Inaugural-Dissertation

zur

Erlangung der Doktorwürde

der

Gesamtfakultät für Mathematik, Ingenieur- und Naturwissenschaften

der

Ruprecht-Karls-Universität Heidelberg

vorgelegt von

Shuhan Xiao, M. Sc.

geboren in Duisburg, Deutschland

Tag der mündlichen Prüfung: _____

Predictive Imaging Biomarker Discovery and Treatment Effect Estimation Using Deep Learning in Randomized Imaging Studies

Primary Supervisor: Prof. Dr. Klaus Maier-Hein
Secondary Supervisor: Prof. Dr. Philipp Vollmuth

Abstract

Personalized medicine aims to tailor treatments to patients based on individual patient characteristics and plays an essential role for advancing healthcare and achieving better patient outcomes. As patients often respond very differently, improving personalized treatment decisions is a key challenge in this field. In clinical practice, such decisions are based on predictive biomarkers that indicate whether a patient might benefit from treatment. While established predictive biomarkers often require invasive procedures, medical imaging offers a non-invasive alternative by providing high-dimensional, spatially resolved information that could reveal patterns relevant for making treatment decisions. However, existing approaches, such as radiomics, rely on handcrafted features rather than directly estimating treatment-specific effects from imaging data.

To address the current gaps, this thesis investigates the task of discovering predictive imaging biomarkers in a data-driven way directly from images and providing treatment recommendations using pre-treatment imaging data without a separate feature extraction step.

In the first part of this thesis, the first approach for discovering predictive imaging biomarkers using deep-learning-based causal models for estimating heterogeneous treatment effects is presented. Its main contribution is an evaluation protocol for assessing identified predictive imaging biomarker candidates and for assessing model performances, which enables quantitative benchmarking and qualitative interpretation of image-based treatment effect estimation models. The proposed protocol specifically makes the important distinction between predictive and prognostic biomarkers, the latter of which can predict patient outcomes independently of treatment, by comparing predictive and prognostic effects.

In the second part, image-based treatment effect estimation methods are applied to both semi-synthetic and real clinical imaging data from a randomized phase II/III trial in glioblastoma by developing an extension of previous models for binary or continuous outcomes adapted to more clinically relevant survival outcomes. Furthermore, it investigates the impact of multimodal integration of clinical tabular data and the use of pre-trained image encoders on the resulting treatment recommendations of the proposed model and patient stratification.

The experimental results demonstrate that image-based treatment effect estimation models can identify predictive imaging biomarkers from semi-synthetic image datasets and provide interpretable insights, although the performance on real clinical data remains limited due to small sample sizes and weak treatment effect signals. Nevertheless, the findings of this work offer valuable insights into the opportunities and current limitations of image-based treatment effect estimation under realistic constraints and highlight key directions for future research. Overall, this work bridges causal inference and medical image analysis, establishing a foundation for future research on radiomics-free predictive imaging biomarker discovery and for advancing image-based methods that support personalized treatment decision-making.

Zusammenfassung

Die personalisierte Medizin hat das Ziel, Behandlungen auf Grundlage individueller Patientenmerkmale auf Patient:innen maßzuschneidern und spielt eine wesentliche Rolle dabei, das Gesundheitswesen weiterzuentwickeln und bessere Patientenergebnisse zu erreichen. Da Patient:innen oft sehr unterschiedlich auf Behandlungen ansprechen, ist die Verbesserung personalisierter Behandlungsentscheidungen eine zentrale Herausforderung in diesem Bereich. In der klinischen Praxis basieren solche Entscheidungen auf prädiktiven Biomarkern, die anzeigen, ob Patient:innen von einer Behandlung profitieren könnten. Während etablierte prädiktive Biomarker oft invasive Eingriffe erfordern, bietet die medizinische Bildgebung eine nicht-invasive Alternative, indem sie hochdimensionale, räumlich aufgelöste Informationen liefert, die Muster erkennen könnten, die für Behandlungsentscheidungen relevant sind. Allerdings verlassen sich bestehende Ansätze wie Radiomics auf manuell entwickelte Merkmale, anstatt behandlungsspezifische Effekte direkt aus Bilddaten abzuschätzen.

Um diese bestehenden Lücken zu schließen, untersucht diese Dissertation die Aufgabe, ohne einen separaten Schritt prädiktive bildgebende Biomarker auf eine datengestützte Weise direkt aus Bildern zu entdecken und Behandlungsempfehlungen anhand von Bildgebungsdaten zu geben, die vor einer Behandlung aufgenommen wurden.

Im ersten Teil dieser Dissertation wird der erste Ansatz zur Ermittlung prädiktiver bildgebender Biomarker vorgestellt, der Deep-Learning-basierte kausale Modelle zur Abschätzung heterogener Behandlungseffekte verwendet. Der zentrale Beitrag ist ein Evaluierungsprotokoll, das dazu dient, identifizierte prädiktive bildgebende Biomarker-Kandidaten zu bewerten und die Leistung eines Modells zu beurteilen, welches ein quantitatives Benchmarking und qualitative Interpretation bildbasierter Modelle zur Schätzung von Behandlungseffekten ermöglicht. Das vorgeschlagene Protokoll unterscheidet ausdrücklich zwischen prädiktiven und prognostischen Biomarkern, wobei letztere Patientenergebnisse unabhängig von der Behandlung vorhersagen können, indem es prädiktive und prognostische Effekte vergleicht.

Im zweiten Teil werden bildbasierte Methoden für die Schätzung von Behandlungseffekten sowohl auf semi-synthetische als auch auf echte klinische Bilddaten aus einer randomisierten Phase-II/III-Studie zu Glioblastomen angewendet, indem eine Erweiterung früherer

Modelle entwickelt werden, die ursprünglich für binäre oder kontinuierliche Ergebnisse konzipiert wurden, auf klinisch relevantere Überlebensergebnisse zu erweitern. Darüber hinaus wird der Einfluss der multimodalen Integration von klinischen tabellarischen Daten und der Verwendung von vortrainierten Bildencodern auf die resultierenden Behandlungsempfehlungen des vorgeschlagenen Modells sowie die Patientenstratifizierung untersucht.

Die experimentellen Ergebnisse zeigen, dass bildbasierte Modelle zur Schätzung von Behandlungseffekten prädiktive bildgebende Biomarker aus semi-synthetischen Bilddatensätzen identifizieren und interpretierbare Einblicke liefern können, obwohl die Leistungsfähigkeit in der Anwendung auf echte klinische Daten aufgrund kleiner Stichprobengrößen und schwacher Signale für die Behandlungswirkung nach wie vor begrenzt bleibt. Dennoch bieten die Ergebnisse dieser Dissertation wertvolle Einblicke in die Möglichkeiten und aktuellen Einschränkungen der bildbasierten Schätzung von Behandlungseffekten unter realistischen Bedingungen und weisen auf wichtige Richtungen für zukünftige Forschung hin. Insgesamt schlägt diese Arbeit eine Brücke zwischen kausaler Inferenz und medizinischer Bildanalyse und schafft damit eine Grundlage für die zukünftige Radiomics-freie Entdeckung prädiktiver bildgebender Biomarker sowie für die Weiterentwicklung bildbasierter Methoden, die personalisierte Behandlungsentscheidungen unterstützen.

Acknowledgments

Over the years, many people have been part of my journey as a PhD student at the German Cancer Research Center (DKFZ) who have supported me in one way or another, and who I would like to acknowledge here.

First and foremost, I would like to express my deepest gratitude to my supervisor, Klaus Maier-Hein, for welcoming me to his division, having faith in me to tackle the research problems he originally envisioned, and giving me the opportunity to grow as an independent researcher. I appreciated the calmness and guidance he provided whenever there were uncertainties, and I will not take the community and environment he shaped for granted.

Secondly, I would like to thank my clinical co-supervisor, Philipp Vollmuth, for the immensely exciting initial research direction he provided and for giving me the opportunity to work on it, which led me to explore many fascinating scientific areas and new perspectives across different fields that I likely would not have had the opportunity to work on otherwise, and I am glad to have followed these paths.

The initial stages of my research were especially shaped by my (internal) thesis advisory committee (iTAC) member Paul Jäger, and I would like to thank him for his initial ideas and advice, which provided clarity and helped me significantly to refine my research directions. Further, my sincere thanks go to all other iTAC members for their collaboration and for supporting me in my PhD projects, Lukas Klein, Jens Petersen, and Jonas Bohn. Thank you also to Nico Disch for reading parts of this thesis.

A big thank-you goes to the SYMIC office, Stefanie Strzysch, Michaela Gelz, Nina Kraft, and Theresa Klocke, and our scientific coordinators Kathrin Brunk, Daniel Walther, and Nina Decker, for their support and for keeping everything running and handling all the organisational hurdles.

Thank you to the rest of the Division of Medical Image Computing (MIC), especially the third floor, for lifting the mood (including ACVL and the former IML group), as well as group 4 within MIC and the SYMIC community in general. I am glad that I could be part of such a collaborative and supportive environment where people were always happy to help or where there was always someone to ask whenever needed. On that

note, Jan Sellner generously shared his LaTeX template of his thesis with me, for which I want to thank him again. In this context, I would also like to thank Yannick Kirchhoff for answering my data-specific questions and Sebastian Ziegler for code-related ones I had in the second part of my thesis.

Throughout my experience as a PhD student, my office mates Bálint Kovács, Maximilian Zenk, Stefan Denner, Ole Johannsen, Alexandra Ertl, Hamideh Haghiri, and Silvia Dias Almeida were a significant factor in creating to such an enjoyable work environment, for which I am very grateful. Working together on challenges with Alex and Bálint, and especially our creative projects (including our Christmas videos), were experiences I really love looking back to. I also like to think that running together with Kim-Celine Kahl, and the (half-marathon) races I did together with Stefan, Michi, Steffi, Markus Bujotzek, Lisa Kausch, and many other SYMIC members contributed to the endurance I needed for finishing writing up this thesis in some way, and I want to thank them for the motivation they provided in that regard.

Last, but not least, I would like to express a heartfelt thank you to my parents for their unconditional support, especially during the last phase of finishing up this thesis, and their understanding, despite all the stress this period might have caused. I really hope I can make up for it soon.

Contents

1 Introduction

1.1 Motivation

Images, whether captured by medical scanners, microscopes, satellites, or industrial sensors, contain a wealth of information useful for making informed decisions. Often, the central goal is not only to analyze images by what is seen to get a better understanding (e.g. by segmenting objects or classifying them into categories), but to decide what actions to take based on predicted consequences, whether to intervene, what treatment to choose, or which process to apply to change the state and condition of a system. Examples for areas of application where such questions are relevant may range from policy-making for poverty relief using satellite imagery (Jerzak et al. 2023) or precision agriculture using aerial images (Tantalaki et al. 2019; Kim et al. 2021) to robotics using video data (Ho et al. 2020; Li 2023; Gupta et al. 2024).

Fundamentally, making good decisions relies on understanding causal relationships: how will an action such as the application of a treatment or a change in policy affect the eventual outcome? One prominent example where visual information plays a critical role in the decision-making process of treatment planning and in investigating the underlying causal mechanisms is the area of personalized medicine. There, the ultimate goal is to achieve the best possible patient outcome by tailoring treatments to individual patients (Radiology (ESR) 2015).

Medical imaging, such as magnetic resonance imaging (MRI) or computed tomography (CT), plays an important part in planning such tailored treatments as it can provide immediate high-dimensional and spatially resolved information about a patient through a non-invasive acquisition process. For instance, in oncology, images are used in clinical practice for radiotherapy planning, and criteria such as the Response Evaluation Criteria in Solid Tumors (RECIST) (Eisenhauer et al. 2009) or the Response Assessment in Neuro-Oncology (RANO) criteria (Wen et al. 2023) are routinely applied to assess treatment

response and guide subsequent treatment decisions based on measurements of tumor size or lesions.

However, since not all patients might benefit equally from a given treatment and since response assessments typically occur after treatment has started, an important aspect of developing personalized treatments is to find criteria that can predict whether a patient will likely benefit from a treatment beforehand. For this reason, personalized medicine increasingly tailors treatments based on so-called *predictive biomarkers*.

While biomarkers are generally measurable patient-specific characteristics indicating the medical state of an individual or which can be associated with clinical outcomes such as disease status, physiological measures or survival (Lohr 1988; Strimbu et al. 2010), predictive biomarkers indicate the likely benefit (or generally effect) of a treatment for a given individual within a wider population, where the treatment effect can vary. Identifying predictive biomarkers is therefore crucial for determining which subgroup of individuals will have a positive treatment effect and ultimately for making informed treatment decisions. As this concept is also relevant to making treatment decisions outside of biomedicine, predictive biomarkers are also referred to as predictive covariates or features in a wider context.

The use of predictive biomarkers has been successfully and widely adopted in clinical practice in oncology to select the most beneficial treatment for specific patient subgroups. For instance, in breast cancer, the human epidermal growth factor receptor 2 (HER2) overexpression and estrogen receptor (ER) status in tumors serve as predictive biomarkers for therapies such as trastuzumab (a targeted therapy) or tamoxifen (a hormone therapy), respectively (Tarighati et al. 2023). These well-established predictive biomarkers illustrate how stratifying patients based on a single covariate can inform treatment and improve patient outcomes. Furthermore, these examples demonstrate that the discovery of predictive biomarkers can go hand-in-hand with the development of novel, targeted treatments and support drug discovery.

While predictive biomarkers, such as HER2 and ER status, are usually determined invasively through the analysis of tumor biopsy samples, ongoing research has explored whether the biomarker status can also be estimated through less invasive imaging techniques such as nuclear imaging (Ulaner et al. 2016; Weaver et al. 2018; Salvatore et al. 2019). This growing interest highlights the potential of predictive imaging biomarkers, i.e. features extracted from images that can inform treatment decisions (O'Connor et al. 2017).

When researching new predictive biomarkers, it is essential to distinguish them from *prognostic biomarkers*, which can predict a patient's outcome independent of a treatment and are therefore not indicative of treatment effects (Ballman 2015). Examples of prognostic biomarkers include tumor size or age, where higher values are typically associated with a worse prognosis, regardless of which treatment a patient receives. It is important

to note that biomarkers can be both prognostic and predictive simultaneously, and are associated with both the outcome itself and the treatment effect. For these reasons, it is not possible to identify predictive biomarkers from studying treatment response data of all patients receiving the same treatment alone, such as by simply comparing the states before and after treatment, as observed responses may stem from prognostic biomarkers.

To discover predictive biomarkers and drive the development of personalized treatments, it is necessary to compare the outcomes of treated individuals against a control (such as placebo or standard treatment) to investigate whether specific covariates are associated with variations in treatment effects across patient subgroups, where the effects are measured with respect to a pre-defined outcome of interest (Mandrekar et al. 2009). In clinical research, such outcome data is typically acquired through randomized controlled trials (RCTs), which are the gold standard of clinical trials to make claims about causal relationships (Zabor et al. 2020). In the initial stages of predictive biomarker research, data from such clinical trials is often analyzed retrospectively (Alymani et al. 2010).

RCTs are conducted by randomly assigning individuals to a treatment or control arm to ensure that patient covariates (e.g. age, tumor volume) are equally distributed across arms to reduce confounding. However, RCT data are often analyzed only at a population level by computing the average treatment effect. For personalized medicine and predictive biomarker discovery, the interest lies instead in the individual treatment effect, that is, how a specific patient would respond to one treatment compared to another. As observing both treatment and control arm outcomes of the same patient and thereby measuring a patient's individual treatment effect is not possible due to the fundamental problem of causal inference (Holland 1986), the problem of treatment effect estimation is inherently different from standard prediction tasks where a ground truth can be obtained.

This problem has been investigated in the field of causal inference, which provides a theoretical framework for estimating treatment effects from randomized trials and observational data. In particular, machine learning methods have been developed to estimate the conditional average treatment effect (CATE), which is a measure for the true individual treatment effect (ITE) that cannot be observed directly and captures how the expected treatment effects vary across individuals based on their characteristics. These methods aim to capture treatment effect heterogeneity within a population and can support personalized decision-making by estimating how much a given patient might benefit from a treatment (Curth et al. 2024).

Motivated by the goals of personalized medicine, these CATE estimation have also been investigated for identifying predictive biomarkers (Sechidis et al. 2018; Bahamyirou et al. 2022; Crabbé et al. 2022; Boileau et al. 2023; W. Zhu et al. 2023; Verhaeghe et al. 2025). While CATE estimation methods are well-established for tabular input data, translating them to high-dimensional inputs such as medical images remains an emerging research topic. Recent work has begun adapting deep-learning-based treatment effect estimation methods to image inputs (Durso-Finley et al. 2022; Durso-Finley et al. 2023; Ma et al. 2023),

but their potential to identify predictive imaging biomarkers has not been systematically explored.

Instead of using treatment effect estimation, the discovery of predictive imaging biomarkers has conventionally relied on handcrafted features such as radiomics features (Chiu et al. 2023). These image features, including features such as tumor intensity, texture, or shape, are extracted in a multi-step process involving segmenting a region of interest (e.g. tumor), feature selection, and a statistical analysis, where biases can easily be introduced (Lambin et al. 2017; Hosny et al. 2019).

This thesis addresses the research gap in this area by developing novel methods for discovering predictive imaging biomarkers in a data-driven manner and estimating treatment effects using deep learning directly from medical images. Motivated by the "Bitter Lesson" of artificial intelligence (AI) Sutton (2019), which emphasizes the success of scalable, data-driven methods over handcrafted solutions with manually in-built knowledge, this thesis explores whether deep learning models can learn predictive imaging biomarkers directly from images, without pre-defined features. Through evaluations on both semi-synthetic and clinical imaging datasets, this work investigates their feasibility, limitations, and potential clinical utility in supporting treatment decision-making based on images.

While this thesis aims to develop generalizable methods across disease areas, a large part of this thesis investigates the application of the newly developed methods to an RCT that studies the treatment of glioblastoma. Glioblastoma is the most common and most aggressive form of malignant brain cancer (Grochans et al. 2022). It remains difficult to treat with standard treatments such as surgery, radiotherapy, and chemotherapy, as they often fail to halt tumor progression, which has motivated a continued effort in researching more effective therapies (Rodríguez-Camacho et al. 2022). One candidate that has been widely studied is bevacizumab (BEV), an anti-angiogenic drug used to target the growth of blood vessels in the brain tumors. However, RCTs with unstratified glioblastoma patient populations have shown only limited or no benefit in terms of overall survival (OS) (Chinot et al. 2014; Gilbert et al. 2014; Wick et al. 2017; Ameratunga et al. 2018). This limited success has raised the question of whether bevacizumab may be beneficial to a certain patient subgroup, and whether these subgroups can be identified non-invasively through predictive imaging biomarkers from routinely acquired imaging data.

Despite recent attempts using radiomics features and other handcrafted features, no reliable predictive imaging biomarker has been established for bevacizumab to date, and the discovery of novel imaging biomarkers still remains a significant challenge (Kickingereder et al. 2015; Kickingereder et al. 2016; Grossmann et al. 2017; Schell et al. 2020; Ammari et al. 2021). The availability of data from a randomized trial in patients with recurrent glioblastoma presents a valuable opportunity to tackle this challenge from a new perspective. This work leverages this RCT dataset to investigate whether deep-learning-based treatment effect estimation methods, utilizing images directly, can

provide a new direction for guiding personalized therapy strategies with the goal of improving the overall survival of patients.

1.2 Objectives and Contributions

Motivated by the potential to advance the field of image-based decision-making, the objective of this thesis is to investigate whether deep-learning-based heterogeneous treatment effect (HTE) estimation methods can support predictive imaging biomarker discovery directly from imaging data and improve the ability to make treatment decisions.

This thesis addresses this topic from two perspectives while answering following research questions (RQs), through research spanning from methodological development of an evaluation protocol for predictive imaging biomarker discovery ("Part 1") to a study investigating methods for the application of treatment effect estimation methods in clinical imaging data ("Part 2"):

Objective of Part 1: Method Development and Evaluation

The first part of the thesis studies the evaluation for predictive imaging biomarker discovery in a semi-synthetic setting, where treatment effects are simulated from known pre-defined image features. At the same time, it focuses on the feasibility and robustness of image-based heterogeneous treatment effect estimation models by benchmarking how well such models perform at this task. Two main RQs are investigated in this part:

RQ 1.1: Can deep-learning-based heterogeneous treatment effect estimation be used to discover predictive imaging biomarkers directly from image data without a separate feature extraction step?

This question investigates whether predictive imaging biomarkers can be learned directly using deep learning in a data-driven way, with the aim of providing complementary and new insights to conventional approaches using feature engineering. It is directly motivated by previously discussed limitations of radiomics-based approaches, including their potential bias and their often laborious feature extraction process. Given the lack of true ground-truth predictive biomarkers in real, non-synthetic data, the question additionally encompasses the development of experiments that can demonstrate the feasibility of the predictive imaging biomarker process using image data.

RQ 1.2: How can the performance and reliability of image-based heterogeneous treatment effect models in discovering predictive imaging biomarkers be evaluated both quantitatively and qualitatively?

To quantify the predictive imaging biomarker candidates identified by the deep-learning-based models and assess their interpretability, this question focuses on establishing robust evaluation strategies. These strategies are designed to be applied in both semi-synthetic settings and real applications to study any newly identified predictive imaging biomarker candidates.

Objective of Part 2: Clinical Imaging Application Study

The second part of the thesis investigates the translation of methodological approaches to imaging data of real clinical patients. It applies the image-based treatment effect estimation to data from a randomized phase II/III trial in recurrent glioblastoma as well as a semi-synthetic lung cancer dataset simulating an RCT. The goal is to investigate the feasibility of making treatment recommendations directly from imaging inputs in realistic clinical settings. Furthermore, this part addresses specific requirements of such studies by extending the models to handle survival (time-to-event) outcomes and multimodal inputs (e.g. clinical tabular data). It investigates the following RQs:

RQ 2.1: Can image-based heterogeneous treatment effect estimation methods be extended from categorical or continuous outcomes to survival (time-to-event) outcomes, and how does their treatment recommendation performance compare to binary-outcome models?

While most prior work on image-based treatment effect estimation focuses on binary or continuous outcomes, clinical applications are often primarily interested in the overall survival, which is the main endpoint (i.e. outcome of interest) in many clinical studies, particularly in oncology (Delgado et al. 2021). This question explores whether survival-specific loss functions can improve treatment effect recommendations by handling censored time-to-event data (i.e. where the event of interest, in this case death, is not observed) with more nuance compared to modeling a binary survival status derived by thresholding the survival time.

RQ 2.2: Can the integration of multimodal inputs or pre-trained image encoders improve treatment effect estimation performance and robustness on clinical imaging data?

As clinical tabular data is almost always available in clinical trial settings, this question investigates whether leveraging multimodal inputs can lead to an improvement in treatment effect estimation and treatment recommendation performance. Furthermore, to incorporate richer (e.g. anatomical) information and to reduce the risk of overfitting in applications with limited data, this question explores the benefit of fine-tuning pre-trained image encoders.

RQ 2.3: To what extent can image-based heterogeneous treatment effect estimation models be applied to glioblastoma MRI data from a randomized clinical trial, and what are their limitations and implications for predictive imaging biomarker discovery?

This question addresses the translation of the developed treatment effect estimation methods to the clinical trial dataset in glioblastoma. It critically examines model performance, potential sources of limitations, and the reliability of predictive imaging biomarker discovery under realistic constraints such as limited sample size, censoring, and dataset imbalances.

Contributions

By addressing the aforementioned objectives, this thesis made the following contributions:

- Introduced a novel task for radiomics-free predictive imaging biomarker discovery using deep-learning-based CATE estimation directly from images, without relying on handcrafted or radiomics features.
- Proposed a new quantitative and qualitative evaluation protocol to assess predictive imaging biomarker discovery and to benchmark the performance and interpretability of deep-learning-based CATE models in discovering predictive imaging biomarkers.
- Conducted a comprehensive evaluation of image-based treatment effect estimation models in both semi-synthetic and clinical settings, including experiments that simulate predictive and prognostic imaging biomarkers from pre-defined imaging features with varying strengths and generate semi-synthetic RCT survival outcomes from real clinical observational outcomes.
- Established the first approach for image-based CATE estimation with survival (time-to-event) outcomes, analyzing the trade-offs between modeling censored survival data and using simple thresholded binary survival outcomes for treatment recommendations.
- Presented the first investigation of integrating and fine-tuning pre-trained image encoders for treatment effect estimation and treatment recommendation from MRI data, particularly in glioblastoma. Conducted the first systematic investigation into the impact of advanced deep learning strategies for improving image-based CATE estimation in clinical imaging settings, including multimodal extensions by integrating clinical tabular data, segmentation masks, and multitask learning.
- Performed the first application of image-based treatment effect estimation to an RCT in glioblastoma, analyzing heterogeneous treatment effects and evaluating predictive imaging biomarker discovery performance for treatment with bevacizumab.

This work provided a rigorous analysis of the translational gap and key challenges in applying such models under real-world clinical constraints, establishing insights for future clinical applications.

1.3 Outline

Most chapters of this thesis, apart from this chapter and Chapter 2, are structured by the two core research focuses corresponding to the research questions presented earlier:

1. "Evaluating heterogeneous treatment effect estimation models for predictive imaging biomarker discovery", addressing RQ1.1 and RQ1.2, and
2. "Image-based heterogeneous treatment effect estimation in clinical imaging studies" addressing RQ2.1, RQ2.2 and RQ2.3.

Following this introduction, the thesis begins with an overview of the relevant clinical and methodological background covering causal inference and medical image analysis in Chapter 2.

Chapter 3 reviews the recent works and state-of-the-art methods related to predictive imaging biomarker discovery and treatment effect estimation, and highlights the main research gaps in the relevant domains that are addressed in this thesis.

Next, Chapter 4 describes the image-based treatment effect estimation and predictive biomarker discovery evaluation methods developed in this thesis and the experimental setup for validating them, including imaging datasets and a strategy for the simulation of RCT outcomes. Chapter 5 presents the results of the experiments and an extensive evaluation conducted to answer the research questions.

The main part concludes with a discussion of the experimental results, their broader implications, main limitations, the overarching themes connecting both studies, and future research directions in Chapter 6. Finally, a summary of the main contributions and insights, as well as a brief outlook are provided in Chapter 7.

Background 2

This chapter introduces the most relevant medical and methodological background for this thesis. Section 2.1 provides the clinical context of predictive and prognostic imaging biomarkers, and introduces glioblastoma as a motivating application. Section 2.2 introduces the formal framework of causal inference and treatment effect estimation, while Section 2.3 covers the fundamentals of survival analysis, both of which are necessary for the methodological developments in later chapters.

2.1 Medical Context

2.1.1 Predictive and Prognostic Imaging Biomarkers

Biomarkers, which is short for biological markers, are measurable characteristics that provide information about disease status, prognosis, or response to therapy of a patient or underlying biological processes, and are used to inform clinical decisions (Strimbu et al. 2010). Motivated by their diverse clinical applications, biomarkers have been categorized into different subtypes, including diagnostic, predictive, and prognostic biomarkers (Califf 2018).

A *prognostic biomarker* provides information about the likely course of a disease independent of treatment and is used to identify patients with higher risks. In contrast, a *predictive biomarker* indicates whether a patient is likely to benefit from a particular therapy (Ballman 2015). These predictive biomarkers play a particularly important part in personalized medicine, as they inform treatment decisions and can be used for enrichment in clinical trials, where the trial design focuses on selecting patients that might respond more positively to a treatment (Renfro et al. 2016). Established examples from oncology include HER2 in breast cancer or epidermal growth factor receptor (EGFR) in non-small-cell lung cancer (NSCLC), which guide the use of targeted therapies such as

trastuzumab or (Šutić et al. 2021; Tarighati et al. 2023). These illustrate how predictive biomarkers can directly inform therapy selection. However, most established biomarkers, as the previously mentioned examples, are derived from molecular profiling, where tumor tissue samples obtained through invasive biopsy are analyzed on a molecular level, and which may not capture the full biological heterogeneity within a lesion or across multiple lesions.

Medical imaging, in contrast, provides a non-invasive alternative for characterizing tumors and other diseases. Radiologists routinely assess MRI or CT scans to evaluate tumor morphology, enhancement patterns, and progression over time. Leveraging this data for biomarker research has led to the development of imaging biomarkers and, more specifically, the field of radiomics (H. Aerts et al. 2014; Parmar et al. 2015; Kickingereder et al. 2016; Lambin et al. 2017; Limkin et al. 2017; Park et al. 2018; Chaddad et al. 2019). Radiomics, as first developed by H. Aerts et al. (2014), aims to extract imaging biomarkers automatically by segmenting a region of interest, extracting pre-defined handcrafted features, and performing statistical analysis of prognostic or predictive effects and often identifying correlations with molecular biomarkers or clinical outcomes. More recently, deep learning methods have been used for the feature extraction step with automatically learned representations from image data using pre-trained networks. A key limitation of radiomics approaches lies in their limited reproducibility, which can be affected by variations in image acquisition, preprocessing, and feature definition across studies (Pfaehler et al. 2021).

2.1.2 Glioblastoma: Clinical Context and Treatment

Glioblastoma is the most aggressive and common malignant primary brain tumor in adults (Grochans et al. 2022). Classified as a World Health Organization (WHO) grade IV glioma, it is associated with a poor prognosis, with median overall survival typically below 15 months despite aggressive therapy (Ballman et al. 2007; Tan et al. 2020). Standard treatment includes maximal safe surgical resection followed by radiotherapy and concurrent chemotherapy, while corticosteroids are often used to alleviate peritumoral edema and neurological symptoms (Caramanna et al. 2022). Recurrence, however, is almost always observed.

In recent years, many computational tools for brain tumor analysis have been developed in medical image computing, including automated tumor segmentation (Kickingereder et al. 2019; Helland et al. 2023), growth modeling (Petersen et al. 2019; Elazab et al. 2020), and survival or outcome prediction (Patel et al. 2021; Li et al. 2022; Poursaeed et al. 2024). These approaches aim to objectively quantify tumor morphology and progression, supporting diagnosis, therapy planning, and treatment monitoring. Their overall clinical goal aligns with the motivation of this thesis, which is to use quantitative imaging information to better guide personalized treatment decisions.

MRI is central to the diagnosis, treatment planning, and follow-up of glioblastoma. Multiple MRI sequences, such as T1-weighted, contrast-enhanced T1-weighted (cT1-w), T2-weighted (T2-w), and FLAIR, provide complementary information about tumor structure, infiltration, necrosis, and edema. Tumor burden and response to therapy are commonly assessed using the Response Assessment in Neuro-Oncology (RANO) criteria (Wen et al. 2010; Wen et al. 2023), which rely on manual two-dimensional measurements of contrast-enhancing lesions.

The anti-angiogenic therapy Bevacizumab (BEV), which is a humanized monoclonal antibody (hence the ending "-zumab") that targets vascular endothelial growth factor (VEGF) (Ferrara et al. 2005), has been investigated for recurrent glioblastoma. While BEV has been found to reduce edema and delay progression, large randomized controlled trials have failed to demonstrate a consistent benefit in overall survival across unstratified patient populations (Gilbert et al. 2014; Wick et al. 2017). This has motivated efforts to identify subgroups of patients who might benefit from BEV using predictive biomarkers.

Previous studies have examined both molecular and imaging biomarkers for BEV response. Imaging-based biomarkers, such as the apparent diffusion coefficient (ADC_{low}) or perfusion parameters derived from dynamic susceptibility contrast MRI, have shown prognostic associations but failed to demonstrate predictive value (Kickingereder et al. 2020; Schell et al. 2020). Molecular biomarkers, including MGMT promoter methylation, NF1 mutation status, and the proneural subtype, have shown promise in some studies (Sandmann et al. 2015; Kessler et al. 2023), but require invasive sampling and are not yet clinically validated for treatment selection. This ongoing lack of reliable predictive imaging biomarkers underscores the need for new approaches that can leverage MRI data to identify patient subgroups that respond positively to treatment non-invasively, which has also been the medical motivation for the methodological developments in this thesis.

2.2 Causal Inference and Treatment Effect Estimation

2.2.1 Potential Outcomes and Treatment Effects

In treatment effect estimation, the relation between outcomes and treatment effects has been formalized by the Neyman-Rubin potential outcome framework (Rubin 2005). In this framework, each individual i with observed pre-treatment features (e.g. clinical features or imaging data) $x_i \in \mathbb{R}^d$ is considered to have potential outcomes $Y_i(T) \in \mathbb{R}$, which are the outcomes that would be observed depending on a treatment assignment T_i. These outcomes can be binary, categorical, continuous or time-to-event data, for example describing the disease status or survival time of a patient. The observed dataset of a population of n individuals is then $\mathcal{D} = \{(x_i, T_i, Y_i)\}_{i=1}^n$.

Under the assumption that the treatment assignment is binary, i.e. $T \in \{0, 1\}$ indicating for example a control treatment or standard therapy $T = 0$ vs. an experimental treatment $T = 1$, the average treatment effect (ATE) can be computed to investigate treatment effects on a population level. It is defined as the difference between the two potential outcomes $Y_i(T = 0)$ and $Y_i(T = 1)$:

$$\text{ATE} := \mathbb{E}\left[Y_i(T = 1) - Y_i(T = 0)\right]. \tag{2.1}$$

More relevant for making treatment recommendations is the ITE for an individual

$$\text{ITE} := Y_i(T = 1) - Y_i(T = 0). \tag{2.2}$$

Depending on the actual applied treatment the observed outcome is known as the factual outcome, whereas the unobserved outcome is known as the counterfactual outcome. Both potential outcomes cannot be observed for the same individual at the same time, which is the fundamental problem of causal inference (Holland 1986) and which is why the ITE cannot be measured directly.

For this reason, only the conditional average treatment effect (CATE) τ

$$\tau_i(x_i) := \mathbb{E}\left[Y_i(T = 1) - Y_i(T = 0)|X = x_i\right] \tag{2.3}$$

can instead be estimated in practice. Whether this is possible, or in other words whether the CATE is identifiable, relies on several standard assumptions (Rosenbaum et al. 1983; Imbens et al. 2009):

- *Overlap (Positivity):* This states that every individual has a non-zero probability of receiving each treatment, i.e. $\Pr(T = T_i|X = x_i) > 0$ for all $T_i \in \{0, 1\}$ and all x_i.
- *Ignorability (Conditional Exchangeability):* Given the observed covariates x, the treatment assignment is independent of the potential outcomes, i.e. $Y(T) \perp T \,|\, x$.
- *Stable Unit Treatment Value Assumption (SUTVA):* This assumes that an individual's outcome is not affected by the treatment assignments of individuals (no interference) and that always the same potential outcome is observed if an individual receives a treatment T, i.e. $Y = Y(T)$ (consistency).

Additionally, especially in observational settings, an assumption is that there are no hidden confounders, meaning that all possible confounders that affect the treatment assignments are observed.

In RCTs, T is randomly assigned, which ensures ignorability by design, while observational data require additional modelling strategies to account for confounding.

