, where a value of 1 corresponds to a perfectly calibrated prediction. It is recommended to be used for Image-level Classification (ImLC), Object Detection (ObD), and Instance Segmentation (InS) problems. Further abbreviations: Semantic Segmentation (SemS).

A.4 Metric recommendations for common biomedical use cases

The framework can not only be applied to challenges, but to any classification-related biomedical image analysis scenario. In Figures A.37-A.40, we present four common use cases for each of the considered problem categories[†]. It can be seen that use cases that are concerned with assessing similar properties yield the same metric recommendations. For example, the semantic segmentation of large objects, such as use cases (a) "lung cancer cell segmentation from microscopy images" [Castilla et al., 2018] and (b) "liver segmentation in Computed Tomography (CT) images" [Antonelli/Reinke et al., 2022; Simpson et al., 2019] in Figure A.38, yield a similar problem fingerprint for very different problems from different domains and modalities. Still, the assessed properties are similar, yielding the same metric recommendations. We present the following biomedical use cases:

Image-level classification (Figure A.37)

- **(a)** Frame-based sperm motility classification based on microscopy time-lapse video containing human spermatozoa [Haugen et al., 2019]
- **(b)** Disease classification in dermoscopic images [Codella et al., 2019]
- **(c)** Classification of the overall autophagy stage for a collection of cells [Zhang et al., 2020; Nagao et al., 2020]
- **(d)** Identification of new lesions in brain multi-modal Magnetic Resonance Imaging (MRI) images of patients with Multiple Sclerosis (MS) [Kofler et al., 2022; Commowick et al., 2018]

Semantic segmentation (Figure A.38)

- **(a)** Lung cancer cell segmentation from microscopy images [Castilla et al., 2018]
- **(b)** Liver segmentation in CT images [Antonelli/Reinke et al., 2022; Simpson et al., 2019]
- **(c)** Aneurysm segmentation in Time of Flight Magnetic Resonance Angiographs (TOF-MRA) images [Timmins et al., 2021]
- **(d)** Labeling of invasive/ non-invasive/ benign lesions on breast Whole Slide Imaging (WSI) [Aresta et al., 2019]

Object detection (Figure A.39)

- **(a)** Cell detection and tracking during the autophagy process in time-lapse microscopy [Zhang et al., 2020; Nagao et al., 2020]
- **(b)** MS lesion detection in multi-modal brain MRI images [Kofler et al., 2022; Commowick et al., 2018]
- **(c)** Polyp detection in colonoscopy videos with predefined sensitivity of 0.9 [Sánchez-Montes et al., 2019; Bernal et al., 2019]
- **(d)** Mitosis detection in histopathology images [Aubreville et al., 2022]

Instance segmentation (Figure A.40)

- **(a)** Instance segmentation of neurons from the fruit fly in 3D multi-color light microscopy images [Tirian and Dickson, 2017; Mais et al., 2020; Meissner et al., 2022]
- **(b)** Surgical instrument instance segmentation in colonoscopy videos [Maier-Hein et al., 2021]
- **(c)** Cell nuclei instance segmentation in time-lapse light microscopy with a subsequent goal of cell tracking [Ulman et al., 2017]
- **(d)** MS lesion segmentation in multi-modal brain MRI images [Kofler et al., 2022; Commowick et al., 2018]

[†]For better readability, we omitted all fingerprints that led to a negative instantiation (FALSE) in all use cases. Images in the figure were taken from the references provided on this page.

Figure A.37: Instantiation of the metric recommendation framework for common biomedical image classification use cases. For every use case, we show exemplary input and output images, the instantiation of the problem fingerprint (with red cells referring to 'no' and green cells referring to 'yes'), and the resulting metric recommendations based on the fingerprint. Used abbreviations: Average Precision (AP), Area under the Receiver Operating Characteristic Curve (AUROC), Balanced Accuracy (BA), Expected Cost (EC), Expected Calibration Error (ECE), Positive Likelihood Ratio (LR+), Kernel Calibration Error (KCE), Magnetic Resonance Imaging (MRI), Multiple Sclerosis (MS), Negative Log Likelihood (NLL), Root Brier Score (RBS).

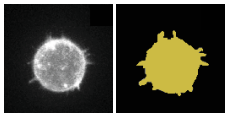
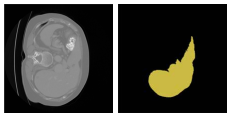
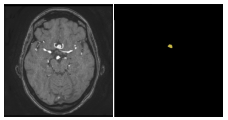
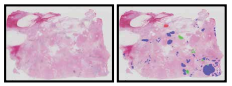
	(a) Lung cancer cell segmentation from microscopy images	(b) Liver segmentation in CT images	(c) Aneurysm segmentation in TOF-MRA images	(d) Labeling of invasive/ non-invasive/ benign lesions on breast WSIs
				
Particular importance of structure boundaries?				
Particular importance of structure volume?				
Unequal interest across classes?				
Unequal severity of class confusions?				
Compensation for class imbalances requested?				
Handling of spatial outliers?	Existence-based	Existence-based	Distance-based with outlier focus	Existence-based
Compensation for annotation imprecisions requested?				
Small size of structures?				
High variability of structure sizes?				
Possibility of multiple labels per unit?				
Possibility of overlapping or touching target structures?				
Possibility of disconnected target structure(s)?				
Presence of class imbalance?				
High inter-rater variability?				
Possibility of spatial outliers in reference annotation?				
Possibility of reference without target structure(s)?				
Possibility of algorithm output not containing target structure(s)?				
Possibility of invalid algorithm output?				
Recommendations	**Overlap-based metric:** DSC **Boundary-based metric:** NSD	**Overlap-based metric:** DSC **Boundary-based metric:** NSD	**Overlap-based metric:** DSC **Boundary-based metric:** HD95	**Overlap-based metric:** F_β Score **No boundary-based metric recommended** (possibility of overlapping or touching structures)

Figure A.38: Instantiation of the metric recommendation framework for common biomedical semantic segmentation use cases. For every use case, we show exemplary input and output images, the instantiation of the problem fingerprint (with red cells referring to 'no' and green cells referring to 'yes'), and the resulting metric recommendations based on the fingerprint. Used abbreviations: Computed Tomography (CT), Dice Similarity Coefficient (DSC), Hausdorff Distance 95 Percentile (HD95), Normalized Surface Distance (NSD), Time of Flight Magnetic Resonance Angiographs (TOF-MRA), Whole Slide Imaging (WSI).

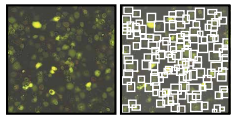
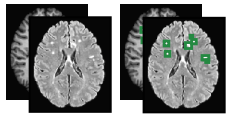
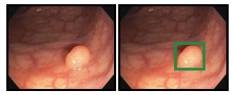
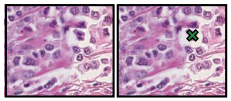
	(a) Cell detection and tracking during the autophagy process in time-lapse microscopy 	(b) MS Lesion detection in multi-modal brain MRI images 	(c) Polyp detection in colonoscopy videos with predefined sensitivity of 0.98 	(c) Mitosis detection in histopathology images 
Required granularity of localization?	Rough outline	Rough outline	Rough outline	Only position
Unequal severity of class confusions?				
Penalization of multiple predictions assigned to same reference object?				
Penalization of multiple reference objects assigned to same prediction?				
Cutoff on predicted class scores?	Target value-based	Target value-based	Target value-based	
Calibration assessment requested?		Comparison of (re-)calibration methods		
Small size of structures?				
High variability of structure sizes?				
Possibility of overlapping or touching target structures?				
Possibility of disconnected target structure(s)?				
Presence of class imbalance?				
High inter-rater variability?				
Granularity of provided reference annotations?	Rough outline	Rough outline	Rough outline	Only position
Non-independence of test cases?				
Possibility of reference without target structure(s)?				
Availability of predicted class scores?				
Possibility of algorithm output not containing target structure(s)?				
Possibility of invalid algorithm output?				
Possibility of overlapping predictions?				
Recommendations	**Localization criterion:** Box IoU **Assignment strategy:** Greedy (by Score) Matching, set double assignments to FP **Multi-threshold metric:** FROC Score **Per-class counting metric:** FPPI@Sensitivity	**Localization criterion:** Box IoU **Assignment strategy:** Greedy (by Score) Matching, set double assignments to FP **Multi-threshold metric:** FROC Score **Per-class counting metric:** FPPI@Sensitivity **Output calibration:** BS	**Localization criterion:** Box IoU **Assignment strategy:** Greedy (by Score) Matching, set double assignments to FP **Multi-threshold metric:** FROC Score **Per-class counting metric:** FPPI@Sensitivity	**Localization criterion:** Center Distance **Assignment strategy:** Greedy (by Center Distance) Matching, set double assignments to FP **No multi-threshold metric needed** (predicted class score not available) **Per-class counting metric:** F_β Score

Figure A.39: Instantiation of the metric recommendation framework for common biomedical object detection use cases. For every use case, we show exemplary input and output images, the instantiation of the problem fingerprint (with red cells referring to 'no' and green cells referring to 'yes'), and the resulting metric recommendations based on the fingerprint. Used abbreviations: Brier Score (BS),False Positive (FP), False Positives per Image (FPPI), Free-Response Receiver Operating Characteristic (FROC), Intersection over Union (IoU), Magnetic Resonance Imaging (MRI), Multiple Sclerosis (MS).