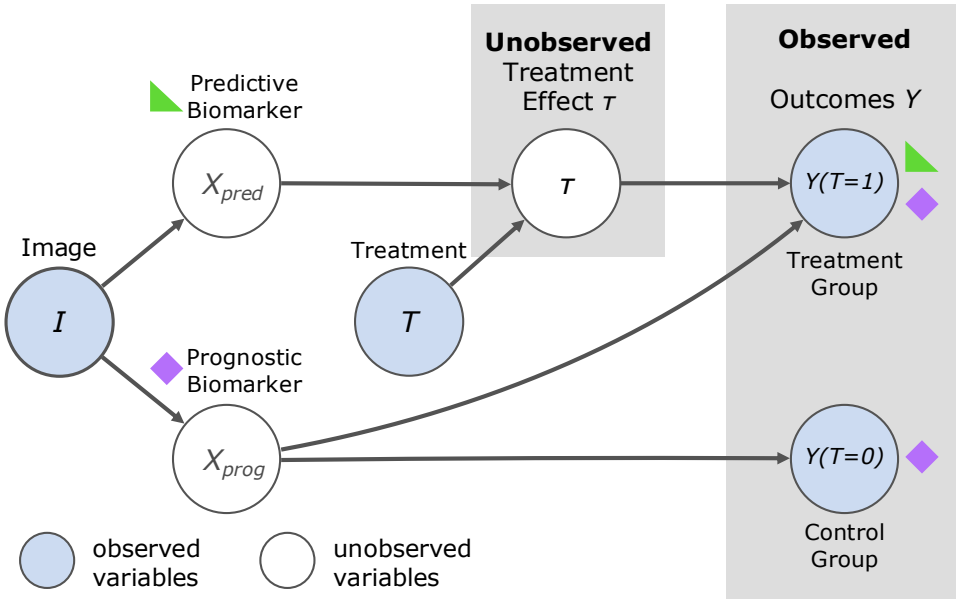


Figure 2.1: Diagram of the relationship between a prognostic biomarker x_{prog} and a predictive biomarker x_{pred} within a causal inference framework. The biomarkers affect the outcomes $Y(T)$ differently: the prognostic biomarker x_{prog} affects the outcome of treatment T, while the predictive biomarker x_{pred} is associated with T, contributing to a treatment effect τ. The figure also illustrates the fundamental challenge of causal inference: the individual treatment effect is not inferable directly in practice, since both potential outcomes $Y_i(T=0)$ and $Y_i(T=1)$ cannot be observed for a given individual at the same time. ©2025 IEEE. Adapted and reprinted with permission from Xiao et al. (2025).

In practice, personalized treatment recommendations can be derived from the estimated CATE by applying a decision threshold. For example, if $\tau(x) > 0$, the treatment $T = 1$ would be recommended, as it is expected to improve an individual's outcome compared to the alternative control treatment $T = 0$.

To make personalized recommendations grounded by pre-treatment data and identifying subgroups that may benefit from treatment, only the treatment effects that vary among individuals and covariates x, i.e. HTEs, are relevant. Building on the clinical introduction of biomarkers in Section 2.1.1, *predictive biomarkers* x_{pred} are formally defined as covariates that directly interact with the treatment and contribute to these HTEs (Ballman 2015). In contrast, *prognostic biomarkers* x_{prog} are defined as covariates that are associated with outcomes independent of which treatment is applied. The relationship between these two types of biomarkers, potential outcomes and treatment effects is illustrated in Figure 2.1, while their effect on observed survival outcomes is shown later in Figure 2.2 in the context of survival analysis.

2.2.2 Estimation of Treatment Effects

Even though individual counterfactual outcomes are unobserved, conditional treatment effects can still be estimated by learning predictive models for the potential outcomes under each treatment. This corresponds to modeling the response surfaces $f_0(x)$ and $f_1(x)$ (or equivalently a single function $f(x, T)$) that map covariates x to the expected outcomes $Y(T = 0)$ and $Y(T = 1)$, respectively, allowing interpolation across individuals with different treatment assignments.

Different strategies have been developed to estimate CATEs. Among the most widely used approaches are the so-called meta-learners (Künzel et al. 2019), which can be applied using standard supervised prediction models, such as regression or deep learning models. The simplest meta-learner is the S-learner, which uses a single model to predict outcomes for all treatment groups by taking the treatment indicator as an additional input feature. It can leverage the full dataset and is stable when sample sizes are limited, but often underestimates the treatment effect when the true effect is small. The T-learner consists of two separate models for each treatment group and is thereby more flexible, but might not be able to capture similarities between treatment groups. The X-learner was proposed by Künzel et al. (2019) to balance these limitations by combining both modeling strategies and directly estimating the CATE. Other extensions, such as Treatment-Agnostic Representation Network (TARNet) (Shalit et al. 2017) and its variants, learn shared feature representations with separate outcome heads to improve balance between treatment groups. Matching, weighting, or regularization techniques (e.g. integral probability metric (IPM)) can further reduce bias due to covariate imbalance, as explored in later deep-learning-based methods of treatment effect estimation.

2.2.3 Evaluation of Treatment Effect Estimators

Evaluating treatment effect estimators is challenging, since the true individual treatment effect cannot be observed directly. When both factual and counterfactual outcomes are available, for example in simulated or semi-synthetic experiments, performance of CATE estimation models can be quantified using the precision of estimating heterogeneous effects (PEHE) (Hill 2011):

$$\epsilon_{PEHE} = \frac{1}{n} \sum_i (\tau_i - \hat{\tau}_i)^2, \tag{2.4}$$

where n is the number of test samples, τ_i the true CATE and $\hat{\tau}_i$ estimated CATE of a model, and lower ϵ_{PEHE} indicate better estimated treatment effects.

For real clinical applications, where counterfactuals are unobservable, evaluation instead focuses on decision-based metrics. The policy value (Kallus et al. 2018; Jesson et al. 2021)

measures the expected outcome under a treatment recommendation policy $\pi(x) \in \{0, 1\}$ (such as the one described in Section 2.2.1):

$$\begin{aligned} V_{Pol} &= \mathbb{E}[Y_{\pi(x)}] \\ &= \Pr(\pi(x) = 1) \cdot \mathbb{E}[Y(1) \mid \pi(x) = 1] \\ &+ \Pr(\pi(x) = 0) \cdot \mathbb{E}[Y(0) \mid \pi(x) = 0] \end{aligned} \tag{2.5}$$

When computed using observed (factual) outcomes only, the policy value corresponds to the *Expected Response Under Proposed Treatments (ERUPT)* (Zhao et al. 2019; Hitsch et al. 2024), which measures the average observed outcome among individuals whose assigned treatment follows the ones recommended by the model.

The policy value in Equation 2.5 can be applied to outcomes of any scale, while the policy risk can be applied when the outcomes are bounded to values in $[0, 1]$, such as in binary classification. In that case, the policy risk provides a normalized performance measure in the range of 0 and 1 for treatment recommendations, representing the expected decrease in outcome when following the given policy:

$$\begin{aligned} R_{Pol} &= 1 - V_{Pol} \\ &= 1 - \Big(\Pr(\pi(x) = 1) \cdot \mathbb{E}[Y(1) \mid \pi(x) = 1] \\ &\quad + \Pr(\pi(x) = 0) \cdot \mathbb{E}[Y(0) \mid \pi(x) = 0] \Big). \end{aligned} \tag{2.6}$$

Another metric for assessing treatment recommendations is the decision accuracy, which quantifies the fraction of correctly predicted treatment recommendations. In its basic form, it requires access to the ground-truth treatment effect and the optimal policy, but extensions have extended this metric to observed data only (Efthimiou et al. 2023).

2.3 Survival Analysis

Survival analysis deals with time-to-event outcomes, where each individual is represented by a set $\{(x_i, Y_i, \delta_i)\}$, which consists of features describing the individual x_i, the observed time until the occurrence of an event Y (e.g. death or failure of a system), referred to as survival time throughout this thesis, and a censoring indicator $\delta \in \{0, 1\}$ denoting whether an event was observed ($\delta = 1$) or censored ($\delta = 0$). Unlike classification or regression, where the outcome is a single categorical or continuous value, survival analysis must account for time-dependent outcomes and censoring, as some patients may still be alive or lost to follow-up at the end of a study or withdraw from it. For this reason, the observed survival time $\tilde{Y} = \min(Y, C)$ is equal to the censoring C when $\delta = 0$.

2.3.1 Survival and Hazard Functions

To describe the survival probability of surviving beyond t, the survival function

$$S(t|x) = P(Y > t|x) \tag{2.7}$$

is defined. Its derivative, the hazard function $\lambda(t|x) = -\frac{d}{dt} \log S(t|x)$, corresponds to the instantaneous risk of an event taking place at time t, given it has not occurred before yet.

The median survival time is defined as the time point t at which $\widehat{S}(t) = 0.5$, and is used for the characterization and comparison of survival curves. Alternatively, the restricted mean survival time (RMST) (Irwin 1949) is defined as the area under the survival curve up to t',

$$\text{RMST}(\tau) = \int_0^{t'} S(t)\, dt. \tag{2.8}$$

Cox Proportional Hazards Model

To model the survival time depending on covariates x, the Cox proportional hazards model (Cox 1972) describes the hazard function using $\lambda(t|x) = \lambda_0(t) \exp(\beta^\top x)$, where it is assumed that only the baseline hazard function $\lambda_0(t)$ varies and is the same for all individuals, while the multiplicative contribution $\exp(\beta^\top x)$ remains constant over time (proportional hazards assumption).

Under this proportional hazards assumption, the ratio between the hazards of two individuals i and j, or hazard ratio (HR), is

$$\text{HR}_{ij} = \frac{\lambda(t|x_i)}{\lambda(t|x_j)} = \exp(\beta^\top (x_i - x_j)), \tag{2.9}$$

which is constant as well over time. This property implies that the survival curves for different subgroups do not cross.

The model parameters β can be estimated by minimizing the negative partial log-likelihood function:

$$\mathcal{L}(\beta) = -\sum_{i:\delta_i=1} \left[\beta^\top x_i - \log\left(\sum_{j:y_j \in \mathcal{R}_i} e^{\beta^\top x_j} \right) \right], \tag{2.10}$$

where the risk set $\mathcal{R}_i = \{j \mid Y_j \geq Y_i\}$ includes all patients who have survived up to time point Y_i.

This formulation is widely used in modern survival models, such as the neural-network-based survival model by Faraggi et al. (1995) or DeepSurv (Katzman et al. 2018) (see Section 3.2 and Section 4.2.2).

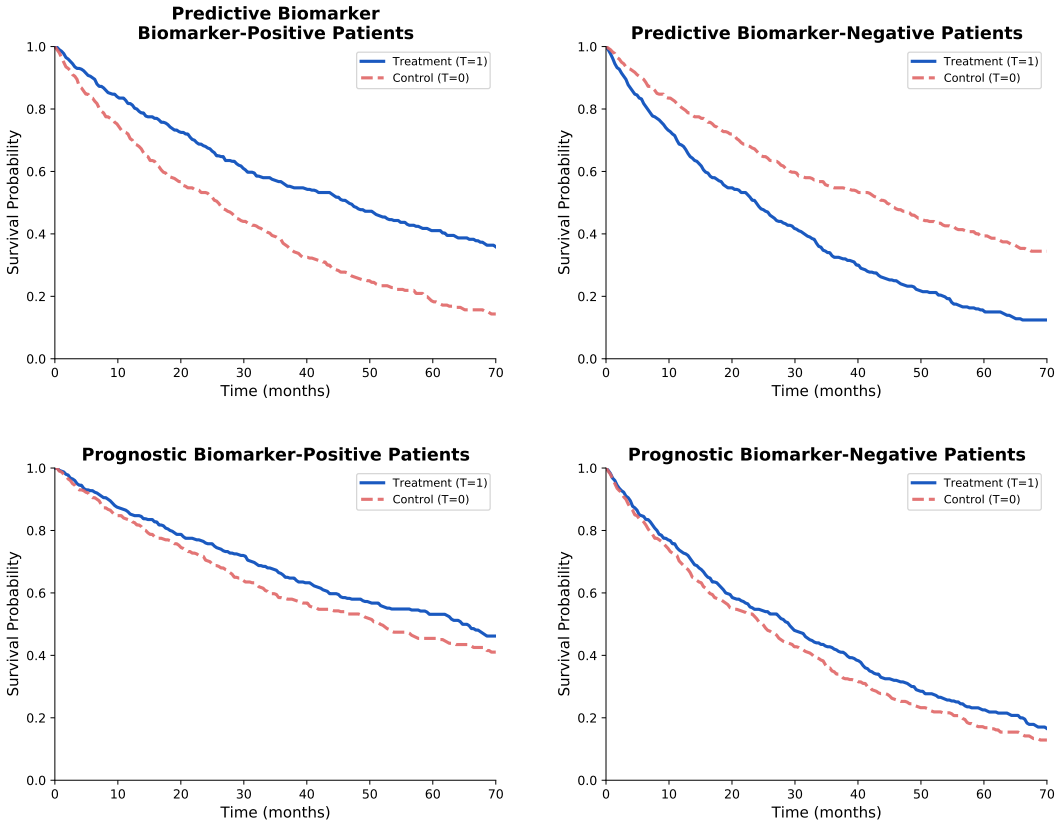


Figure 2.2: Illustration of how predictive and prognostic biomarkers affect survival probability, shown using simulated Kaplan-Meier curves (adapted from Ballman (2015)). When stratified by predictive biomarkers (top), treatment effects differ between biomarker subgroups, with only biomarker-positive patients benefiting from the treatment (e.g. here represented as the difference in median survival time). In contrast, when stratified by prognostic biomarkers (bottom), the biomarker influences overall survival prognosis but not the treatment effect, which remains constant across subgroups.

Kaplan-Meier Estimator

While the Cox model provides a semi-parametric approach to modeling effects of covariates on survival, the Kaplan-Meier estimator (Kaplan et al. 1958) offers a fully non-parametric estimate of the survival function. It estimates the probability of surviving beyond time t as

$$\hat{S}(t) = \prod_{t_i \leq t} \left(1 - \frac{d_i}{n_i}\right), \tag{2.11}$$

where d_i is the number of events and n_i the number of individuals that are at risk at time t_i.

Kaplan-Meier curves are often used to visualize differences in survival between subgroups, such as treatment versus control or biomarker-positive and biomarker-negative subgroups. An example for such a visualization is shown in Figure 2.2, which illustrates how predictive and prognostic biomarkers can lead to different effects in survival curves when comparing treatment and control groups.

To statistically compare two survival curves and assess differences, the log-rank test (Mantel et al. 1966) is commonly used to test the null hypothesis that there is no difference between the survival distributions of two groups. Alternatively, the HR obtained from a Cox model can quantify the relative risk between groups, where HR< 1 indicates a survival benefit from treatment.

2.3.2 Evaluation of Survival Analysis

C-index

The performance of survival models is commonly evaluated using Harrell's concordance index (C-index) (Harrell et al. 1982; Harrell Jr et al. 1996), which measures the proportion of correctly ordered pairs of predicted risk scores η_i, η_j and observed survival times Y_i, Y_j:

$$C = \frac{\sum_{i,j} \mathbb{1}[\eta_i < \eta_j] \, \mathbb{1}[Y_i > Y_j] \, \delta_j}{\sum_{i,j} \mathbb{1}[Y_i > Y_j] \, \delta_j}. \tag{2.12}$$

A value of $C = 0.5$ corresponds to random ranking and $C = 1$ to perfect concordance. Time-dependent variants (Antolini et al. 2005) extend the metric to handle censored data more accurately.

Handling censoring

Naively, censoring data can be handled by excluding censored observations during training and evaluation. However, this approach can lead to a bias, as patients with longer survival times are more likely to be censored (e.g. due to study end or dropout), so that the outcome distribution skews towards shorter survival times.

To account for the impact of censoring, inverse probability of censoring weighting (IPCW) was introduced by (Robins et al. 1992; Robins et al. 2000). This approach weights the contribution of uncensored observations inversely by their probability of remaining uncensored, which can, for example, be estimated non-parametrically using a Kaplan-Meier estimator or through covariate-dependent models, such as Cox regression, to model the censoring process (Robins et al. 2000; Satten et al. 2001). Uncensored observations with a lower probability remaining uncensored thus have a higher contribution, ensuring that they better represent censored cases. IPCW can be applied both during training of survival models (Vock et al. 2016) and evaluation, including the computation of metrics

such as the C-index (Uno et al. 2011) or Brier score (Gerds et al. 2006) or the restricted mean survival time (Tian et al. 2014).

Related Work

3

The areas investigated by this thesis cover multiple research topics, including causal inference (specifically treatment effect estimation), predictive biomarker discovery, medical imaging, and survival analysis. Table 3.1 summarizes representative works in these areas, structured to highlight the key gaps addressed by this thesis.

This chapter is structured by two main parts of this thesis, with Section 3.1 providing an overview of works related to heterogeneous treatment effect estimation models for predictive imaging biomarker discovery and Section 3.2 providing an overview of works related to image-based heterogeneous treatment effect estimation models in clinical imaging studies with a focus on those concerning survival outcomes. While this structure is used for clarity, many of the topics discussed are inherently cross-cutting and relevant to both main areas of investigation of the thesis.

3.1 Evaluating Heterogeneous Treatment Effect Estimation Models for Predictive Imaging Biomarker Discovery

3.1.1 Heterogeneous Treatment Effect Estimation

The estimation of heterogeneous treatment effects aims to quantify how treatment efficacy varies across individuals or subgroups, supporting personalized treatment recommen-

Table 3.1: Comparison of works related to this thesis, including treatment effect estimation, survival prediction, and predictive biomarker discovery from image data. Methods that only partially address a task that is not the main focus are marked with (✓).

a Related to heterogeneous treatment effect estimation for predictive imaging biomarker discovery.

Method	Input		Task		
	Image data	Multi-modal	Treatment effect estimation	Survival prediction	Predictive biomarker discovery
Predictive biomarker discovery					
Sechidis et al. (2018)	✗	✗	✗	✗	✓
Hermansson et al. (2021)	✗	✗	✓	✗	✓
Bahamyirou et al. (2022)	✗	✗	✓	✗	✓
Crabbé et al. (2022)	✗	✗	✓	✗	✓
W. Zhu et al. (2023)	✗	✗	✗	✗	✓
Boileau et al. (2023)	✗	✗	✓	✗	✓
Svensson et al. (2025)	✗	✗	✓	✗	✓
Vollenweider et al. (2025)	✗	✗	✓	✗	✓
Verhaeghe et al. (2025)	✗	✗	✓	✗	✓
Bo et al. (2025)	✗	✗	✓	✗	✓
Z. Liu et al. (2025)	✗	✗	✓	✗	✓
Arango-Argoty et al. (2025)	✗	✓	✗	✓	✓
Image-based treatment effect estimation					
Medical imaging data					
Durso-Finley et al. (2022), Durso-Finley et al. (2023)	✓	✓	✓	✗	✗
Ma et al. (2023), Ma et al. (2024)	✓	✓	✓	✗	✗
Herzog et al. (2025)	✓	✓	✓	✗	✗
Jiang et al. (2023)	✓	✓	✓	✗	✗
Non-medical imaging data					
Takeuchi et al. (2021)	✓	✗	✓	✗	✗
Deshpande et al. (2022)	✓	✓	✓	✗	✗
Jerzak et al. (2023)	✓	✓	✓	✗	✗
Cadei et al. (2024)	✓	✗	✓	✗	✗
F. W. Zhu et al. (2025)	✓	✗	✓	✗	(✓)
Xiao et al. (2025)	✓	✗	✓	✗	✓

b Related to image-based heterogeneous treatment effect estimation in clinical imaging studies.

Method	Input		Task		
	Image data	Multi-modal	Treatment effect estimation	Survival prediction	Predictive biomarker discovery
Deep-learning-based survival treatment effect estimation					
Katzman et al. (2018)	✗	✗	(✓)	✓	✗
Curth, Lee, et al. (2021)	✗	✗	✓	✓	✗
Schrod et al. (2022)	✗	✗	✓	✓	✗
Chapfuwa et al. (2021)	✗	✗	✓	✓	✗
Frauen et al. (2025)	✗	✗	✓	✓	✗
Kim (2025)	✗	✗	✓	✓	✗
Survival prediction using medical imaging data					
Haarburger et al. (2019)	✓	✗	✗	✓	✗
Li et al. (2022)	✓	✗	✗	✓	✗
Vale-Silva et al. (2021)	✓	✓	✗	✓	✗
Wolf et al. (2022)	✓	✓	✗	✓	✗
Meng et al. (2022)	✓	✗	✗	✓	✗
Hao et al. (2022)	✓	✓	✗	✓	✗
Huo et al. (2025)	✓	✓	✗	✓	✗
This thesis	✓	✓	✓	✓	✓

dations. The PATH (Predictive Approaches to Treatment effect Heterogeneity) Statement (Kent et al. 2020) provided a unifying framework for such analyses in clinical trials, distinguishing between risk-modeling and effect-modeling strategies and emphasizing their role in patient-centered treatment recommendations.

Different deep learning methods have been developed for heterogeneous treatment effect estimation from tabular input data, addressing, for example, observational data, data coming from imbalanced classes, or multiple types of treatments (Alaa et al. 2017; Shalit et al. 2017; Amsterdam et al. 2019; Künzel et al. 2019; Shi et al. 2019; Jin et al. 2021).

In contrast to treatment effect estimation models on tabular input data, image-based models remain less widely explored, but are an emerging field (see Table 3.1a). In medical imaging, early approaches adapted multi-headed deep neural networks based on the TARNet architecture by Shalit et al. (2017) to estimate treatment effects from MRI scans in multiple sclerosis, incorporating lesion segmentation masks and clinical tabular covariates as additional inputs (Durso-Finley et al. 2022; Durso-Finley et al. 2023).

For head CT scans, more recent work has proposed a multimodal architecture incorporating clinical tabular data as well, for example, using representation learning and distribution balancing in aneurysmal subarachnoid hemorrhages (Ma et al. 2023; Ma et al. 2024), or for outcome prediction using a pre-trained SwinUNETR image encoder (Herzog et al. 2025). Jiang et al. (2023) further explored deep ensemble models for treatment effect estimation from chest X-ray images.

Outside medical imaging, image-based treatment effect estimation has been applied to diverse domains, including satellite imagery (Jerzak et al. 2023; F. W. Zhu et al. 2025), spatial crowd movements (Takeuchi et al. 2021), or video data (Cadei et al. 2024).

For example, Cadei et al. (2024) conducted benchmark experiments on causal inference tasks with image and video data, introducing the ISTAnt benchmark to compare different neural architectures.

Jerzak et al. (2023) proposed a probabilistic model to cluster images based on similar estimated treatment effects, providing interpretable subgroups for anti-poverty policies.

Although not explicitly framed as predictive biomarker discovery, F. W. Zhu et al. (2025) applied image-based CATE estimation to a similar remote-sensing dataset and investigated the features driving heterogeneous treatment effects. Although these features could be interpreted as predictive biomarkers, they did not explicitly perform biomarker discovery, focusing instead on representation-level analyses of treatment-effect heterogeneity.

All of these works address categorical or continuous outcomes (for example, predicting a survival probability at a fixed time point rather than a time-to-event outcome (Ma et al. 2024)), and none explicitly aim to identify predictive biomarkers.

Evaluating HTE estimation models (i.e. CATE estimation models) by their ability to learn predictive biomarkers has been stressed by works of Curth, Svensson, et al. (2021) and Crabbé et al. (2022), as this is often the more relevant downstream task for real-world applications such as personalized medicine. They also emphasized the importance of utilizing semi-synthetic data to benchmark these methods, which motivated the use of simulated imaging biomarkers in this thesis.

3.1.2 Predictive Biomarker Discovery Using Causal Inference

Building on the concept of predictive imaging biomarker discovery introduced in Section 2.1.1, which has relied on radiomics features and correlations with molecular biomarkers in medical imaging traditionally, recent work has explored this task using causal inference methods.

Many approaches have been proposed to evaluate whether predictive effects can be identified in a data-driven way as listed in Table 3.1a, which typically rank features by their contribution to treatment effect heterogeneity Hermansson et al. (2021), Bahamyirou et al. (2022), Crabbé et al. (2022), Boileau et al. (2023), Svensson et al. (2025), and Verhaeghe et al. (2025). In these studies, predictive biomarker discovery is often formulated as a downstream task of HTE or CATE estimation, using explainable artificial intelligence (XAI) techniques or other variable-importance measures. Other related works, such as Sechidis et al. (2018) and W. Zhu et al. (2023), approach predictive biomarker discovery from a more general feature-importance perspective without explicitly modeling treatment effects. This ranking approach however limits their applicability to high-dimensional inputs such as imaging data.

A key challenge for predictive biomarker discovery is distinguishing predictive from prognostic covariates, since both can influence outcome prediction, but only the former contribute to heterogeneous treatment effects. Several studies have shown that CATE estimators can mistakenly identify prognostic as predictive biomarkers (Sechidis et al. 2018; Hermansson et al. 2021; Crabbé et al. 2022), which can lead to treatment recommendations that are potentially ineffective or even harmful.

It is therefore essential to ensure that these methods can distinguish the two types of biomarkers, which has motivated the development of methods that explicitly separate them (Sechidis et al. 2018; Hermansson et al. 2021; Arango-Argoty et al. 2025; Verhaeghe et al. 2025).

Several deep-learning-based predictive biomarker discovery approaches have mentioned their potential applicability to imaging data as well, although none of them have explicitly demonstrated it. The Predictive Biomarker Modeling Framework (PBMF) proposed by Arango-Argoty et al. (2025) employs a contrastive learning objective based on survival differences within treatment groups to directly predict biomarker status and confidence

scores. So far, it has only been shown on clinical and omics-type features, and its scalability to mini-batch training remains to be evaluated. Similarly, Z. Liu et al. (2025) introduced DeepRAB, a deep-learning-based framework for subgroup identification and predictive biomarker discovery that incorporates a biomarker selection layer to output feature-importance scores.

Closely related to heterogeneous treatment effect is the concept of digital twins (also known as virtual twins), which aim to build individualized computational models that simulate a patient, or systems in general, to predict patient-specific treatment outcomes or disease progression under alternative conditions such as different therapies. Such approaches have also been explored for subgroup identification and predictive biomarker discovery (Foster et al. 2011; Hermansson et al. 2021; Susilo et al. 2023), although many approaches rely on mechanistic modeling rather than purely data-driven learning.

Despite these advances, to date, there is currently no well-validated method for predictive imaging biomarker discovery that directly leverages raw image data without a separate handcrafted feature extraction step.

3.2 Image-Based Heterogeneous Treatment Effect Estimation in Clinical Imaging Studies

3.2.1 Treatment Effect Estimation Methods for Survival Outcomes

The goal of personalized therapy is often to extend the survival time of patients, which is why overall survival is one of the most common clinical endpoints and outcomes of interest in many clinical studies, particularly in oncology (Delgado et al. 2021). Therefore, an important requirement for treatment recommendation methods based on CATE estimation is to model time-to-event outcomes, and to account for its specific challenges such as censoring.

Classical survival models such as the Cox proportional hazards model (Cox 1972) and its generalization to deep learning, DeepSurv (Katzman et al. 2018), or DeepHit (Lee et al. 2018), can in principle be integrated into standard treatment effect frameworks, such as S- or T-learners, to model treatment-specific survival functions. DeepSurv reformulates the Cox model using the negative partial log-likelihood loss and has been applied to treatment recommendation scenarios. While the original work does not explicitly estimate heterogeneous treatment effects or perform causal evaluation, it has been used as baselines for CATE estimation for survival outcomes, but typically without explicitly accounting for confounding or covariate imbalance.

Recent research has proposed deep learning architectures that explicitly extend CATE estimation to censored survival data, which are summarized in Table 3.1b.

Curth, Lee, et al. (2021) introduced SurvITE, which estimates heterogeneous treatment effects based on the restricted mean survival time by learning discrete-time treatment-specific conditional hazard functions. The model uses separate outcome heads for each time interval and applies covariate balancing to mitigate distributional shifts between treatment arms.

Schrod et al. (2022) proposed Balanced Individual Treatment Effect for Survival Data (BITES), which incorporates an IPM loss term into a Cox-based survival model to balance latent representations among treatment arms and reduce confounding bias during training.

Similarly, Chapfuwa et al. (2021) developed a generative framework for individualized treatment effect estimation on survival outcomes, using planar flow-based latent transformations to jointly account for selection bias and censoring bias.

For the evaluation of survival-outcome treatment effect estimation, Efthimiou et al. (2025) proposed performance metrics tailored to treatment recommendations, extending existing measures such as the C-for-benefit and decision accuracy to time-to-event outcomes.

Beyond methodological developments in deep learning, related work has also investigated predictive biomarker discovery and using treatment effect estimation models for survival data. For example, Ternes et al. (2017) compared a range of approaches based on the Cox model for identifying predictive biomarkers from high-dimensional tabular data in RCTs.

Although these works have been successfully applied to make treatment recommendations, all have only been applied to structured tabular inputs and have not been extended to imaging data. A straightforward extension of survival prediction models, such as DeepSurv or BITES, could in theory provide treatment-specific survival modeling, but such adaptations have not yet been validated or benchmarked on clinical imaging datasets.

3.2.2 Survival Prediction from Imaging Data

Many survival prediction methods for medical imaging data have relied on traditional approaches such as the Cox proportional hazards model (Cox 1972) or classical machine learning methods such as random survival forests (Ishwaran et al. 2008) and gradient-boosted survival trees (Chen et al. 2016). These methods typically operate on handcrafted or radiomics-derived image features and thus depend on a separate feature extraction step. While these methods are straightforward and effective on limited dataset sizes, the reliance on handcrafted feature-based pipelines is not well suited for discovering predictive imaging biomarkers and cannot easily integrate multimodal data.

Deep learning has enabled end-to-end survival prediction directly from images (also see Table 3.1b) and has the advantage of learning subtle patterns in a data-driven way. Several models extend the DeepSurv framework by Katzman et al. (2018) to convolutional neural

network (CNN) architectures, such as DeepConvSurv (X. Zhu et al. 2016), DeepMTS (Meng et al. 2022) or the work by Haarburger et al. (2019). These models replace the handcrafted feature step with learned representations and optimize the Cox partial log-likelihood loss for time-to-event prediction.

Haarburger et al. (2019) further used a hybrid approach combining CNN-derived with radiomics features and compared direct survival modeling against binary median-survival classification and highlighted the practical challenges of mini-batch training in survival settings due to censored samples. Similarly, SurvCNN (Hao et al. 2022) combines CT imaging with radiomics features, while DeepMTS (Meng et al. 2022) jointly learns tumor segmentation and survival risk prediction, showing that such multitask and multimodal designs can outperform purely radiomics-based baselines.

Several studies have explored architectures specifically for brain cancer survival prediction. For example, Li et al. (2022) proposed the DeepRisk model, which incorporates spatial and channel attention mechanisms into a residual neural network (ResNet) backbone to identify high-risk regions across whole-brain MRI, illustrating the potential of fully convolutional attention-based networks to identify prognostic imaging patterns.

Multimodal integration has been shown to enhance survival prediction further. The DAFT framework (Wolf et al. 2022) and MultiSurv (Vale-Silva et al. 2021) integrate imaging and clinical covariates through learned feature fusion, with the latter employing a discrete-time survival formulation to model time-to-event outcomes.

Recent studies have also leveraged pre-training and transfer learning to mitigate data scarcity in medical imaging, for example by pre-training time-aware models on longitudinal imaging data for improved survival prediction (Huo et al. 2025). Similarly, Dancette et al. (2025) introduced the CURIA multimodal foundation model for radiology, showing that large-scale pre-training on observational imaging data can improve downstream survival prediction performance.

Finally, some studies compared discrete binary risk classification with continuous time-to-event modeling. For instance, Haarburger et al. (2019) used median-survival thresholds as binary endpoints, while Zhou et al. (2023) found that quantized survival categories improved patient stratification robustness using multimodal whole slide imaging, clinical, and genomic data. However, other analyses have shown that directly modeling time-to-event outcomes yields more accurate and interpretable risk estimates on multimodal (chest X-ray and demographic) data (M. Liu et al. 2024).

While these studies demonstrate substantial progress in prognostic survival prediction from medical images, they do not address treatment effect estimation or predictive imaging biomarker discovery, which remain open challenges motivating this thesis.

3.2.3 Transfer and Self-Supervised Learning for Treatment Effect Estimation

Transfer and self-supervised learning have emerged as important strategies to address data scarcity in deep-learning-based treatment effect estimation. Several studies demonstrated that pre-training can enhance generalization and stability of CATE models, even though most prior work focused on tabular data. For instance, Aloui et al. (2023) investigated transfer learning for tabular treatment-effect estimation, and R. Liu et al. (2024) introduced a large-scale foundation model (CURE) trained on electronic health records to predict individualized treatment responses. More recently, Zhang et al. (2024) proposed a causally aware foundation model that unifies observational and interventional representations, indicating a step toward integrating causal reasoning into pre-trained architectures. In the imaging domain, Herzog et al. (2025) applied a pre-trained SwinUNETR encoder for outcome prediction on CT data, illustrating the potential of leveraging pre-trained backbones for image-based CATE estimation.

Despite these advances, no prior work has systematically applied pre-trained image encoders to treatment effect estimation from brain MRI data, which is investigated in this thesis.

Conclusion

This chapter has provided a detailed review of the state-of-the-art in heterogeneous treatment effect estimation, predictive biomarker discovery, and image-based survival analysis. As summarized in Table 3.1, while prior work has advanced treatment effect estimation, survival prediction, and predictive biomarker discovery individually, no approach jointly addresses all three tasks for making treatment recommendations from imaging data and thereby advancing personalized medicine. This thesis therefore proposes methods that integrate these tasks into an approach for discovering and evaluating predictive imaging biomarker discovery (Section 4.1) and a unified model for image-based CATE estimation for survival outcomes (Section 4.2).

Materials and Methods

4

This chapter presents the methods that have been developed in this thesis to address the research questions outlined in Section 1.2, as well as the datasets, model architectures and training strategies used throughout the experiments.

In Section 4.1 the proposed approach for discovering predictive imaging biomarkers (RQ1.1) is described, along with a novel evaluation protocol to assess the performance of heterogeneous treatment effect estimation models in predictive imaging biomarker discovery (RQ1.2). Section 4.2 introduces methods for image-based heterogeneous treatment effect estimation to deal with survival outcomes (RQ2.1) and for integrating tabular inputs and pre-trained image encoders (RQ2.2) while supporting their application to clinical imaging study data (RQ2.2).

4.1 Evaluating Heterogeneous Treatment Effect Estimation Models for Predictive Imaging Biomarker Discovery

This section establishes the task of radiomics-free predictive imaging biomarker discovery by explaining how it is connected to heterogeneous treatment effect estimation (specifically CATE estimation), which is done in Section 4.1.1, and presents the deep-learning-based model used in this thesis for this purpose in Section 4.1.2. One of the main contributions, the proposed evaluation protocol for predictive imaging biomarker discovery including the quantitative statistical evaluation and qualitative evaluation using XAI

methods, is presented in Section 4.1.3. The section closes with the experimental setup in Section 4.1.3, where a strategy for simulating outcomes using pre-defined predictive and prognostic imaging biomarkers is proposed, to enable the benchmarking of predictive imaging biomarker discovery. *The following subsections in this section were adapted from (Xiao et al. 2025), which was originally written by the author of this thesis.*

4.1.1 Treatment Heterogeneity and Predictive Biomarkers

After establishing the fundamentals of causal inference and treatment effect estimation in Section 2.2, this subsection describes the relationship between estimating heterogeneous treatment effects and identifying predictive imaging biomarker, which is a central topic to RQ1.1.

In heterogeneous treatment effect estimation, the goal is to estimate the CATE from observable pre-treatment covariates $x \in X$. When the CATE of the observed outcome Y directly depends on such a covariate, it is considered to be a biomarker, and when this biomarker can be extracted from images I and measure image features, it is referred to as an "imaging biomarker". In this thesis, the term "biomarker" is used to describe such covariates with an established relation to the CATE of the observed outcome Y in general, independent of whether used in the biomedical context, where the term was originally established, or not. For making treatment decisions or selecting subgroups of patients that may benefit from treatment, only heterogeneous treatment effects are relevant, which vary among individuals and covariates x within the whole observed population. Therefore, identifying predictive biomarkers, which in this context are covariates that directly contribute towards the heterogeneous treatment effect, i.e. CATE and interact with the treatment, is highly relevant for the aforementioned tasks.

A common assumption in literature, e.g. by Sechidis et al. (2018), Künzel et al. (2019), Curth, Svensson, et al. (2021), and Hermansson et al. (2021), is that treatment-independent prognostic effects f_{prog} and treatment-related predictive effects f_{pred} are additive. Under this assumption, the expected outcome can be written as

$$\mathbb{E}\left[Y^T(x)\right] = f_{prog}(x) + f_{pred}(x)T, \tag{4.1}$$

where $f_{pred}(x)$ only depends on predictive biomarkers x_{pred}, but not prognostic ones (x_{prog}). In this formulation, the CATE as defined in Equation 2.3 directly corresponds to $f_{pred}(x)$ as a constant average treatment effect is not modeled explicitly in more general formulations one could instead write the contribution of the treatment effect as $f_{treat}(x) = \tau_0 + f_{pred}(x)$ with τ_0 representing a constant average treatment effect.

Thus, prognostic and predictive effects can automatically, in principle, be separated by treatment effect estimation, which in turn identifies predictive biomarkers x_{pred}.

It is important to note that since a biomarker can be both prognostic and predictive at the same time, the same covariates can contribute to both $f_{prog}(x)$ and $f_{pred}(x)$.

4.1.2 Deep Learning Model for Treatment Effect Estimation from Imaging Data

To implement the discovery of predictive imaging biomarkers using treatment effect estimation, deep neural network-based CATE estimation models are leveraged and adapted with a convolutional image encoder to process image inputs. The experiments presented in this thesis used a modified version of the TARNet model (Shalit et al. 2017), which was originally designed for tabular input data.

The network, illustrated in Figure 4.1, is adapted similarly to (Durso-Finley et al. 2022), utilizing shared convolutional layers (in this case, ResNet blocks (He et al. 2016)) as an image encoder. Its purpose is to learn the similarities between the control and treatment arms that contribute to prognostic effects (Curth and Schaar 2021). These shared layers are followed by two treatment-specific heads (in this case, fully connected layers) for predicting the potential outcomes $Y(T)$.

In the training phase of the image-based CATE estimation model, as shown in Figure 4.1a, the loss is computed only for the output head corresponding to the treatment that was actually received and is therefore specific to the treatment arm. The weights of both network heads are updated jointly in every training step based on the total loss, which is obtained by summing up the loss of the control group head output and the treatment group head output within a mini-batch.

The inference step is depicted in Figure 4.1b, where the CATE is estimated using

$$\hat{\tau}_i = \hat{Y}_i(T = 1) - \hat{Y}_i(T = 0), \tag{4.2}$$

i.e. by subtracting the model's predicted control group outcome from the predicted treatment group outcome.

Baseline. To assess whether the CATE estimation model truly learns treatment-dependent effects and can successfully facilitate the discovery of predictive imaging biomarkers, it is compared against a single-headed baseline with the same shared image encoder. It is expected that such a model likely predicts the average outcome across both treatment arms from predictive and prognostic biomarkers, without explicitly separating treatment-specific contributions from predictive imaging biomarkers and treatment-independent contributions from prognostic biomarkers.

4.1.3 Proposed Evaluation Protocol for Predictive Imaging Biomarker Discovery

The evaluation protocol detailed in the following subsections was designed to directly address RQ1.2 by describing how the performance and reliability of image-based CATE estimation models can be assessed both quantitatively and qualitatively.

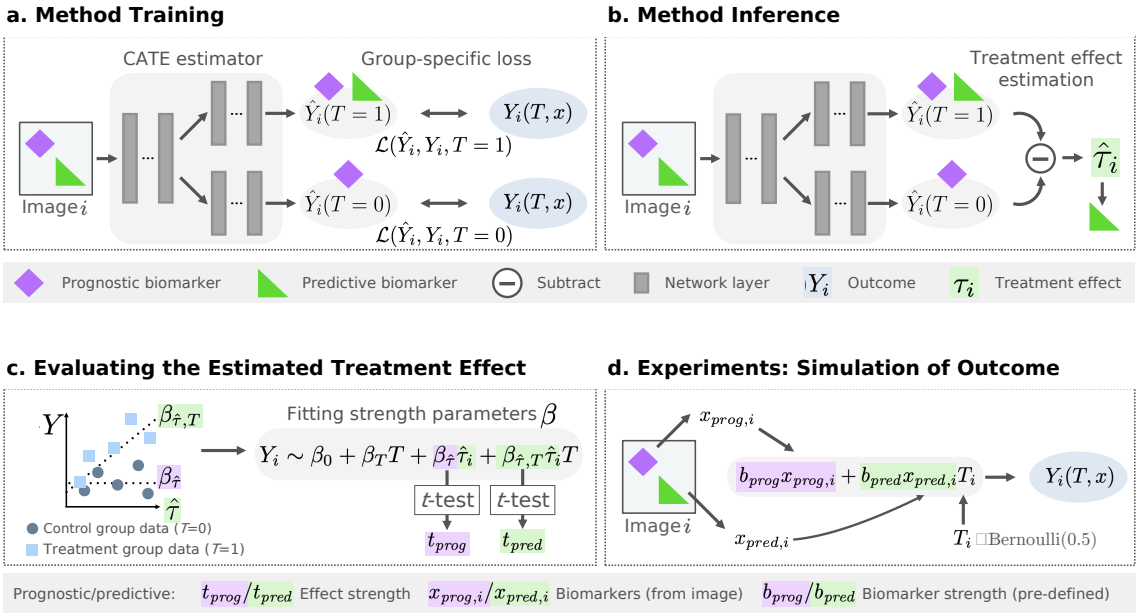


Figure 4.1: Overview of the identification of predictive biomarkers approach from pre-treatment images. During the training phase (**a**) a two-headed TARNet-like architecture (Shalit et al. 2017) is used to predict the potential outcomes $\hat{Y}_i(T=0)$ and $\hat{Y}_i(T=1)$, optimized using a group-specific loss applied to only predictions with available factual outcomes. In the inference step (**b**), these outcomes are used to estimate the treatment effect CATE $\hat{\tau}$ from images. In the evaluation step (**c**), the estimated $\hat{\tau}$ is treated as a predictive biomarker candidate and used to quantify the predictive strength with a regression analysis. In the simulation experiments (**d**), the synthetic outcomes Y_i are generated using image features from ground truth annotations, which are designated to be prognostic or predictive biomarkers, and random treatment assignments T. ©2025 IEEE. Adapted and reprinted with permission from Xiao et al. (2025).

Statistical Evaluation of the Predictive Biomarker Strength

The goal of the statistical evaluation, as illustrated in Figure 4.1c, is to assess whether the trained CATE estimation models have recovered a heterogeneous treatment effect, i.e. predictive effect, by testing whether the estimated CATE $\hat{\tau}$ is indeed predictive and can be considered a predictive biomarker candidate.

To achieve this, a test for biomarker-by-treatment interactions is performed, which is also performed in clinical biomarker validation studies (Polley et al. 2013; Ballman 2015).

Here, a linear relationship between biomarkers and outcome is assumed (as in Equation 4.5) when a linear regression of the outcomes Y is performed with coefficients β_i

according to

$$\beta_0 + \beta_T T + \beta_{\hat{\tau}} \hat{\tau} + \beta_{\hat{\tau},T} \hat{\tau} T \sim Y, \tag{4.3}$$

where the term $\beta_{\hat{\tau},T} \hat{\tau}$ represents the biomarker-by-treatment interaction term.

To assess this term, the null hypothesis that the corresponding interaction coefficient is $\beta_{\hat{\tau},T} = 0$ is tested using a Student's t-test. The resulting t-value test statistic $t_{\beta_{\hat{\tau},T}}$ is proportional to the estimated parameter $\hat{\beta}_{\hat{\tau},T}$ and indicates whether the estimated biomarker has a statistically significant predictive effect. This test is repeated for all other coefficients β_i to assess the contributions of the other terms.

To quantify the predictive strength of the estimated CATE $\hat{\tau}$ compared to its prognostic strength, the t-value ratio of the corresponding test statistics is computed:

$$t_{\beta_{\hat{\tau},T}} / t_{\beta_{\hat{\tau}}} =: t_{pred} / t_{prog}. \tag{4.4}$$

Finally, the experimental lower bound (indicating a purely prognostic biomarker) and upper bound (indicating a purely predictive biomarker) of the relative predictive strength is determined. This is done by conducting the same evaluation, replacing $\hat{\tau}$ in Equation 4.3 with either the purely prognostic or a purely predictive ground truth biomarker x_{prog}, x_{pred}.

While the evaluation described here uses linear regression under the assumption of a linear relationship between biomarkers and outcomes, the analysis could also, in principle, be repeated using a Cox proportional hazards regression model and survival outcomes, as applied in the experiments described in Section 5.2.2.

Interpretation of Biomarkers using Feature Attribution Methods

The second part of the predictive imaging biomarker discovery evaluation protocol focuses on interpreting which parts of the images likely contributed to the prognostic or predictive effects learned by the model. This is achieved by investigating which input image features the trained model attends when predicting the CATE $\hat{\tau}$, which in turn likely correspond to candidate predictive imaging biomarkers. For assessing the model performance when using semi-synthetic data, this evaluation additionally allows comparing the identified features to the ground truth predictive imaging biomarkers.

While predictive biomarkers identified by CATE estimation models for tabular input data can easily be assessed quantitatively using feature attribution methods, as for example applied by Crabbé et al. (2022) and Verhaeghe et al. (2025), such an assessment is generally not straightforward for image input data. To enable the discovery and clinical adoption of novel biomarkers, the identified features must be interpretable. For this reason, the assessment relies on a qualitative analysis using visual explanations via attribution maps (Simonyan et al. 2014).

To this end, the XAI methods Expected Gradients (Expected Gradients) (Erion et al. 2021) and Guided Gradient-weighted Class Activation Mapping (GGrad-CAM) (Springenberg et al. 2015; Selvaraju et al. 2017) are employed to generate attribution maps from the trained model and input images.

Using the attribution maps of the control group head prediction $\hat{Y}(T = 0)$, it is then possible to assess the contribution of individual pixels to the prognostic effect on the one hand. On the other hand, the attribution map of the estimated CATE $\hat{Y}(T = 1) - \hat{Y}(T = 0)$ enables the analysis of individual pixels' contribution to the predictive effect.

Simulation of Imaging Biomarkers and Outcomes for Validation

To validate the proposed approach using CATE estimators for predictive imaging biomarker discovery, experiments were performed with the aim of assessing and interpreting the predictive imaging biomarker that the CATE estimation model was able to identify, and also to investigate how well the model performs at the predictive imaging biomarker discovery task while disentangling them from prognostic imaging biomarkers.

The experiments were conducted on image datasets under controlled conditions, specifically with synthetic outcomes, which were necessary as ground truth counterfactual outcomes are not available in real datasets.

For this reason, synthetic datasets with simulated treatment effects and ground truth counterfactual outcomes of varying predictive and prognostic biomarker strengths and were generated to experimentally verify the model, as illustrated in Figure 4.1d.

In this study, an approach to simulate outcomes from image data is proposed. In contrast to tabular-data simulations, where outcomes are directly sampled from predefined covariates, the present approach simulates outcomes based on imaging biomarkers by linking specific image features to biomarker values $x_{prog,pred}$. These features are selected from available image information and can represent attributes from available metadata, class labels, or radiomics features.

Examples of such features used in the experiments of the first part of the thesis are illustrated in Figure 4.2. In the experiments, the biomarkers were defined to be either purely prognostic or predictive and could take binary or continuous values depending on the type of dataset.

The simulated outcomes Y were then generated according to a simple linear function from only two biomarkers:

$$Y(T, x) = \underbrace{b_{prog} x_{prog}}_{\text{Prognostic Effect}} + \underbrace{b_{pred}\, x_{pred} T}_{\text{Predictive Effect}}. \tag{4.5}$$

For simplicity, it was assumed here that no offset b_0 and constant treatment effect b_T are present following the setup of Krzykalla et al. (2020).

This experimental setup using simulated outcomes allows direct control over the relative magnitude of prognostic or predictive effects by adjusting the parameters b_{prog} and b_{pred}. Thus, the ratio of the parameters b_{pred}/b_{prog}, which is also referred to as the biomarker parameter strength ratio in this thesis, can be interpreted as a measure of the signal-to-noise ratio of the predictive effect in the input data. The contributions of prognostic imaging biomarkers can hereby be considered to represent noise, as the models need to visually disentangle them from predictive contributions, and as there is a risk of assuming a purely prognostic biomarker to be predictive.

In all simulations, the treatment assignment $T \in \{0, 1\}$ was randomized with equal probabilities $p(T) = 0.5$ to simulate RCT data.

4.1.4 Experimental Setup

Datasets and Imaging Biomarker Features for Outcome Simulation

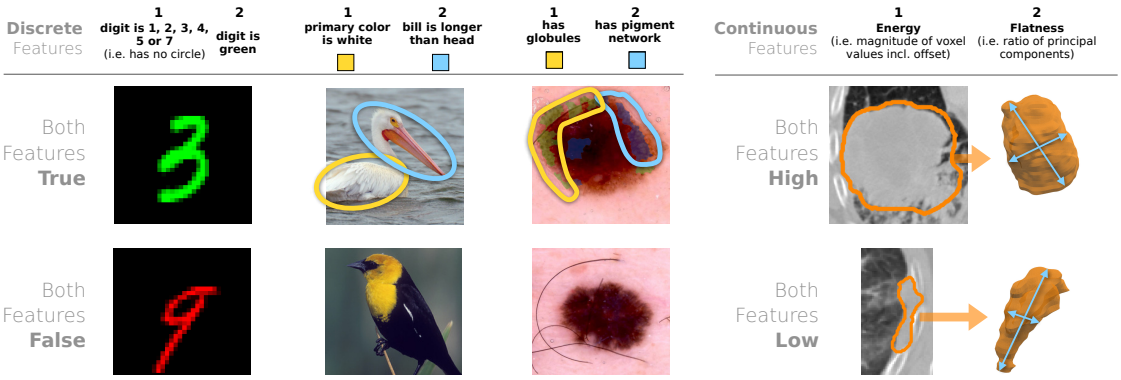


Figure 4.2: Image features from the four datasets that were used to simulate the outcomes, where either feature 1 or 2 is designated as predictive or prognostic biomarkers. For the ISIC 2018 dataset, the skin lesion features are shown with ground-truth masks. Globules (highlighted with a light green mask) appear as darker dots, whereas pigment networks (shown with a dark blue mask) exhibit dark grid-like patterns of streaks interspersed with lighter regions or "holes". For the NSCLC-Radiomics lung CT images, features are extracted from the segmented tumor regions outlined for a 2D slice (left) and in the corresponding reconstructed 3D volumes (right). Images shown in the top row depict images where both biomarkers features are either present (CMNIST, CUB-200-2011, ISIC 2018) or have a high value (NSCLC-Radiomics), whereas the images in the bottom row show examples where the features are absent or low. ©2025 IEEE. Adapted and reprinted with permission from Xiao et al. (2025).

Parts of the following paragraphs in this subsection were taken from (Xiao et al. 2025) with minor adaptation, which was originally written by the author of this thesis.

For the experiments in this part of the thesis, four diverse publicly available image datasets were used, which are shown in Figure 4.2, including the features used as biomarkers. The datasets are: a colored version of the Modified National Institute of Standards and Technology database (MNIST) dataset (CMNIST) (Deng 2012; Arjovsky et al. 2019), a dataset of bird images from Caltech-UCSD Birds (CUB), CUB-200-2011 (Wah et al. 2011), a skin lesion dataset from the International Skin Imaging Collaboration (ISIC), ISIC 2018 (Tschandl et al. 2018; Codella et al. 2019), and a 3D dataset with lung cancer CT scans of NSCLC patients NSCLC-Radiomics (H. J. W. L. Aerts et al. 2014).

Colored MNIST (CMNIST). The MNIST dataset is adapted by introducing color as an image feature. The color of the digits is determined based on the random variable x_i sampled from a binomial distribution (with $p = 0.5$). The following binary features are defined as imaging biomarkers $x_{pred,prog} \in \{0, 1\}$: (a) the color (green or not green) as prognostic feature and whether digits lack or contain a circle or loop (i.e. $\{1, 2, 3, 4, 5, 7\}$ vs. $\{0, 6, 8, 9\}$) as the predictive feature or (b) vice versa.

For intuition, a setting where image-based treatment effect estimation could be relevant for this dataset could involve a treatment such as the application of an image filter to alter the digit's appearance. In this scenario, it could be of interest to assess how an outcome, such as a digit classifier's confidence score, changes when a treatment is applied, depending on which color or shape of the digit is present.

Bird species dataset (CUB-200-2011). The dataset includes images of 200 bird species, 5,794 for testing and 5,994 for training, which is further split into training and validation data with an $80\%/20\%$ split. From the binary attributes of the birds, two visually distinct biomarkers $x_{pred,prog} \in \{0, 1\}$ with high annotator certainty are selected: (a) "*has primary color: white*" as prognostic and "*has bill length: longer than head*" as the predictive feature or (b) vice versa.

In this scenario, an illustrative example for the relevance of treatment effect estimation could be investigating predictive imaging biomarkers for the modification of a habitat, such as snow, serving as the treatment. The outcome of interest in this case might relate to a bird's future observed behavior, which is to be extracted from pre-treatment observations of birds in the form of imaging input data and might depend on both the color of the bird and the length of the bill.

Skin lesion dataset (ISIC 2018). The ISIC 2018 dataset contains skin lesion images with a designated training dataset of 2,594 images, which is split into a training and validation set of sizes 2,075 and 519, respectively. Final evaluations were performed on the designated validation set with 100 images.

Dermoscopic attributes, i.e. visual skin lesion patterns, are identified using ground truth segmentation masks and assigned their presence to biomarkers. In feature set (a), the presence of globules is prognostic and the presence of a pigment network is predictive, or in (b) vice versa. Both features have been evaluated as imaging biomarkers for diagnosing melanoma (Gareau et al. 2017; Gareau et al. 2020), making them realistic examples of biomarkers. Unlike the features of the previous datasets, these features are based on the presence of patterns rather than localized features or color values.

Lung cancer CT dataset (NSCLC-Radiomics).
This dataset comprises 415 3D CT volumes of pre-treatment scans from NSCLC patients and ground truth segmentation masks of the lung tumors. The volumes were cropped to the largest connected tumor volume bounding box. The dataset was divided such that 332 samples were used for 5-fold cross-validation, and 83 samples were reserved for testing. Two continuous, uncorrelated radiomics features described in (Zwanenburg et al. 2020) were defined as biomarkers, which have both been evaluated for their prognostic or predictive value before (H. Aerts et al. 2014; Bortolotto et al. 2021): (a) the shaped-based feature "flatness" describing the ratio between the smallest and largest principal tumor components as a prognostic feature and the first-order statistics feature "Energy" characterizing the sum of squares of tumor intensity values as a predictive feature or (b) vice versa. The flatness feature is inverse to the actual flatness of the tumor. Values close to 0 indicate flat shapes, whereas values close to 1 indicate sphere-like shapes. Energy depends strongly on both volume and minimum pixel intensity, as the minimum intensity value is added as an offset. The radiomics features were extracted from the annotated ground truth tumor segmentation volumes with PyRadiomics (Van Griethuysen et al. 2017).

All datasets were randomly split into two equally sized treatment arm subsets, a control ($T = 0$) and a treatment group dataset ($T = 1$). Treatment group-specific outcomes $Y(T, x)$ were then generated according to Equation 4.5. For each CMNIST feature, the biomarker strength parameters $b_{pred,prog} \in \{0.0, 0.1, \dots, 1.0\}$ were chosen, resulting in training 121 models. For the remaining datasets, the biomarker strength parameters $b_{pred,prog} \in \{0.0, 0.2, 0.4, 0.6, 0.8, 1.0\}$ were chosen, resulting in 36 different trained models.

Implementation Details

Model Architecture and Training. In the experiments, the two-headed CATE estimation models were all based on the ResNet (He et al. 2016) architecture tailored to each dataset. For the CMNIST experiments, a MiniResNet (ResNet-14) was utilized, which had 14 layers, 0.20 M parameters, and only three building blocks. In the CUB-200-2011 and ISIC 2018 experiments, a two-headed ResNet-18 with 11.18 M parameters, and for the NSCLC-Radiomics a two-headed 3D ResNet with 33.30 M parameters were used. In all

architectures, the treatment-specific heads consisted of either the last fully connected layer or the last four fully connected layers for NSCLC-Radiomics experiments. Its preceding convolutional layers learn shared presentations of control and treatment group data. The corresponding ResNet architectures with a single output head were used as baseline models.

The models for CMNIST were trained for 400 epochs with a mini-batch size of 1000. For CUB-200-2011 and ISIC 2018, the models were trained with a mini-batch size of 64 and for 1000 or 2000 epochs respectively. The NSCLC-Radiomics models were trained with a batch size of 8 and 2000 epochs.

For all datasets, the mean squared error (MSE) loss function, a learning rate of $\mathrm{lr} = 0.001$, and the stochastic gradient descent (SGD) optimizer were used.

Data preprocessing and augmentation. For preprocessing, zero padding of size 2 was applied to each edge of the CMNIST images. The CUB-200-2011 images were resized so that their smaller edge had the size 256. The data was augmented by performing random crop and horizontal flips so that all final images have the size 224×224. The ISIC 2018 images were resized to 224 for the shorter edge, cropped to between 40% and 100% of their previous size, and resized again to size 224×224. This dataset was augmented with random horizontal and vertical flips, randomly applied rotations by 90 degrees and color jitters. During the inference of both CUB-200-2011 and ISIC 2018 images, center crops were used. All 2D images were normalized by subtracting the mean and dividing by the standard deviation of the respective channel from the training dataset. For the NSCLC-Radiomics dataset, padding of value -1024 (HU) was added so that all padded 3D patches had the size $162 \times 162 \times 54$. All radiomics features were normalized by subtracting the mean and dividing by the standard deviation of each feature. 3D image augmentations were implemented using the MONAI deep-learning framework (Cardoso et al. 2022) and included random flipping, random rotation by 90 degrees along the xy-axis, and random zooming with probability 0.5 by a factor in the range $[0.9, 1.1]$. Resampling to the median spacing of the dataset [0.9765625, 0.9765625, 3.0] mm was based on Isensee et al. (2021) and uses a third-order spline in-plane and nearest-neighbor interpolation out-of-plane.

Evaluation details. For the statistical evaluations, linear regression using ordinary least squares and t-tests for the fit coefficients as described in Section 4.1.3 was performed using the statsmodels python module (Seabold et al. 2010). To create attribution maps, expected gradients (EG) (Erion et al. 2021) for CMNIST and Guided Grad-CAM (Springenberg et al. 2015; Selvaraju et al. 2017) for the other three datasets were used. Using EG enabled determining the attribution of each color channel in contrast to CAM methods, which is vital for discovering the color-related CMNIST biomarkers. Both methods

were implemented using Captum (Kokhlikyan et al. 2020) and enhanced by SmoothGrad (Smilkov et al. 2017) to make the attribution maps less noisy and more robust.

4.2 Image-Based Heterogeneous Treatment Effect Estimation in Clinical Imaging Studies

Building on the methodological foundations for predictive imaging biomarker discovery introduced in the previous section, the methodological developments in this section aim to enhance the applicability of image-based heterogeneous treatment effect estimation methods to clinical practice. The overarching goal is to experimentally assess whether such methods can provide more useful treatment recommendations and can support the identification of potential predictive imaging biomarker candidates in such a setting.

This section begins with Section 4.2.1 describing the two pre-treatment clinical imaging datasets used in the experiments, namely the lung cancer CT dataset previously described in Section 4.1 (NSCLC-Radiomics) and a brain cancer MRI dataset, and the motivation behind selecting them for the experiments. The outcomes of interest provided by both datasets are time-to-event, specifically survival, data. However, for the image-based heterogeneous treatment effect estimations used so far, the assumption has been that the outcomes of interest are either continuous of categorical, which is also a limitation in literature that has not been addressed so far as noted in Chapter 3. For this reason, this part of the thesis adapts existing treatment effect estimation methods that handle survival outcomes to accommodate image-based inputs, as detailed in Section 4.2.2. The utility of these models for treatment recommendations is compared experimentally against an alternative approach that reformulates the survival prediction task as a simple binary classification task (also see Section 5.2). Additionally, it is described how multimodal data consisting of both image and tabular inputs is integrated. The subsection is followed by Section 4.2.3, which presents a strategy to incorporate pre-trained image encoders into the image-based treatment effect estimation pipeline along with regression-based CATE estimation baselines for tabular-only inputs. The section closes with Section 4.2.4, where the evaluation setup of the models and the evaluation metrics are described.

4.2.1 Clinical Imaging Study Datasets

The treatment effect estimation experiments for this part of the thesis investigates the lung cancer CT dataset NSCLC-Radiomics and a brain cancer MRI dataset of glioblastoma patients, both of which are chosen as they provide survival outcome data in addition to 3D images and tabular clinical information. For the lung cancer CT dataset, which is also used in Section 4.1, additional survival outcomes are simulated based on the available

factual outcomes to enable controlled validation of estimated individual treatment effects and help bridge the gap to clinical applications. The brain cancer MRI dataset, in contrast, contains data from an RCT and is therefore used for the retrospective application study to assess how suitable the methods developed in this thesis are for clinical trial settings with real patient outcomes.

Lung Cancer (NSCLC) Dataset and Simulation of Semi-Synthetic Survival Outcomes

The publicly available lung cancer CT dataset, NSCLC-Radiomics, was initially used by H. Aerts et al. (2014) to study the prognostic association between radiomics features obtained from pre-treatment scans and the overall survival time of NSCLC patients retrospectively and is referred to as "Lung 1" in that publication.

A general description of the image preprocessing steps can be found in paragraph 4.1.4, including the details regarding the image preprocessing steps (resampling to a consistent spacing, intensity normalization, cropping to the tumor bounding box, padding and computation of radiomics features), which are the same for this part of the thesis. The cropping was done so that the model could focus on relevant information related to the treatment effect, which was only related to the tumor shape itself according to the implemented simulation. As some experiments use the provided tumor segmentation masks as a second image channel to provide additional anatomical information, these segmentation maps are preprocessed in the same way as the CT images, except by omitting the intensity normalization and by padding with pixels having the value of the background label. Also, the same number of cases are excluded, with the 415 cases remaining, which are split into the same splits used for 5-fold cross-validation (332 cases) and testing (83 cases) as before.

The experiments in this section build upon the previous simulations described in paragraph 4.1.4 using radiomics features as imaging biomarkers to create outcomes with known ground truth individual treatment effects, which serves as a valuable baseline to validate the proposed approaches. In addition, the available survival outcome data for this dataset are utilized, which contain the overall survival time measured from the start of treatment, of which 47 are censored cases. Seven covariates of this dataset's tabular clinical data were used in the experiments: the tumor stage (T stage), nodal stage (N stage), metastatic stage (M stage), overall stage, histological subtype, patient age and gender. These clinical variables have been shown to be prognostic for survival prediction using the NSCLC-Radiomics dataset (H. Aerts et al. 2014; Braghetto et al. 2022; F.-H. Tang et al. 2023). The tabular data were preprocessed as follows: The tumor, nodal, metastatic, and overall stages were kept as categorical covariates with integer values. The histological subtype was one-hot encoded by converting it into four binary covariates: "adenocarcinoma", "large cell", "squamous cell carcinoma", and "not otherwise specified"

(NOS). Missing patient age values were imputed using the mean age of the training set with an additional binary covariate "age missing" introduced to flag imputed data points, before applying z-score normalization.

Here, all patients are considered to have received the "standard" or baseline treatment ($T = 0$) are treated as the control group when simulating semi-synthetic outcomes, as the NSCLC-Radiomics dataset is observational rather randomized (i.e. the treatment was not randomly assigned). The idea behind simulating semi-synthetic outcomes for the experiments presented in Section 5.2.1 is to then simulate the treatment effect using the real outcomes, i.e. additional survival times for a counterfactual "experimental" treatment group ($T = 1$) are generated. This approach has been described by (Curth, Svensson, et al. 2021) as a slightly more realistic alternative to using fully synthetic outcomes for benchmarking.

While the focus of the experiments using this dataset is on semi-synthetic outcomes with simulated treatments, it is worth noting the NSCLC patients underwent different types of treatment: all received radiotherapy, with only a subset additionally receiving chemo-radiation and the remaining undergoing radical radiotherapy alone. Although two types of treatments were applied, the observed outcome data is not suitable for individualized treatment effect estimation. The reason is that the type of treatment was not randomized, but was made in clinical practice based on the tumor and nodal (lymph node) stage as a criterion. Specifically, patients with more advanced stages were more likely to receive chemotherapy. Therefore, the distribution of clinical stages differs between the two treatment groups, violating the assumption that the distributions of covariates overlap between both groups, which is usually a key assumption causal effect estimation relies on (see Section 2.2.1). This limitation, however, makes the dataset well suited for simulating hypothetical treatments as the non-randomized assignments are confounded and can be safely disregarded in the simulation setup so that only known synthetic treatment effects remain.

The generation of survival outcomes follows a similar procedure as in Section 4.1 by employing radiomics features as imaging biomarkers x and by randomly assigning patients to treatments $T \in \{0, 1\}$ with equal probability $p(T) = 0.5$. However instead of assuming a linear biomarker–outcome relationship with continuous outcomes, the assumption here is that the underlying hazard $\lambda(t)$ of the survival outcome Y follows a Cox proportional hazards model (Figure 2.3.1).

For simulating semi-synthetic counterfactual outcomes under a hypothetical experimental treatment $T = 1$, it is additionally assumed the hazard of the observed control group $\lambda_{T=0}(t)$ at time t scales by a proportional-hazard shift $\exp\left(Tb_T + Tb_{pred}\, x_{pred}\right)$ depending on a predictive biomarker x_{pred}:

$$\lambda\left(t \mid x_{\text{pred}}, T\right) = \underbrace{\lambda_{T=0}(t)}_{\text{observed hazard under } T=0} \exp\left(Tb_T + Tb_{pred}\, x_{pred}\right). \tag{4.6}$$

A survival time Y with such a hazard constant in time, can then be generated according to Bender et al. (2005) using $Y = -\frac{\log(u)}{\lambda(t)}$ with a random uniformly distributed variable $u \sim \mathcal{U}(0, 1)$, which is equivalent to drawing Y from an exponential distribution $Y \sim \text{Exp}(\lambda(t))$. Therefore, to semi-synthetically simulate the survival time for those patients that have been randomly assigned to $T = 1$, the observed control group survival time $Y^{T=0}$ is scaled using

$$Y^{T=1} = \frac{Y^{T=0}}{\exp(b_T + b_{pred} x_{pred})}. \tag{4.7}$$

The outcomes of censored cases, *i.e.* the observed time of the last follow-up, are also scaled in this simulation to at least partially preserve the information about the treatment-effect and as it is assumed that the censoring is non-informative and treatment independent. In the experiments for this thesis, the parameters were set to $b_T = 0.0$ and $b_{pred} = -0.8$.

In addition to the time-to-event outcomes, binary survival outcomes are simulated by thresholding the survival time at a fixed cutoff of 365 days. This results in an imbalanced label distribution: 268 patients with $Y^T > 365$ are assigned to $Y^T_{\text{binary}} = 1$ ("long survival"), while 147 patients with $Y^T \leq 365$ are assigned to $Y^T_{\text{binary}} = 0$. Despite the imbalance, the 365-day cutoff is retained, as the 12-month (1-year) survival rate is a commonly reported threshold and endpoint in oncology (Ballman et al. 2007; Antonia et al. 2018). Censoring is disregarded in this binarization, as it is generally assumed that all patients are observed up to the cutoff time. Only one censored patient in the hold-out test set had a last recorded follow-up time below the cutoff (314 days) and was therefore labelled as "short survival" (0) for evaluation purposes.

To simulate a synthetic predictive imaging biomarker, the z-score standardized version of the shape-based radiomics feature "flatness" is used. This feature, which is also used in Section 4.1, was chosen for simplicity, as it showed minimal prognostic influence in a Cox proportional hazards model fitted on real survival outcomes when included individually alongside clinical tabular features, making it suitable for simulating isolated predictive effects without confounding prognostic effects.

Brain Cancer (Glioblastoma) Dataset from a Randomized Clinical Trial

One of the main areas of investigation in this part of the thesis, as detailed in RQ2.3, is the application of the CATE estimation models proposed in Section 4.2.2 to a real non-synthetic RCT dataset.

Specifically, experiments were performed on the brain cancer dataset "EORTC", which comprises pre-treatment MRI scans and the tabular clinical data of 427 glioblastoma patients. It is a subset of data initially acquired as part of the phase 3 portion of the EORTC-26101 trial (ClinicalTrials.gov identifier: NCT01290939), a large-scale, multi-center randomized phase II and III clinical trial conducted by the European Organisation

for Research and Treatment of Cancer exploring the treatment of glioblastoma patients using the drug bevacizumab (BEV) in combination with lomustine (Wick et al. 2016; Wick et al. 2017).

Due to its randomized design, the dataset is well-suited for studying the individual treatment effects between the two treatments groups. In the experiments in Section 5.2.2 two groups of patients are compared according to their assigned treatment at time of randomization: those who received bevacizumab in their initial treatment (experimental arm $T = 1$ with 160 patients) and those who have not received bevacizumab but lomustine alone (control arm $T = 0$ with 267 patients). While some patients from the control arm received bevacizumab at a later stage, but not during initial treatment, the number of patients the number of patients treated with bevacizumab (323) and the number of those not treated with any at all (104) was imbalanced, as noted by Kickingereder et al. (2020). This distinction is however not made in the experiments, as the focus is only on the type of the initial treatment.

The overall survival time, which is the primary outcome of interest of the trial and the treatment effect estimation experiments, was recorded from the time of randomization until the last follow-up or death, and was censored for 84 cases. Although the trial also collected follow-up scans, they are not used in this thesis, as it is assumed that only the pre-treatment data is relevant for treatment decision-making and predictive imaging biomarker discovery. As for the NSCLC-radiomics lung cancer dataset, binary survival outcomes are additionally computed by thresholding the time-to-event survival outcomes at a fixed cutoff of 365 days, resulting in an imbalanced distribution with 343 patients having a "long survival" ($Y^T_{\text{binary}} = 1$) and 84 patients having a "short survival" $Y^T_{\text{binary}} = 0$.