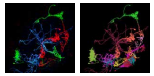
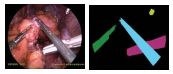
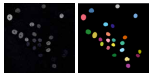
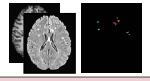
	(a) Instance segmentation of neurons from the fruit fly in 3D multi-color light microscopy images 	(b) Surgical instrument instance segmentation in colonoscopy videos 	(c) Cell nuclei instance segmentation in time-lapse light microscopy 	(d) MS Lesion segmentation in multi-modal brain MRI images 
Particular importance of structure boundaries?				
Particular importance of structure center?				
Unequal severity of class confusions?			Detection / Segmentation	
Handling of spatial outliers?	Existence-based	Existence-based		Existence-based
Compensation for annotation imprecisions requested?				
Penalization of multiple predictions assigned to same reference object?				
Penalization of multiple reference objects assigned to same prediction?				
Cutoff on predicted class scores?	Argmax-based	Argmax-based		Target value-based
Calibration assessment requested?				Comparison of (re-)calibration methods
High variability of structure sizes?				
Target structures feature tubular shape?				
Possibility of multiple labels per unit?				
Possibility of overlapping or touching target structures?				
Possibility of disconnected target structure(s)?				
Presence of class imbalance?				
High inter-rater variability?				
Possibility of spatial outliers in reference annotation?				
Possibility of reference without target structure(s)?				
Availability of predicted class scores?				
Possibility of algorithm output not containing target structure(s)?				
Possibility of invalid algorithm output?				
Possibility of overlapping predictions?				
Recommendations	**Localization criterion:** Mask IoU **Assignment strategy:** Greedy (by Score) Matching, set double assignments to FP **Multi-threshold metric:** AP **Per-class counting metric:** F_β Score **Overlap-based metric:** clDice **Boundary-based metric:** NSD	**Localization criterion:** Boundary IoU **Assignment strategy:** Greedy (by Score) Matching, set double assignments to FP **Multi-threshold metric:** AP **Per-class counting metric:** F_β Score **Overlap-based metric:** clDice **Boundary-based metric:** NSD	**Localization criterion:** IoR **Assignment strategy:** Matching via IoR > 0.5 Set double assignments to FP **No multi-threshold metric needed** (predicted class score not available) **Per-class counting metric:** F_β Score **Overlap-based metric:** IoU **No boundary-based metric needed** (no interest in structure boundaries)	**Localization criterion:** Boundary IoU **Assignment strategy:** Greedy (by Score) Matching, set double assignments to FP **Multi-threshold metric:** FROC Score **Output calibration:** BS **Per-class counting metric:** FPPI@Sensitivity **Overlap-based metric:** DSC **Boundary-based metric:** NSD

Figure A.40: Instantiation of the metric recommendation framework for common biomedical instance segmentation use cases. For every use case, we show exemplary input and output images, the instantiation of the problem fingerprint (with red cells referring to 'no' and green cells referring to 'yes'), and the resulting metric recommendations based on the fingerprint. Used abbreviations: Average Precision (AP), Brier Score (BS), Centerline Dice Similarity Coefficient (clDice), Dice Similarity Coefficient (DSC), False Positive (FP), False Positives per Image (FPPI), Free-Response Receiver Operating Characteristic (FROC), Intersection over Reference (IoR), Intersection over Union (IoU), Magnetic Resonance Imaging (MRI), Multiple Sclerosis (MS), Normalized Surface Distance (NSD).

A.5 Overview of the descriptions from the RobustMIS teams

In Section 4.4, we reimplemented the participants of the Robust Medical Instrument Segmentation (RobustMIS) 2019 challenge. In this section, we provide a detailed overview of the submitted methods along with assumptions that were taken for the reimplementation.

Team caresyntax The overall description of team *caresyntax'* method was very short and did not contain lots of details, therefore leaving some room for interpretation[‡]. The team described that they utilized a Mask R-CNN [He et al., 2017] architecture with a ResNet-50 [He et al., 2016] backbone. From their description, it could be assumed that the team used the reference Mask R-CNN implementation [FacebookResearch, 2019] for their model, written in `PyTorch` [Paszke et al., 2019]. We therefore filled in missing information with the defaults from this reference.

For data pre-processing, team *caresyntax* cropped the images to remove black parts of the images and used a minimum size of 800 and maximum size of 1,333 similar to the sizes described in the reference implementation, resulting in upscaling the images by a factor of roughly 1.5. Finally, they applied a normalization of the channel mean and standard deviations similar to the one used in the popular ImageNet competition [Deng et al., 2009]. The team stated that they pre-trained the network on the Microsoft-Common Objects in COntext (COCO) 2017 challenge training data set [Lin et al., 2014]. However, the actual weights in the reference implementations were only available as pre-trained on the ImageNet data set. As team *caresyntax* also used the ImageNet normalization and did not describe that they actually did the pre-training on Microsoft-COCO by themselves, we followed the reference implementation with a pre-training on ImageNet. The ResNet-50 backbone used in the reference implementation was labeled as "COCO-InstanceSegmentation", which might have confused. Yet, this naming referred to an example use-case, not to the actual weights of the ResNet-50.

For training, team *caresyntax* split the provided challenge training data set into 4,884 training (80%) and 103 validation cases (20%). The team described a large manual effort to ensure a similar distribution of cases, such as smoke, in the training and validation data sets. This step was not described in detail, we therefore chose the same amount of cases in both data sets and made sure to have a representable split, however, we could not fully reproduce the split procedure. The team described a data augmentation with random horizontal flips with a probability of 50%. Batch sizes of 2 (training data set) and 1 (validation data set) were specified as well as an initial learning rate of 5e-3, which was halved every five epochs and a minimum learning rate of 5e-4. Team *caresyntax* described 15 epochs for training and model selection based on the mean Dice Similarity Coefficient (DSC) score over instruments on the validation set. During our training process for the reimplementation, we found that the model was not converged after 15 epochs. We therefore extended the training to 30 epochs. The team used an Stochastic Gradient Descent (SGD) optimizer with a momentum of 0.9 and weight decay of 5e-4. In our follow-up survey, the team described that they used the smooth L1 loss, the Cross Entropy (CE) loss, and the Binary Cross Entropy (BCE) loss. Finally, although the team underlined the importance of tuning the confidence thresholds for their predictions, they did not specify their actual threshold values. We therefore defined confidence thresholds of 0.8. We picked the best model for team *caresyntax* at iteration 63.7k, which achieved an Average Precision (AP) score of 34.66%.

Team Casia SRL Although the description of team *Casia SRL* was longer than the one from team *caresyntax*, many details were missing or misleading. In fact, we were not able to identify the concrete network architecture. Throughout their method description, the team utilized three different names for the used architecture: Dense Pyramid Attention Network (DPANet), Residual attention U-Net (RAUNet) [Ni et al., 2019], and Bilinear Attention Network (BARNet) [Ni et al., 2020], from which RAUNet and BARNet refer to architectures developed by this team. We did not find any matching architecture for the DPANet in secondary literature,

[‡]It should be noted that this team was from industry and purposefully withheld information.

therefore assuming that this specific architecture was newly developed for the challenge participation. From the method description, we could clearly identify the basic architecture of a U-Net [Ronneberger et al., 2015] with additional components. The team mentioned the usage of attention modules [Vaswani et al., 2017], bilinear pooling [Lin et al., 2015], and an ImageNet-pre-trained ResNet-34 as an encoder for the U-Net. However, exact details for those components that were necessary for our reimplementation were missing in the description. In addition, the concrete upsampling scheme for the U-Net was not provided. We therefore needed to rely on secondary literature to implement this method. We could complement the missing information from the BARNet paper and a paper on bilinear pooling cited by the team in their method description [Lin et al., 2015]. As specified by the team, we implemented the network in `PyTorch`.

For data pre-processing, team *Casia SRL* described a manual split of the provided challenge training data set of 3,643 training and 532 validation cases. To increase the diversity of the data, the team reported the application of data augmentation (random rotation, shifting, and flipping) to get an additional 1,680 cases. Data augmentation was applied based on the number of instruments visible in the image frames to account for underrepresented cases with many instrument instances. While the number of augmentations per instrument count was provided by the team, details were missing regarding the concrete augmentations. We thus chose to sequentially apply horizontal and vertical flips with a probability of 50%, horizontal and vertical shifts in the range of [-20%, 20%], and rotation in the range of [-25°, 25°]. It should be noted that the reported instrument count was not correct. For instance, the team mentioned 4,175 cases with at least one instrument instance, but the real number was 4,988. However, as we agreed to stick with the method description, we truncated the data set to match 4,175 cases with at least one visible instrument. The team did not describe any resizing or cropping operation, although their network architecture implied that the dimension of cases should be evenly divisible at least five times. We therefore cropped and padded the images to a size of 544×928.

The team did neither report the training duration nor the model selection criterion. We therefore used 120 epochs for training based on the convergence of the reimplemented model. The team described batch sizes of 16 and the usage of the Adaptive Moment Estimation (Adam) optimizer. For the latter, a momentum of 0.9 and weight decay of 1e-4 were given. The initial learning rate of 2e-5 was described to be multiplied by 0.8 every 30 iterations. A hybrid loss consisting of the CE and the Jaccard loss was applied, with the parameter α set to 0.5:

$$\mathcal{L} = CE - \alpha \log(Jaccard) \tag{A.1}$$

Finally, given our chosen resizing of the images, we applied a sliding window approach to match the defined output size of 540×960. This was not described by the team, although this step was necessary to run the described network architecture and match the required output image size. The network architecture provided only binary segmentation masks. The team did not mention how the instance segmentation was approached. We therefore chose to apply connected component analyses [Chen et al., 2006] to transfer the binary masks into instance segmentations. We picked the best model for team *Casia SRL* at epoch 14, which achieved the lowest validation loss of 0.13.

Team fisensee Team *fisensee* carefully described their network architecture in sufficient detail with only minor ambiguities. They used a U-Net, whose outputs were post-processed via a connected component analysis to achieve instance segmentation outputs. The U-Net-typical downsampling was described such that spatial dimensions were exactly halved and convolutional operations were padded to match the input spatial dimensions. The padding procedure was not described in sufficient detail for reproduction. We therefore used the default zero padding from the `PyTorch` library, which was used by the team for the general model implementation. This step also impacted the skip connections of the U-Net, for which the feature map retrieved from the encoder and the upsampled output are concatenated. Given that the dimensions of the decoder blocks were not provided by the team, the first convolution of each decoder block was assumed, similar to the original publication, to keep the number of output

channels as the one of input channels. In addition, for upsampling, we assumed that the output of the transposed convolution has the same number of channels as the feature map from the corresponding resolution level of the U-Net. The team described the usage of the sum of the DSC and CE as the loss function.

For data pre-processing, the team processed all images at half resolution to ensure faster training. They reported scaling, rotation, mirroring, additive Gaussian noise, brightness, contrast, gamma, and elastic deformations as data augmentations. We used the described package in Isensee et al. [2020] for the reimplementation of the data augmentation step. The training procedure was described in sufficient detail. Team *fisensee* reported the use of the SGD optimizer with a momentum of 0.9 and a continuous *PolyLr* learning rate scheduler [Liu et al., 2015]. However, the hyperparameter controlling the polynomial degree of the learning rate decay was not provided. We set this parameter to 0.9, similar to how it was introduced in the original publication. The team described an initial learning rate of 1, for which we could not reach convergence in the reimplementation. As the *PolyLr* publication described learning rates in the range of 1e-4 and 1e-9, we assumed a typo and decreased the learning rate to a standard value of 1e-2.

The team reported an 8-fold cross-validation of an ensemble of networks. The data split into eight validation folds was described such that one surgery per surgery type was reserved for validation, while the others were used for training. In this challenge, three surgery types were used: proctocolectomy, rectal resection, and sigma resection (see Section 5.3 for details). While reproducing this procedure, we noticed a large imbalance in the training and validation proportions. This is because some surgeries had far fewer examples than to others, as most images were reserved for the challenge test data set. As an imbalanced proportion across splits contradicted the consensus on how to conduct k-fold cross-validation [Kohavi et al., 1995], we made sure the sizes of each split were approximately equal by either removing some samples or loading images several times. No details were provided concerning the connected component analysis for the instance segmentation because U-Nets only provide binary masks as outputs. We assumed that a 1-connected component analysis was employed, which suppressed predictions in the black border regions, in which no medical instruments were to be expected. In addition, we assumed a minimum size of 2,048 connected pixels for the final image, as the ensembles sometimes resulted in many tiny false positive instances.

We trained all ensembles for a maximum of 2,000 epochs, as specified by the team. The exact model selection criteria were not described by the team, we thus selected the final models based on the loss on the respective validation fold. The final models were selected between epochs 139 and 1,456 and on average in epoch 552.

Team Squash Team *Squash* carefully described their network architecture in sufficient detail. They employed a Mask R-CNN in combination with a classification network, implemented as ResNet-50. The purpose of the latter was to train the network to identify whether an image frame contained instrument instances or not. More specifically, the team used the provided video data to estimate the probability that the last frame contained an instrument. Based on their description, we derived that the team used the `PyTorch` ResNet implementation for the classification network [Torch, 2017]. The classification network was first trained alone without training the other parts of the Mask R-CNN network and the convolutional layers were then used for the Mask R-CNN backbone. Only if a case contained at least one instrument, this frame was considered for producing a segmentation output. From their detailed description, we could clearly identify the architecture as the reference `PyTorch` implementation [FacebookResearch, 2019], as most of the hyperparameters exactly matched. The team only reported two major changes: First, they used a focal loss instead of the BCE loss for the classification network, and second, they considered the 2,000 best candidates produced by the Region Proposal Network (RPN) instead of the default value of 1,000 candidates. Although the team provided many details in their description, inconsistencies were present throughout the document. It was, for example, not entirely clear, how the training set was composed for the classification network.

For the split of the training data into training and validation, the team sorted the data by the number of instrument instances. They then randomly sampled from the data pool to achieve a 70%/30% split for all numbers of instruments. For example, from the total amount of images with one instrument instance, 70% were used for training, and 30% were used for validation. Data augmentation was applied to 35% in the classification network and 25% in the segmentation part of the network only for the cases that contained instrument instances for every epoch. The team described the usage of Gaussian blur with $\sigma \in [0.0, 3.0]$, gamma contrast enhancement [Poynton, 2012] in the range of [0.5, 2.0], and sharpening in the range of [0.0, 1.0]. Another 35% of those images with data augmentation were mirrored along both axes. In addition, the team described that the minority class was upsampled by horizontal translation with 5% of image width in combination with an upscaling with a factor of 1.5.

Team *Squash* reported the usage of an SGD optimizer with learning rates of 0.2 for the classification and 0.02 for the segmentation networks. As our reimplemented models did not converge when utilizing those learning rates, we reduced them to 0.02 and 0.002 respectively. The team described a decay in the learning rate of 50% after 40 epochs, but it was not entirely clear whether they were referring to a steady decrease in every epoch or a hard decrease at epoch 40. Based on their phrasing, we decided to implement a step-wise decrease at every epoch. While the team mentioned the usage of the DSC and AP validation metrics for picking the final model, it did not become clear if and how the metrics were weighted. The team did not specify the total number of epochs trained. We therefore chose to train for a maximum of 120 epochs. We picked the best model at epoch 68, yielding an AP of 46.54%.

Team Uniandes Team *Uniandes* clearly specified that they used the official reference implementation of the Mask R-CNN [FacebookResearch, 2019] and the method description did not leave much room for interpretation. Similar to the reference implementation, they used a ResNet-101 with a Feature Pyramid Network (FPN) as the backbone.

For data preprocessing, the team split the provided challenge training data set into training and validation sets. This was achieved by assigning all cases from one patient to one of the two splits. In particular, patients 3 and 5 from the proctocolectomy and patients 3 and 6 from the rectal resection surgery were assigned to the validation set, other patients were used for training. Furthermore, the team reported that they tried to incorporate the temporal data from the provided videos for data augmentation. However, along the same lines, the team described that this approach was not sufficient and was not used in the final submission. We thus also skipped this step in our reimplementation.

Team *Uniandes* trained for 30 epochs, by applying the SGD optimizer with a learning rate of 0.075 and a weight decay of 1e-4. In particular, they reported a 500-step linear warm-up with a factor of 0.33. Specifically, the reported additional steps at epochs 15 and 22.5. The standard Mask R-CNN losses were used by the team [He et al., 2017] (see Section 2.3). The team stated that the final model was chosen based on a combination of the metrics from both, the RobustMIS and the Microsoft-COCO, challenges. Thus, we decided to train the reimplementation until the algorithm clearly converged. We picked the best model at iteration 72,120, which achieved an AP score of 61.76%.

Team VIE Team *VIE* was the only participating team who extensively used the provided video data in their method. It was used for an optimal flow calculation in combination with a Mask R-CNN. Although this approach would have benefited from a detailed description, the team only provided a very brief overview with some incorrect assumptions, requiring many interpretations to be taken for our reimplementation.

While it becomes clear that the team implemented the optical flow for additional input of the Mask R-CNN model, they only stated that the python library OpenCV [Bradski, 2000] was used for implementation. Based on the output format of a pixel-wise vector field, we considered the dense optical flow from the OpenCV

library [Farnebäck, 2003]. In their overview figure, team *VIE* showed five frames used to calculate the optical flow, although optical flow is typically computed over two consecutive frames. Thus, we calculated the optical flow for five successive image frames, resulting in four vector-fields. As this resulted in a comparatively large input, we assumed that summary statistics were calculated over the four vector-fields, resulting in a single averaged vector field. We closely followed an example implementation of the OpenCV library [OpenCV, 2000] for this step. The summarized optical flow was then fed into the Mask R-CNN network. Yet, the data set contained several image frames without predecessor frames, for example, frames at the very beginning of a procedure. The team did not specify how to handle those cases and we defined the optical flow component to be zero for such occurrences.

Team *VIE* specified the model as a Mask R-CNN with a ResNet-101 backbone, but did not provide details on the network weights. Using randomized weights in our reimplementation yielded no convergence in the training. Consequently, we followed the Mask R-CNN reference implementation [FacebookResearch, 2019] and used a pre-trained network on ImageNet. The team did not describe any further details, so we used the default parameters of the reference implementation.

For data pre-processing, team *VIE* reported that they resized all image frames to the same size of 1024×1024. We assumed that the team did not notice that the frames were already of the same size of 540×940. They did not further specify how the resizing took place, so we chose to resize the longest side to match the given pixel size of 1024 and padded the remaining pixels with zeros. The team further described that the bounding boxes provided along the challenge data set were discarded and newly generated based on the segmentation masks. However, no bounding boxes were provided to the participants at all. Based on the description, we assumed that data augmentation techniques were applied but they were not further specified. We therefore used the default augmentation of the Mask R-CNN reference.

Similarly, the team did not specify many details of the training process besides the original Mask R-CNN loss, the learning rate (1e-4), and momentum (0.9). For the reimplementation, we were required to choose all other parameters by ourselves. We closely followed the Mask R-CNN reference implementation. In addition, we defined that 20% of the provided challenge data was used for validation purposes and trained the model for a maximum of 100 epochs, following standard techniques for such networks. Given the complexity of the optical flow, the training stopped after two days with an AP score of 14.21% at iteration 27,860.