The tabular clinical information additionally included several covariates recorded at baseline: contrast-enhancing tumor volume, age, sex, corticosteroid use (yes vs. no), and WHO performance status (>0 vs. 0). These covariates have been established as known prognostic confounders with, for instance, larger tumor volumes, higher age, male sex, corticosteroid use and a poor WHO performance status (>0) being associated with a shorter overall survival (Kickingereder et al. 2019).

For each MRI scan, four MRI modalities were acquired: T1-w, contrast-enhanced T1-weighted after administration of a contrast agent (cT1-w), fluid-attenuated inversion recovery (FLAIR) and T2-w 3D images, which are treated as four separate input channels when fed into a neural network. The initial preprocessing of the provided MRI scans is described by (Kickingereder et al. 2019): the images had been co-registered to the T1-w image and skull-stripped, with the background value set to 0, and cropped to the foreground bounding box. Additionally, segmentation masks labeled by experts are available, containing the classes edema, contrast-enhancing tumors and background.

For the experiments in Section 5.2.2, the images and segmentation maps were further preprocessed by resampling to the median voxel spacing $1 \times 1 \times 1$mm, followed by

performing a z-score normalization of the intensity the of MRI scans, which subtracted the mean and divided by the standard deviation computed on foreground pixels only, following the procedure by Isensee et al. (2021). Two scans and segmentation maps with incorrectly cropped bounding boxes required manual correction of their bounding boxes to exclude non-connected foreground regions and remove large sections of empty background slices.To ensure compatibility with the residual encoder (ResEnc)-L image encoders described in Section 4.2.3, the images used the experiments with pre-trained image encoders in Section 5.2.2 were automatically preprocessed using the preprocessing pipeline derived from nnU-Net (Isensee et al. 2021), which resampled all images to the same constant shape of $160 \times 192 \times 160$ (instead of the same spacing) after normalizing the MRI images using the same z-score normalization.

The dataset was randomly split into 341 (80 %) cases used for 5-fold cross-validation and 86 (20 %) for testing.

4.2.2 Model for Multimodal Inputs and Survival Outcomes

The treatment effect estimation model developed for clinical imaging RCT data in this second part of the thesis specifically addresses the fact that in clinical studies, especially in oncology, the most important outcome of interest is the overall survival. The survival time often represents the primary endpoint that defines treatment success in a clinical trial, which is why image-based treatment effect estimation for continuous or categorical outcomes (see Section 4.1.2) is extended to handle survival outcomes. Additionally, multimodal data is typically acquired from each patient, including tabular clinical data (as mentioned in Section 4.2.1). As this data often offers important prognostic information, it is integrated as an additional input modality to the model.

An overview of the full proposed CATE estimation method and evaluation strategy is presented in Figure 4.3, which illustrates the three key objectives it aims to address simultaneously: using multimodal inputs (A) to estimate heterogeneous treatment effects (B) by predicting survival outcomes (C), with the goal of obtaining treatment recommendations (D).

The details of the combined model, representing the first CATE estimation approach for survival outcomes based on clinical imaging data and using multimodal integration, are described in the following.

Survival Modeling and Loss Function

As illustrated in Figure 4.3, the goal is to estimate the causal effect of a treatment on survival (or more generally, time-to-event) outcome data and make treatment recommendations based on individualized predictions. While this task could be reformulated

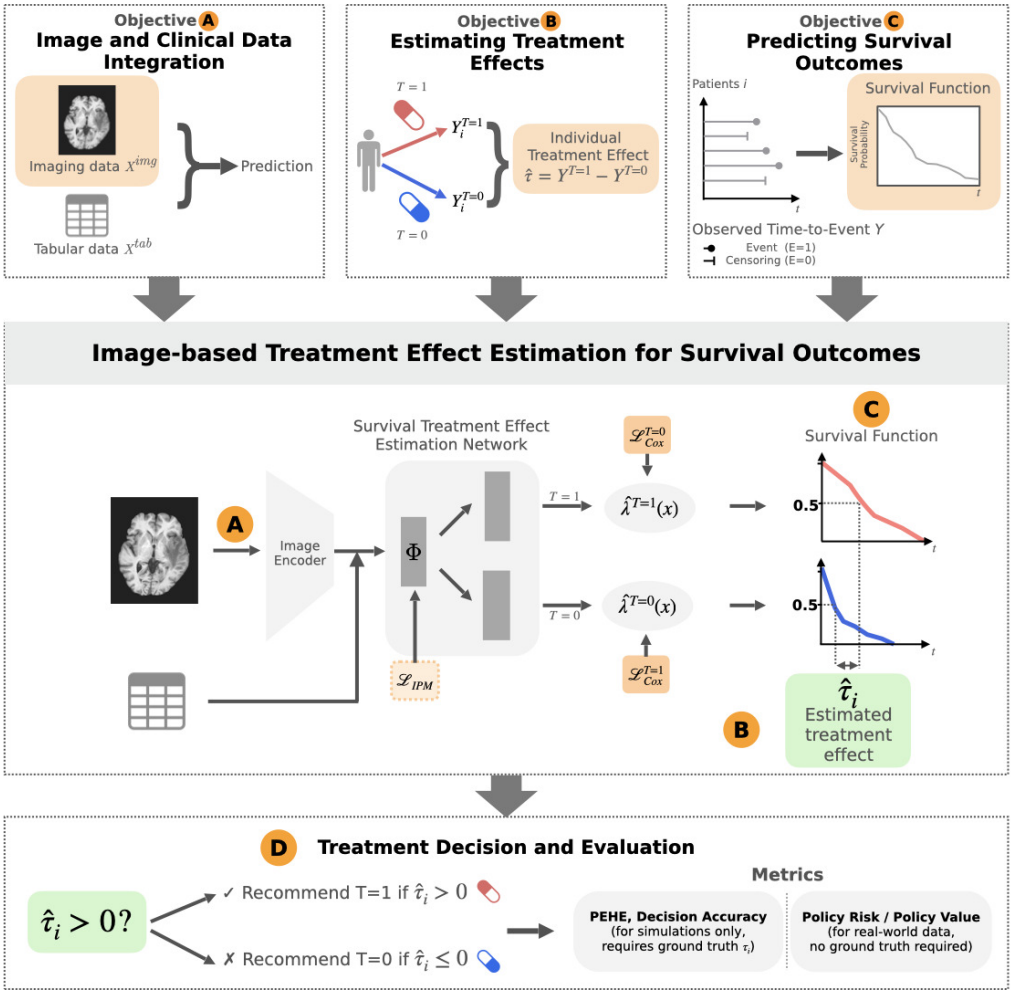


Figure 4.3: Overview of the proposed image-based treatment effect estimation approach for survival outcomes, which addresses three key objectives simultaneously: **A** multi-modal data integration for 3D imaging and tabular (clinical) data, **B** estimation of heterogeneous treatment effects, and **C** modeling survival outcomes. This is implemented using a TARNet-like architecture with a shared CNN-based image encoder, trained using the BITES loss function (Schrod et al. 2022), which combines the Cox partial log-likelihood and a balancing term IPM to account for covariate imbalance. The network predicts treatment-specific hazards $\hat{\lambda}^T$, which are used to compute the survival functions. Here, the estimated CATE is defined as the difference in time points at which the predicted survival probability is 50% for treated vs. control. Treatment decisions (i.e. the model's recommended treatment) in **D** are made by comparing the estimated CATE $\hat{\tau}_i$ to a given threshold: if $\hat{\tau}_i > 0$, which indicates that a more favorable survival outcome is predicted for $T = 1$, the model recommends $T = 1$ and $T = 0$ otherwise. The evaluation assesses how well the model recommends the optimal treatment. Both "oracle metrics" (PEHE, decision accuracy) requiring counterfactual outcomes and factual-outcome-based metrics (policy value, policy risk) based on observed outcomes are used.

into a binary classification problem by thresholding the survival times at a given cutoff time, it could potentially lead to loss of information. For this reason, the direct modeling of survival outcomes using survival analysis techniques is proposed in this thesis to potentially allow a more detailed and nuanced evaluation of heterogeneous treatment effects and subsequently more accurate treatment recommendations.

The key difference in the survival modeling approach is that the model output is a patient-specific hazard λ^T for a treatment T, from which the baseline hazard function and the survival function describing the survival probability at time t are derived (see Section 2.3). To estimate these hazards using deep learning, an established approach using the DeepSurv model was employed (Katzman et al. 2018), where the network is trained using the Cox model's negative partial log-likelihood loss function $\mathcal{L}_{Cox}$

$$\mathcal{L}_{\text{Cox}}^{T}(\mathbf{x}) = -\sum_{i:\delta_i=1}\left[h^T\left(\Phi(\mathbf{x}_i)\right) - \log\left(\sum_{j:y_j\in\mathcal{R}_i} e^{h^T(\Phi(\mathbf{x}_j))}\right)\right]. \tag{4.8}$$

This function is a version of the Cox partial log-likelihood Equation 2.10 from the Cox proportional hazards model, which is modified according to Faraggi et al. (1995) so that it can be applied to deep learning by replacing $\beta^\top \mathbf{x}_i$ with the output of a neural network $h\left(\Phi(\mathbf{x}_i)\right)$. Here, $\Phi(\mathbf{x}_i)$ represents the shared intermediate feature representation (see Figure 4.3) derived from the multimodal input $\mathbf{x}_i$ and h^T is the treatment-specific network head prediction also known as the relative risk score. By minimizing this loss function, the likelihood that subjects who experienced an event (death) have a higher predicted risk than all other subjects at risk (given by $\mathcal{R}_i$) at that specific time point gets maximized.

For reliable treatment effect estimation, it is important that the distributions of learned representations are balanced for both treatment arms and covariate shifts are minimized, as imbalances can lead to biases in the counterfactual predictions. While balancing techniques are especially crucial for observational datasets where the treatment assignment may be confounded, they can also be important in an RCT setting such as the EORTC study, where the size of the two treatment groups is imbalanced and the number of samples in the dataset is limited. Additionally, training deep learning models with 3D image inputs requires the use of mini-batches (instead of the full dataset at once) due to memory constraints, which can additionally cause instabilities.

For this reason, the network was trained using the BITES loss (Schrod et al. 2022), an extension of the Cox negative partial log-likelihood loss with an additional balancing term originally designed for observational data, but which could also be used to improve the robustness in an RCT setting. The balancing term $\mathcal{L}_{\text{IPM}}$ is based on IPM (Müller 1997) for regularizing the learned shared representations $\Phi(\mathbf{x}_i)$. Schrod et al. (2022) defines the BITES loss as

$$\begin{aligned}\mathcal{L}_{\text{BITES}} =& q\mathcal{L}_{\text{Cox}}^{T=0}(h^{T=0}\left(\Phi(\mathbf{x}_i)\right)) + (1-q)\mathcal{L}_{\text{Cox}}^{T=1}(h^{T=1}\left(\Phi(\mathbf{x}_i)\right)) \\ &+ \alpha\cdot\mathcal{L}_{\text{IPM}}(\Phi^{T=0}, \Phi^{T=1}),\end{aligned} \tag{4.9}$$

where q denotes the fraction of control group samples in the mini-batch and α a hyper-parameter for adjusting the strength of the IPM regularization term.

Following Schrod et al. (2022), the ITE for this study is defined based on the median survival time Y_{median}, i.e. the time at which the survival function $S(Y_{\text{median}}|\mathbf{x})$ reaches the probability of 50%., as illustrated in Figure 4.3. The median survival time is computed by solving $S(Y_{\text{median}}|\mathbf{x}) = 0.5$ using the survival functions obtained from the predicted hazard function.

Thus, the treatment recommendations are informed by the estimated CATE, which is computed from the difference in expected median survival time for the treated and control outcomes

$$\widehat{\tau}(\mathbf{x}_i) = \widehat{Y}^{T=1}_{\text{median}}(\mathbf{x}_i) - \widehat{Y}^{T=0}_{\text{median}}(\mathbf{x}_i). \tag{4.10}$$

The exact decision rule and corresponding evaluation metrics are detailed in Section 4.2.4.

Model Architecture and Integration of Tabular and Image Data

Similar to the architecture from Section 4.1.2, a TARNet-like architecture with a shared ResNet-18 image encoder with four ResNet blocks to learn treatment-independent common representations Φ and two treatment-specific fully connected layer heads was employed for the CATE estimation model (see Figure 4.3). The output of each of the heads represents the treatment-specific hazard $\widehat{\lambda}^T$ as described earlier.

To leverage all available data for making survival predictions and address RQ2.2, both clinical tabular data and 3D multi-channel image data were integrated to obtain Φ by simply concatenating flattened image representations (after adaptive average pooling) with tabular data directly, similar to Durso-Finley et al. (2022).

Additionally, integration using a Dynamic Affine Feature Map Transform (DAFT) block by Wolf et al. (2022) was investigated, which was used to replace the fourth ResNet block, where the image feature maps are dynamically transformed conditioned on the tabular data, to allow the tabular data to directly influence the learned features, rather than being treated separately.

Experimental Details

Training setup and sampling strategy. In contrast to survival CATE estimation models for tabular inputs, which are commonly trained with batches consisting of the entire training dataset (i.e. whole batch) or at least very large mini-batches (e.g. models by Curth, Lee, et al. (2021) and Schrod et al. (2022)), a model for 3D image inputs trained using SGD needs much smaller mini-batch sizes to manage memory requirements and to mitigate overfitting. Although the Cox negative partial log-likelihood loss function

requires the risk set over the full dataset, the approximation using mini-batches has been theoretically justified by Kvamme et al. (2019), Tarkhan et al. (2024), and Zeng et al. (2025).

To ensure stable training with small mini-batch sizes, the CATE estimation model was trained using a stratified mini-batch sampling strategy, where it was ensured that each batch contained at least one patient per treatment group.

In the experiments, all models for this part of the thesis (unless specified otherwise) were trained using 5-fold cross-validation, a mini-batch size of 10 with a constant learning rate of $\text{lr} = 10^{-4}$, and the SGD optimizer for 1000 epochs. Due to the smaller batch sizes, instance normalization (Ulyanov et al. 2016) was used in the ResNet blocks instead of the standard batch normalization (Ioffe et al. 2015). The fully connected heads consisted of two layers with 16 and 8 hidden units, respectively. The dropout rate of these heads was set to 0.1 only for the NSCLC-Radiomics dataset to reduce overfitting. On the EORTC dataset, where dropout had no benefit, it was set to 0.

The BITES loss was only applied to the EORTC dataset, where the weight of the balancing IPM loss term was set to $\alpha = 0.01$, whereas the model for the NSCLC-Radiomics was trained using the standard Cox loss terms in Equation 4.8.

All deep learning methods were implemented using the PyTorch framework (Paszke et al. 2019). The training pipelines were built with PyTorch Lightning (Falcon et al. 2019) and performed on a single graphics processing unit (GPU).

Image augmentation. The image augmentation scheme for both datasets included padding to a uniform patch size ($54 \times 162 \times 162$ for NSCLC-Radiomics and $164 \times 192 \times 162$ for EORTC), random zooming, and random rotation. Additionally, dataset-specific image augmentation was applied.

For the NSCLC-Radiomics dataset, the augmentation scheme included random flipping along all three spatial axes, and the background padding value was set to -1024 HU to maintain consistency with the physical background voxels (air). For the EORTC dataset, random flipping was only applied along the left–right axis and the background padding value was set to 0.

Stronger intensity-based or non-linear deformation augmentations were deliberately avoided to prevent the loss of subtle image information (e.g. texture or intensity distribution) that could potentially serve as important information for survival prediction and imaging biomarkers.

Binary Model Comparison. For direct comparison against a CATE estimation model for binary survival classification and to address RQ2.1, a binary-outcome model was trained using the same architecture and hyperparameters as the survival-outcome models.

The survival loss (BITES or the Cox loss) was replaced with the Binary Cross-Entropy with Logits Loss (BCEWithLogitsLoss), as implemented in PyTorch, which consists of the binary cross-entropy (CE) combined with an internal sigmoid final activation layer. The balancing IPM term was kept when the balancing parameter was set to $\alpha > 0$, so that $\mathcal{L}_{\text{binary}} = q\mathcal{L}_{\text{BCE}}^{T=0} + (1-q)\mathcal{L}_{\text{BCE}}^{T=1} + \alpha \cdot \mathcal{L}_{\text{IPM}}$.

Hyperparameter tuning and model selection. As the ground truth treatment effects are not accessible in real datasets, hyperparameter tuning of CATE estimation models needs to rely on alternative performance metrics (Machlanski et al. 2023).

For the preliminary tuning, the hyperparameters, including the selection of the learning rate, learning rate scheduling, balancing parameter α, or number of epochs, were primarily selected based on the best validation C-index obtained through 5-fold cross-validation using the image-only survival-outcome CATE estimation model. The C-index was prioritized to ensure a reliable survival predictions and as it is more stable compared to metrics directly related to the treatment effect.

To ensure consistency across experiments, the best-performing configuration of the image-only survival model was kept fixed for all subsequent multimodal and binary-outcome experiments. This strategy was necessary to isolate the influence of the choice of hyperparameters, which can otherwise dominate model performance in CATE estimation tasks as emphasized in (Machlanski et al. 2023).

The selection of the final model architecture and input modalities (e.g. binary vs. survival and image-only vs. multimodal inputs) was based on the validation set's policy value or policy risk metrics, as detailed in Section 5.2. This ensured that the final model was chosen for its performance on the relevant downstream task of informing optimal treatment decisions.

Multitask learning with auxiliary classification head. The idea behind training a multitask learning model for CATE estimation was to leverage the faster convergence and additional supervision provided by the binary classification task to guide the model during training and regularization, thereby potentially leading to improved learned representation.

In the multitask learning experiments, two treatment-specific auxiliary classification heads were added to CATE estimation model. These heads predict the binary survival outcome from the same shared representations Φ as the survival-outcome prediction heads using the following combined loss function:

$$\mathcal{L}_{\text{multitask}} = q\mathcal{L}_{\text{Cox}}^{T=0} + (1-q)\mathcal{L}_{\text{Cox}}^{T=1} + q\mathcal{L}_{\text{BCE}}^{T=0} + (1-q)\mathcal{L}_{\text{BCE}}^{T=1} + \alpha \cdot \mathcal{L}_{\text{IPM}}. \tag{4.11}$$

S-Learner architecture variant. The S-Learner CATE estimation architecture (Künzel et al. 2019) was used in the experiments to study whether using a single combined prediction head could better leverage shared similarities between treatment and control arm during training, especially when treatment effects are close to zero. In the S-Learner models, treatment indicators T are given as an additional input, which are concatenated to the latent representations Φ. Predictions for the counterfactual potential outcomes ($Y^{T=0}$ and $Y^{T=1}$) were then generated by passing the input through the head twice with different treatment indicators, i.e. once with $T = 0$ and once with $T = 1$.

4.2.3 Baselines and Pre-trained Encoder Extension

This subsection presents two variants of the CATE estimation model from Section 4.2.2 that are used in the experiments for further comparisons: First, the regression models for are considered as baselines, as they only leverage mainly known prognostic covariates. Secondly, the image-based CATE estimation models are extended using pre-trained image encoders to investigate whether their image representations can improve treatment effect estimation over models trained from scratch.

Tabular-only Regression Baselines

To put the performance of the previously proposed deep-learning-based CATE estimation models (Section 4.2.2) into context and assess the added value of image inputs, these models are compared against simple regression models for tabular input data only.

Similar to Schrod et al. (2022), the T-learner was employed as the metaalgorithm for the regression-based CATE estimation (Künzel et al. 2019), where separate models are trained for the treated $T = 1$ and the control group $T = 0$. The S-Learner version, which only requires performing regression using a single model by taking the treatment directly as an input variable, was not pursued further due to its poorer performance in making individual treatment recommendations, as noted by Schrod et al. (2022). This was also supported by preliminary experiments on the tabular data of both datasets, which consistently showed a lower decision accuracy for treatment recommendations compared to the T-learner on the semi-synthetic NSCLC-Radiomics dataset, despite inconclusive or inconsistent results with respect to the factual survival prediction performance metrics (e.g. C-index) across both datasets.

For binary outcomes, logistic regression models were implemented using *scikit-learn* (Pedregosa et al. 2011) and fitted with a maximum of 1000 iterations. For survival outcomes, Cox proportional hazards models were implemented using the *lifelines* package (Davidson-Pilon 2019). All regression models were trained using 5-fold cross-validation with the same data splits as those for the deep-learning-based models.

For NSCLC-Radiomics dataset, the clinical variable M stage was excluded from the input covariates for the Cox proportional hazards regression. This exclusion was due to its near-constant values for almost all samples (M stage$=0$ in 98.5% of cases), which caused instability in the convergence.

Finally, the evaluation metrics and procedures were identical to those used for the deep-learning-based models.

Leveraging Pre-trained Image Encoders

The comparatively small number of labeled samples in medical imaging datasets, particularly in clinical trial datasets, is often one of the main limiting factors in making accurate survival predictions and subsequently identifying heterogeneous treatment effects. Especially in cases when heterogeneous treatment effects are small compared to anatomical variations and other noise factors in the dataset, deep learning models trained from scratch are often prone to overfitting and fail to capture robust features.

To mitigate the challenges of data scarcity, semi-supervised learning and transfer learning have been adopted for not only standard medical image analysis tasks such as image segmentation or classification to improve their performance, but also most recently for tasks related to treatment effect estimation and survival prediction (see discussed in Section 3.2.3).

This thesis therefore explored whether integrating pre-trained image encoders could provide a benefit for image-based CATE estimation methods and reduce overfitting, which addresses RQ2.2. To this end, publicly available encoders from Wald et al. (2025) were used, which were pre-trained on a large-scale public dataset comprising 114k 3D MRI volumes of brains using different semi-supervised learning strategies. To leverage the information specific to anatomical region and modalities, the transfer learning experiments for this thesis focused on their application to the EORTC dataset, as it includes images from a similar imaging modality (MRI sequences) and anatomical region (brain).

For the image encoders, the ResEnc-L architecture from Isensee et al. (2021) and Isensee et al. (2024), a CNN-based encoder derived from the U-Net (Ronneberger et al. 2015) with residual connections, was chosen because it showed higher average segmentation and classification performance in (Wald et al. 2025).

Similar to Wald et al. (2025), the implementation of the image-based CATE estimation models followed the *Image Classification framework* (Ziegler et al. 2024) but used two instead of one single-layer classification heads to form a TARNet. As the framework was initially designed for classification outputs and to specifically isolate the impact of pre-training, only binary-outcome CATE estimation models are assessed for the comparison between encoders trained from scratch and pre-trained ones. Tabular data was integrated

the same way as described in Section 4.2.2 by concatenating the tabular covariates with the representations after the image encoder before the classification heads.

Experimental details for fine-tuning pre-trained encoders. The publicly available pre-trained ResEnc-L encoders used in the experiments were previously trained using the SwinUNETR (Y. Tang et al. 2022) and masked autoencoder (MAE) (He et al. 2022) pre-training scheme. They were chosen for these experiments due to their best validation results in preliminary classification experiments on the EORTC dataset. As a baseline, they were compared against a model using a ResEnc-L encoder trained from scratch on the EORTC dataset.

All encoders were either fully fine-tuned (i.e. without any frozen weights) or trained from scratch for 200 epochs using a batch size of 2, gradient accumulation for 12 batches, the AdamW optimizer (Loshchilov et al. 2017) with a weight decay coefficient set to 0.01, a maximum learning rate of 1×10^{-4} with a cosine annealing scheduler where the learning rate is increased for 20 epochs during the warm-up phase, similar to (Wald et al. 2025). Also, the same image augmentation scheme as used by Wald et al. (2025) for their downstream classification task was applied. Additionally, label smoothing (Szegedy et al. 2016) with a value of 0.1 was used in combination with the binary CE loss function as a regularization technique to reduce overfitting and increase the robustness to noisy labels. The evaluation procedure of the binary-outcome CATE estimation models with pre-trained encoders was identical to the previously described models and is specified in the following subsection, Section 4.2.4.

4.2.4 Evaluation Setup

Factual Outcome Prediction Metrics

On both datasets, factual metrics were used to assess the model performance at predicting observed (non-counterfactual) outcomes as a secondary task to ensure a fair comparison. The advantage of these metrics is that the observed outcomes required to compute the factual outcome prediction metrics are always available, unlike the counterfactual outcomes needed for assessing treatment effect estimation.

The binary-outcome models were evaluated with standard classification metrics for binary classification, including the balanced accuracy, F1, average precision, and area under the receiver operating characteristic curve. The predicted binary classification labels needed to compute the balanced accuracy and the F1-score were obtained by thresholding the predictions at 0.5. The balanced accuracy was reported instead of the accuracy due to class imbalances in both datasets, and for its intuitive interpretability, AUROC summarizes overall discrimination performance, while F1 and AP were included for reference to capture whether the model tends to always predict the majority class.

When averaging these metrics, the samples were weighted using IPCW to adjust for possible biases introduced by censoring (see Equation 2.3.2).

For the evaluation of the survival-outcome models, the survival curves were obtained following the evaluation of (Schrod et al. 2022) and using the implementation by the PyCox package (Kvamme et al. 2019; Kvamme et al. 2021), where the survival curves are obtained using the computed treatment group specific baseline hazard functions and the relative risk score obtained directly from the model's outputs. The resulting survival curves were evaluated with Antolini's C-index (Antolini et al. 2005), a version of Harrell's C-index (Harrell et al. 1982; Harrell Jr et al. 1996) computed only on uncensored pairs without reweighting.

Evaluation of Treatment Effect Recommendations and Decision Rules

As illustrated in Section 4.2.2, both metrics requiring counterfactual outcomes (i.e. decision accuracy and PEHE) and metrics only requiring observed outcomes (i.e. policy risk and policy value) were computed to directly assess the performance of the models with respect to the primary task investigated in this part of the thesis, treatment effect estimation performance and treatment recommendation. Since counterfactual outcomes were only available for the semi-synthetic NSCLC-Radiomics dataset, the evaluation differed from that of the EORTC dataset. All the mentioned metrics were again adjusted using IPCW to account for censoring.

The PEHE was computed as defined in Equation 2.4, whereby the respective definition of the CATE was used depending on the outcome type of the model. The CATE of the binary-outcome model was defined as the difference in predicted probabilities, whereas the CATE of the survival-outcome model was defined as the difference in the estimated median survival times ($\hat{Y}_{\text{median}}$) computed from the survival curves. The ground truth CATE (ITE) was defined analogously using the actual (observed) survival times from both potential outcomes.

Treatments $T = 1$ were recommended based on the predicted median survival times for treated and control (see Equation 4.10) if $\hat{\tau}_i > 0$ and $T = 0$ otherwise. In other words, treatments were recommended if the estimated treatment effect was more than 0 d, indicating a longer survival associated with that treatment. These recommendations are then used to compute the fraction of correctly assigned treatments, i.e. decision accuracy (Efthimiou et al. 2023), by comparing the treatment recommendations obtained from the ground truth treatment effect and the estimated CATE values.

To evaluate the treatment recommendations without ground truth, the treatment recommendations based on the estimated CATE were also used for the calculation of the observed policy risk $\widehat{R}_{Pol}$ (Equation 2.6) for the binary-outcome models or policy value

$\hat{V}_{Pol}$ for the survival-outcome models, as defined by Equation 2.6 and Equation 2.5 (see Section 2.2.3).

Kaplan-Meier Curves and Patient Stratification

To further interpret and qualitatively validate the impact of the treatment recommendations on patient stratification, Kaplan-Meier survival curves were generated using the lifelines Python package (Davidson-Pilon 2019).

Following the approach of Katzman et al. (2018) and Schrod et al. (2022), patients were grouped based on whether their actual received treatment matched the model's recommendation ("recommendation-followed group") or not ("anti-recommendation followed group"). Additionally, for assessing the impact of treatment within the patient subgroups with a positive or negative estimated CATE, the patients were also grouped by the treatment they received.

The survival differences were assessed using the Kaplan-Meier curves using log-rank tests and Cox proportional hazards regression. The resulting HR indicates an improved survival if HR < 1, no difference if HR$\sim$ 1, and worse survival outcomes if HR > 1.

Additionally, patients were stratified into tertile subgroups according to the estimated treatment effect magnitude to explore the heterogeneity of the difference between the average observed survival times of treated and control group patients within each tertile subgroup (see Section 5.2.2), following Durso-Finley et al. (2022).

When generating the treatment recommendations for the Kaplan-Meier curves, model ensembling was applied by averaging the predictions of all five models trained during cross-validation before computing the estimated CATE. For survival-outcome models, the individual median survival times were averaged to obtain the ensemble prediction before computing treatment effects. For binary-outcome models, predicted probabilities (after the sigmoid activation) were averaged, as Ju et al. (2018) found it to be slightly more beneficial than averaging logits. The treatment effect (CATE) was then derived from these ensembled predictions.

Conversion of Survival Predictions for Binary-Outcome Model Comparisons

To enable a direct comparison of the results with the binary-outcome model, the empirical policy risk was computed using the observed thresholded outcomes at 365 d. Additionally, all other metrics, including the PEHE, observed policy risk $\hat{R}_{Pol}$, and classification metrics for factual predictions were computed after binarizing the predictions from the survival-outcome model.

The binarized predictions used for the Balanced Accuracy and F1 score were obtained by thresholding the predicted median survival time $\hat{Y}^T_{0.5}$ at the threshold time of 365 d. For

AUROC and AP, the survival-outcome model's predictions were converted into classification probabilities using the predicted survival function $\widehat{S}^{T}(365\,\mathrm{d})$, i.e. the survival probability of surviving beyond one year. Similarly, for PEHE on the survival model, the treatment effect is defined using $\widehat{\tau}^{\text{prob}} = \widehat{S}^{T=1}(365\,\mathrm{d}) - \widehat{S}^{T=0}(365\,\mathrm{d})$, which differs from the median-time CATE used for recommendations but allows the comparison to binary ground-truth treatment effects on the probability scale. All metrics were IPCW adjusted to account for biases introduced by censoring.

5 Experiments and Results

This chapter presents the description and results of the experiments, which have the general aim of assessing predictive imaging biomarker discovery using deep-learning-based treatment effect estimation methods for imaging data. It is divided into two main sections: Section 5.1, where the feasibility and proposed evaluation protocol of predictive imaging biomarker discovery is studied on image datasets in a controlled setting with simple semi-synthetic outcomes, and Section 5.2, where the deep-learning-based treatment effect estimation approach previously used in Section 5.1 and extended to survival outcomes and tabular inputs is evaluated on two more complex clinical imaging datasets, one with semi-synthetic outcomes, and the other with outcomes from real outcomes from a randomized controlled trial.

5.1 Evaluating Heterogeneous Treatment Effect Estimation Models for Predictive Imaging Biomarker Discovery

The first major goal of this thesis was to study how predictive imaging biomarkers can be directly discovered using image-based methods for estimating heterogeneous treatment effects, i.e. CATE estimation models. The underlying hypothesis is that methods trained to estimate heterogeneous treatment effects also automatically learn to identify relevant imaging features that are predictive. To recover, quantify, and interpret these features so that they can be further verified and used to guide treatment decisions in practice, this thesis proposed an evaluation protocol in Section 4.1.3.

The experiments studied the image-based CATE estimation models with the proposed evaluation protocol for predictive imaging biomarker discovery in a controlled setting with known treatment effects and with varying strengths of predictive and prognostic imaging biomarkers. For this purpose, four different semi-synthetic image datasets were used (Section 4.1.4). They were guided by the following research questions, as outlined in Section 1.2:

RQ 1.1: Can deep-learning-based heterogeneous treatment effect estimation be used to discover predictive imaging biomarkers directly from image data without a separate feature extraction step?

RQ 1.2: How can the performance and reliability of image-based heterogeneous treatment effect models in discovering predictive imaging biomarkers be evaluated both quantitatively and qualitatively?

RQ1.1 is an overarching question addressed throughout this section, whereas the section is structured according to the two aspects of RQ1.2. The quantitative evaluation from Section 5.1.1 is presented in Section 5.1.1, and the qualitative evaluation is presented in Section 5.1.2.

5.1.1 Predictive Strength of the Estimated CATE

Predictive imaging biomarker discovery performance. The purpose of this subsection is to demonstrate how the strength of a predictive imaging biomarker identified by a CATE estimation model can be quantified using the evaluation protocol presented in Figure 5.1 and in Figure 4.1. This is done by computing the relative predictive strength $|t_{pred}/t_{prog}|$ using the estimated CATE as a predictive biomarker candidate, which does not require the ground truth treatment effect and is therefore in principle not limited to synthetic data.

At the same time, this subsection investigates whether the image-based deep learning CATE estimation model outlined in Section 4.1.2 is capable of identifying predictive imaging biomarkers (relating to RQ1.1) and, if that is the case, its performance at doing so. Answering these questions relies on the predictive biomarker strength being known, which is why the experiments employed image datasets with outcomes simulated according to Section 4.1.3.

The performance is measured by how robust the model is to varying strengths of prognostic biomarkers present at the same time as predictive biomarkers in the images. This

is captured by the relative size of the true predictive effect b_{pred}/b_{prog} used in the data simulation. A model that identifies predictive imaging biomarkers correctly should ideally also be sensitive to the strength of the predictive imaging biomarker, here given by b_{pred}, while not being affected by the prognostic imaging biomarker, with strength given here by b_{prog}.

To assess this, the relative predictive strength $|t_{pred}/t_{prog}|$ of the model estimations (i.e. CATE) along with that of the baseline model prediction are summarized as boxplots. The results are shown in Figure 5.1 with respect to b_{pred}/b_{prog}, which represent the experimental parameter setting here. For comparison, the figure includes the results of a one-headed baseline model used for simply regressing the outcome regardless of treatment instead of estimating the CATE (see Section 4.1.2). Additionally, $|t_{pred}/t_{prog}|$ of the ground truth predictive biomarker is plotted as the experimental upper bound along the $|t_{pred}/t_{prog}|$ of the prognostic biomarker as the experimental lower bound for comparison. The results are shown for different choices of predictive and prognostic imaging biomarkers, denoted with (a) or (b), to assess if there is a difference in performance in that regard.

The following paragraphs from this section are adapted from the article by Xiao et al. (2025), originally written by the author of this thesis, and therefore resemble the text of the original manuscript.

A common observation across all four datasets, CMNIST, CUB-200-2011, ISIC 2018, and NSCLC-Radiomics, is that the relative predictive strength $|t_{pred}/t_{prog}|$ of the CATE estimation model increases with increasing relative predictive biomarker signal strength b_{pred}/b_{prog}. As expected, the predictive biomarker discovery performance is lower in terms of $|t_{pred}/t_{prog}|$ for $b_{pred}/b_{prog} > 1$, indicated by lower $|t_{pred}/t_{prog}|$, likely because the prognostic biomarker effects dominate over the predictive biomarker effects. In most cases, the $|t_{pred}/t_{prog}|$ values surpass those of the baseline models, especially in the range for lower b_{pred}/b_{prog}.

CMNIST results. The comparably high $|t_{pred}/t_{prog}|$ values indicate that the models were able to identify strong predictive imaging biomarkers for different settings with varying b_{pred}/b_{prog}. In addition, the smaller gap to the upper bound and the notably larger gap from both the baseline and the lower bound indicate that the models perform best on CMNIST among all four datasets. For example, the gap between baseline and CATE estimation model result reaches a factor of 10^2 for b_{pred}/b_{prog} in the range of 0 to 1. Comparing the results for models trained on biomarker feature set (a) ("model (a)") to the ones trained on feature set (b) ("model (b)") reveals that the results were more similar between models (a) and (b) than the other three datasets, where some of the results for $|t_{pred}/t_{prog}|$ vary considerably.