Team www The overall description of team *www*'s method was very vague and did not cover many aspects of the network architecture. The team described the usage of a Mask R-CNN combined with a Dense Atrous Convolution (DAC) block [Gu et al., 2019]. Throughout their description, the team often cited the general Mask R-CNN architecture [He et al., 2017]. We therefore derived missing information from this reference. The team described the usage of a ResNet-101, pre-trained on ImageNet, as the backbone. However, it remained unclear how the team actually integrated the DAC block. They did not specify whether the residual backbone was integrated as an FPN or in the C4 configuration, as detailed in He et al. [2017]. For our reimplementation and based on He et al. [2017], we assumed that the backbone was implemented using an Feature Pyramid Network (FPN) configuration, as this was found to perform better than the C4 configuration. DAC blocks were originally proposed for U-Nets, not in the Mask R-CNN context, making the adaption without further details difficult to interpret. In consideration of Gu et al. [2019]'s work, we decided to place the DAC block at the end of the final convolutional ResNet-101 block before it was fed into the corresponding feature pyramid convolution. However, other interpretations would have been possible, such as the DAC block being at the end of the final convolutional layer with the ResNet-101 in C4 configuration. The team specified that the method was written in `Keras` [Chollet et al., 2015] and we assumed that they used the most popular Mask R-CNN `Keras` implementation [Matterport, 2017] and derived missing information from this repository, such as the anchor scales and ratios.

For training, team *www* used a batch size of 2 and randomly augmented images by rotations of $[0^\circ,\ 10^\circ]$ in

addition to horizontal and vertical flips. The Adam optimizer was applied with an initial learning rate of 1e-4, which was divided by 10 after every 10,000 iterations. A weight decay of 5e-3 was used. The team described a mixture of the smooth L1 loss, the focal loss, and the BCE loss. The team did not reveal information on the number of training epochs but mentioned that the training ran for one day. We therefore mimicked this number by training for a maximum of 70 epochs, which was roughly equivalent to one day of training in our experimental setup. We picked the best model for team *www* at epoch 44 with a loss of 0.68.

A.6 RobustMIS challenge design document

This section contains the complete design document for the Robust Medical Instrument Segmentation (RobustMIS) challenge[§]. Updates after acceptance of the challenge are highlighted in green. To protect personal information, we will not show the challenge organization section of the challenge.

Summary

Title Robust Medical Instrument Segmentation Challenge 2019

Acronym RobustMIS 2019

Abstract Intraoperative tracking of laparoscopic instruments is often a prerequisite for computer and robot assisted interventions. Although previous challenges have targeted the task of detecting, segmenting and tracking medical instruments based on endoscopic video images, key issues remain to be addressed:

1. The methods proposed still tend to fail when applied to challenging images (e.g. in the presence of blood, smoke or motion artifacts) and
2. algorithms trained for a specific intervention in a specific hospital typically do not generalize.

The goal of this challenge is, therefore, the benchmarking of algorithms for medical instrument segmentation with a specific emphasis on the robustness and generalization capabilities of the methods. The challenge is based on the biggest annotated data set made (to be made) available in the field, comprising more than 10,000 annotated images that have been extracted from a total of 30 surgical procedures from three different surgery types.

Keywords Instrument segmentation, multiple instance segmentation, instrument detection, minimally invasive surgery, robustness, generalization

Mission of the challenge

Field of application Intervention assistance.

Task categories 1) Binary segmentation, 2) Multiple instance segmentation, 3) Multiple instance detection

Target cohort Patients undergoing minimally invasive surgery.

Challenge cohort Patients undergoing rectal resection, proctocolectomy or UNKNOWN SURGERY (will be made public after the docker submission deadline).

Imaging modality Laparoscopic video.

Context information corresponding to the image data A whole surgical video and information on the type of surgery performed is provided for each training case. No additional information is given for the test cases.

[§]Please refer to `https://www.synapse.org/Portal/filehandle?ownerId=syn18779624&ownerType=ENTITY&fileName=RobustMIS2019_Design.pdf&preview=false&wikiId=591266` for reference.

Context information corresponding to the patient in general None.

Data origin An abdomen shown in laparoscopic video data.

Algorithm target An elongated rigid object put into the patient and then manipulated directly from outside the patient. Examples: grasper, scalpel, trocar. Counterexamples: non-rigid tubes, bandage, needle (not directly manipulated from outside but manipulated with an instrument). Please refer to `https://www.synapse.org/Portal/filehandle?ownerId=syn18779624&ownerType=ENTITY&fileName=LabelingInstructions.pdf&preview=false&wikiId=592660` for further details.

Assessment aims 1) Identify robust methods for instrument detection, binary instrument segmentation and multiple instance segmentation. 2) Assess generalization capabilities of the methods proposed. 3) Identify which image properties (e.g. smoke, bleeding, motion artifacts) makes images particularly challenging to process.

Challenge data sets

Acquisition devices Laparoscopic camera Karl Storz Image 1 with a 30° optic (Karl Storz SE & Co KG). As a light source Karl Storz Xenon 300 was used.

Acquisition protocol Data acquisition took place during daily routine procedures with the integrated operating room (Karl Storz OR1 FUSION ®). Video data was then anonymized by excluding parts of the video displaying parts outside the abdomen. Image resolution was downscaled from 1920 $\times$ 1080 (HD) in the primary video to 960 $\times$ 540.

Center Institute: Heidelberg University Hospital, Department of Surgery

Characteristics of the subjects No characteristics available due to fully anonymized data.

Definition of cases A training cases encompasses a 10 second video snipped in form of 250 endoscopic image frames and a reference annotation for the last frame. In the annotated frame, a "0" indicates the absence of a medical instrument and

- (Binary segmentation) the number "1" represents a medical instrument.
- (Multiple instance segmentation and detection) numbers "1", "2", ... represent different instances of medical instruments.

The test cases are identical in format but do not include a reference annotation. For training images, the entire corresponding video is provided as context information along with information on the surgery type.

Total number of cases Videos from 30 surgical procedures corresponding to three different surgery types (10 rectal resection, 10 proctocolectomy, 10 UNKNOWN SURGERY) served as a basis for this challenge. From these 30 procedures, a total of 10,040 training and test cases were extracted according to the following procedure:

Algorithm 4 Frame extraction for the Robust Medical Instrument Segmentation (RobustMIS) challenge

```
for surgery type s in {rectal resection, proctocolectomy, UNKNOWN SURGERY} do
    for procedure 1, ..., 10 in surgery type s do
        Extract frame
        if the frame is blue (modified for anonymization reasons) then
            Ignore it
        else
            Add its ID to the IDs of the challenge data set
        end if
    end for
end for
```

This resulted in a total of 4,456 frames (corresponding to the extracted IDs) to be annotated. To obtain at least 10,000 annotated frames in total, additional frames in interesting parts (phase transitions) were obtained using the following protocol:

Algorithm 5 Frame extraction (phase transition) for the Robust Medical Instrument Segmentation (RobustMIS) challenge

```
for surgery type s in {rectal resection, proctocolectomy, UNKNOWN SURGERY} do
    for procedure 9 and 10 in surgery type s (8, ..., 10 if s== rectal resection) do
        for each phase transition in the video do
            Extract frame
            if the frame is blue (modified for anonymization reasons) then
                Ignore it
            else
                Add its ID to the IDs of the challenge data set
            end if
        end for
    end for
end for
```

This procedure led to 10,040 frames in total. The extracted frames were annotated (see parameter 23) and complemented by the 249 preceding frames to form the training and test cases for the challenge. The performance assessment for the challenge will be performed in three stages.

- Stage 1: The test data is taken from the procedures (patients) from which the training data were extracted.
- Stage 2: The test data is taken from the exact same type of surgery as the training data but from procedures (patients) not included in the training data.
- Stage 3: The test data is taken from a different but similar type of surgery (and different patients) compared to the training data.

To achieve this, the data of all 10 procedures from on surgery type (UNKNOWN SURGERY) was reserved for testing in Stage 3. From the remaining 20 procedures, 80% were reserved for training and 20% (i.e. two procedures from each type) for testing in Stage 2. More specifically, the two patients with the lowest number of annotated frames were taken as test data for Stage 2 (for both, rectal resection and proctocolectomy). For Stage 1, every 10th annotated case from the remaining 2·(10 - 2) = 16 procedures was used.

UPDATE: While all training and test cases were used for the multiple instance detection task, cases not showing an instrument in the image were removed from training and test sets for the binary and multiple instance segmentation tasks. No validation cases are provided by the organizers; hence, it is up to the challenge participants to split the training and validation data. This all led to a total of:

- *Training cases:* 5,983 cases in total (2,943 cases for the proctocolectomy surgery and 3,040 cases for the rectal resection surgery)
- *Test data:*
  - Stage 1: 663 cases in total (325 cases for the proctocolectomy surgery and 338 cases for the rectal resection surgery)
  - Stage 2: 514 cases (225 cases for the proctocolectomy surgery and 289 cases for the rectal resection surgery)
  - Stage 3: 2,880 cases for the UNKNOWN SURGERY

Explanation of total number All data from a specific surgery type was reserved for testing (Stage 3) to test generalization capabilities of the methods. The ratio 80%/20% (Stage 2) is commonly used in challenges.

Further important characteristics of the cases No further important characteristics.