CUB-200-2011 results. The separation between the relative predictive strength $|t_{pred}/t_{prog}|$ of the CATE estimation model and the baseline is slightly smaller for the CUB-200-2011

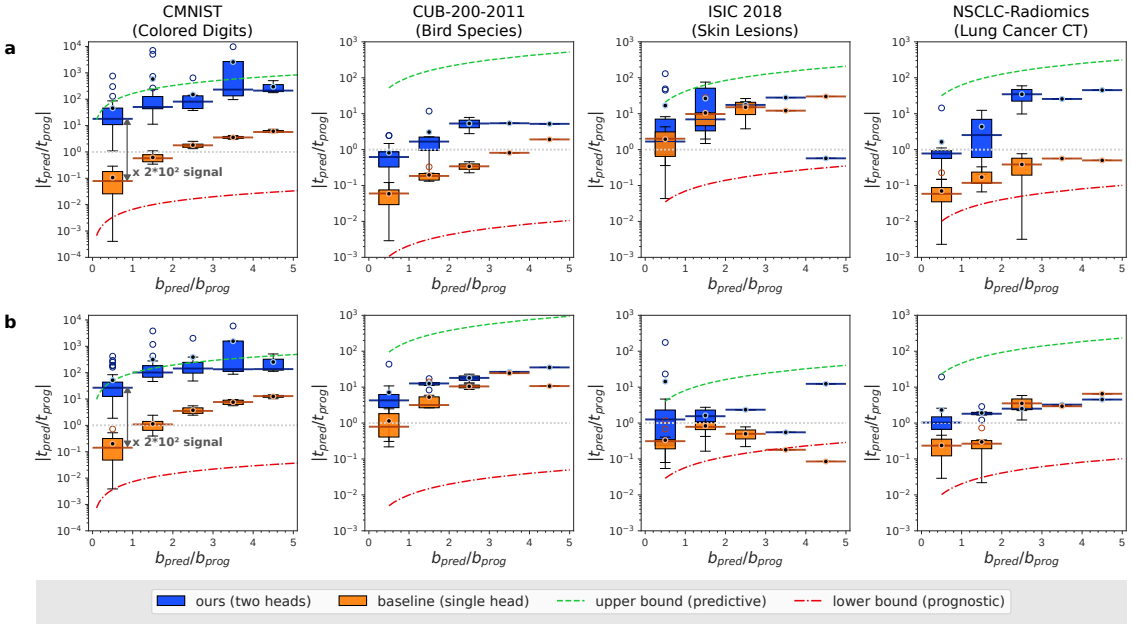


Figure 5.1: Model performance based on the relative predictive strength t_{pred}/t_{prog} of the estimated CATE, shown on a logarithmic scale. The two-headed TARNet-like CATE estimator is compared to a one-headed baseline model trained to predict the outcome regardless of treatment. The results are shown across different values of simulation parameters b_{pred}/b_{prog}. This ratio is the relative strength of the predictive effect versus the prognostic effect, both of which influence the simulated outcome. Boxplots summarize data averaged over b_{pred}/b_{prog}-bin widths, as indicated by the horizontal error bars over the median line. Rows (**a**) and (**b**) correspond to different sets of prognostic and predictive features used for generating the data (see Section 4.1.4 and Figure 4.2). Note that the variance of the boxplots is affected by the differing number of samples each bin contains. The horizontal gray dotted line marks where $t_{pred}/t_{prog} = 1$. ©2025 IEEE. Reprinted with permission from Xiao et al. (2025).

dataset than the CMNIST dataset, indicating that the higher complexity of the predictive and prognostic imaging biomarkers chosen for the bird species dataset ("primary color is white" and "bills is longer than head") also impacted the performance. For example, the median $|t_{pred}/t_{prog}|$ is a factor of 10 or 5 of the medians of the baseline for models (a) and (b) respectively for b_{pred}/b_{prog} between 0 and 1. The relative predictive strengths remain much closer to the upper bound than the lower bound and mostly above 1. Generally, even though the gap to the baselines is slightly larger for models (a) than (b), the absolute values are smaller for (a), which is further evidence of the dependency on the biomarker choice.

ISIC 2018 results. The relative predictive strength values $|t_{pred}/t_{prog}|$ of the skin lesion dataset ISIC 2018 results show a higher variability across b_{pred}/b_{prog} bins compared to the CUB-200-2011 dataset. Their mean values remain above the lower bound and mostly above 1, except for two outliers at high b_{pred}/b_{prog}, which are based on a single sample. The results also depend on which image features were chosen to be predictive or prognostic biomarkers. The models trained on feature set (a), where the presence of globules (i.e. "has globules") was predictive, had higher $|t_{pred}/t_{prog}|$ values that were much closer to the upper bound, compared to the models trained on feature set (b), where the predictive biomarker was the presence of pigment networks (i.e. "has pigment networks"). This indicated that the models (a) were able to identify a stronger predictive imaging biomarker. However, the overlap of the boxplots with the baseline was also greater for models (a) compared to (b), especially for low b_{pred}/b_{prog}. The large $|t_{pred}/t_{prog}|$ values of the baseline models suggest that their outcome predictions also strongly relied on the predictive biomarker. For models (b), the medians of the $|t_{pred}/t_{prog}|$ values differed by a factor of 4 for relative b_{pred}/b_{prog} in the range of 0 to 1.

NSCLC-Radiomics results. For the NSCLC-Radiomics, the image-based CATE estimation model showed an inconsistent behavior when trained with data generated with feature set (a), where the radiomics feature "energy" is predictive, compared to (b), where the radiomics feature "flatness" is predictive. While the relative predictive strengths $|t_{pred}/t_{prog}|$ were generally large and increasing with larger b_{pred}/b_{prog} for models (a), the gaps decreased for models (b). The smaller gaps indicate that the predictive imaging biomarkers of the NSCLC-Radiomics, especially the feature "flatness" were the most difficult features to extract for the models among all four datasets. The medians of the models (a) and (b) for b_{pred}/b_{prog} in the range of 0 to 1 were larger than the baseline by a factor of 13 and 4, respectively.

Heterogeneous treatment effect estimation performance. While the primary focus of this part of the thesis is to assess CATE estimation models by how well they perform at predictive imaging biomarker discovery, the models can also be evaluated by how accurately they estimate heterogeneous treatment effects. This is commonly assessed using metrics such as the root PEHE $\sqrt{\epsilon_{PEHE}}$ and the root mean squared error (RMSE) for

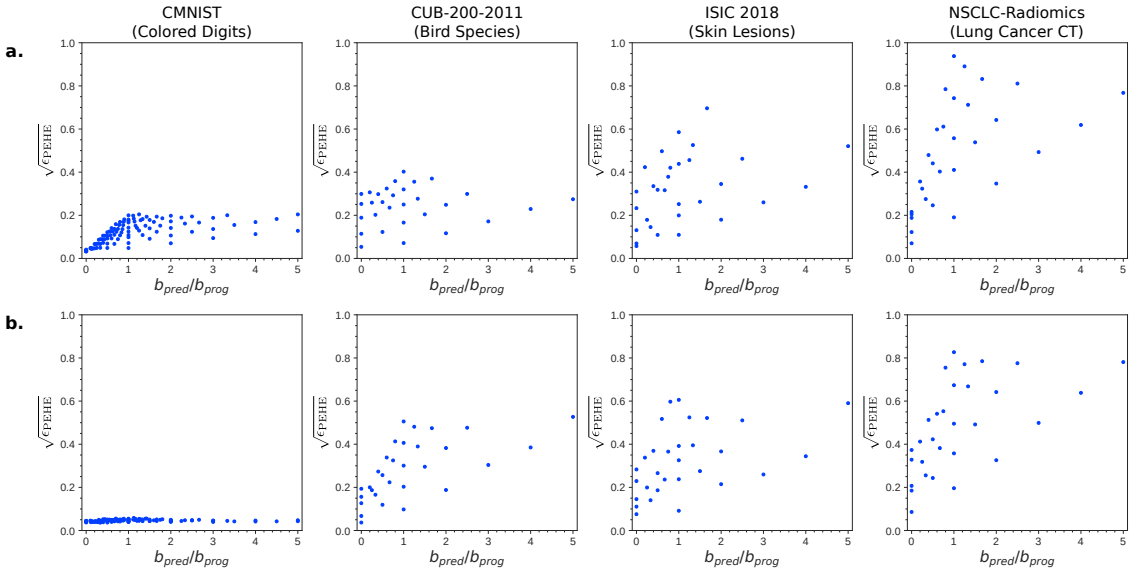


Figure 5.2: Performance of the CATE estimation models, evaluated with respect to the root precision of estimating heterogeneous effects (PEHE), denoted as $\sqrt{\epsilon_{PEHE}}$, across different simulation parameters b_{pred}/b_{prog}, which indicate the relative size of the predictive effect in the simulated outcomes. Lower values of $\sqrt{\epsilon_{PEHE}}$ indicate a better performance. The row (**a**) and (**b**) indicate the different sets of biomarker features used for generating the data. ©2025 IEEE. Adapted and reprinted with permission from Xiao et al. (2025).

Table 5.1: Performance of CATE estimation models trained with biomarkers from feature set (a) or (b) per dataset, evaluated on simulated outcomes. Metrics shown are the root PEHE $\sqrt{\epsilon_{PEHE}}$ for treatment effect estimation and RMSE for the prediction of factual outcomes only.

Dataset	Feature Set	$\sqrt{\epsilon_{PEHE}} \downarrow$	RMSE $\downarrow$
CMNIST	(a)	0.121	0.094
	(b)	0.045	0.115
CUB-200-2011	(a)	0.227	0.304
	(b)	0.277	0.261
ISIC 2018	(a)	0.304	0.352
	(b)	0.308	0.362
NSCLC-Radiomics	(a)	0.475	0.561
	(b)	0.469	0.633

factual outcome prediction, as introduced in Chapter 2. For completeness, these metrics are reported in Figure 5.2 and Table 5.1. This analysis aims to investigate whether the performance of predictive biomarker discovery, which is found to depend heavily on the dataset, can be linked to either treatment effect estimation performance or factual outcome prediction performance. Although these metrics are only of secondary interest, they help contextualize the observed differences in predictive biomarker discovery across datasets, and will also be relevant in the subsequent part of the thesis (Section 5.2), where they are related to the model performance in terms of making treatment recommendations.

The results show a lower PEHE and RMSE for the CMNIST compared to the other three datasets, which also corresponds to the higher performance for the relative predictive strength as observed in Figure 5.1. The root PEHE and RMSE are the highest for the NSCLC-Radiomics dataset, and the PEHE also has the highest variance, which matches the worst performance in identifying predictive biomarkers. Figure 5.2 and Table 5.1 also illustrate the variations within the same dataset between models (a) and (b). While the models trained on CMNIST and NSCLC-Radiomics have lower PEHE for feature set (b) compared to (a), the difference is smaller for CUB-200-2011 and ISIC 2018. The RMSE values, however, show that the factual prediction performance is better for models trained on feature set (a) compared to (b) for those two datasets. This suggests that the models trained on feature set (a) were worse at predicting counterfactual outcomes than those trained on (b), while the opposite is the case for CUB-200-2011.

Unlike the predictive strength shown in Figure 5.1, the PEHE increases (i.e. worsens) on average for increasing b_{pred}/b_{prog} as shown in Figure 5.2. The reason for this is likely the different scale of the actual CATE, which automatically changes with the absolute values of parameters b_{pred} and b_{prog} as noted by Crabbé et al. (2022), resulting also in changes in the scale of the root PEHE and RMSE. As the metrics depend on the scale of the outcomes, this limits the comparability across different settings b_{pred}/b_{prog} and further highlights the shortcomings of solely using PEHE or RMSE for the evaluation.

5.1.2 Interpreting Predictive Imaging Biomarker Candidates

The following paragraphs from this section are adapted from the article by Xiao et al. (2025), and therefore resemble the text of the original manuscript.

As mentioned RQ1.2, another goal of this part of the thesis is to demonstrate how the image features identified by an image-based CATE estimation model can be qualitative assessed and interpreted. This part of the evaluation is done using the XAI-based evaluation protocol described in Section 4.1.3, where attribution maps (Springenberg et al. 2015; Selvaraju et al. 2017; Sundararajan et al. 2017; Erion et al. 2021) are computed to highlight the predictive and prognostic features learned by the model for different input images.

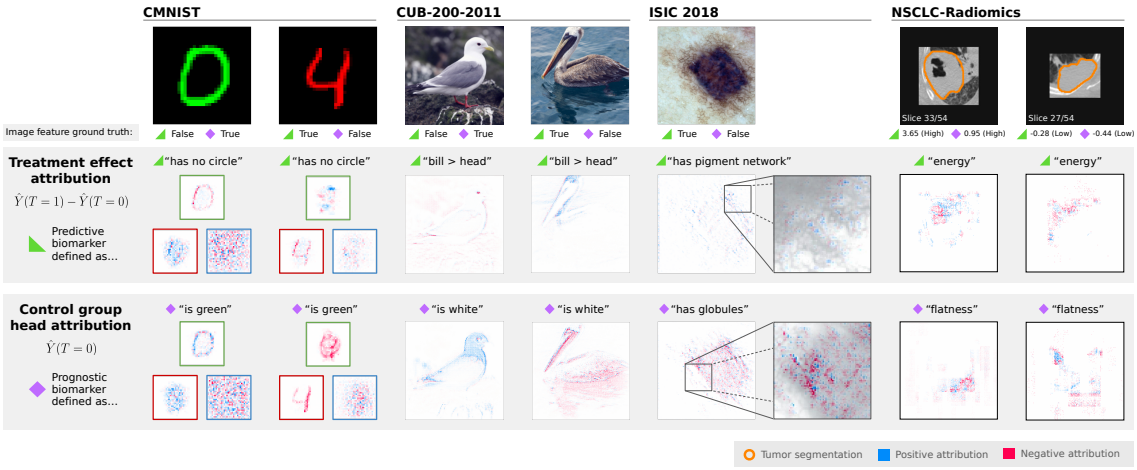


Figure 5.3: Attribution maps for the control group prediction head (last row) and the estimated CATE output (middle row) for different example images from each dataset (top row). For the CMNIST dataset, the attribution is shown for each RGB color channel (red: left, green: top, blue: right), as the color information is important for the biomarker prediction. An additional zoomed-in patch of the ISIC 2018 attribution map is overlaid with a grayscale version of the original image. For the NSCLC-Radiomics dataset, sagittal slices of the 3D patches are shown with segmented tumors outlined in orange. Here, results are based on models trained with $b_{pred}, b_{prog} = 1.0$. ©2025 IEEE. Adapted and reprinted with permission from Xiao et al. (2025).

To assess the performance of image-based CATE estimation model at identifying the correct predictive and prognostic imaging biomarkers, the analysis for this part of the thesis again relies on semi-synthetic datasets with known ground-truth imaging biomarkers initially used to simulate the outcomes. These ground-truth imaging biomarkers are then compared to the attribution maps, indicating positive (blue) and negative (red) contributions to the prediction.

Examples of such attribution maps, computed with respect to two different model outputs for the four datasets, are shown in Figure 5.3: the attribution maps for the predicted CATE $\hat{Y}(T=1) - \hat{Y}(T=0)$ ("treatment effect attribution"), which is expected to be sensitive only to the predictive biomarker (Figure 4.1b), and the attribution maps for control group head $\hat{Y}(T=0)$ ("control group head attribution"), which is expected to be sensitive to the prognostic biomarker.

CMNIST results. The qualitative results for the first example from the CMNIST dataset, showing a green digit zero, reveal that the treatment effect attribution maps are mostly negative for the green channel in the same ring-like shape as the object, and mostly positive for the red channel. For the second example image, showing a red digit four, the outline of the object can be observed as well, but in a different color channel (i.e. red). This suggests that, according to the CATE estimation model, the strokes of the

digits themselves did not contribute to the predictive imaging biomarker. While the red channel shows mostly negative attribution values, along with some noisy positive attribution in the background, more localized positive attribution values can be seen in the green channel, indicating that the model identified the gaps between the strokes of the digits as possible contributions of a predictive imaging biomarker. These observations for all channels combined mostly correspond to the ground truth predictive imaging biomarker "has no circle" being absent in the first example and present in the second one. However, the positive attribution from the red channel in the first example and the negative attribution from the red channel in the second example appear to contradict this observation when only considering a single channel in isolation.

The attribution maps of the control group head for channels green and red are mostly positive for the first CMNIST example and mostly negative for the second one, strongly indicating that the model found a prognostic image biomarker to be present or absent, respectively, and that it depends on the presence or absence of the green or red color channel. This observation directly corresponds to the prognostic imaging biomarker "digit is green", which also indicates that the CATE estimation model correctly identified the correct feature from the respective color channel.

For both input images and both outputs, the attribution maps are mostly noisy for the blue channel, suggesting that the CATE estimation model did not use this channel for predictions and did not identify relevant image features from it.

CUB-200-2011 results. For the CUB-200-2011 dataset, the treatment effect attribution map of the first input image shows mostly diffuse negative attributions with some heatmap pixels focused around the eye, as well as the outlines of the throat and breast of the bird. In contrast, the attribution map shows mostly positive and localized values outlining the area of the head, neck, and bill, and indicates that the other areas of the bird were mostly ignored. The attribution map patterns suggest that the image-based CATE estimation model identified the absence of the predictive imaging biomarker in the first example from the eye and general shape of the bird, while it identified the presence of the predictive imaging biomarker in the second example by the outline of the bill and head of the bird. Especially the attribution map for the second example image directly matches the ground truth predictive imaging biomarker "bill longer than head" and suggests that the model was sensitive to the correct region.

The control group head attribution map of the first input image is overall positive, indicating that features of the head, neck, and breast region were primarily used for the predictions, while the wings were largely ignored. The attribution of the second bird, in contrast, is mostly negative, particularly in the wing, main body, and pouch region, but also in the area of the reflections in the water. The regions with a strong positive attribution overlap with the regions where the bird is primarily white, and the regions with a strong negative attribution overlap with the regions where the bird (or its reflection) is dark. This suggests that the CATE estimation model identified a prognostic

imaging biomarker related to the color and brightness of the bird, which indeed directly corresponds to the ground truth prognostic imaging biomarker "is white".

ISIC 2018 results. Overlaying the treatment effect attribution map with the corresponding image from the image-based ISIC 2018 dataset and zooming in shows that positive attributions are given to the area surrounding the dark center of the skin lesion, especially the less pigmented gaps between the dark network-like structures. This suggests that the CATE estimation model identified the predictive imaging biomarker to be related to the gaps in the darker network or grid-like structure. The ground truth mask for the predictive imaging biomarker "has pigment network" reveals that the model was indeed able to identify the pigment network in the correct area and from the correct patterns.

The control group head attribution map displays strong and predominantly negative attributions to the darker center of the skin lesion, marked by red and blue spots. This indicates that the model identified the absence of prognostic imaging biomarkers from the darker patterns within the skin lesion, which matches the fact that the prognostic imaging biomarker "has globules" is indeed absent in the shown image example. However, due to the higher complexity of the imaging biomarkers, the attribution maps only provide a limited insight into what the corresponding image feature looks like, especially when the imaging biomarker is absent.

NSCLC-Radiomics results. The treatment effect attribution map of the first NSCLC-Radiomics example image slice in Figure 5.3 shows the highest absolute values within the tumor area, with negative attributions to the darker tumor regions and a larger region with positive attributions to the surrounding areas. Mostly negative attributions are observed for the second example, particularly in the upper left region of the tumor outline. These two attribution maps suggest that the predictive imaging biomarker identified by the image-based CATE estimation model is negatively correlated with regions of low image intensity but positively correlated with areas of very high image intensities, such as the very bright structures on the left side of the first example image. While the observations are consistent with the ground-truth predictive biomarker "energy", which has a higher value in the first example than in the second, attributions are also given to areas outside the tumor volume. This suggests that the model had some difficulty in correctly identifying the exact tumor boundary.

The control group head attribution maps show strong negative and positive attributions, primarily to areas outside the tumor outline, with a focus on specific regions, such as the bottom right region for the first example image slice and the top left region for the second example image slice. This indicates that the model identified the prognostic imaging biomarker from some localized areas around the tumor, but not from the whole outline. Additionally, attributions are assigned to regions surrounding the image patch, indicating that the patch shapes also partially contributed to the model predictions.

Attribution maps for additional lung tumor CT slices are provided for completeness in Figure 5.4 to further demonstrate how the CATE estimation model behaves for different

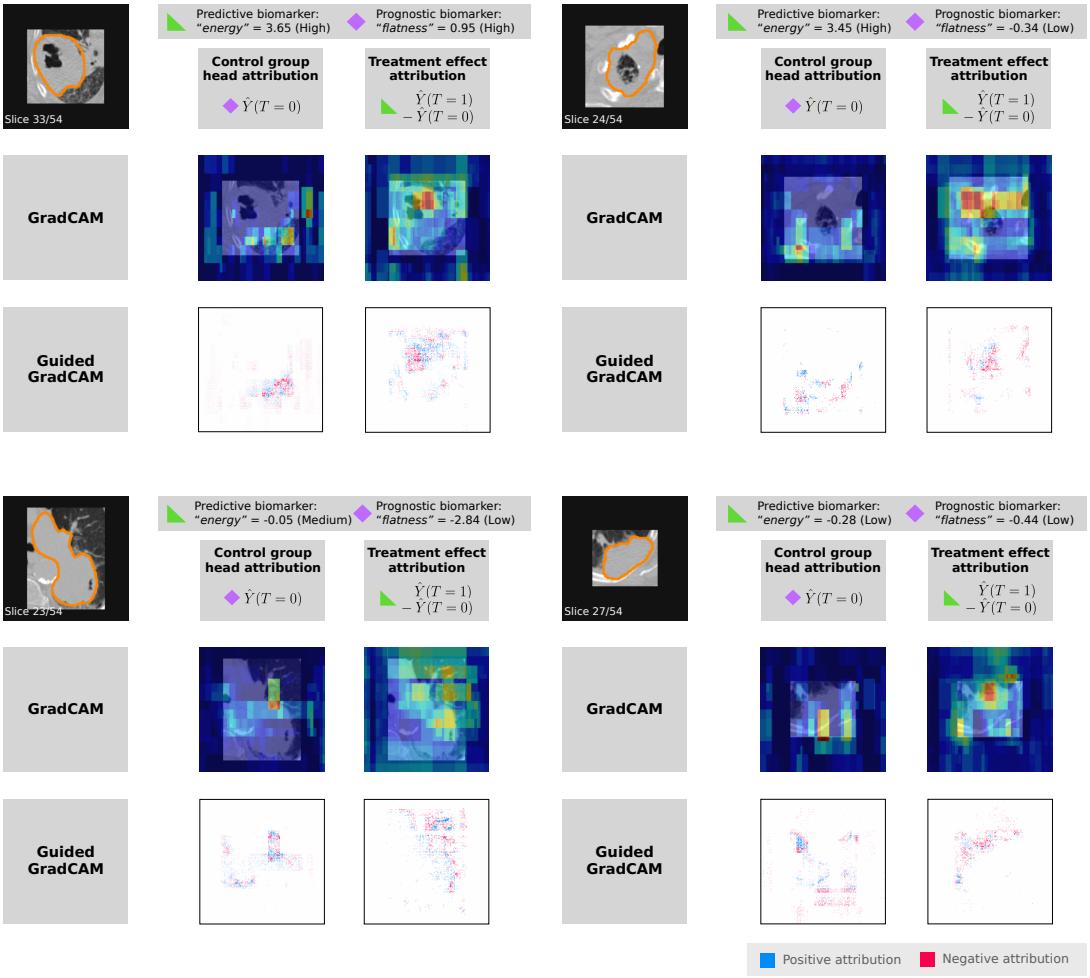


Figure 5.4: Attribution maps generated using Grad-CAM and Guided-Grad-CAM for the control head, and CATE prediction of the trained CATE estimation model for one sagittal slice of each of the four NSCLC-Radiomics dataset samples, showcasing the predictive and prognostic biomarker with varying strengths. The tumor segmentation outlines are shown in orange. ©2025 IEEE. Adapted and reprinted with permission from Xiao et al. (2025).

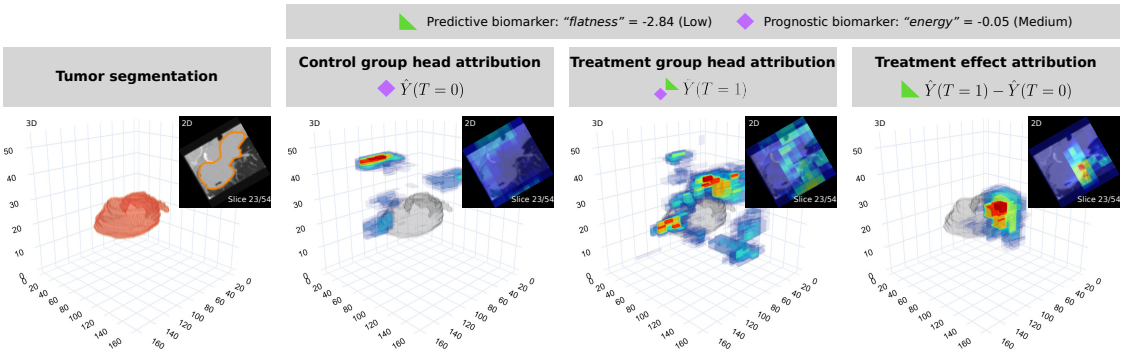


Figure 5.5: Three-dimensional Grad-CAM attribution maps of the trained CATE estimation model for the control head, treatment group head, and CATE prediction, illustrated for a 3D patch from the NSCLC-Radiomics dataset. Additionally, a 3D render of the segmented tumor and a corresponding 2D sagittal slice are shown for reference in gray. ©2025 IEEE. Adapted and reprinted with permission from Xiao et al. (2025).

types of lung tumors. The Grad-CAM treatment effect attribution maps, where the color scale is based on absolute attribution scores, show that the model focuses more on regions with low intensities in the image slices. The corresponding Guided-Grad-CAM attribution maps again reveal strong negative attributions to areas within the tumor with very low intensities and strong positive attributions to the lighter areas neighboring those low-intensity regions, similar to the examples shown previously in Figure 5.3. Both Grad-CAM and Guided-Grad-CAM attribution maps again show that the highest importance is given to localized areas surrounding the tumors.

To provide deeper insights otherwise not captured by the 2D slices, attribution maps are also shown in 3D in Figure 5.5, in addition to a 3D render of the tumor volume itself. It should be noted that the image features for the prognostic and predictive biomarkers are reversed compared to the previous figures. Here, the "energy" is prognostic and "flatness" is predictive. The 3D treatment attribution maps show strong attribution to the upper right side of the tumor, likely erroneously, while the 3D control group head attribution maps show attribution to areas outside the tumor boundaries.

Overall, the observations for the attribution maps highlighting areas with the lowest intensity and neighboring areas with higher intensities largely align with the fact that the minimum pixel intensity value contributes strongly to the image feature "energy". However, the observations also demonstrate the model's difficulty in accurately localizing the tumor. The observation that the control group head attribution in Figure 5.3 and Figure 5.4 or treatment effect attribution in Figure 5.5 only highlights localized areas on the tumor border is consistent with the fact that only the principal components contribute to the image feature "flatness". The qualitative results emphasize the challenges in interpreting predictive and prognostic imaging biomarkers when the ground-truth imag-

ing biomarkers are unknown. The figures also highlight the importance of considering both 2D slices and 3D attribution maps to support the assessment of identified imaging biomarkers.

5.2 Image-Based Heterogeneous Treatment Effect Estimation in Clinical Imaging Studies

To bridge the gap between the image-based CATE estimation approaches investigated in the previous part Section 5.1 and their application in real clinical imaging studies, this section's experiments assess the methodological extensions of the earlier deep-learning-based models (Section 4.1.2) to support survival outcomes and multimodal inputs (Section 4.2). Here, the underlying goal is to answer the following research questions previously presented in Section 1.2:

RQ 2.1: Can image-based heterogeneous treatment effect estimation methods be extended from categorical or continuous outcomes to survival (time-to-event) outcomes, and how does their treatment recommendation performance compare to binary-outcome models?

RQ 2.2: Can the integration of multimodal inputs or pre-trained image encoders improve treatment effect estimation performance and robustness on clinical imaging data?

RQ 2.3: To what extent can image-based heterogeneous treatment effect estimation models be applied to glioblastoma MRI data from a randomized clinical trial, and what are their limitations and implications for predictive imaging biomarker discovery?

The experiments use two different clinical imaging study datasets (see Section 4.2.1 for details). The first is the semi-synthetic lung cancer CT dataset NSCLC-Radiomics with simulated ground-truth treatment effects, the second is the brain cancer MRI dataset EORTC from a glioblastoma RCT with unknown treatment effects. Because the interpretation of the results for both datasets differs noticeably, the results are presented separately for clarity in two different subsections: Section 5.2.1 and Section 5.2.2.

Although presented separately, both subsections follow a similar parallel structure. First, RQ2.1 is addressed by directly comparing the results of a binary-outcome CATE estimation

model with a survival-outcome model with binarized predictions ("Using Binary vs. Survival Outcomes"). Then, for RQ2.2, the impact of using different input configurations, for example, with or without clinical tabular data, is assessed ("Value of Multimodal Integration for Treatment Effect Estimation"). The results for pre-trained image encoders are presented only in Section 5.2.2. Both subsections conclude with an assessment of model robustness and limitations to address RQ2.3, for example, by comparing to simple regression baselines.

Most evaluations focus on assessing the performance of heterogeneous treatment effect estimation and the resulting treatment recommendations (using empirical estimates whenever the ground-truth treatment effect is unavailable), as well as whether the resulting subgroup stratification is expected to improve the overall survival outcomes. Additionally, factual prediction quality metrics for survival prediction or classification are also presented. Finally, for an exploratory analysis, a limited predictive imaging biomarker analysis is provided specifically for Section 5.2.2.

5.2.1 Baseline Experiments on Semi-Synthetic Survival Data

The experiments in this subsection used the NSCLC-Radiomics dataset with simulated ground-truth individual treatment effects to establish a performance baseline for the proposed image-based CATE estimation approach for survival outcomes before moving on to RCT outcomes of real patients in Section 5.2.2. To this end, the treatment effects were simulated using the radiomics feature "flatness" as a predictive imaging biomarker extracted from the tumor region of the images. In this setup, information about the treatment effect was only contained in the image data. The semi-synthetic outcomes were then generated from the real survival outcomes by scaling them with a predictive imaging biomarker-dependent factor, as described in more detail in Section 4.2.1. Since the ground-truth ITEs and therefore also the resulting optimal treatment recommendations were available here, oracle metrics such as the PEHE and decision accuracy, which require the ground-truth ITEs, could be computed. This also enabled a more detailed quantitative performance assessment, providing more accurate insights compared to empirical metrics like observed policy risk or policy value, which were computed from observed outcomes alone.

The experiments first compare the performance of a model trained on binary outcomes versus survival outcomes after binarizing the predictions, then they assess whether integrating tabular clinical data or ground-truth tumor segmentation masks as an input provides a benefit compared to a model for image inputs, and finally, they provide insights regarding the reliability and limitations of the proposed model and its treatment recommendations.

Using Binary vs. Survival Outcomes

The proposed image-based CATE estimation models for binary survival classification outcomes, trained with a binary cross-entropy loss function, were directly compared to survival-outcome models trained using the negative Cox partial log-likelihood loss (Section 4.2.2) to investigate RQ2.1. This comparison evaluates whether performing treatment effect estimation on survival outcomes yields better and more nuanced information for making treatment recommendations compared to framing it as a binary classification task, which is often simpler to optimize but discards not only time-to-event information but also censoring information.

The general evaluation setup is described in Section 4.2.4. It also includes the specific details for evaluating the survival-outcome models, for which the predictions were converted to binary outcomes to match the metric scale of the binary-outcome models, enabling a direct comparison.

Table 5.2: Comparison of CATE models trained on survival versus binary outcomes on the NSCLC-Radiomics dataset. Reported are the fraction of correctly assigned treatments (Decision Accuracy), root PEHE ($\sqrt{\epsilon_{PEHE}}$) and the observed policy risk $\hat{R}_{Pol}$, as well as Balanced Accuracy, F1, AP and AUROC, with mean $\pm$ SD across folds. To enable a direct comparison, all metrics except Decision Accuracy are computed using binarized survival outcomes (threshold: 365 d). An asterisk (*) indicates models trained on continuous survival outcomes whose predictions were post hoc binarized for evaluation. All metrics are IPCW-adjusted.

a Treatment effect estimation and recommendation performance.

Split	Model Type	Decision Acc ↑	$\sqrt{\epsilon_{PEHE}}$ ↓	$\hat{R}_{Pol}$ ↓
Val.	Binary	0.48 ± 0.09	**0.49 ± 0.11**	0.71 ± 0.03
	Survival*	**0.56 ± 0.10**	0.57 ± 0.10	**0.64 ± 0.05**
Test	Binary	0.50 ± 0.01	**0.47 ± 0.05**	0.67 ± 0.00
	Survival*	**0.53 ± 0.02**	0.51 ± 0.02	**0.64 ± 0.03**

b Factual outcome prediction performance.

Split	Model Type	Balanced Acc	AUROC ↑	F1 ↑	AP ↑
Val.	Binary	**0.55 ± 0.04**	0.59 ± 0.07	**0.789 ± 0.017**	0.71 ± 0.05
	Survival*	0.54 ± 0.06	**0.62 ± 0.06**	0.751 ± 0.026	**0.75 ± 0.04**
Test	Binary	0.51 ± 0.01	0.49 ± 0.03	**0.790 ± 0.009**	0.68 ± 0.02
	Survival*	**0.52 ± 0.02**	**0.57 ± 0.02**	0.788 ± 0.005	**0.74 ± 0.01**

The results in Table 5.2 show that the survival-outcome models consistently achieve a higher decision accuracy and lower empirical policy risk than the binary-outcome models across validation and hold-out test sets, even though the decision accuracy is only slightly above chance level (0.5). Both models have a high variability across folds, especially in terms of validation decision accuracy and root PEHE.

The root PEHE, which has a theoretical upper bound of 2 on the probability scale (since $\tau, \hat{\tau} \in [-1, 1]$), remains moderate for both model types and is lower for the binary-outcome than the survival-outcome model. However, this difference is difficult to interpret as the models use different definitions of the probability scale to compute the treatment effect. As noted in the evaluation details Section 4.2.4, the binary-outcome models use classification probabilities, whereas the binarized survival-outcome models use the survival probabilities at one year.

There is no consistent winner for the factual prediction metrics. While the AUROC and AP scores are higher for the survival-outcome model, the F1 score is higher for the binary-outcome model. The balanced accuracy score is similar on both splits and only slightly exceeds 0.5 in all cases. The relatively high F1 and AP scores, in contrast, are likely driven by the strong class imbalance in the data and the resulting tendency of both models to predominantly predict the positive class "long survival". While some metrics (balanced accuracy, AUROC, and AP) show a drop in performance from validation to test set results, other metrics remain similar.