Annotation procedure Each case was annotated according to the following procedure:

1. Initialization: The company UnderstandAI extracted video frames as described in parameter 22b and did an initial segmentation for the extracted frames.
2. Refinement:
   (a) The challenge organizers analyzed the annotations provided, identified inconsistencies and agreed on an annotation protocol.
   (b) A team of 15 engineers analyzed all annotations again and refined them according to the annotation protocol if necessary. In ambiguous/unclear cases, a team of two engineers and one medical student generated a consensus annotation.
3. Quality control:
   (a) A medical expert went through the refined segmentations and reported potential errors.
   (b) An engineer did the refinement according to the instructions of the medical expert.
4. Final check: After refinement, a medical expert checked the refined annotations together with the engineer for possible final correction.

Annotation protocol See separate document: `https://www.synapse.org/Portal/filehandle?ownerId=syn18779624&ownerType=ENTITY&fileName=LabelingInstructions.pdf&preview=false&wikiId=592660`

Data pre-processing methods For each frame that had to be segmented, 249 previous frames were also extracted. All frames were saved as .ong using the `python-opencv2` library. The 249 video frames were then compressed using .zip and must be extracted before use. The annotation masks were stored as binary images in .png format. This process was identical for both training and test data.

Sources of error As each case was analyzed by multiple annotators, errors in annotation can mainly be attributed to image quality (i.e. the inherent ambiguity of the problem when relying solely on the image data).

Assessment methods

Metrics The following metrics were calculated:

- Binary segmentation: Dice Similarity Coefficient (DSC) and Normalized Surface Distance (NSD)
- Multiple instance segmentation: Multi-Instance Dice Similarity Coefficient (MI DSC) and Multi-Instance Normalized Surface Distance (MI NSD)
- Multiple instance detection (**UPDATE**): The F_1 Score will be computed according to the following procedure:

Algorithm 6 Computation of the F_1 Score for the multiple instance detection task of the Robust Medical Instrument Segmentation (RobustMIS) challenge.

1: Compute matching matrix: For each instrument instance i in the reference output (represented by a specific ID > 0) and each instance j in the participant's output, m_{ij} is set to the Intersection over Union (IoU).
2: **if** m_{ij} < 0.3 **then**
3: m_{ij} = 0 ("no match")
4: **end if**
5: Apply the Hungarian algorithm to assign participant's instances to reference instances. Instruments with matches are considered True Positive (TP). Reference instances without a match are considered False Negative (FN). Participant's instances without a reference match are considered False Positive (FP).
6: Compute the F_1 Score.

Justification of metrics We based our design choice regarding the metrics (MI) DSC and (MI) NSD on the Medical Segmentation Decathlon (MSD) for the binary and multiple instance segmentation tasks. **UPDATE:** We could not use standard object detection metrics such as the Average Precision (AP) for the multiple instance detection task, as we did not require participants to submit predicted class scores. Thus, we chose another common per-class counting metric, namely the F_1 Score.

Ranking scheme **UPDATE:** An accuracy ranking and a robustness ranking will be computed for the (MI) DSC and (MI) NSD metrics for the binary and multiple instance segmentation tasks. Rankings will be computed for each metric m separately as follows:

Algorithm 7 Accuracy and robustness ranking computation for the Robust Medical Instrument Segmentation (RobustMIS) challenge.

1: Let T=$\{t_1, \ldots, t_N\}$ be the test cases for a given task.
2: **for** all participating teams a_i for each test case t_j **do**
3: **if** $m(a_i, t_j)$ == `Not a Number (NaN)` **then**
4: $m(a_i, t_j)$ =0
5: **end if**
6: Aggregate metric values $m(a_i, t_j)$ with the following two aggregation methods
1. Accuracy: Compute the test-based ranking described in Section 2.1.2. This yields the accuracy rank $r_a(a_i)$.
2. Robustness: Compute the 5% percentile of all $m(a_i, t_j)$ to obtain a robustness rank $r_r(a_i)$ for algorithm a_i
7: **end for**

For the multiple instance detection task, the computation of the F_1 Score is done over the complete data set. This results in one F_1 Score per participant which will determine the ranking. Please be aware that the number of test cases as well as the number of algorithms generally differ for each task and stage. This procedure will lead to four separate rankings for both, the binary and multiple instance segmentation tasks (accuracy and robustness rankings for the (MI) DSC and (MI) NSD) and one ranking for the multiple instance detection task.

Submissions with missing results Missing cases are set to the worst possible value, namely 0 for all metrics.

Justification of ranking method **UPDATE:** To address multiple aspects of the challenge purpose, separate rankings for accuracy and robustness will be computed for Stage 3 of the challenge. We decided to use a ranking scheme that tends to group algorithms in case of minor performance differences. The computation of the F_1 Score already involves an indirect ranking scheme because only one value per participant is generated.

Statistical analyses Stability will be investigated via bootstrapping and hypothesis testing, as it was identified as an appropriate approach to investigate ranking variability [Maier-Hein et al., 2018].

Further analyses Performance gain based on ensembling the algorithms will be investigated. Common problems of the submitted methods will be identified.

A.7 MSD challenge design document

This section contains the complete design document for the Medical Segmentation Decathlon (MSD) challenge. To protect personal information, we will not show the challenge organization section of the challenge.

Summary

Title Medical Segmentation Decathlon

Acronym MSD

Abstract International challenges have become the de facto standard for comparative assessment of image analysis algorithms. Although segmentation is the most widely investigated medical image processing task, the various challenges have been organized to focus only on specific clinical tasks. We organize the Medical Segmentation Decathlon (MSD) – a biomedical image analysis challenge, in which algorithms compete in a multitude of both tasks and modalities to investigate the hypothesis that a method capable of performing well on multiple tasks will generalize well to a previously unseen task and potentially outperform a custom-designed solution.

The challenge is composed of two tasks: the development and the mystery phase. The development phase serves for model development and includes seven open training data sets: Brain, heart, hippocampus, liver, lung, pancreas and prostate. The mystery phase aims to investigate whether algorithms are able to generalize to three unseen segmentation tasks: Colon, hepatic vessel and spleen.

Keywords Image segmentation; Medical image; Deep learning; Grand challenge

Mission of the challenge

Field of application Treatment planning, Assistance, Intervention planning, CAD, Diagnosis, Surgery, Prognosis, Research.

Task categories Segmentation.

Target cohort *Development phase:*

- Brain: Brain tumor patients with diagnostic Magnetic Resonance Imaging (MRI) scans, including T1-weighted 3D acquisitions, T1-weighted contrast-enhanced (gadolinium contrast) 3D acquisitions and T2-weighted Fluid-attenuated Inversion Recovery (FLAIR) 3D acquisitions
- Heart: Patients undergoing heart scans
- Hippocampus: Adults with a non-affective psychotic disorder
- Liver: Liver cancer patients with visible tumors
- Lung: Patients with non-small cell lung cancer
- Pancreas: Pancreas cancer patients with visible tumors and cystic lesions
- Prostate: Prostate cancer patients

Mystery Phase:

- Colon: Patients undergoing resection of primary colon cancer.
- Hepatic Vessel: Patients with a variety of primary and metastatic liver tumors.
- Spleen: Patients undergoing chemotherapy treatment for liver metastases.

Challenge cohort *Development phase:*

- Brain: Patients diagnosed with either glioblastoma or lower-grade glioma from Multiparametric Magnetic Resonance Imaging (mp-MRI)
- Heart: Patients from MRI scans of the entire heart acquired during a single cardiac phase (free breathing with respiratory and Electrocardiogram (ECG) gating)
- Hippocampus: Healthy adults and adults with a non-affective psychotic disorder undergoing a T1-weighted Magnetization Prepared Rapid Acquisition Gradient Echo (MPRAGE)
- Liver: Patients with primary cancers and metastatic liver disease, as a consequence of colorectal, breast, and lung primary cancers
- Lung: Patients with non-small cell lung cancer
- Pancreas: Patients undergoing resection of pancreatic masses
- Prostate: Prostate cancer patients undergoing mp-MRI

Mystery Phase:

- Colon: Patients undergoing resection of primary colon cancer.
- Hepatic Vessel: Patients with a variety of primary and metastatic liver tumors.
- Spleen: Patients undergoing chemotherapy treatment for liver metastases.

Imaging modality *Development phase:* Brain, Prostate: mp-MRI, Heart, Hippocampus: MRI; Liver, Lung, Pancreas: Computed Tomography (CT). *Mystery Phase:* Colon, Hepatic Vessel, Spleen: CT.

Context information corresponding to the image data None.

Context information corresponding to the patient in general None.

Data origin *Development phase:* Brain, Heart, Hippocampus, Liver, Lung, Pancreas, Prostate. *Mystery Phase:* Colon, Hepatic Vessel, Spleen.

Algorithm target *Development phase:* Brain tumors, left atrium, anterior and posterior of hippocampus, liver and liver tumor, lung tumor, pancreas and pancreatic tumor mass, prostate Peripheral Zone (PZ) and Transition Zone (TZ). *Mystery Phase:* Colon cancer primaries, hepatic vessel and hepatic tumor, spleen.

Assessment aims Generalizability.

Challenge data sets

A detailed description of the challenge data sets is available in Simpson et al. [2019].