Overall, the survival-outcome models demonstrate a small but consistent advantage in treatment recommendation metrics, suggesting that modeling survival outcomes can yield slight improvements in treatment recommendations at least for the semi-synthetic setting of the NSCLC-Radiomics dataset. However, the evidence in favor of RQ2.1 is limited, since neither approach produces highly accurate predictions.

Value of Multimodal Integration for Treatment Effect Estimation

Motivated by the limited treatment recommendation and factual prediction performance of image-based CATE estimation models presented in the previous experiments, as seen by the low decision accuracy and balanced accuracy, this study also explored whether additional input data could improve these metrics, addressing RQ2.2.

As the treatment effects for the semi-synthetic NSCLC-Radiomics outcomes are generated directly using a single image feature (i.e. the radiomics feature "flatness"), all treatment effect information is contained in the images by construction. Nevertheless, the available clinical tabular data might provide further prognostic information, as explained in Section 4.2.1, which is relevant for predicting factual outcomes. Additionally, the available ground-truth segmentation maps for the lung tumors may provide both prognostic and treatment effect information. While segmentation masks highlight the relevant regions,

they can also help guide the model in learning the shape of the tumor and thus identify the predictive imaging biomarker "flatness".

To test these hypotheses, both binary-outcome and survival-outcome image-based CATE estimation models were trained with different combinations of tabular data and segmentation masks as additional inputs using the same hyperparameter settings as described in Section 4.2.2.

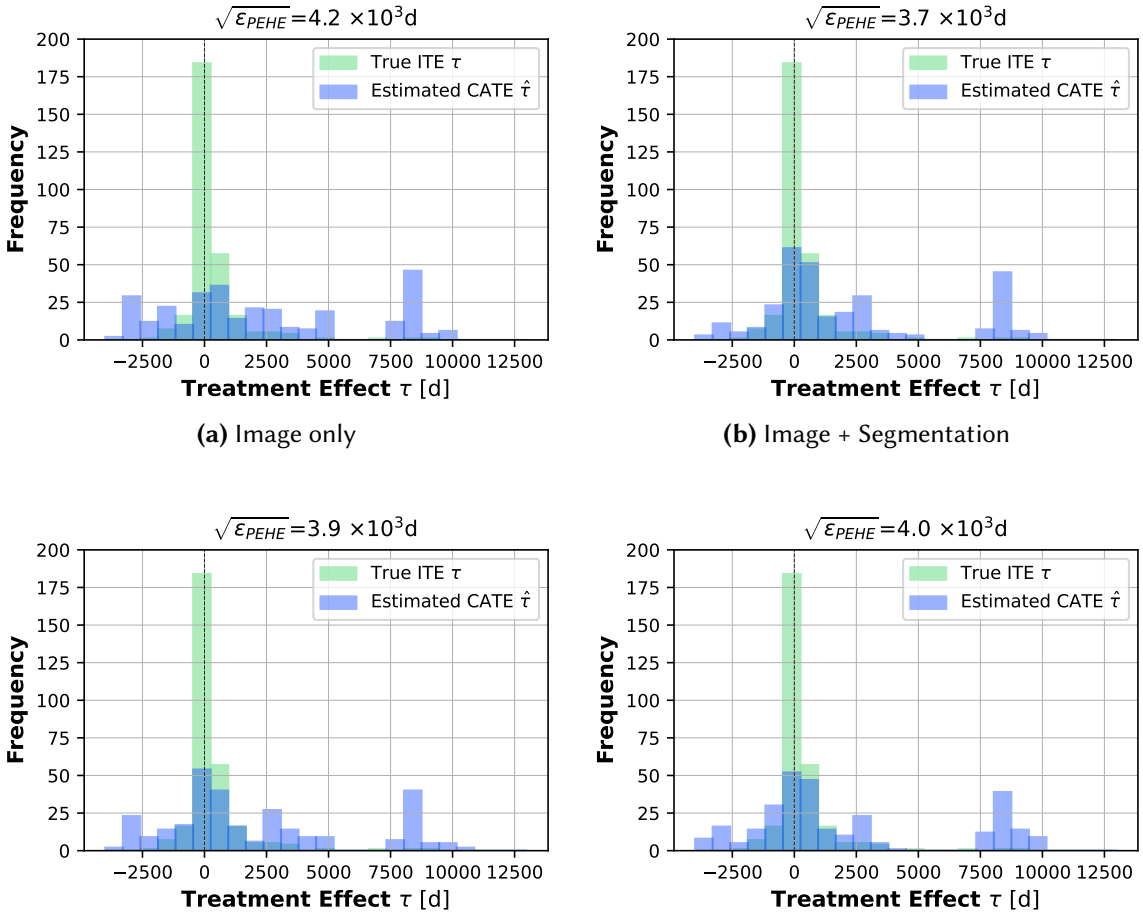
Table 5.3: Comparison of CATE estimation models trained with different combinations of input modalities (tabular data and tumor segmentation mask "Seg.") on the NSCLC-Radiomics dataset. Reported are the fraction of correctly assigned treatments (Decision Accuracy), root PEHE ($\sqrt{\epsilon_{PEHE}}$), observed policy risk $\hat{R}_{Pol}$ or policy value $\hat{V}_{Pol}$, as well as the Balanced Accuracy and Antolini's C-index, with mean $\pm$ SD across folds. All metrics are IPCW-adjusted, except for the C-index.

a Performance of binary-outcome CATE estimation models.

Split	Modalities			Decision Acc ↑	$\sqrt{\epsilon_{PEHE}}$ ↓	$\hat{R}_{Pol}$ ↓	Balanced Acc ↑
	Image	Segm.	Tabular				
Val.	✓	-	-	0.48 ± 0.09	0.49 ± 0.11	0.71 ± 0.03	0.552 ± 0.045
	✓	-	✓	0.46 ± 0.07	**0.46 ± 0.10**	**0.69 ± 0.06**	0.548 ± 0.021
	✓	✓	✓	**0.50 ± 0.08**	0.49 ± 0.10	0.70 ± 0.01	**0.557 ± 0.041**
Test	✓	-	-	0.50 ± <0.01	0.47 ± 0.05	0.67 ± 0.00	0.508 ± 0.015
	✓	-	✓	**0.51 ± 0.02**	**0.43 ± <0.01**	**0.65 ± 0.03**	**0.510 ± 0.006**
	✓	✓	✓	0.50 ± 0.01	0.44 ± <0.01	0.67 ± 0.00	0.499 ± 0.008

b Performance of survival-outcome CATE estimation models.

Split	Modalities			Decision Acc ↑	$\sqrt{\epsilon_{PEHE}}$ ↓ $[10^3\text{d}]$	$\hat{V}_{Pol}$ ↑ $[10^3\text{d}]$	C-Index ↑
	Image	Segm.	Tabular				
Val.	✓	-	-	**0.56 ± 0.10**	4.3 ± 1.8	0.37 ± 0.13	0.564 ± 0.029
	✓	✓	-	0.55 ± 0.09	**3.9 ± 2.0**	0.38 ± 0.14	**0.583 ± 0.035**
	✓	-	✓	0.54 ± 0.09	4.0 ± 1.8	**0.46 ± 0.14**	0.568 ± 0.034
	✓	✓	✓	0.53 ± 0.10	4.1 ± 2.2	0.32 ± 0.05	0.579 ± 0.028
Test	✓	-	-	0.53 ± 0.02	**2.1 ± 0.3**	0.36 ± <0.01	0.474 ± 0.017
	✓	✓	-	0.55 ± 0.03	2.2 ± 0.5	**0.52 ± 0.37**	0.500 ± 0.003
	✓	-	✓	0.50 ± 0.04	2.5 ± 0.7	0.35 ± 0.03	0.456 ± 0.026
	✓	✓	✓	**0.58 ± <0.01**	2.7 ± 0.2	0.35 ± <0.01	**0.508 ± 0.007**



(a) Image only

(b) Image + Segmentation

(c) Image + Tabular

(d) Image + Segmentation + Tabular

Figure 5.6: Histograms showing the distribution of true individual treatment effects τ (green) versus the estimated CATEs $\hat{\tau}$ (blue) for four different combinations of input modalities on the NSCLC-Radiomics dataset. The vertical dashed line indicates the zero point, where no treatment benefit is expected. Results are shown for the validation splits, aggregated across all cross-validation folds. The root PEHE ($\sqrt{\epsilon_{PEHE}}$) scores are reported above the plots for reference.

The results are reported for the metrics related to treatment effect estimation, as well as the factual prediction performance metrics for reference, including balanced accuracy for classification and Antolini's C-index for survival analysis, in Table 5.3.

For the binary-outcome models, adding tabular information resulted in a slight improvement across all metrics in both the validation and hold-out test sets, except for the validation balanced accuracy (Table 5.3a). Including the segmentation mask as an extra channel does not generally help. While it led to a minor increase in validation decision accuracy from 0.48 ± 0.09 to 0.50 ± 0.08, the test balanced accuracy even decreased in performance from 0.508 ± 0.015 to 0.499 ± 0.008. Generally, both decision accuracy and balanced accuracy exhibit performance close to random chance (0.5), indicating limited treatment recommendation and survival classification performance. Some standard deviations in the results are zero due to identical predictions across folds, which can occur with small datasets and thresholded outputs or identical treatment policies.

Integrating additional modalities into the survival-outcome models does not consistently improve results across metrics and splits, as shown in Table 5.3b. For example, the image-only model achieves the best decision accuracy on the validation set. On the test set, however, the model that additionally takes segmentation and tabular inputs achieves the best decision accuracy score. In addition, the PEHE and its variation across folds are high with values up to 4.3×10^3 d, corresponding to an error for the treatment effect of around 12 years.

Although both the balanced accuracy and the C-index have a validation score above chance, larger than 0.5, the drop from validation to test results suggests that both the binary-outcome and the survival-outcome models generalize poorly.

To give a better picture of the treatment effect distribution, histograms of the estimated CATE by the survival-outcome models and the corresponding ground-truth ITE are plotted in Figure 5.6. While the true ITE has a large peak around 0 d, the estimated CATE is more spread out, with outliers at high values around 8000 d. This indicates that the models tend to highly overestimate the treatment effect for certain cases, while underestimating it for some others. The histogram for the image-only has the lowest overlap with the ground-truth ITE and is more spread out compared to, for example, the model trained on image and segmentation inputs, which shows a higher peak at 0 d. This observation is also consistent with the lower PEHE score of models trained on multimodal inputs, suggesting integrating tabular or segmentation data has an impact on the estimated CATE for cases with very high or low treatment effects, even though this might not directly affect the decision accuracy if the sign of the estimated CATE stays the same.

Additional ablation results presented in Appendix Table B.1 indicate that replacing the simple multimodal fusion method of concatenating image representations and tabular data with a more complex fusion method using the DAFT module (Wolf et al. 2022) leads

to even worse results. Additionally, setting the regularization hyperparameter α used for the BITES loss to 0.01 instead of 0.0 resulted in worse general performance, albeit a better policy value, but with increased variation across folds.

Overall, the results provide only weak support for RQ2.2 and indicate that for the NSCLC-Radiomics dataset, the models are not able to utilize the additional information from tabular and segmentation data well for treatment effect estimation. While binary-outcome models benefit slightly from including tabular data, survival-outcome models show inconsistent gains and no clear improvement. Moreover, the significant decrease in performance from validation to test indicates that generalizability remains an important challenge, particularly for the survival-outcome models, which perform less reliably than the binary-outcome models.

Model Reliability and Baseline Comparison

The results from previous experiments indicated limitations in performance, motivating a more detailed investigation to make an assessment for RQ2.3 before moving on to RCT data. In the following experiments, the limitations of the proposed CATE estimation model are further examined. For this purpose, the performance of both binary-outcome and survival-outcome models is compared against a tabular-only regression baseline, analyzing the reliability of the resulting treatment recommendations and conducting a limited investigation into whether the models can identify the ground-truth predictive imaging biomarker.

Comparison to regression and alternative baselines. The proposed image-based CATE estimation models are compared against simple regression models (Section 4.2.3) to evaluate whether using image inputs with a deep-learning approach provides a benefit over using clinical tabular data only with a regression approach. Table 5.4 shows the results for the best-performing image-based deep-learning model with a TARNet-like architecture (denoted as "Bin-TARNet" for binary outcomes or "Surv-TARNet" for survival outcomes), selected based on the best validation decision accuracy. These are presented alongside the logistic regression T-learner results ("Logistic Reg.") for binary outcomes and the Cox proportional hazards regression T-learner results ("Cox PH"), fitted using 11 clinical covariates, which are expected to provide only prognostic information by simulation design, with and without the predictive imaging biomarker "flatness" as an additional covariate.

The results in Table 5.4 show that regression using the clinical tabular data plus "flatness" consistently outperforms the deep-learning-based TARNet, even though the TARNet has access to the same predictive imaging biomarker feature in theory and, in the Bin-TARNet case, additional information from imaging data and segmentation masks beyond the clinical tabular covariates. The regression model using clinical tabular data and "flatness" also

Table 5.4: Comparison of the proposed deep-learning-based CATE estimation models using image inputs (configuration selected based on validation performance) with a regression-based T-Learner CATE estimation baseline trained on tabular clinical data only, with and without the predictive imaging biomarker "flatness" included in the tabular inputs. The "flatness" radiomics feature was used to simulate the treatment effects on the NSCLC-Radiomics outcomes. "Bin-TARNet" denotes the binary-outcome CATE estimation model, whereas "Surv-TARNet" denotes the survival-outcome CATE estimation model. Reported are the fraction of correctly assigned treatments (Decision Accuracy), root PEHE ($\sqrt{\epsilon_{PEHE}}$), observed policy risk $\widehat{R}_{Pol}$ or policy value $\widehat{V}_{Pol}$, as well as the Balanced Accuracy and Antolini's C-index, with mean $\pm$ SD across folds. All metrics are IPCW-adjusted, except for the C-index.

a Performance of binary-outcome CATE estimation models.

Split	Method	Modalities	Decision Acc ↑	$\sqrt{\epsilon_{PEHE}}$ ↓	$\widehat{R}_{Pol}$ ↓	Balanced Acc ↑
Val.	Logistic Reg.	Tabular	0.54 ± 0.04	0.44 ± 0.09	0.68 ± 0.06	0.49 ± 0.05
	Logistic Reg.	Tab. + Flatness	**0.84 ± 0.02**	**0.40 ± 0.08**	**0.64 ± 0.04**	**0.56 ± 0.06**
	Bin-TARNet	Img. + Seg. + Tab.	0.50 ± 0.08	0.49 ± 0.10	0.70 ± 0.01	**0.56 ± 0.04**
Test	Logistic Reg.	Tabular	0.46 ± 0.03	0.42 ± 0.02	0.66 ± 0.02	0.56 ± 0.04
	Logistic Reg.	Tab. + Flatness	**0.74 ± 0.06**	**0.38 ± 0.02**	**0.64 ± 0.02**	**0.59 ± 0.02**
	Bin-TARNet	Img. + Seg. + Tab.	0.50 ± 0.01	0.44 ± <0.01	0.67 ± 0.00	0.50 ± <0.01

b Performance of survival-outcome CATE estimation models.

Split	Method	Modalities	Decision Acc ↑	$\sqrt{\epsilon_{PEHE}}$ ↓ [10^3d]	$\widehat{V}_{Pol}$ ↑ [10^3d]	C-Index ↑
Val.	Cox PH	Tabular	0.51 ± 0.13	2.0 ± 0.9	0.68 ± 0.30	0.55 ± 0.03
	Cox PH	Tab. + Flatness	**0.74 ± 0.11**	**1.6 ± 0.7**	**1.13 ± 0.32**	**0.61 ± 0.02**
	Surv-TARNet	Image only	0.56 ± 0.10	4.3 ± 1.8	0.37 ± 0.13	0.56 ± 0.03
Test	Cox PH	Tabular	0.47 ± 0.05	1.8 ± 0.1	0.83 ± 0.44	0.55 ± 0.04
	Cox PH	Tab. + Flatness	**0.70 ± 0.06**	**1.6 ± 0.1**	**1.21 ± 0.37**	**0.59 ± 0.04**
	Surv-TARNet	Image only	0.53 ± 0.02	2.1 ± 0.3	0.36 ± <0.01	0.47 ± 0.02

clearly outperforms the clinical tabular-only baseline, which is expected since "flatness" is the true predictive biomarker in this setting.

For binary-outcome models, even the clinical tabular-only regression baseline achieves a better score for decision accuracy, PEHE, and policy risk compared to the image-based model on the validation set, as well as for the PEHE, policy risk, and the balanced accuracy on the test set. The only metric where the TARNet performs better than the clinical tabular-only baseline is the validation balanced accuracy, which it also matches with the clinical tabular data plus "flatness" regression model.

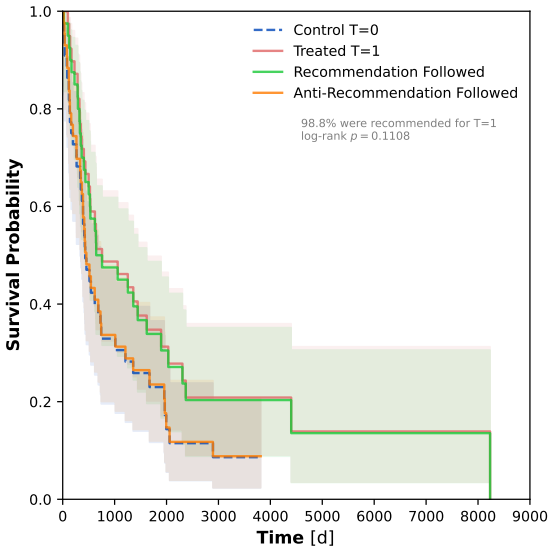
For the survival-outcome models, the clinical tabular-only regression baseline outperforms the Surv-TARNet in terms of PEHE and policy value, but not in terms of decision accuracy. The C-index of the survival-outcome TARNet is slightly higher than the Cox PH baseline on the validation set but drops below 0.5 on the test set, again indicating poorer generalization.

For binary and survival outcomes, the regression models show a better generalization for factual prediction metrics (i.e. balanced accuracy and C-index) compared to the deep-learning models. It is also notable that the decision accuracy is higher for binary-outcome regression (logistic regression) compared to survival-outcome regression (Cox PH), likely reflecting that regression using simplified survival class labels has a greater stability. This trend is reversed for the deep-learning models, which is consistent with the finding from the direct comparison between binary-outcome and survival-outcome models that the image-based TARNet appears to benefit from a richer supervision from time and censoring information for treatment recommendations, although these gains do not translate to factual prediction metrics.

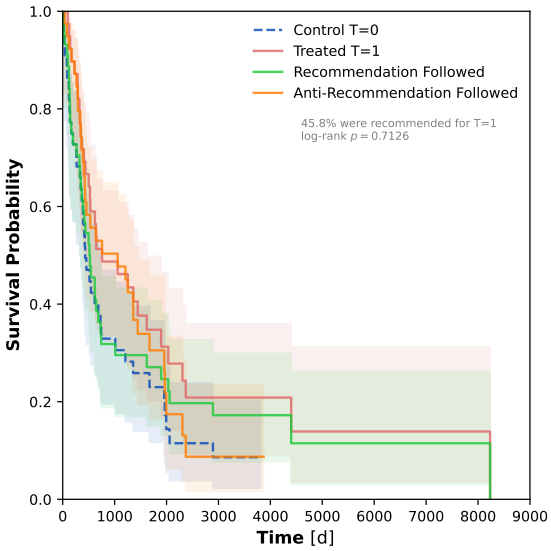
Both balanced accuracy and C-index remain only slightly above 0.5 for all regression models, indicating that the available clinical covariates do not provide sufficient information for strong factual outcome prediction performance, and that other relevant prognostic factors may be missing.

Overall, the results show no clear performance benefit of image-based deep-learning CATE estimation models over regression CATE estimation models, even when the deep-learning model should, in theory, have access to the predictive imaging biomarker through the image data. This suggests that, despite having access to all relevant information, deep-learning models are unable to effectively capture the predictive imaging biomarker signal in this setting for the NSCLC-Radiomics dataset.

Reliability of survival analysis and treatment recommendations. For an analysis of the quality of the final treatment recommendations derived from the ensembled estimated CATE by the best-performing image-based CATE estimation models (determined by validation decision accuracy), Kaplan-Meier curves were plotted to illustrate the ex-



(a) Recommendations by binary-outcome CATE estimation model.



(b) Recommendations by survival-outcome CATE estimation model.

Figure 5.7: Kaplan-Meier curves on the NSCLC-Radiomics dataset comparing the survival probability for patients who received the treatment recommended by the estimated CATE (green) versus those who did not (orange). For reference, the curves for the treated group ($T = 1$, red) and control group ($T = 0$, blue) are also shown. Results are shown on the hold-out test set using ensembled cross-validation models selected based on validation performance. Log-rank p-values are reported for reference.

pected impact of the recommendations on the survival probability of patients in the test set (see Section 2.3.1)

Patients with recommended and anti-recommended treatments. Figure 5.7 shows the Kaplan-Meier curves of the observed subset of patients who received the same treatment as the one recommended by the model ("recommendation followed"), as well as the curves of the subset of patients who received the opposite treatment ("anti-recommendation followed"), following the evaluation procedure of Katzman et al. (2018) and Schrod et al. (2022). In an ideal scenario, where the recommendations lead to an improvement in survival probability compared to randomly assigned treatments in an RCT, effective recommendations would lead to a "recommendation followed" Kaplan-Meier curve that lies above the treated and control group curves, whereas the "anti-recommendation curve" would lie below.

The binary-outcome model (Figure 5.7 (a)) recommended the treatment ($T = 1$) for almost all patients (98.8%), resulting in the curve for the "recommendation followed" subset to almost coincide with the curve of the treated group and the one of the "anti-recommendation followed" subset to almost coincide with the control group curve. As the treatment and control group curves already showed a separation, this likely indicates that the model made recommendations purely based on the average treatment effect, rather than the true predictive imaging biomarker.

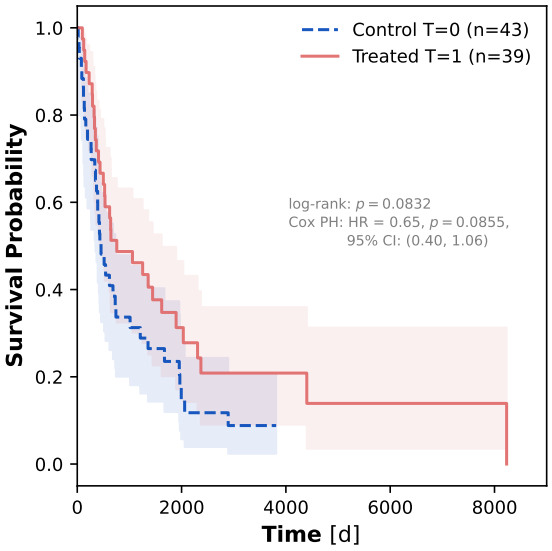
The survival-outcome model (Figure 5.7 (b)) recommended the treatment for only 45.8% of patients. This led to a worse separation between the "recommendation followed" curve and the"anti-recommendation followed" curve with an intersection between them, and the "recommendation followed" subset showing a lower survival probability than the originally treated group.

The log-rank tests did not yield significant differences between the "recommendation followed" and "anti-recommendation" groups for either model ($p = 0.1108$ for binary-outcome, $p = 0.7126$ for survival-outcome model). Further, the wide confidence intervals of the curves, reflecting the small sample size, limit the interpretability of these results.

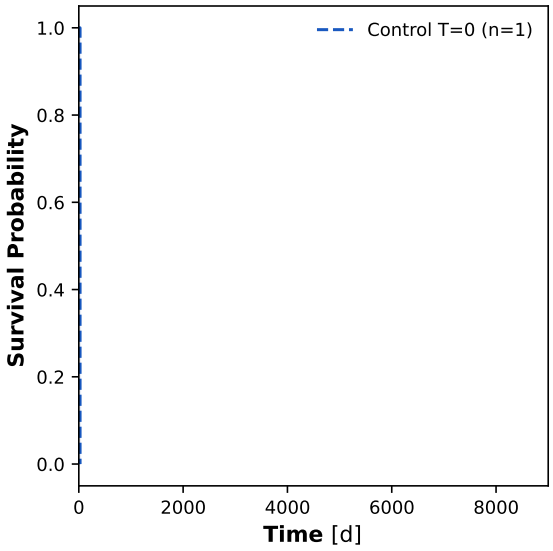
Recommended subgroups. To further investigate if the patient subgroups identified by the models truly benefit from their respective recommended treatment (i.e. $T = 1$ for the $\hat{\tau}_i > 0$ subgroup and $T = 0$ for $\hat{\tau}_i \leq 0$), additional Kaplan-Meier curves for the treatment arms are shown for each subgroup in Figure 5.8.

In the ideal scenario, the $\hat{\tau}_i > 0$ subgroup benefits from the treatment $T = 1$, so that the Kaplan-Meier curve of the treated subset of that subgroup would lie above the one of the control subset. In contrast, the $\hat{\tau}_i \leq 0$ would not benefit from $T = 1$, leading to a lower treated subset curve compared to the control subset.

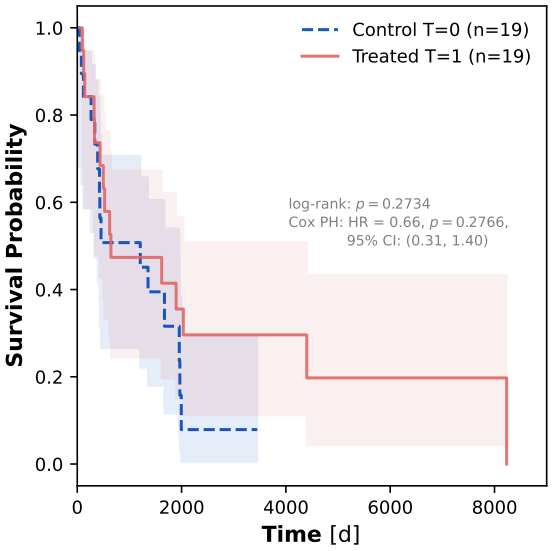
The results indicate that the treated subset in the predicted $\hat{\tau}_i > 0$ subgroup shows a slightly higher survival probability compared to the control group for both binary-outcome (Figure 5.8 (a)) and survival-outcome model (Figure 5.8 (c)). Additionally, the Cox



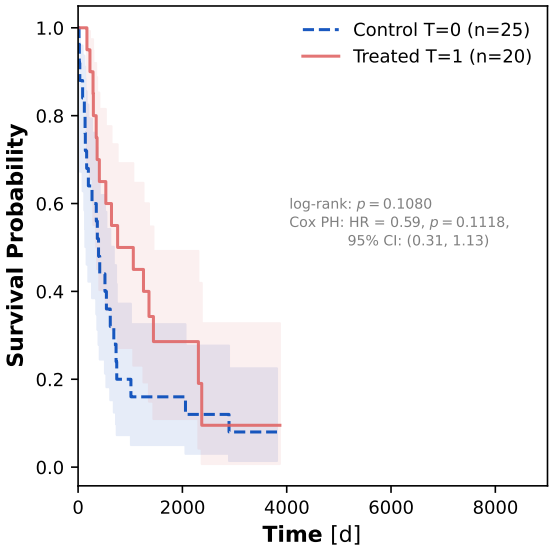
(a) $\hat{\tau}_i > 0$ subgroup, binary-outcome model



(b) $\hat{\tau}_i \leq 0$ subgroup, binary-outcome model



(c) $\hat{\tau}_i > 0$ subgroup, survival-outcome model



(d) $\hat{\tau}_i \leq 0$ subgroup, survival-outcome model

Figure 5.8: Kaplan-Meier curves on the NSCLC-Radiomics dataset for observed patient subgroups stratified by the sign of the estimated CATE ($\hat{\tau}_i > 0$ vs. $\hat{\tau}_i \leq 0$), where a positive CATE indicates that the patient is predicted to benefit from treatment $T = 1$. Within each subgroup, curves compare the survival probability for patients who were actually treated ($T = 1$, red) versus the control group patients ($T = 0$, blue). Results are shown on the hold-out test set using ensembled cross-validation models selected based on validation performance. Log-rank p-values and Cox proportional hazards results are reported for reference.

proportional hazards regression using treatment as the only covariate (see Section 4.2.4) reported a hazard ratio below 1, indicating a lower hazard or better survival under treatment, but with relatively large confidence intervals. As the binary-outcome model recommended treatment $T = 1$ for almost all patients except for one, the remaining data point for the $\hat{\tau}_i \leq 0$ subgroup ($n = 1$) offered no insight into the expected treatment response (Figure 5.8 (b)), highlighting the extreme imbalance in predicted treatment effect. For the survival-outcome model, the $\hat{\tau}_i \leq 0$ subgroup (Figure 5.8 (d)) also showed a hazard ratio below 1 with a relatively large confidence interval. Neither of the log-rank tests for the four Kaplan-Meier curves indicated a significant separation between curves.

For completeness, additional Kaplan-Meier curves are provided in the Appendix Section B.1 for the results on a single validation split. The curves in Figure B.2 show stronger separation between "recommendation followed" and "anti-recommendation followed" groups than in the ensembled evaluation, with the "recommendation followed" curve lying slightly above both the treated and control curves. For the subgroups, the curves in Figure B.3 indicated a slight benefit for the treated subset compared to the control subset in the $\hat{\tau}_i > 0$ subgroup, and the opposite tendency in the $\hat{\tau}_i \leq 0$ subset. However, neither the log-rank test nor the Cox PH results were statistically significant. An analysis using the ground truth ITE as policy on the same validation fold as shown in Figure B.5 and Figure B.4 did not yield clear separation either. Possible explanations include small subgroup sizes, wide confidence intervals, and small (simulated) treatment effect sizes relative to the noisy prognostic effects and survival variability of real patients.

Overall, both the analysis of the "recommendation" versus "anti-recommendation" followed subset and the subgroup analysis results provide no strong evidence that either model can reliably identify patient groups that truly benefit from treatment $T = 1$.

Relationship between treatment effect and predictive biomarker. Because the predictive imaging biomarker ("flatness") used to simulate the treatment effects in semi-synthetic NSCLC-Radiomics experiments is known, it enables the direct investigation of whether the CATE estimation model is able to identify and recover it. To this end, the z-score normalized "flatness" feature is plotted against the estimated CATE and the true ITE in Figure 5.9 for the validation fold with the highest decision accuracy (0.68, image-only survival-outcome model). As expected from the simulation, the true ITE values follow a clear monotonic trend with increasing "flatness", with larger positive treatment effects for larger positive "flatness" values and vice versa. In contrast, the estimated CATE does not show a clear association with the "flatness" and only a very weak correlation (Spearman correlation coefficient $\rho = 0.10$), despite the model's comparably high decision accuracy among all other models. The sign also often mismatches, indicating that the recommended treatment arm do not reflect the true treatment effect determined by the predictive imaging biomarker. This suggests that in many cases, model makes correct recommendations even without capturing the underlying predictive imaging

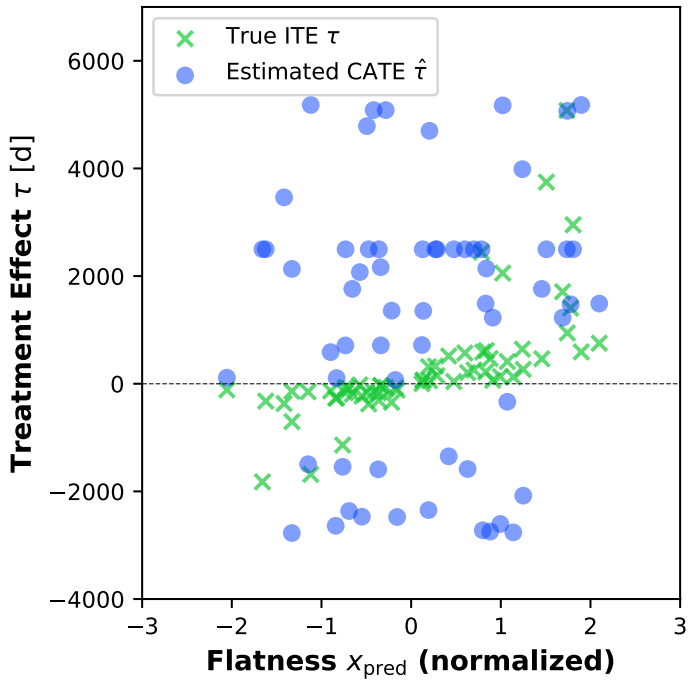


Figure 5.9: Scatter plot of the predictive imaging biomarker (radiomics feature "flatness") x_{pred} used to simulate the treatment effects of NSCLC-Radiomics outcomes versus the estimated CATEs $\hat{\tau}$ (blue) and the true individual treatment effect τ (green). Results are shown for the validation split using the cross-validation fold with the best validation performance.

biomarker signal. A full overview of the combined scatter plots across folds is provided in Appendix Figure B.6, which also shows systematic overestimation of treatment effects for some samples, similar to what was observed in Figure 5.6, and in Figure B.1 for the ensembled results on the test set, which exhibit a similarly weak correlation.

In summary, semi-synthetic experiments presented in this section provided a controlled setting to validate the methodology and to assess whether the models could recover a known predictive imaging biomarker. The results provided insights into the proposed CATE estimation model behavior, and also highlighted its limitations, such as a weak correlation with the true treatment effect despite seemingly high decision accuracy. However, poor performance in a simulated setting does not necessarily imply failure on real clinical data. Therefore, the subsequent experiments on the EORTC dataset focus on the ultimate application of interest, that is, the image-based treatment effect estimation in a real RCT, and aim to gain further insight into the sources of model limitations observed here.

5.2.2 Application Study on Glioblastoma Imaging from a Randomized Controlled Trial

The results in this subsection summarize the findings of the experiments on the EORTC brain cancer MRI dataset. Their overall purpose was to address RQ2.3 by investigating whether the proposed image-based CATE estimation model for either binary or survival outcomes is able to identify a heterogeneous treatment effect of an experimental glioblastoma treatment from this retrospective RCT study dataset that could be leveraged for making treatment recommendations from pre-treatment images and for identifying possible predictive imaging biomarkers.

To investigate dataset-specific behaviors of the model, similar experiments that were conducted on the NSCLC-Radiomics are also repeated for the EORTC dataset, including comparisons of the performance of binary-outcome and survival-outcome models and different input modalities or multitask learning, where binary-outcome and survival-outcome heads are trained jointly. Additionally, variations of the model and strategies to possibly mitigate the limited dataset size and signal are studied, such as the leveraging pre-trained image encoders.

As the ground truth treatment effect estimation is unknown for this case, it is impossible to compute the same oracle metrics that were used for the evaluation of the semi-synthetic NSCLC-Radiomics experiments, decision accuracy and PEHE. The evaluation is therefore restricted to metrics that only require observed outcomes, such as the observed policy risk and policy value and other factual prediction metrics. Instead, an assessment of the strength of the predictive biomarker signal identified by the model is made using part of

the proposed evaluation from Section 5.1.1, combined with an analysis of the treatment response levels and the information content of the images is made.

Using Binary vs. Survival Outcomes.

Table 5.5: Comparison of CATE models trained on survival versus binary outcomes on the EORTC-Radiomics dataset. Reported are the fraction of correctly assigned treatments (Decision Accuracy), root PEHE ($\sqrt{\epsilon_{PEHE}}$) and the observed policy risk $\widehat{R}_{Pol}$, as well as Balanced Accuracy, F1, AP and AUROC, with mean $\pm$ SD across folds. To enable a direct comparison, all metrics except Decision Accuracy are computed using binarized survival outcomes (threshold: 365 d). An asterisk (*) indicates models trained on continuous survival outcomes whose predictions were post hoc binarized for evaluation. All metrics are IPCW-adjusted.