Acquisition devices *Development phase:*

- Brain: Various devices have been used to acquire the data. The scanners varied from 1T to 3T
- Heart: 1.5T Achieva scanner (Philips Healthcare, Best, The Netherlands)
- Hippocampus: All images were collected on a Philips Achieva scanner (Philips Healthcare, Inc., Best, The Netherlands)
- Liver: Various devices have been used to acquire the data
- Lung: Various devices have been used to acquire the data
- Pancreas: Various devices have been used to acquire the data
- Prostate: Various devices have been used to acquire the data

Mystery Phase:

- Colon: Various devices have been used to acquire the data
- Hepatic Vessel: Various devices have been used to acquire the data
- Spleen: Various devices have been used to acquire the data

Acquisition protocol *Development phase:*

- Brain: The MRI scans were acquired during routine clinical practice, using different equipment and acquisition protocols, among 19 different institutions and pooled to create a publicly available benchmark dataset for the task of segmenting brain tumour sub-regions (i.e., edema, enhancing, and non-enhancing tumour). All scans were co-registered to a reference atlas space using the SRI24 brain structure template [Rohlfing et al., 2010], resampled to isotropic voxel resolution of $1mm^3$, and skull-stripped using various methods followed by manual refinements.
- Heart: Voxel resolution: $1.25 \times 1.25 \times 2.7mm^3$
- Hippocampus: Structural images were acquired with a 3D T1-weighted MPRAGE sequence (Inversion Time (TI)/Repetition Time (TR)/Echo Time (TE), 860/8.0/3.7 s; 170 sagittal slices; voxel size, $1.0mm^3$). Manual tracing of the head, body, and tail of the hippocampus on images was completed following a previously published protocol. For the purposes of this dataset, the term hippocampus includes the hippocampus proper (CA1-4 and dentate gyrus) and parts of the subiculum, which together are more often termed the hippocampal formation. The last slice of the head of the hippocampus was defined as the coronal slice containing the uncal apex.
- Liver: Some images contained metal artifacts, consistent with real-world clinical scenarios for abdominal CT. The images were provided with an in-plane resolution of 0.5 to 1.0 mm, and slice thickness of 0.45 to 6.0mm.
- Lung: Pre-operative thin-section CT scans were obtained with the following acquisition and reconstruction parameters: section thickness, <1.5mm; 120kVp; automatic tube current modulation range, 100–700mA; tube rotation speed, 0.5s; helical pitch, 0.9-1.0; and a sharp reconstruction kernel.
- Pancreas: Pitch/table speed 0.984–1.375/39.37–27.50mm; automatic tube current modulation range, 220–380mA; noise index, 12.5–14; 120kVp; tube rotation speed, 0.7–0.8ms; scan delay, 80–85s; and axial slices reconstructed at 2.5mm intervals.
- Prostate: Manual segmentation of the whole prostate from transverse T2-weighted scans with resolution $0.6 \times 0.6 \times 4mm$ and the apparent diffusion coefficient map ($2 \times 2 \times 4mm$) was used.

Mystery Phase:

- Colon: 100–140kVp; exposure time, 500–1,782ms; and tube current, 100–752mA. Reconstruction parameters were: slice thickness, 1 to 7.5mm; and reconstruction diameter, 274–500mm.
- Hepatic Vessel: 120kVp; exposure time, 500–1,100ms; and tube current, 33–440mA. Images were reconstructed at a section thickness varying from 2.5 to 5 mm with a standard convolutional kernel and with a reconstruction diameter range of 360–500mm. Iodinated contrast material (150mL, Omnipaque 300, GE Healthcare, Chicago, IL, USA) was administered intravenously for each CT at a rate between 1 and 4cc/s.
- Spleen: 120kVp; exposure time, 500–1,100ms; and tube current, 33–440mA. Images were reconstructed at a section thickness varying from 2.5 to 5 mm with a standard convolutional kernel and with a reconstruction diameter range of 360–500mm. Iodinated contrast material (150mL, Omnipaque 300, GE Healthcare, Chicago, IL, USA) was administered intravenously for each CT at a rate between 1 and 4cc/s.

Center *Development phase:*

- Brain: 1) Center for Biomedical Image Computing and Analytics (CBICA), University of Pennsylvania, PA, USA, 2) University of Alabama at Birmingham, AL, USA, 3) Heidelberg University, Germany, 4) University Hospital of Bern, Switzerland, 5) University of Debrecen, Hungary, 6) Henry Ford Hospital, MI, USA, 7) University of California, CA, USA, 8) MD Anderson Cancer Center, TX, USA, 9) Emory University, GA, USA, 10) Mayo Clinic, MN, USA, 11) Thomas Jefferson University, PA, USA, 12) Duke University School of Medicine, NC, USA, 13) Saint Joseph Hospital and Medical Center, AZ, USA, 14) Case Western Reserve University, OH, USA, 15) University of North Carolina, NC, USA, 16) Fondazione IRCCS Istituto Neurologico Carlo Besta, Italy, 17) Washington University School of Medicine in St. Louis, MO, USA, and 18) Tata Memorial Centre, Mumbai, India. Data from institutions 6-16 describe data from `http://www.cancerimagingarchive.net/`.
- Heart: King's College London.
- Hippocampus: Vanderbilt University Medical Center.
- Liver: IRCAD Hôpitaux Universitaires.
- Lung: Several centers (The Cancer Imaging Archive).
- Pancreas: Memorial Sloan Kettering Cancer Center.
- Prostate: Radboud University, Nijmegen Medical Centre.

Mystery Phase:

- Colon: Memorial Sloan Kettering Cancer Center.
- Hepatic Vessel: Memorial Sloan Kettering Cancer Center.
- Spleen: Memorial Sloan Kettering Cancer Center.

22a) Definition of cases *Development phase:*

- Brain and Prostate: Training and test cases both represent a 4D mp-MRI image.
- Heart and Hippocampus: Training and test cases both represent a 3D MRI image.
- Hippocampus: Vanderbilt University Medical Center.
- Liver, Lung and Pancreas: Training and test cases both represent a 3D CT image.

Mystery Phase: Colon, Hepatic Vessel and Spleen: Training and test cases both represent a 3D CT image.

Total number of cases *Development phase:*

- Brain: 750 4D volumes (484 Training + 266 Testing)
- Heart: 30 3D volumes (20 Training + 10 Testing)
- Hippocampus: 394 3D volumes (263 Training + 131 Testing)
- Liver: 201 3D volumes (131 Training + 70 Testing)
- Lung: 96 3D volumes (64 Training + 32 Testing)
- Pancreas: 420 3D volumes (282 Training +139 Testing)
- Prostate: 48 4D volumes (32 Training + 16 Testing)

Mystery Phase:

- Colon: 190 3D volumes (126 Training + 64 Testing)
- Hepatic Vessel: 443 3D volumes (303 Training + 140 Testing)
- Spleen: 61 3D volumes (41 Training + 20 Testing)

Explanation of total number 70% training data and 30% test data.

Further important characteristics of the cases Full annotation (pixel level).

Annotation process *Development phase:*

- Brain: Reference annotations for all tumor sub-regions in all scans were approved by expert board-certified neuroradiologists.
- Heart: The left atrium appendage, mitral plane, and portal vein end points were segmented by an expert using an automated tool followed by manual correction.
- Hippocampus: Annotations were performed by expert raters.
- Liver: Annotations of the liver and tumors were performed by radiologists.
- Lung: The tumor region was denoted by an expert thoracic radiologist on a representative CT cross section using OsiriX [Rosset et al., 2004].
- Pancreas: The pancreatic parenchyma and pancreatic mass (cyst or tumor) were manually segmented in each slice by an expert abdominal radiologist using the Scout application [Dawant et al., 2007].
- Prostate: Annotations were performed by expert raters.

Mystery Phase:

- Colon: The colon was manually segmented using ITK Snap [Yushkevich et al., 2016] by an expert radiologist in body imaging.
- Hepatic Vessel: The liver vessels were semi-automatically segmented using the Scout application. Briefly, a seed point was drawn on the region of interest and grown using a level-set based approach. Contours were manually adjusted by an expert abdominal radiologist.
- Spleen: The spleen was semi-automatically segmented using the Scout application. A spline was drawn on the region of interest and grown using a level-set based approach. Contours were manually adjusted by an expert abdominal radiologist.

Annotation protocol Specific instructions from each center.

Data pre-processing methods To pre-process data we used `FSL` library functions. In particular, all images were transposed (without resampling) to the most approximate right-anterior-superior coordinate frame, ensuring the data matrix x-y-z direction was consistent using `fslreorient2std`. Lastly, non-quantitative modalities (e.g., MRI) were robust min-max scaled to the same range by means of a mixture of `fsl_maths` and `fsl_stats`.

Sources of error Data from several centers/devices and raters may cause inconsistencies.

Assessment methods

Metrics Overlap-based metric: Dice Similarity Coefficient (DSC); boundary-based metric: Normalized Surface Distance (NSD). For the NSD, tolerance values were based on clinical feedback and consensus, and were chosen by the clinicians segmenting each organ. NSD was defined at task level and was the same for all the targets of each task. The value represented what they would consider an acceptable error for the segmentation they were performing. The following values have been chosen for the individual tasks (in mm): Brain – 5; Heart – 4; Hippocampus – 1; Liver – 7; Lung – 2; Prostate – 4; Pancreas – 5; Colon – 4; Hepatic vessel – 3; Spleen – 3.

Justification of metrics The metrics DSC and NSD were chosen due to their popularity, rank stability, and smooth, well-understood and well-defined behavior when Region of Interest (ROI)s do not overlap. Having simple and rank-stable metrics also allows the statistical comparison between methods. It is important to note that the proposed metrics are not task-specific nor task-optimal, and thus, they do not fulfill the necessary criteria for clinical algorithmic validation of each task.