Split	Model Type	$\widehat{R}_{Pol} \downarrow$	Balanced Acc ↑	AUROC ↑	F1 ↑	AP ↑
Val.	Binary	0.86 ± 0.04	**0.54 ± 0.09**	**0.50 ± 0.14**	**0.40 ± 0.12**	**0.40 ± 0.16**
	Survival*	**0.85 ± 0.06**	0.50 ± 0.09	**0.50 ± 0.09**	0.18 ± 0.20	0.38 ± 0.09
Test	Binary	0.82 ± 0.00	0.50 ± 0.00	0.50 ± <0.01	**0.42 ± <0.01**	**0.47 ± <0.01**
	Survival*	**0.78 ± 0.01**	**0.52 ± 0.02**	**0.55 ± <0.01**	0.24 ± 0.07	**0.47 ± <0.01**

Following the experiments on the NSCLC-Radiomics dataset (Section 5.2.1), the proposed model trained on different outcome formulations, i.e. binary survival outcomes thresholded at 365 d and time-to-event survival outcomes, using different loss functions, are compared to provide more evidence to address RQ2.1. Using these experiments, it is investigated whether the same trend of an improvement in treatment recommendations with regard to policy risk can also be seen in this dataset for the survival-outcome model.

To make the performance of the two models directly comparable, the results are again presented for the binarized survival outcomes thresholded using the same threshold of 365 d as described in Section 4.2.4.

The results shown Table 5.5 only report a slight, but not significant improvement of the policy risk on the validation set for the survival-outcome model (0.85 ± 0.06) over the binary-outcome model (0.86 ± 0.04), which turned out to be slightly larger on the hold-out test set (0.78 ± 0.01 vs. 0.82 ± 0.00). Here, some standard deviations are again zero due to identical predictions across folds, which can occur with small datasets and thresholded outputs or identical treatment policies.

For the other factual outcome classification metrics, the difference between the models is varied. While it is very clear for the F1 score that the binary model outperforms on both the validation and test set, the other metrics, balanced accuracy, AUROC and AP are all very similar, but with a tendency of the binary-outcome model to have slightly higher

scores on the validation set and the survival-outcome model to have slightly higher scores on the test set.

All factual outcome classification metrics show values very close to random (0.5), or even below, and minor variation between validation and test results, suggesting a limited binary survival classification performance for both models. This also indicates that the observed improvement in treatment recommendations in policy risk with the survival-outcome model may have limited robustness and could be influenced by the variability arising from the limited sample size.

Value of Multimodal Integration and Multitask Learning for Treatment Effect Estimation

The clinical tabular data in the EORTC dataset include several known prognostic factors, as mentioned in Section 4.2.1, which motivates further investigation of RQ2.2. Therefore, the following experiments evaluate whether incorporating these tabular covariates might provide additional information to improve both treatment recommendations and survival predictions, similar to the experiments on the NSCLC-Radiomics dataset. They also explore the added value of a segmentation mask with delineated tumors as an additional input channel and the benefits of using multitask learning, where both the binary-outcome and survival-outcome treatment effect estimations are trained simultaneously (Section 4.2.2).

As shown in Table 5.6, integrating tabular inputs into the binary outcome models did not yield a significant difference in the results. Combining tabular input data with multitask learning lowered the policy risk on both the validation set (0.858 ± 0.038 to 0.831 ± 0.047) and hold-out test set (0.821 ± 0.000 to 0.811 ± 0.022) and increased the AUROC (0.51 ± 0.14 to 0.61 ± 0.05) on the validation data only, but at the cost of a decrease in the balanced accuracy. Adding a segmentation channel on top of tabular inputs and multitask learning resulted in the lowest test policy risk (0.801 ± 0.027), but did not lead to a consistent improvement in the other scores.

For the survival-outcome model, integrating tabular data and additionally a segmentation mask channel resulted in a higher policy value on the validation set ($(0.137 \pm 0.041) \times 10^3$d to $(0.166 \pm 0.045) \times 10^3$d), but resulted in no change or a slight decrease on the test set, with the best policy value remaining at $(0.165 \pm 0.014) \times 10^3$d. Multitask learning reduced the policy value on the validation set and did not help to surpass the best non-multitask model on the test set. In contrast, it contributed to achieve a higher C-index on both splits, with the best C-index achieved by the model trained on image, segmentation and tabular inputs and higher values for the test compared to the validation set (0.536 ± 0.032 for the validation set and 0.582 ± 0.004 for the test set). Integrating tabular input data and a segmentation mask channel also improved both the validation and test C-index compared to the respective model trained without these modalities.

Table 5.6: Comparison of CATE estimation models trained with different combinations of input modalities (tabular data and tumor segmentation mask "Seg.") on the EORTC dataset. Reported are the observed policy risk $\hat{R}_{Pol}$ or policy value $\hat{V}_{Pol}$, as well as the Balanced Accuracy, AUROC, and Antolini's C-index, with mean $\pm$ SD across folds. All metrics are IPCW-adjusted, except for the C-index.

a Performance of binary-outcome CATE estimation models.

Split	Modalities			Multitask	$\hat{R}_{Pol} \downarrow$	Balanced Acc ↑	AUROC ↑
	Image	Segm.	Tabular				
Val.	✓	-	-	-	0.858 ± 0.038	**0.541 ± 0.092**	0.50 ± 0.14
	✓	-	✓	-	0.858 ± 0.038	**0.541 ± 0.092**	0.51 ± 0.14
	✓	-	✓	✓	**0.831 ± 0.047**	0.518 ± 0.041	**0.61 ± 0.05**
	✓	✓	✓	✓	0.837 ± 0.059	0.496 ± 0.038	0.54 ± 0.11
Test	✓	-	-	-	0.821 ± 0.000	**0.501 ± 0.000**	0.50 ± <0.01
	✓	-	✓	-	0.821 ± 0.000	**0.501 ± 0.000**	**0.51 ± 0.02**
	✓	-	✓	✓	0.811 ± 0.022	0.500 ± <0.001	0.49 ± 0.02
	✓	✓	✓	✓	**0.801 ± 0.027**	0.500 ± <0.001	**0.51 ± 0.05**

b Performance of survival-outcome CATE estimation models.

Split	Modalities			Multitask	$\hat{V}_{Pol} \uparrow [10^3 \text{d}]$	C-Index ↑
	Image	Segm.	Tabular			
Val.	✓	-	-	-	0.137 ± 0.041	0.497 ± 0.027
	✓	-	✓	-	0.147 ± 0.046	0.501 ± 0.026
	✓	✓	✓	-	**0.166 ± 0.045**	0.503 ± 0.048
	✓	-	✓	✓	0.109 ± 0.007	0.527 ± 0.035
	✓	✓	✓	✓	0.109 ± 0.007	**0.536 ± 0.032**
Test	✓	-	-	-	**0.165 ± 0.015**	0.546 ± 0.004
	✓	-	✓	-	**0.165 ± 0.014**	0.556 ± 0.017
	✓	✓	✓	-	0.159 ± 0.006	0.561 ± 0.014
	✓	-	✓	✓	0.161 ± 0.000	0.576 ± 0.011
	✓	✓	✓	✓	0.161 ± 0.000	**0.582 ± 0.004**

In summary, multimodal integration and multitask learning resulted in, at best, some small but inconsistent gains on the EORTC dataset. While some configurations of input modalities and multitask learning improved the policy risk or policy values, as well as AUROC and C-index, these effects were not robust across dataset splits or types of outcome formulation. Additionally, as the balanced accuracy and AUROC remain close to chance (0.5), and the C-index scores only slightly exceed it, this suggests an overall weak survival prediction performance on the factual data.

The results also highlight that optimizing for factual prediction metrics, such as C-index, does not automatically also improve metrics related to treatment effect estimation, such as policy risk or policy value. However, even though policy risk or policy value are the more relevant for the task at hand, i.e. treatment recommendations, they are also insensitive to minor changes in the model predictions, making them not the ideal choice either for hyperparameter tuning. The fact that experiments with the same policy risk or value still yielded different AUROC or C-index values, as observed in the experiments, shows that the models may make similar treatment decisions even when the underlying model predictions differ.

Impact of Leveraging Pre-trained Image Encoders

A common observation from all the previous experiments was the consistently poor performance of the deep-learning CATE estimation models in making factual predictions. This was also seen in the NSCLC-Radiomics experiments (Section 5.2.1), despite being semi-synthetic with a clearly defined treatment effect, but with a similar number of samples.

The purpose of the experiments presented in the following was to assess whether a possible strategy to mitigate the limited amount of data using transfer learning could lead to an improvement of CATE estimation models compared to a model trained from scratch. The experiments thereby address RQ2.2 by examining the potential benefits of using pre-trained encoders and tabular data integration. Transfer learning was implemented by using image encoders pre-trained on a large-scale brain MRI dataset (OpenMind) and fine-tuned on the EORTC dataset, as detailed in Section 4.2.3.

The results summarized in Table 5.7 show that the best metrics (policy risk, balanced accuracy and AUROC) are achieved by models with pre-trained encoders on both the validation and hold-out test sets, although not consistently by the same model across all metrics.

On the validation set, the model with the MAE pre-trained encoder, fine-tuned on both image and tabular inputs, had the lowest policy risk and highest AUROC among all models, whereas the highest balanced accuracy was achieved by the model with the same pre-trained encoder, but fine-tuned without tabular inputs. On the test set, the lowest

Table 5.7: Comparison of binary-outcome CATE estimation models trained from scratch or with different pre-trained ResEnc-L encoders released by Wald et al. (2025), fine-tuned on the EORTC dataset, with and without clinical tabular inputs. Reported are the observed policy risk $\widehat{R}_{Pol}$, as well as the Balanced Accuracy and AUROC, with mean $\pm$ SD across folds. All metrics are IPCW-adjusted.

Split	Pre-training Method	Tabular	$\widehat{R}_{Pol} \downarrow$	Balanced Acc ↑	AUROC ↑
Val.	From Scratch	-	0.827 ± 0.026	0.533 ± 0.074	0.54 ± 0.09
		✓	0.869 ± 0.046	0.528 ± 0.076	0.57 ± 0.10
	SwinUNETR	-	0.850 ± 0.044	0.517 ± 0.033	0.53 ± 0.05
		✓	0.835 ± 0.032	0.520 ± 0.062	0.54 ± 0.07
	MAE	-	0.827 ± 0.058	**0.539 ± 0.053**	0.56 ± 0.04
		✓	**0.821 ± 0.034**	0.529 ± 0.020	**0.60 ± 0.03**
Test	From Scratch	-	0.801 ± 0.037	0.503 ± 0.028	0.51 ± 0.05
		✓	0.811 ± 0.022	0.518 ± 0.015	0.55 ± 0.04
	SwinUNETR	-	0.801 ± 0.029	**0.554 ± 0.060**	0.60 ± 0.06
		✓	0.811 ± 0.032	0.549 ± 0.061	**0.62 ± 0.08**
	MAE	-	0.814 ± 0.032	0.505 ± 0.018	0.54 ± 0.03
		✓	**0.788 ± 0.011**	0.514 ± 0.019	0.57 ± 0.07

policy risk was again achieved by the model with the MAE pre-trained encoder and with tabular data integration. In contrast, the SwinUNETR pre-trained models exhibited better generalization on the test set than the MAE pre-trained model for the factual prediction metrics, with a notably higher balanced accuracy and AUROC compared to both models trained from scratch, though the models did not outperform MAE pre-trained models in terms of policy risk.

Integrating the clinical tabular data led to an increased AUROC for all models. However, this integration did not consistently improve policy risk or balanced accuracy, which is consistent with earlier observations from the multimodal integration comparison. This also highlights that integrating tabular data impacts treatment recommendations and factual predictions differently.

As the experimental setup differed slightly from the previous experiments, including the architecture (ResEnc-L instead of ResNet encoder) as well as data preprocessing and augmentation scheme, the results are only comparable to a limited extent. Nevertheless, the models trained from scratch with the ResEnc-L encoder already achieved a lower policy risk and higher AUROC compared to their counterpart from the earlier ResNet-based setup (see Table 5.6a), indicating that the architecture and data preprocessing choice already had an impact itself. On top of this, the best policy risk achieved by the ResEnc-L CATE estimation models (0.821 ± 0.034 on the validation set and 0.788 ± 0.011 on the

test set, both for MAE pre-trained encoders) showed a further improvement over the ResNet-based models (0.831 ± 0.047 on the validation set and 0.801 ± 0.027 on the test set).

To sum up, the models with pre-trained encoders demonstrated an improved performance over models trained from scratch, providing some evidence in favor of RQ2.2 that pre-training could improve image-based treatment effect estimation. However, the performance gains were modest relative to the SDs across folds. Even the best balanced accuracy remained close to chance (0.5), and the best AUROC only slightly exceeded it. The improvement also depended on the type of pre-training method and, to some extent, on whether tabular data was included during fine-tuning.

Model Reliability and Predictive Signal of Images

The previous experiments assessed possible strategies to improve the performance of the proposed image-based CATE estimation models for survival outcomes on an RCT imaging dataset, including multimodal integration, multitask learning, and transfer learning, which yielded modest performance gains at best. To better understand the possible root causes as well as the limitations and robustness of the proposed model and also to address RQ2.3, a similar analysis to that in Section 5.2.1 for the NSCLC-Radiomics dataset is presented here for the EORTC dataset. This includes comparisons against regression baselines and an analysis of the treatment recommendations. As it remains unknown whether a predictive imaging biomarker, which could be leveraged in making treatment recommendations, is truly present in the EORTC images, the predictive biomarker signal strength in the estimated treatment effects is assessed as in Section 5.1.1. Additionally, the information content of the imaging data is investigated by testing whether tabular clinical covariates can be predicted from the baseline MRI scans.

Comparison to regression and alternative baselines. To further assess whether a deep-learning approach using image input data provides a benefit for CATE estimation on the EORTC dataset over using tabular clinical covariates, the proposed image-based CATE estimation models with a TARNet-like architecture were compared against regression baselines, similar to the experiments for the NSCLC-Radiomics dataset in Section 5.2.1. The presented CATE estimation regression baselines include the logistic regression T-learner ("Logistic Reg.") for binary outcomes and the Cox proportional hazards regression T-learner ("Cox PH") for survival outcomes, which were fitted using five known prognostic clinical covariates (see Section 4.2.3). Their performance is presented in Table 5.8, along with the results of the best-performing image-based models ("Bin-TARNet" and "Surv-TARNet") trained from scratch, which were selected based on their validation policy risk or policy value, respectively.

Table 5.8: Comparison of the proposed deep-learning-based CATE estimation models using image inputs (configuration selected based on validation performance) with a regression-based T-Learner CATE estimation baseline trained on clinical tabular data only. "Bin-TARNet" denotes the binary-outcome CATE estimation model, "Surv-TARNet" the survival-outcome CATE estimation model, and "Surv-CNN" a survival outcome prediction model with the same backbone as Surv-TARNet but trained to predict only factual outcomes rather than performing CATE estimation. Both the "Bin-TARNet" and "Surv-CNN" were trained using multitask learning in this case. Reported are the observed policy risk $\widehat{R}_{Pol}$ or policy value $\widehat{V}_{Pol}$, as well as the Balanced Accuracy, AUROC and Antolini's C-index, with mean $\pm$ SD across folds. All metrics are IPCW-adjusted, except for the C-index.

a Performance of binary-outcome CATE estimation models.

Split	Method	Modalities	$\widehat{R}_{Pol} \downarrow$	Balanced Acc ↑	AUROC ↑
Val.	Logistic Reg.	Tabular	**0.83 ± 0.06**	0.515 ± 0.037	**0.62 ± 0.11**
	Bin-TARNet	Img. + Tab.	**0.83 ± 0.05**	**0.518 ± 0.041**	0.61 ± 0.05
Test	Logistic Reg.	Tabular	**0.78 ± 0.01**	**0.554 ± 0.016**	**0.69 ± 0.02**
	Bin-TARNet	Img. + Tab.	0.81 ± 0.02	0.500 ± 0.000	0.49 ± 0.02

b Performance of survival-outcome CATE estimation and survival prediction models.

Split	Method	Modalities	CATE Est.	$\widehat{V}_{Pol} \uparrow [10^3 \text{d}]$	C-Index ↑
Val.	Cox PH	Tabular	✓	**0.173 ± 0.065**	**0.60 ± 0.09**
	Surv-TARNet	Img. + Seg. + Tab.	✓	0.166 ± 0.045	0.50 ± 0.05
	Surv-CNN	Img. + Seg. + Tab.	-	-	0.58 ± 0.06
Test	Cox PH	Tabular	✓	**0.201 ± 0.018**	**0.68 ± 0.01**
	Surv-TARNet	Img. + Seg. + Tab.	✓	0.159 ± 0.006	0.56 ± 0.01
	Surv-CNN	Img. + Seg. + Tab.	-	-	0.64 ± 0.01

For binary-outcome models, the performance of both the tabular-only regression baseline and Bin-TARNet turned out to be very similar on the validation set, with the Bin-TARNet having a slightly higher balanced accuracy, but a slightly lower AUROC than the regression model, and both having almost the same policy risk. The test results, however, show that the Bin-TARNet generalizes poorly, with a lower AUROC and balanced accuracy compared to the validation set, both of which are close to chance, indicating that it strongly overfits and thus that it fails to capture robust image features for predicting the outcomes. The regression baseline model outperformed the Bin-TARNet on all three metrics, exhibiting even better performance on the test set compared to the validation set.

For the survival-outcome models, the tabular-only regression model always outperformed the Surv-TARNet across both splits and both policy value and C-index. The table also shows the results for a CNN model ("Surv-CNN") for factual survival prediction only instead of CATE estimation, which was also selected based on the best validation policy value and trained with the same image encoder architecture, but with a single output head and without any information about the treatment indicator. It showed a clearly higher C-index on both splits compared to the Surv-TARNet. This suggests that the individual treatment effects are weak in this dataset compared to the overall prognostic signal.

Table 5.9: Comparison of the proposed two-headed TARNet-like CATE estimation model (configuration selected based on validation performance) with a single-headed S-Learner architecture sharing the same backbone and configuration. Reported are the observed policy risk $\hat{R}_{Pol}$ or policy value $\hat{V}_{Pol}$, as well as the Balanced Accuracy, AUROC and Antolini's C-index, with mean $\pm$ SD across folds. All metrics are IPCW-adjusted, except for the C-index. For full results, please refer to Appendix Table B.2.

a Performance of binary-outcome CATE estimation models.

Split	Model	$\hat{R}_{Pol} \downarrow$	Balanced Acc $\uparrow$	AUROC $\uparrow$
Val.	S-Learner (Img. + Tab., Multitask)	**0.810 ± 0.053**	0.500 ± 0.000	0.58 ± 0.06
	TARNet (Img. + Tab., Multitask)	0.831 ± 0.047	**0.518 ± 0.041**	**0.61 ± 0.05**

b Performance of survival-outcome CATE estimation models.

Split	Model	$\hat{V}_{Pol} \uparrow [10^3 \text{d}]$	C-Index $\uparrow$
Val.	S-Learner (Img. + Tab.)	**0.200 ± 0.073**	**0.506 ± 0.030**
	TARNet (Img. + Tab.)	0.147 ± 0.046	0.501 ± 0.026

S-Learner. Given that the Surv-CNN clearly outperformed Surv-TARNet in terms of C-index, this also raised the question of whether the shared information and struc-

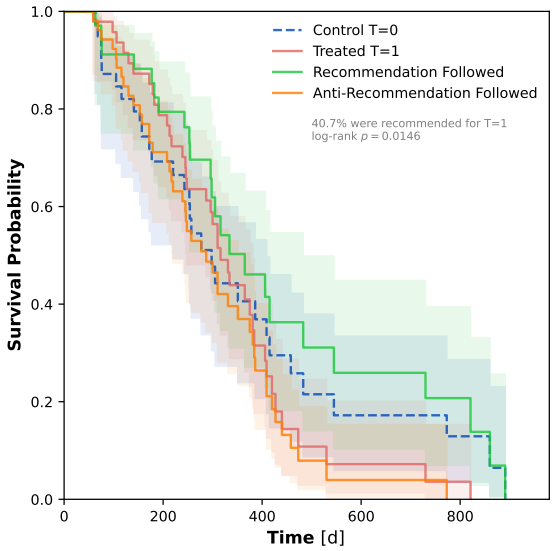
ture across treatment groups could be further exploited. For this reason, the proposed Surv-TARNet is compared against an S-Learner version, which only has one output head (as the Surv-CNN) to induce a stronger inductive bias through more shared layers, but incorporates treatment assignment information as an additional tabular input covariate (see Section 4.2.2). The results are shown in Table 5.9 for the S-Learner with the best validation policy risk or value and the TARNet with the same configuration. While the binary-outcome S-Learner achieved a lower policy risk, it performed worse in terms of balanced accuracy and AUROC. For the survival-outcome models, however, the S-Learner outperformed the TARNet model in both policy value and C-index. The additional results in Appendix Table B.2 reveal that these differences were not consistent across splits and metrics, indicating that neither architecture offers a systematic advantage on the EORTC dataset.

Reliability of survival analysis and treatment recommendations. The final treatment recommendations by the ensembled CATE estimation predictions on the test set of the EORTC dataset and the resulting subgroups are assessed using Kaplan-Meier plots in the same way as for the NSCLC-Radiomics dataset (see Section 5.2.1). The results are only shown for the best-performing image-based CATE estimation models selected by the best policy risk or value.

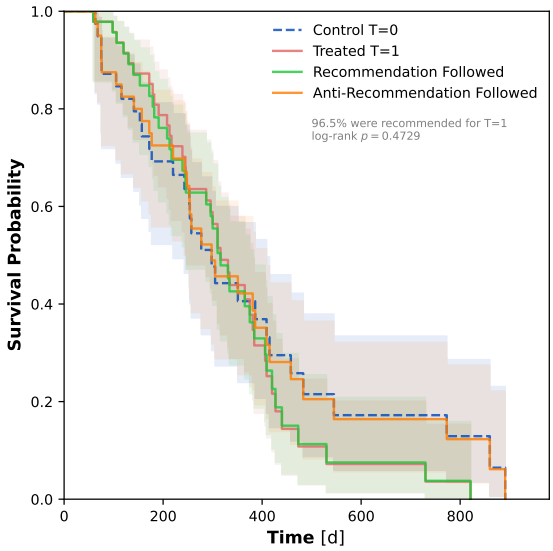
Patients with recommended and anti-recommended treatments. The plots in Figure 5.10 show the Kaplan-Meier curves of the patients stratified by whether they have received the same treatment as recommended by the model according to the estimated CATE or the opposite treatment, denoted as "recommendation followed" and "anti-recommendation followed". As outlined for the NSCLC-Radiomics dataset, in an ideal scenario the "recommendation followed" curve would show a larger survival benefit compared to the "anti-recommendation followed" curve respectively.

The plot for the binary-outcome model (Figure 5.10 (b)) shows a "recommendation followed" Kaplan-Meier curve (green) with a higher survival probability compared to the other curves for almost all time points. A log-rank test comparing the curve to the "anti-recommendation followed" curve (orange) indicates a statistically significant separation ($p = 0.0146$). Latter "anti-recommendation followed" curve lies closer to the control group for times below 400 d and closer to the treated group for times above 400 d. The results suggest that the ensembled models could provide useful treatment recommendations to improve the overall survival outcome. Nevertheless, due to the limited sample size and resulting large confidence however, the level of uncertainty remains high.

As the survival-outcome model recommended the treatment for almost all patients (96.5%, see Figure 5.10 (b)), the "recommendation followed" curve almost coincide with the treated group (T=1) curve, and similarly for the "anti-recommendation followed" curve and the control group curve. All four curves lie very close to each other and intersect, indicating

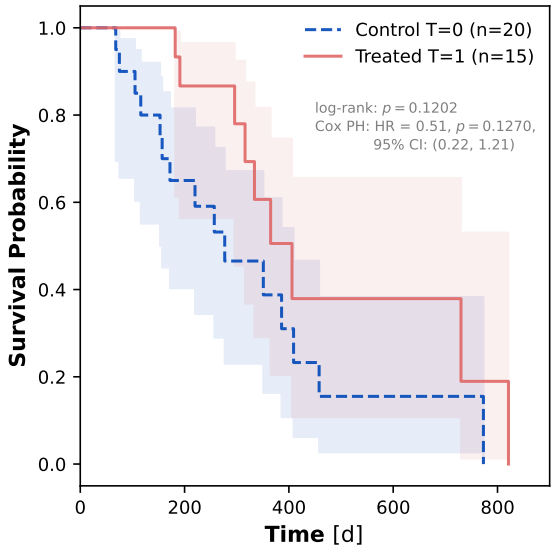


(a) Recommendations by binary-outcome CATE estimation model.

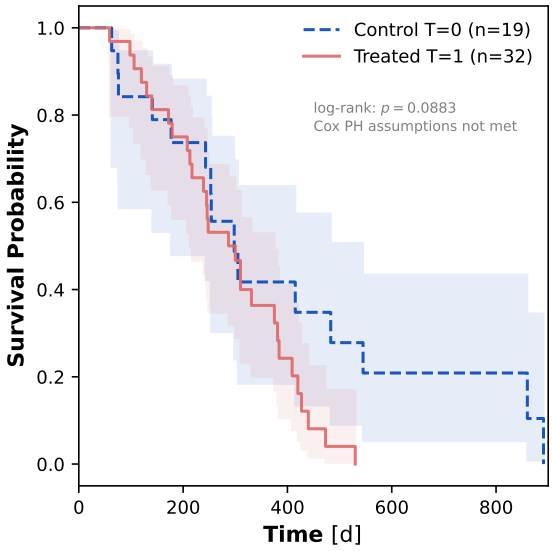


(b) Recommendations by survival-outcome CATE estimation model.

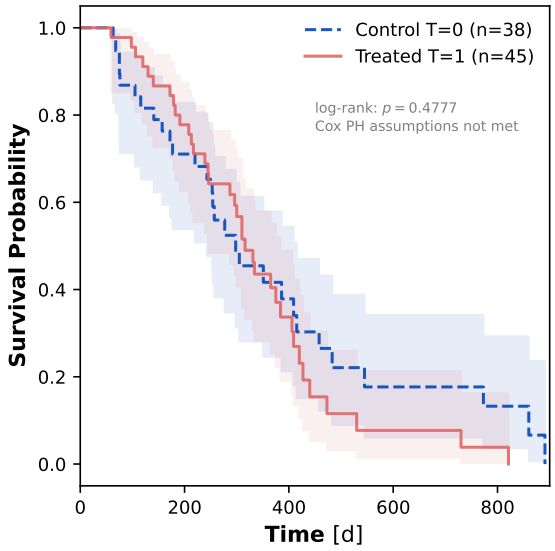
Figure 5.10: Kaplan-Meier curves on the EORTC dataset comparing the survival probability for patients who received the treatment recommended by the estimated CATE (green) versus those who did not (orange). For reference, the curves for the treated group ($T = 1$, red) and control group ($T = 0$, blue) are also shown. Results are shown on the hold-out test set using ensembled cross-validation models selected based on validation performance. Log-rank p-values are reported for reference.



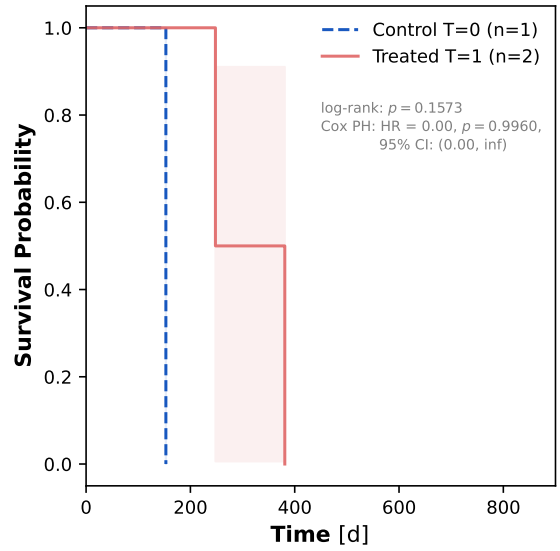
(a) $\hat{\tau}_i > 0$ subgroup, binary-outcome model



(b) $\hat{\tau}_i \leq 0$ subgroup, binary-outcome model



(c) $\hat{\tau}_i > 0$ subgroup, survival-outcome model



(d) $\hat{\tau}_i \leq 0$ subgroup, survival-outcome model

Figure 5.11: Kaplan-Meier curves on the EORTC dataset for observed patient subgroups stratified by the sign of the estimated CATE ($\hat{\tau}_i > 0$ vs. $\hat{\tau}_i \leq 0$), where a positive CATE indicates that the patient is predicted to benefit from treatment $T = 1$. Within each subgroup, curves compare the survival probability for patients who were actually treated ($T = 1$, red) versus the control group patients ($T = 0$, blue). Results are shown on the hold-out test set using ensembled cross-validation models selected based on validation performance. Log-rank p-values and Cox proportional hazards results are reported for reference.

that there is no significant benefit to be expected if the recommendations from the models are followed.

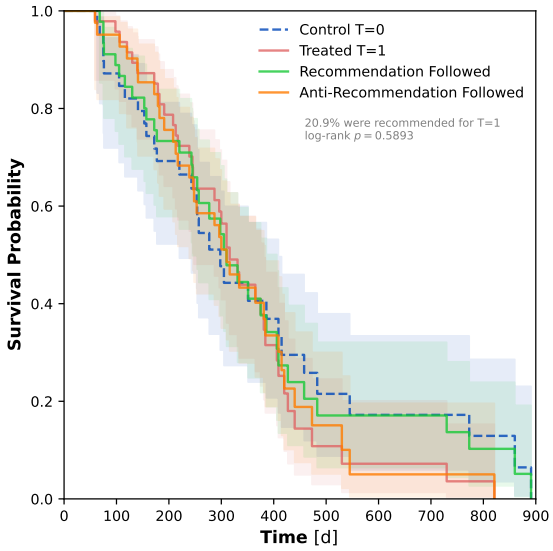
Recommended subgroups. The next step of the reliability analysis addresses the question: How do the patient subgroups identified by the model (i.e. the $\hat{\tau}_i > 0$ or the $\hat{\tau}_i \leq 0$ subgroup) respond to treatment? For this purpose, Figure 5.11 displays the Kaplan-Meier curves of patients within these subgroups, stratified by the treatment actually received, similar to Figure 5.8 for the NSCLC-Radiomics dataset.

The subplots for the binary-outcome model, Figure 5.11 (a) and Figure 5.11 (b), show a behavior close to what would be desired: within the $\hat{\tau}_i > 0$ subgroup, the curves indicate a higher survival probability for the treated patients $T = 1$ (red) compared to the control group patients $T = 0$ (blue) with a HR of 0.51 (95% CI: 0.22–1.21, $p = 0.1270$), whereas within the $\hat{\tau}_i \leq 0$ subgroup, the survival probability of the $T = 0$ is similar or higher, especially for later time points past around 400 d. However, neither the log-rank test nor the Cox proportional hazards results report a significant difference between treated and control patients within the two subgroups, as also indicated by the wide confidence intervals for the Kaplan-Meier curves.

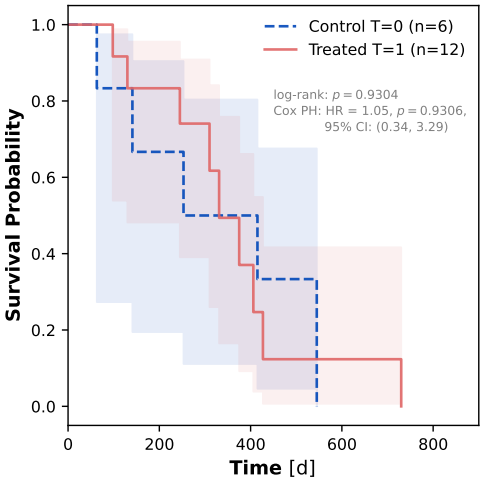
Since the survival-outcome model estimated a treatment effect $\hat{\tau}_i > 0$ for almost all patients, the resulting Kaplan-Meier curves for $T = 1$ and $T = 0$ in Figure 5.11 (c) almost directly correspond to the curves for the whole test set cohort, as seen in Figure 5.10 (b), and do not exhibit a significant separation or benefit from the treatment. For the only three remaining patients in the $\hat{\tau}_i \leq 0$ subgroup, Figure 5.11 (d), no useful conclusion could be made from the plot either.

For completeness, the Kaplan-Meier plots for the best-performing ResEnc-L CATE estimation model for binary outcomes, which was fine-tuned using a MAE-pre-trained encoder, are shown in Figure 5.12. Even though the recommended subgroups for $T = 1$ and $T = 0$ are more balanced compared to the survival-outcome model, the plot comparing the Kaplan-Meier curves for the "recommendation followed" with the "anti-recommendation followed" patients (Figure 5.12 (a)) did not show a statistically significant separation, as is also supported by log-rank test ($p = 0.5893$). The two plots for the treated and control patients within the recommended subgroups (Figure 5.12 (b) and Figure 5.12 (c)) also showed wide confidence intervals and no separation, with log-rank test p-values of 0.9306 and 0.5990, respectively. These findings indicate that the recommendations by the model with the MAE-pre-trained encoder could not provide a clear benefit for the patients either.

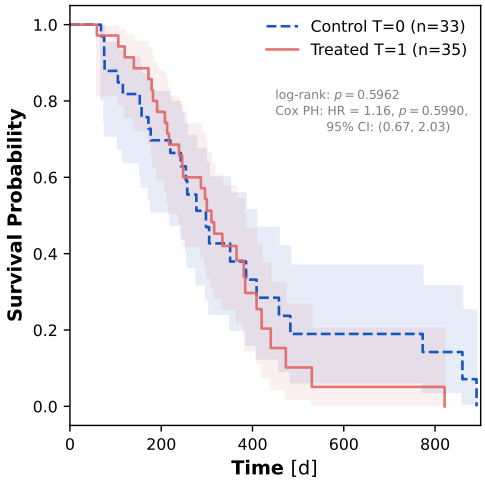
Estimated CATE and predicted outcomes. To further understand the behavior of the CATE estimation models in making their individual predictions and to identify possible weaknesses, histograms of the estimated CATE values $\hat{\tau}$ for the test set are provided in Appendix Figure B.7. Additionally, scatter plots of the individual predicted outcomes $\hat{Y}(T)$, which were used to compute the estimated CATE, are provided in Appendix



(a) Recommendations vs. anti-recommendations



(b) $\hat{\tau}_i > 0$ subgroup



(c) $\hat{\tau}_i \leq 0$ subgroup

Figure 5.12: Kaplan-Meier curves for the recommendation results by a binary-outcome CATE estimation model with a pre-trained ResEnc-L encoder by Wald et al. (2025) using MAE (selected based on validation performance), fine-tuned on EORTC dataset. (a) Curves for patients who received the treatment recommended by the estimated CATE (green) versus those who did not (orange), and the treated group ($T = 1$, red) and control group ($T = 0$, blue) shown for reference. (b) and (c) Survival probability for patient subgroups stratified by the sign of the estimated CATE. Within each subgroup, curves compare the survival probability for patients who were actually treated ($T = 1$, red) versus the control group patients ($T = 0$, blue). Results are shown on the hold-out test set using ensembled cross-validation models. Log-rank p-values and Cox proportional hazards results are reported for reference are reported for reference.