Ranking scheme A so-called significance score was determined for each algorithm a, separately for each task/target ROI c_i and metric $m_j \in \{\text{DSC}, \text{NSD}\}$ and referred to as $s_{i,j}(a)$. The significance score was computed according to the following four-step process:

Algorithm 8 Significance ranking for the Medical Segmentation Decathlon (MSD) challenge.

1: Performance assessment per case: Determine performance $m_j(a_l, t_{ik})$ of all algorithms a_l, with $l = \{1, ..., N_A\}$, for all test cases t_{ik}, with $k = \{1, ..., N_i\}$, where N_A is the number of competing algorithms and N_i is the number of test cases in competition c_i. Set $m_j(a_l, t_{ik})$ to 0 if its value is undefined.
2: Statistical tests: Perform a Wilcoxon signed-rank pairwise statistical test between algorithms $(a_l, a_{l'})$, with values $m_j(a_l, t_{ik}) - m_j(a_{l'}, t_{ik})$, for all $k = \{1, ..., N_i\}$.
3: Significance scoring: $s_{i,j}(a_l)$ then equals the number of algorithms performing significantly worse than a_l, according to the statistical test (per comparison α of 0.05, not adjusted for multiplicity).
4: Significance ranking: The ranking is computed from the scores $s_{i,j}(a_l)$, with the highest score (rank 1) corresponding to the best algorithm. Note that shared scores/ranks are possible. If a task has multiple target ROI, the ranking scheme is applied to each ROI separately, and the final ranking per task is computed as the mean significance rank.

The final score for each algorithm over all tasks of the development phase (the seven development tasks) and over all tasks of the mystery phase (the three mystery tasks) was computed as the average of the respective task's significance ranks. The full validation algorithm was defined and released prior to the start of the challenge, and available on the decathlon website (`http://medicaldecathlon.com/files/MSD-Ranking-scheme.pdf`).

Submissions with missing results Missing cases are set to the worst possible value, namely 0 for all metrics.

Justification of ranking method The ranking scheme was successfully applied for other well-known challenges. In addition, this ranking scheme favors groups of algorithms rather than single ones. It takes into account the hierarchical structure of the data.

Statistical analyses To investigate ranking uncertainty and stability, bootstrapping methods will be applied with 1,000 bootstrap samples. The statistical analysis will be performed using the open-source R toolkit challengeR [Wiesenfarth et al., 2021, 2019] for analyzing and visualizing challenge results. The original rankings computed for the development and mystery phases will be compared to the ranking lists based on the individual bootstrap samples. The correlation of pairwise rankings was determined via Kendall tau, which provides values between -1 (for reverse ranking order) and 1 (for identical ranking order).

A.8 BIAS reporting guideline

BIAS Reporting Guideline

Section/ Topic	Parameter name	Item No	Checklist Item	Reported on page No
TITLE, ABSTRACT, KEYWORDS	**Title**	**1**	Use the title to convey the essential information on the **challenge mission**. The title should ... • ... identify the paper as biomedical image analysis challenge. • ... indicate the image modality(ies) applied with a commonly used term in the title. • ... indicate the task and/or task category (e.g. classification, segmentation; see parameter 18) with a commonly used term in the title. • ... (optionally) include information on the biomedical target application. • ... (optionally) include the year for repeated challenges with fixed cycle.	
	Abstract	**2**	Provide a **summary** of the challenge purpose, design and results and report the main conclusion(s).	
	Keywords	**3**	List the primary **keywords** that characterize the challenge.	
INTRO-DUCTION	**Challenge motivation and objective**	**4a**	Provide a general introduction to the topic from a **biomedical point of view**. This should include the envisioned biomedical impact (short-term and/or long-term).	
		4b	Provide a general introduction to the topic from a **technical point of view**. This should include an overview of the state of the art along the envisioned technical/methodological impact.	
		4c	Based on the biomedical and technical motivation, provide a concise statement of the **primary challenge objective**. This should include a **statement of the task**.	
METHODS Challenge organi-zation	**Challenge name**	**5a**	Provide a **representative name** of the challenge. *Example:* MICCAI Endoscopic Vision Challenge 2015	
		5b	Provide the **acronym** of the challenge (if any). *Example:* EndoVis15	
	Organizing team	**6**	Provide information on the **organizing team** (names and affiliations).	
	Life cycle type	**7**	Define the intended **submission cycle** of the challenge. Include information on whether/how the challenge has been/will be continued after the present study. *Examples:* • One-time event with fixed submission deadline • Open call • Repeated event with annual fixed submission deadline	
Challenge venue and platform		**8a**	Report the **event** (e.g. conference) that was **associated** with the challenge (if any).	
		8b	Report the **platform** (e.g. grand-challenge.org) used to run the challenge.	
		8c	Provide the **URL** for the challenge website (if any).	
Participation policies		**9a**	Define the **allowed user interaction** of the algorithms assessed (e.g. only (semi-) automatic methods allowed).	
		9b	Define the policy on the **usage of training data**. The data used to train algorithms may, for example, have been restricted to the data provided by the challenge or to publicly available data including (open) pre-trained nets.	
		9c	Define the **participation policy for members of the organizers' institutes**. For example, members of the organizers' institutes could participate in the challenge but were not eligible for awards.	
		9d	Define the **award policy**. In particular, provide details with respect to challenge prizes.	
		9e	Define the policy for **results announcement**. *Examples:* • Top three performing methods were announced publicly. • Participating teams could choose whether the performance results will be made public.	
		9f	Define the **publication policy**. In particular, provide details on ... • ... who of the participating teams/the participating teams' members qualified as author • ... whether the participating teams could publish their own results separately, and (if so) • ... whether an embargo time was defined (so that challenge organizers can publish a challenge paper first).	
Submission method		**10a**	Describe the method used for result submission. If available, provide a link to the **submission instructions**. *Examples:* • Docker container on the Synapse platform. Link to submission instructions: <URL> • Algorithm output was sent to organizers via e-mail. Submission instructions were sent by e-mail.	
		10b	Provide information on the possibility for participating teams to **evaluate** their **algorithms before submitting** final results. For example, many	

BIAS Reporting Guideline

		challenges allow submission of multiple results, and only the last run is officially counted to compute challenge results.
Challenge schedule	**11**	Provide a **timetable** for the challenge. Preferably, this should include • the release date(s) of the training cases (if any) • the registration date/period • the release date(s) of the test cases and validation cases (if any) • the submission date(s) • associated workshop days (if any) • the release date(s) of the results
Ethics approval	**12**	Indicate whether **ethics approval** was necessary for the data. If yes, provide details on the ethics approval, preferably institutional review board, location, date and number of the ethics approval (if applicable). Add the URL or a **reference to the document** of the ethics approval (if available).
Data usage agreement	**13**	Clarify how the data can be used and distributed by the teams that participate in the challenge and by others. This should include the explicit **listing of the license** applied. *Examples:* • CC BY (Attribution) • CC BY-SA (Attribution-ShareAlike) • CC BY-ND (Attribution-NoDerivs) • CC BY-NC (Attribution-NonCommercial) • CC BY-NC-SA (Attribution-NonCommercial-ShareAlike) • CC BY-NC-ND (Attribution-NonCommercial-NoDerivs)
Code availability	**14a**	Provide information on the **accessibility of the organizers' evaluation software** (e.g. code to produce rankings). Preferably, provide a link to the code and add information on the supported platforms.
	14b	In an analogous manner, provide information on the **accessibility of the participating teams' code.**
Conflicts of interest	**15**	Provide information related to conflicts of interest. In particular provide information related to **sponsoring/funding** of the challenge. Also, state explicitly who had **access to the test case labels** and when.
Author contributions	**16**	List the **contributions** of all authors to the paper (preferably in the appendix).
METHODS Mission of the challenge	**Field(s) of application**	**17** State the **main field(s) of application** that the participating algorithms target. *Examples:* • Diagnosis • Education • Intervention assistance • Intervention follow-up • Intervention planning • Prognosis • Research • Screening • Training • Cross-phase
	Task category(ies)	**18** State the **task category(ies)**. *Examples:* • Classification • Detection • Localization • Modeling • Prediction • Reconstruction • Registration • Retrieval • Segmentation • Tracking
	Cohorts	We distinguish between the *target cohort* and the *challenge cohort*. For example, a challenge could be designed around the task of medical instrument tracking in robotic kidney surgery. While the challenge could be based on *ex vivo* data obtained from a laparoscopic training environment with porcine organs (challenge cohort), the final biomedical application (i.e. robotic kidney surgery) would be targeted on real patients with certain characteristics defined by inclusion criteria such as restrictions regarding gender or age (target cohort).
		19a Describe the **target cohort**, i.e. the subjects/objects from whom/which the data would be acquired in the final biomedical application.
		19b Describe the **challenge cohort**, i.e. the subject(s)/object(s) from whom/which the challenge data was acquired.
	Imaging modality(ies)	**20** Specify the **imaging technique(s)** applied in the challenge.
	Context information	Provide additional **information given along with the images**. The information may correspond ...
		21a ... directly to the **image data** (e.g. tumor volume). If necessary, differentiate between target and challenge cohort.
		21b ... to the **patient** in general (e.g. gender, medical history). If necessary, differentiate between target and challenge cohort.
		21c ... to the **acquisition process** (e.g. medical device data during endoscopic surgery, calibration data for an image modality). If necessary, differentiate between target and challenge cohort.
	Target entity(ies)	**22a** Describe the **data origin**, i.e. the region(s)/part(s) of subject(s)/object(s) from whom/which the image data would be acquired in the final

BIAS Reporting Guideline

			biomedical application (e.g. brain shown in computed tomography (CT) data, abdomen shown in laparoscopic video data, operating room shown in video data, thorax shown in fluoroscopy video). If necessary, differentiate between target and challenge cohort.
		22b	Describe the **algorithm target**, i.e. the structure(s)/subject(s)/object(s)/component(s) that the participating algorithms have been designed to focus on (e.g. tumor in the brain, tip of a medical instrument, nurse in an operating theater, catheter in a fluoroscopy scan). If necessary, differentiate between target and challenge cohort.
	Assessment aim(s)	**23**	Identify the **property(ies) of the algorithms to be optimized** to perform well in the challenge. If multiple properties were assessed, prioritize them (if appropriate). The properties should then be reflected in the metrics applied (parameter 29), and the priorities should be reflected in the ranking when combining multiple metrics that assess different properties. • *Example 1:* Find liver segmentation algorithm for CT images that processes CT images of a certain size in less than a minute on a certain hardware with an error that reflects inter-rater variability of experts. • *Example 2:* Find lung tumor detection algorithm with high sensitivity and specificity for mammography images. Corresponding metrics are listed below (parameter 29).
METHODS Challenge data sets	**Data source(s)**	**24a**	Specify the **device(s)** used to acquire the challenge data. This includes details on the device(s) used to acquire the imaging data (e.g. manufacturer) as well as information on additional devices used for performance assessment (e.g. tracking system used in a surgical setting).
		24b	Describe relevant details on the imaging process/**data acquisition** for each acquisition device (e.g. image acquisition protocol(s)).
		24c	Specify the **center(s)/institute(s)** in which the data was acquired and/or the **data providing platform/source** (e.g. previous challenge). If this information is not provided (e.g. for anonymization reasons), specify why.
		24d	Describe relevant **characteristics** (e.g. level of expertise) **of the subjects** (e.g. surgeon)/objects (e.g. robot) involved in the data acquisition process (if any).
	Training and test case characteristics	**25a**	State what is meant by one **case** in this challenge. A case encompasses all data that is processed to produce one result that is then compared to the corresponding reference result (i.e. the desired algorithm output). *Examples:* • Training and test cases both represented a CT image of a human brain. Training cases had a weak annotation (tumor present or not and tumor volume (if any)) while the test cases were annotated with the tumor contour (if any). • A case refers to all information that is available for one particular patient in a specific study. This information always includes the image information as specified in *data source(s)* (parameter 24) and may include context information (parameter 21). Both training and test cases were annotated with survival (binary) 5 years after (first) image was taken.
		25b	State the **total number** of cases as well as the number of training, validation and test cases separately.
		25c	Explain **why a total number** of cases and **the specific proportion** of training, validation and test cases was chosen.
		25d	Mention **further important characteristics** of the training, validation and test cases (e.g. class distribution in classification tasks chosen according to real-world distribution vs. equal class distribution) and justify the choice.
	Annotation characteristics	**26a**	Describe the **method for determining the reference annotation**, i.e. the desired algorithm output. Provide the information separately for the training, validation and test cases if necessary. Possible methods include *manual image annotation*, *in silico ground truth generation* and *annotation by automatic methods*. If human annotation was involved, state the **number of annotators**.
		26b	Provide the **instructions given to the annotators** (if any) prior to the annotation. This may include description of a training phase with the software. Provide the information separately for the training, validation and test cases if necessary. Preferably, provide a link to the **annotation protocol**.
		26c	Provide **details on the subject(s)/algorithm(s) that annotated** the cases (e.g. information on **level of expertise** such as number of years of professional experience, medically-trained or not). Provide the information separately for the training, validation and test cases if necessary.
		26d	Describe the **method(s) used to merge multiple annotations** for one case (if any). Provide the information separately for the training, validation and test cases if necessary.
	Data pre-processing method(s)	**27**	Describe the **method(s) used for pre-processing** the raw training data before it is provided to the participating teams. Provide the information separately for the training, validation and test cases if necessary.

BIAS Reporting Guideline

	Sources of error	**28a**	Describe the most relevant **possible error sources related to the image annotation**. If possible, **estimate the magnitude** (range) of these errors, using inter-and intra-annotator variability, for example. Provide the information separately for the training, validation and test cases, if necessary.
		28b	In an analogous manner, describe and quantify **other relevant sources of error**.
METHODS Assess-ment methods	**Metric(s)**	**29a**	Define the **metric(s) to assess a property of an algorithm**. These metrics should reflect the desired algorithm properties described in *assessment aim(s)* (parameter 21). State which metric(s) were used to compute the ranking(s) (if any). • *Example 1:* Dice Similarity Coefficient (DSC) and run-time • *Example 2:* Area under curve (AUC)
		29b	**Justify why** the metric(s) was/were chosen, preferably with reference to the biomedical application.
	Ranking method(s)	**30a**	Describe the **method used to compute a performance rank** for all submitted algorithms based on the generated metric results on the test cases. Typically the text will describe how results obtained per case and metric are aggregated to arrive at a final score/ranking.
		30b	Describe the method(s) used to manage **submissions with missing results** on test cases.
		30c	**Justify why** the described ranking scheme(s) was/were used.
	Statistical analyses	**31a**	Provide **details for all statistical methods** used in the scope of the challenge analysis. This may include • description of the **missing data handling**, • details about the assessment of **variability of rankings**, • description of any method used to assess **whether the data met the assumptions**, required for the particular statistical approach, or • indication of any **software product** that was used for data analysis.
		31b	**Justify why** the described statistical method(s) was/were used.
RESULTS Challenge outcome	**Challenge submissions**	Provide **summarizing information** on ...	
		32a	… the **number of registrations**.
		32b	… the **number of participating teams** that provided valid submissions (if applicable in each phase).
		32c	… the **number of participating teams** that the **paper** refers to (with **justification**).
	Information on selected participating teams	Provide the following information for the participating teams that are included in the paper:	
		33a	**Team identifier**.
		33b	A **method description** including parameter instantiation and/or a reference/URL to a document containing this information.
	Metric values	**34**	Provide raw and/or **aggregated metric values** (including measure of variability) for all participating teams and each metric (if applicable) as well as the **numbers of test set submissions** (the last one was used to compute metric(s)) for each participating team.
	Ranking(s)	**35a**	Report the **ranking(s)** (if any) including the number of test set submissions for each participating team.
		35b	Provide the results of the **statistical analyses**.
	Further Analyses	**36**	Present results of further analyses (if applicable), e.g. related to • **combining algorithms** via ensembling, • **inter-algorithm variability**, • **common problems/biases** of the submitted methods, or • **ranking variability**.
DISCUS-SION	**Summary**	**37**	Summarize the **main results** of the challenge.
	Impact	**38a**	Describe the (expected) **biomedical impact** of the challenge in the context of the state of the art with reference to the challenge motivation (parameter 4a).
		38b	Describe the (expected) **technical impact** of the challenge in the context of the state of the art with reference to the challenge motivation (parameter 4b).
	Discussion of challenge results	**39a**	Provide a detailed discussion and conclusion whether the **task is now solved** in a satisfactory way (e.g. the remaining errors are comparable to inter-annotator variability).
		39b	Provide a detailed **analysis of individual cases**, in which the majority of algorithms performed poorly (if any).
		39c	Provide a **discussion on advantages and disadvantages of the submitted methods**. Include time and memory consumption comparison if time and memory were not among the metrics.
	Limitations of the challenge	**40**	Discuss **limitations** related to the challenge design and execution.
	Future work	**41**	Provide **recommendations for future work** and maintenance plans for the challenge and its website (if any).
	Conclusions	**42**	Provide a **concise conclusion** based on the results of the study.

BIAS Reporting Guideline

A.9 Search terms related to mixed model analysis

Aiming for an overview of how many publications in the biomedical image analysis domain make use of Linear Mixed Model (LMM), we reviewed all 5,390 Medical Image Computing and Computer Assisted Interventions (MICCAI) papers in the period of 2004 to 2021. The publications were searched for 44 search terms that are related to mixed model analysis. The search terms are presented in Table A.1 along with the frequency with which the individual terms have been used among all publications.

Table A.1: List of mixed model analysis-related search phrases, together with how often they appeared in the 5,390 papers that were submitted to the Medical Image Computing and Computer Assisted Interventions (MICCAI) 2004—2021 conferences.

Search term	Frequency [%]	Search term	Frequency [%]
hierarchical model	0.61%	nonlinear mixed-effects model	0.04%
random effects	0.33%	random effect test	0.04%
fixed effects	0.19%	bayesian spatial generalized linear mixed model	0.02%
mixed-effects model	0.16%	clustered data	0.02%
mixed effects model	0.10%	generalized linear mixed model	0.02%
fixed effect	0.09%	LMM	0.02%
linear mixed effects model	0.07%	mixed effect	0.02%
random effect	0.07%	mixed-effect	0.02%
random effect analysis	0.07%	mixed-effect generative model	0.02%
linear mixed effect model	0.06%	mixed-effect linear model	0.02%
mixed-effect model	0.06%	mixed-effect regression	0.02%
mixed-effects analysis	0.06%	mixed effects	0.02%
nonlinear mixed effects model	0.06%	mixed-effects	0.02%
random-effects	0.06%	mixed-effects linear model	0.02%
random effects analysis	0.06%	mixed-effects regression	0.02%
fixed-effects	0.04%	multilevel model	0.02%
linear mixed-effects model	0.04%	nested model	0.02%
LME	0.04%	random effect map	0.02%
mixed effect model	0.04%	random effects model	0.02%
mixed-effect-model	0.04%	random-effects regression	0.02%
NLME	0.04%	random-effects statistics	0.02%
		variance components model	0.02%