Figure B.8. The histogram and scatter plot of the model with the MAE-pre-trained ResEnc-L encoder are summarized in Appendix Figure B.9.

The plots indicate that the behavior differs widely between the types of models. On the one hand, the predicted CATE $\hat{\tau}$ and the predicted outcomes $\hat{Y}(T)$ of the binary-outcome model within a fold have a low variance compared to the differences across folds with little overlap. Even though the CATE of the ensembled model has a mean close to 0, indicating its recommended treatments are fairly balanced between $T = 0$ and $T = 1$, the plots suggest that there is little agreement between the models for each fold, which could be due to a lack of robustness towards variations in the dataset and instabilities. On the other hand, for the survival-outcome model, the agreement of the models for the individual folds is far greater and the histograms of all five models have a similar peak around the mean average treatment effect of 41.6 d, indicating that all model recommendations including the one of the ensembled model tend to favor $T = 1$, as seen in the previous Kaplan-Meier plots. As for the survival-outcome model with the ResNet encoder, the predictions of the model with the MAE-pre-trained ResEnc-L for binary outcomes show a large overlap between folds and a larger variance within each fold, which is possibly an effect of regularization using label-smoothing. As the histograms of the individual folds differ widely with no clear peak of the $\hat{\tau}$, this indicates that there is a similar poor agreement between the cross-validation folds as for the ResNet encoder binary-outcome model. Overall these plots highlight that the CATE estimation models either substantially disagree across folds (both ResNet and ResEnc-L encoder binary-outcome models) or are systematically biased (survival-outcome model), further suggesting that the models were not able to recover a reliable predictive signal from the EORTC dataset.

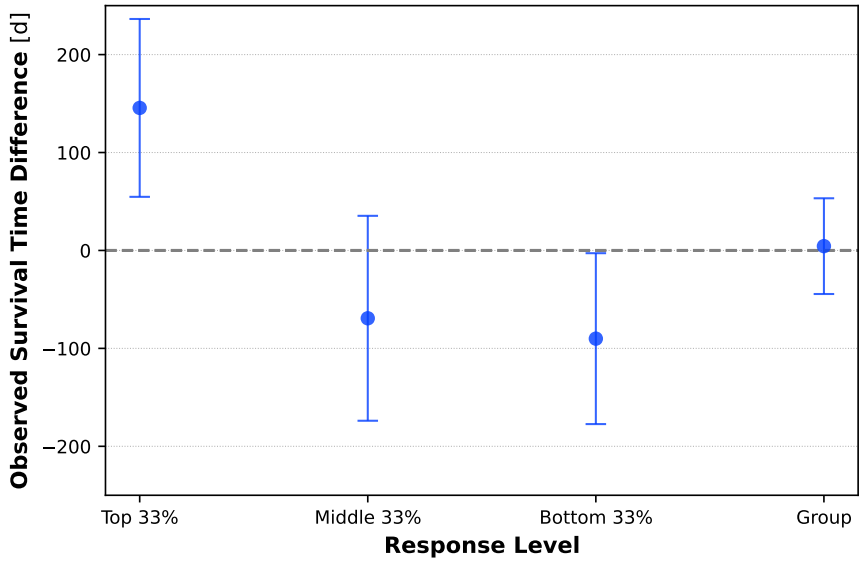
Table 5.10: Evaluation results of the predictive strength of the estimated CATE from binary-outcome "Bin-TARNet" and survival-outcome "Surv-TARNet" CATE estimation models (configuration selected based on validation performance) on the EORTC dataset. To enable a direct comparison, the policy value ($\hat{V}_{Pol}$) is computed using survival outcomes for both models. Reported are the results from a Cox regression testing the estimated CATE as a predictive biomarker candidate (see Section 4.1.3): the p-value from the Wald test for the biomarker-by-treatment interaction term, the ratio of absolute Wald z-statistics $|z_{pred}/z_{prog}|$ for the predictive vs. prognostic term (analogous to the relative predictive strength in Section 4.1.3), and the p-value of the likelihood ratio test comparing the full versus reduced model.

Split	Model	$\hat{V}_{Pol} \uparrow [10^3\text{d}]$	Wald p	$\lvert z_{pred}/z_{prog} \rvert$	LR p
Test	Bin-TARNet	0.167 ± 0.011	0.012	5.13	0.013
	Surv-TARNet	0.159 ± 0.006	0.065	0.94	0.067

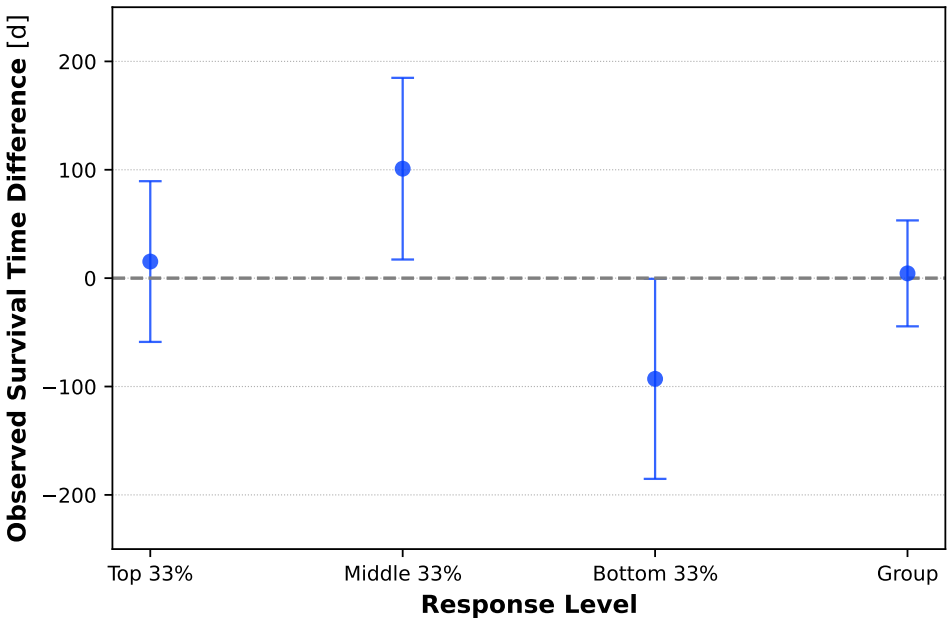
Predictive biomarker signal strength analysis. To investigate whether the image-based CATE estimation models were able to identify a predictive (imaging) biomarker and to answer RQ2.3, the statistical analysis from the previously proposed evaluation protocol in Section 4.1.3 was applied for the EORTC experiments. As the outcomes of interest are survival outcomes for this dataset, the results were obtained by regressing the survival outcomes using a Cox proportional hazards regression model instead of a linear regression model as used previously in Section 5.1.1.

Evaluation results of the predictive strength. The results for this evaluation on the test set are summarized in Table 5.10. In the direct comparison, the binary-outcome model shows a slightly higher policy value ($\widehat{V}_{Pol} = (0.167 \pm 0.011) \times 10^3$d) compared to the survival-outcome model ($(0.159 \pm 0.006) \times 10^3$d), and the results of the statistical test further provide evidence that the binary-outcome model was better able to identify a predictive biomarker signal compared to the survival-outcome model: Both the Wald test for the coefficient of the biomarker-by-treatment interaction term and the likelihood ratio test comparing the full regression model with the biomarker-by-treatment interaction term and the reduced regression model without report significant results with p=0.012 and p=0.013 respectively, whereas this is not the case for the survival-outcome model. Moreover, the ratios of the Wald z-statistics, also described as "relative predictive strength", indicate that the binary-outcome model was able to identify a stronger predictive than prognostic signal in the data with $|z_{pred}/z_{prog}| = 5.13 > 1$ compared to the survival-outcome model, for which $|z_{pred}/z_{prog}| = 0.94$. Results for the binary-outcome ResEnc-L models are included in the Appendix Table B.3. Even though both the ResEnc-L trained from scratch and fine-tuned using a MAE-pre-trained encoder show a higher policy value ($\widehat{V}_{Pol} = (0.173 \pm 0.010) \times 10^3$ d and $\widehat{V}_{Pol} = (0.185 \pm 0.017) \times 10^3$ d respectively), the statistical tests could not indicate that the models were able to identify a strong predictive biomarker signal.

Observed survival treatment effect by tertiles. Additionally, plots for the observed survival treatment effect stratified by the estimated CATE tertiles, or also referred to as uplift bins (Ascarza 2018), are provided in Figure 5.13 to get a better understanding of how useful the model recommendations are. Since an assessment of the treatment effect on an individual level is not possible for real data without access to counterfactuals, the patients are split into three equally sized groups based on their estimated CATE $\hat{\tau}$ and evaluated on a group level. The figures plot the difference between the average observed survival times of treated and control group patients within each tertile subgroup, similar to (Durso-Finley et al. 2022). In an ideal case, the upper tertile, i.e. patients that are predicted to benefit the most from the treatment, would have a greater observed treatment effect than the bottom tertile. While the plot for the binary-outcome model (Figure 5.13 (a)) shows this general trend with a higher observed treatment effect for the upper tertile compared to the middle and the bottom tertile, the predictions for each tertile have high uncertainty, resulting in wide error bars and only a small difference between the middle and the bottom tertile. For the survival-outcome model, the middle tertile even shows a higher



(a) Observed treatment effect of binary-outcome model.



(b) Observed treatment effect of survival-outcome model.

Figure 5.13: Observed survival treatment effect by tertiles of the estimated CATE from the (a) binary-outcome and (b) survival-outcome CATE estimation model (selected based on validation performance) on the EORTC dataset. For each tertile subgroup, the plot shows the average difference in the observed survival time between the patients who were actually treated ($T = 1$) and the control group patients ($T = 0$) within that tertile. The rightmost point shows the average treatment effect across all patients. Error bars represent the standard deviation within each subgroup.

observed survival treatment effect than the top tertile, and all three values again have similarly wide confidence intervals. Both figures indicate that the models were unable to reliably capture the actual heterogeneous treatment effects, which prevented them from accurately ranking the patients into subgroups.

In summary, and to address RQ2.3, while the binary-outcome model showed some statistically significant results in the predictive strength analysis, combined evidence from the quantitative analysis and the observed survival treatment effect by tertiles plot (uplift bins) provide limited evidence that the models were able to robustly identify a predictive imaging biomarker.

Predicting clinical covariates for assessing image information content. While the previous analyses revealed that both the image-based CATE estimation methods trained from scratch and using pre-trained encoders were unable to reliably estimate heterogeneous treatment effects and identify a predictive imaging biomarker signal, the question remains whether this is due to a limitation of the model itself or due to an inherent (lack of) treatment effect signal in the data. For this reason, to examine more closely if this data-related factor could be the case and to assess the information content of the imaging data, models with the same backbone architecture as the previously employed ResNet encoder are trained to individually predict the clinical covariates and treatment indicators of the EORTC dataset using the pre-treatment MRI scans alone as inputs. The purpose of the analysis is twofold: first, to assess whether the model has sufficient capacity to learn from image data by regressing a known image feature, such as tumor volume, and second, to determine if the images contain information about other confounding clinical factors or spurious associations with the treatment assignment.

Experimental details. The same one-headed model with the ResNet encoder architecture as the previously mentioned Bin-CNN and Surv-CNN was employed. The models were trained with the same settings as the ResNet-based CATE estimation models, except with no dropout. The loss functions were replaced with the MSE loss for the regression tasks and binary CE loss for the classification tasks.

Results. The results for this regression and classification analysis are presented in Table 5.11. The regression results (Table 5.11a) show that, as expected, the model struggles more to regress the age compared to the tumor volume, as indicated by the higher mean absolute error, MSE, and the negative R^2 values on both the validation and test set. While the results were better for regressing the tumor volumes from images, adding the ground truth tumor segmentation mask as an extra channel to the image input generally provides either only a slight improvement or no additional improvement for the mean absolute error on the test set. This suggests that the model has some capacity to extract some simple information (i.e. the image feature tumor volume) directly from the images, but it is not able to achieve a strong improvement even with access to the ground truth segmentation masks from which the tumor volumes were computed. The limited performance on a

simple task like tumor volume regression is an indication that the size of the dataset may also be a limiting factor for the model to accurately identify relevant image features.

The classification results for the three clinical characteristics (Table 5.11b) largely show performance close to chance, especially with respect to balanced accuracy. While the model predicting the patient sex had a slightly above chance performance for the AUROC, F1, and AP, the performance drop from validation set to test set suggests that the prediction might not be entirely robust, and that class imbalances might be present.

Classifying the two treatment schemes AnyBEV or BEV versus control (Table 5.11c) consistently shows a balanced accuracy and AUROC close to random, confirming that the treatment assignment in the RCT for the EORTC dataset was not informed by features present in the images but random, ruling out possible confounding.

These results suggest that while the model was able to identify the image-derived feature tumor volume, the images do not contain enough information about prognostic factors such as the patients' age, corticosteroids use, WHO performance status, and also sex as well as the assigned treatment for the model to learn. On the other hand, the results support the hypothesis that integrating the available clinical tabular data might provide additional information complementary to the image data.

Table 5.11: Results from experiments assessing whether EORTC images contain information about baseline clinical characteristics and treatment assignments. Reported are regression performance metrics (mean absolute error (MAE), MSE and R^2) or classification performance metrics (Balanced Accuracy, AUROC, F1 score and AP) for models using images only, which share the same backbones as the Bin-CNN/Surv-CNN models from earlier experiments.

a Regression of age and baseline tumor volume, with or without segmentation masks for tumor volume as additional input. All variables shown are z-scored.

Split	Variable	MAE ↓	RMSE ↓	R^2 ↑
Val.	Age	0.92 ± 0.09	1.11 ± 0.09	-0.28 ± 0.16
	Tumor Volume from Img.	0.56 ± 0.12	0.79 ± 0.11	0.31 ± 0.23
	Tumor Volume from Img. + Seg.	0.53 ± 0.06	0.73 ± 0.07	0.43 ± 0.08
Test	Age	0.78 ± 0.05	0.98 ± 0.06	-0.07 ± 0.14
	Tumor Volume from Img.	0.54 ± 0.11	0.76 ± 0.08	0.37 ± 0.12
	Tumor Volume from Img. + Seg.	0.54 ± 0.10	0.73 ± 0.07	0.43 ± 0.10

b Classification of baseline clinical characteristics (corticosteroid use, sex and WHO performance status)

Split	Variable	Balanced Acc ↑	AUROC ↑	F1 ↑	AP ↑
Val.	Corticosteroids	0.50 ± 0.02	0.51 ± 0.04	0.29 ± 0.22	0.52 ± 0.04
	Sex	0.55 ± 0.04	0.71 ± 0.04	0.75 ± 0.04	0.82 ± 0.03
	WHO PS	0.50 ± <0.01	0.59 ± 0.06	0.79 ± 0.04	0.73 ± 0.08
Test	Corticosteroids	0.55 ± 0.03	0.58 ± 0.02	0.36 ± 0.23	0.57 ± 0.02
	Sex	0.55 ± 0.04	0.66 ± 0.03	0.69 ± <0.01	0.70 ± 0.04
	WHO PS	0.50 ± <0.01	0.48 ± 0.03	0.78 ± <0.01	0.63 ± <0.01

c Classification of applied treatments (anyBEV vs. control and BEV vs. control).

Split	Variable	Balanced Acc ↑	AUROC ↑	F1 ↑	AP ↑
Val.	AnyBEV	0.50 ± 0.00	0.46 ± 0.08	0.87 ± 0.05	0.76 ± 0.08
	BEV	0.50 ± 0.00	0.39 ± 0.03	0.78 ± 0.05	0.60 ± 0.05
Test	AnyBEV	0.50 ± 0.00	0.52 ± 0.03	0.81 ± 0.00	0.71 ± 0.03
	BEV	0.50 ± 0.00	0.54 ± 0.04	0.71 ± 0.00	0.61 ± 0.04

6 Discussion

This chapter places the experimental findings from the previous chapter within a broader context by discussing the resulting insights, implications, and limitations of this thesis for predictive imaging biomarker discovery and heterogeneous treatment effect estimation for clinical imaging studies with survival outcomes. After discussing both parts individually (Section 6.1 and Section 6.2) and then jointly to provide an overarching perspective (Section 6.3), the chapter closes with Section 6.4 by outlining future research directions for which this thesis lays the foundations.

6.1 Evaluating Heterogeneous Treatment Effect Estimation Models for Predictive Imaging Biomarker Discovery

Disclosure: Parts of this discussion section are based on previously published work (Xiao et al. 2025).

The work presented in the first main area of investigation of this thesis (Section 5.1) proposed a novel approach to predictive imaging biomarker discovery, which is to leverage deep-learning-based CATE estimation models for image inputs without relying on a separate feature extraction step and pre-defined handcrafted biomarker candidates. A primary goal of the experiments was to investigate whether image-based CATE estimation models can be leveraged for this task (RQ1.1). To test this, experiments were conducted to assess whether the performance of such image-based CATE estimation models in discovering predictive imaging biomarkers can be reliably evaluated (RQ1.2) using a protocol proposed in this thesis.

The results show that it is indeed feasible for an image-based CATE estimation model to identify predictive imaging biomarkers under the experimental conditions (RQ1.1), specifically when trained on semi-synthetic datasets with real pre-treatment images and simulated RCT outcomes. To support this finding, an evaluation protocol was developed and proposed as a solution to RQ1.2. One of its purposes was to enable the quantification of the strengths of unknown predictive imaging biomarkers using the relative predictive strength t_{pred}/t_{prog}. It further allowed their visual interpretation using attribution maps. Comparing these evaluation results with the strength and appearance of the known ground truth predictive and prognostic imaging biomarkers used in the simulation experiments addressed the second purpose of the protocol: to provide a reliable and reproducible method for assessing model performance in identifying predictive imaging biomarkers.

Predictive Strength of the Estimated CATE

The quantitative evaluation approach introduced in Section 4.1.3 relies on the estimated CATE as a measure for the predictive imaging biomarker candidate identified by the trained model.

The results presented in Section 5.1.1 showed that the measured relative predictive strength t_{pred}/t_{prog} was generally positively correlated with the true relative predictive effects b_{pred}/b_{prog}. This indicates that the estimated CATE is indeed a reliable measure for both the ground truth predictive effect and the ground truth predictive biomarker itself, albeit under the assumption of a linear biomarker–outcome relation, as used in the simulation setup.

The ability of an image-based CATE estimation model to identify predictive imaging biomarkers while not being affected by the presence of prognostic imaging biomarkers was assessed as follows: The relative predictive strength measure t_{pred}/t_{prog} was compared to the experimental baseline, where the regressed outcome is used as the predictive biomarker candidate, and additionally to the experimental upper bound for a purely predictive biomarker and lower bound for a purely prognostic biomarker. This comparison was done across multiple models trained on outcomes simulated using varying predictive and prognostic effects b_{pred} and b_{prog} as well as four types of image datasets and two types of image biomarker features each. The results highlighted that even in scenarios where predictive effects are smaller than prognostic effects, i.e. $b_{pred}/b_{prog} < 1$, which is often observed in clinical data such as the EORTC dataset discussed later in Section 6.2, the model's ability to identify predictive imaging biomarkers could still be demonstrated by showing relative predictive strengths $((\mathrm{t}_{pred}/t_{prog}))greaterthan1$.

The heavy variation in performance across image datasets also highlights the limitations of the specific image-based CATE estimation model and datasets employed in the experiments. Especially for CUB-200-2011, ISIC 2018, and NSCLC-Radiomics, and particularly for cases where b_{pred}/b_{prog} was high, the t_{pred}/t_{prog} was lower and closer to the

baseline, therefore indicating a weaker performance. A possible explanation for this is the lower accuracy of the model in predicting factual outcomes (Table 5.1) when facing more abstract imaging biomarker features, which also translated into a poorer treatment effect estimation performance indicated by a lower PEHE score (Table 5.1, Figure 5.2). In addition to the higher complexity of imaging biomarker features, the imbalance and distribution of image features in the datasets, as well as the dataset size, could have contributed. For example, the NSCLC-Radiomics dataset only had 332 training samples and was therefore much smaller than the CMNIST dataset (60,000 training samples), as well as the CUB-200-2011 dataset (5,794 training samples), and ISIC 2018 dataset (2,075 training samples).

Interpreting Predictive Imaging Biomarker Candidates

While the quantitative part of the evaluation protocol offers insight into the estimated strength of an identified predictive imaging biomarker candidate, the qualitative part of the evaluation protocol introduced in Section 4.1.3 again serves two purposes as mentioned earlier: (1) enabling the interpretation of predictive and prognostic image features identified by an image-based CATE estimation model, as well as (2) empirically validating the model's performance at this task by comparing the learned features to the known ground truth imaging biomarkers.

The experimental results in Section 5.1.2 demonstrated the practical utility of using the attribution maps that were described by the evaluation protocol for the interpretation of the identified imaging biomarkers, even without access to the ground truth. Especially for the CMNIST, CUB-200-2011, and also to a lesser extent for the ISIC 2018 dataset, it was possible to directly infer the predictive and prognostic imaging biomarkers from the treatment effect attribution maps and control group head attribution maps. For the CMNIST dataset, for example, it was possible to infer that the predictive imaging biomarker was a shape-based feature and that the prognostic imaging biomarker was a color-based one with the help of attribution maps from different examples and color channels, as well as the information about the sign of the attribution in different regions. Using this information, it was similarly possible to infer from the CUB-200-2011 dataset that the predictive imaging biomarker was associated with the head, bill, and outline of the bird, while the prognostic biomarker was related to the brightness or color of the bird's main body. This also highlights the importance of using multiple input image examples to interpret the identified biomarkers. For the NSCLC-Radiomics dataset, the 3D attribution maps complemented the 2D attribution maps, providing further insights into which regions contributed positively or negatively to the predictive or prognostic effects.

However, the results were also more ambiguous compared to the other three datasets, and the interpretation was not immediately apparent on its own. The attribution maps for both the treatment effect and the control group head focused on the same pixels, making it difficult to discern whether an image feature that is both predictive and prognostic was

present, or if two independent imaging biomarkers with distinct meanings were spatially overlapping.

By comparing the attribution maps to the ground truth imaging biomarker values and appearances, it was nevertheless possible to validate the model's performance. The results confirmed that the CATE estimation models mostly learned to identify the correct features as biomarkers, including localized features based on color and shape (CMNIST, CUB-200-2011, NSCLC-Radiomics), as well as first-order statistics (NSCLC-Radiomics) or patterns (ISIC 2018).

The comparison to the ground truth provided further evidence for the variations of the model performance across datasets, in addition to the quantitative results. The noisier heatmaps for the NSCLC-Radiomics dataset and attributions to areas outside the region of interest (i.e. the tumor volume), from which the actual ground truth imaging biomarker feature was computed, indicated that the model had more difficulty in identifying the correct imaging biomarkers and possibly learned spurious correlations instead.

Overall, the observations represent a key limitation of the qualitative evaluation using XAI methods: Attribution maps, in general, can only support the explanation but are prone to bias in human interpretations and often require domain-specific knowledge.

General Strengths and Limitations of the Experimental Setup

The experimental design of this part of the thesis relied on simulated outcomes from predefined image features as biomarkers. The main advantage is that it enabled the reliable and reproducible assessment of the image-based CATE estimation model performance, which would have otherwise been impossible without access to the complete ground truth. The disadvantage of this approach, however, is the limited realism of simulated scenarios. This included the imaging biomarkers themselves, which were mostly visually obvious, thereby simplifying the interpretation of the attribution maps and the simple linear biomarker–outcome relationship. As RCT outcomes were simulated, no confounders were present, which could have impeded the model's performance.

As this work primarily served as a methodological proof-of-concept for the evaluation protocol, the same hyperparameter settings were used for all models trained on the same image dataset across all outcome simulation parameters and not tuned individually. While this ensured consistency and comparability, the models may not have achieved the best possible performance. Additionally, 5-fold cross-validation could have helped to assess the robustness of the CATE estimation models to dataset variations. Furthermore, the attribution maps in the qualitative evaluation were shown only for models trained on data from a single parameter setting for the predictive and prognostic effects ($b_{pred}, b_{prog} = 1.0$). A more comprehensive visual interpretation of the identified prognostic and predictive imaging biomarkers would involve comparing the attribution maps across different parameter combinations. Despite the challenges, the XAI analysis using attribution maps remains essential for interpreting and distinguishing predictive and prognostic imaging

biomarkers, particularly since, unlike tabular data, where variable importance scores can be computed directly for pre-defined features, images lack discrete input variables.

Broader Implications

While the experiments were conducted in a controlled setting using semi-synthetic image datasets, the evaluation protocol proposed in this work was designed to be applicable to any dataset with unknown biomarkers to assess if a predictive imaging biomarker candidate may be potentially present. Section 5.1 outlines how such an evaluation can be carried out, without relying on handcrafted features such as radiomics features. In such applications where a single model is trained on data with unknown predictive effects, the same regression and the t-tests on the resulting parameters would be performed, as described in Section 4.1.3 to get a quantitative measure of the predictive biomarker effects.

A second important use case for the proposed evaluation protocol is benchmarking: specifically, to compare the performance of image-based CATE estimation models for model selection. This use case does require simulated outcomes based on known image features, but these can be arbitrarily defined. For instance, radiomics features were only used in the NSCLC-Radiomics dataset to simulate both prognostic and predictive imaging biomarkers.

The focus of this work was to establish a generalizable evaluation protocol for predictive imaging biomarkers that can, in principle, be applied to benchmark image-based CATE estimation methods on any RCT image datasets. While the presented experiments demonstrate the feasibility of this approach, translating these methods to clinical imaging datasets of real patients requires further adaptation to address data-specific challenges, such as more complex biomarker–outcome relationships. These are addressed in the second part of this thesis.

6.2 Image-Based Heterogeneous Treatment Effect Estimation in Clinical Imaging Studies

After addressing predictive imaging biomarker discovery using image-based CATE estimation models in controlled experiments, the second main area of investigation of this thesis (Section 5.2) focused on applying these models to clinical imaging datasets and addressing the challenges this entails. Specifically, the goal was to adapt and evaluate CATE estimation methods for use in realistic clinical settings. To this end, a solution was proposed to extend these methods to handling survival outcomes (RQ2.1), which are often the primary endpoint or outcome of interest in clinical imaging studies, as opposed to simpler categorical or continuous outcomes. The performance of a survival-outcome model was compared to that of a binary-outcome model by evaluating their

ability to recommend the optimal treatment. The second main goal was to investigate if incorporating additional multimodal information from the clinical imaging study or pre-trained encoders could improve the model's performance (RQ2.2). Finally, in addition to assessing the limitations of the image-based CATE estimation model and its methodological adaptations, it was investigated if the model could be used to gain further insights about heterogeneous treatment effects and possible predictive imaging biomarkers in the EORTC dataset, a RCT dataset with real patient outcomes (RQ2.3).

To investigate the research questions, experiments were conducted on both the EORTC dataset (Section 5.2.2) and also the NSCLC-Radiomics dataset (Section 5.2.1), which served as a semi-synthetic baseline to accurately assess the treatment effect estimation performance. This time, the semi-synthetic outcomes included real control group survival outcomes and simulated treatment effects instead of fully simulated continuous outcomes used in Section 5.1.

Using Binary vs. Survival Outcomes

The underlying hypothesis behind studying RQ2.1 was that using survival instead of binary outcomes to train CATE estimation models would allow them to learn more meaningful and nuanced insights for making treatment recommendations. While the binary-outcome model only predicts the probability for a patient to survive past a chosen threshold for a given treatment, i.e. coarse labels, the survival-outcome model predicts the survival probability of a patient over time, i.e. entire survival functions for each individual.

In the direct comparison of binary-outcome and survival-outcome models on both the NSCLC-Radiomics and the EORTC dataset, the survival-outcome model yielded slightly better results in terms of policy risk and also decision accuracy for the NSCLC-Radiomics dataset. These findings supported the answer to RQ2.1, demonstrating that image-based CATE estimation methods could indeed be extended to survival outcomes and that they could offer a slight benefit in heterogeneous treatment effect estimation.

However, the factual outcome prediction metrics and the Kaplan-Meier curves assessing the impact of the recommended treatments did not show a consistent benefit of training on either type of outcome. This observation is also in line with the findings of Haarburger et al. (2019), who reported almost identical C-index results for their two-stage median survival classification model and their hazard-based survival model. This implied that the presumed advantage of using survival-outcome models was balanced out by other disadvantages compared to binary-outcome models. One possible disadvantage is that survival outcomes introduce additional complexity to the outcome prediction task, resulting in less stable training. Binary-outcome CATE estimation models, on the other hand, may be less sensitive to outliers and easier to train, as the binary CE loss does not require large batch sizes as opposed to the Cox proportional hazards loss, which was also noted by Haarburger et al. (2019). Another aspect is that they do not require

population-level evaluation metrics such as the C-index, which is only meaningful when computed on a sufficiently large dataset. However, this binary outcome formulation is sensitive to the choice of the survival threshold. The 1-year survival cutoff chosen in this thesis differed from the median survival of both datasets and resulted in class imbalance, which may have further contributed to overfitting and unstable performance.

It should be noted that the comparison of the factual prediction performance relied on metrics computed on binarized survival predictions, which inevitably discards temporal information and may put the survival-outcome models at a disadvantage, as they were not directly trained to optimize the binary classification objectives. The results should therefore be interpreted with caution when drawing conclusions about whether one outcome formulation should be favored over the other, and the downstream treatment recommendation performance needs to be considered.

In the experiments, the training of binary-outcome models did not explicitly account for censoring, whereas the survival-outcome models were trained with a censoring-aware Cox proportional hazards loss function. To further assess the impact of censoring on the comparison between binary-outcome and survival-outcome models, the training data could also be IPCW-adjusted (Vock et al. 2016), which was only employed for the evaluation metrics in this thesis.

Given the slightly better treatment recommendation performance of survival-outcome models across both clinical imaging datasets, and the fact that these models inherently handle survival-risk modeling over time and censoring, employing the proposed survival-outcome extension of the image-based CATE estimation models is still recommended for future applications, despite the aforementioned challenges.

Binary modeling of the survival outcomes nevertheless remains valuable when only rough treatment recommendations are relevant, for example, when a treatment policy is desired where only patients with an estimated CATE larger than a treatment effect threshold are considered. To leverage potentially higher training stability, binary outcome prediction could also be integrated as an auxiliary task in a multitask learning model, as demonstrated in Section 5.2.2 and discussed in the next paragraph.

Value of Multimodal Integration for Treatment Effect Estimation

The results for investigating the value of integrating clinical tabular data and tumor segmentation masks showed inconsistent changes compared to the results for the image-only CATE estimation model across datasets. At most, the improvements were modest, as seen for the EORTC dataset, where the combination of image, segmentation, and clinical tabular data together with multitask learning led to a higher C-index for the survival-outcome across validation and hold-out test sets, but did not improve the policy value. Based on these results, the experiments did not provide a clear indication for RQ2.2.

The results were also inconsistent across validation and test splits, and different input configurations generalized differently. This points to overfitting as a possible factor contributing to less conclusive results, in addition to the small validation and test set sizes. Another explanation is provided by the differences in generalization performance between the two datasets: The multimodal survival-outcome model with tabular inputs showed much better generalization on the EORTC dataset than on the NSCLC-Radiomics dataset, as indicated by a higher C-index. This implies that the degree of overfitting may largely depend on the dataset and on whether the additional modalities also provide complementary information.

The clinical tabular-only regression models achieved a comparably strong factual prediction performance as the multimodal deep-learning models on both datasets (Table 5.4, Table 5.8). This suggests that the tabular data already provides strong prognostic information, which the multimodal models may not have sufficiently prioritized over the imaging input, and that the model may have failed to sufficiently capture the tabular-outcome relationship.

The similar or even worse performance of a multimodal model compared to one with fewer modalities is also in line with the findings of the review by Cui et al. (2023), who mentioned the introduction of biases and increased model complexity as possible reasons, though they also note that prior studies more commonly report performance gains from multimodal models.

A possible way to overcome overprioritizing one modality and a subject of future work is using different multimodal fusion strategies. Even though the concatenation approach used in the experiments has been reported to only learn limited interactions between the image and tabular representations, as stated by Wolf et al. (2022), it outperformed the alternative approach using the DAFT layer proposed by Wolf et al. (2022), which attempted to overcome this issue.

Despite the ambiguous results, the histograms comparing the distribution of the estimated CATE to the ground truth treatment effect of the NSCLC-Radiomics validation split showed that integrating a segmentation mask channel and clinical tabular data had some impact by slightly increasing the overlap and reducing the prediction of CATE outliers. However, this reduction of outliers could not translate to a direct improvement in treatment recommendation performance, which illustrates again that improving the treatment effect estimation performance does not necessarily translate to an improvement in treatment recommendations. While metrics such as policy risk and policy value are directly related to the relevant downstream task, they are insensitive to small changes in the model predictions. This reiterates that they are not an ideal choice for model hyperparameter tuning.

Based on the aforementioned insights of this thesis, incorporating additional modalities does not guarantee an improved performance of CATE estimation models and should therefore be evaluated on a case-by-case basis to assess their added value.

Impact of Leveraging Pre-trained Image Encoders

Integrating pre-trained image encoders to the image-based CATE estimation models led to modest improvements over the models trained from scratch, and consistently outperformed both versions with or without clinical tabular inputs, as described in Section 5.2.2. Improvements were observed across all three metrics — policy risk, balanced accuracy, and AUROC — however, not consistently for the same pre-training method. Even though the large-scale brain MRI dataset used by Wald et al. (2025) for their self-supervised pre-training only contained anatomical and not disease-related or treatment-related information, especially the MAE pre-trained encoder led to a performance gain in the treatment-related policy risk. This suggests that the anatomical or structural information from the pre-trained encoders provided valuable additional information beyond the EORTC dataset that helped to improve factual outcome predictions, which in turn could have contributed to making the CATE estimations more stable.

Integrating tabular data did not yield a consistent gain for different pre-trained encoders, which suggests that the image representations from the different pre-training strategies (such as SwinUNETR or MAE) may have interacted differently with the concatenated tabular representations, and that multimodal integration depends considerably on the fine-tuning strategy.

These observations for the binary-outcome models fine-tuned on the relatively smaller EORTC dataset supported part of the underlying hypothesis of RQ2.2, which was that leveraging encoders that were pre-trained on another large-scale image dataset of the same anatomical region can improve the treatment recommendation and factual prediction performance.

One caveat is that this study was only performed on the EORTC dataset and using binary-outcome models with a ResEnc-L encoder, which may not be entirely comparable with the previously employed ResNet-based models due to the different architecture, preprocessing, and data augmentation scheme. For this reason, further studies are necessary to investigate if the insights also translate to survival-outcome models or datasets of different anatomical regions and imaging modalities than the one used for pre-training, such as NSCLC-Radiomics.

Nevertheless, the slight improvements in the results suggest that employing a pre-trained image encoder to image-based CATE estimation models should be generally preferred over training from scratch, especially when dealing with limited dataset sizes. The experiments provided a proof-of-concept for this recommendation, serving as a basis for further investigating the potential of pre-training, transfer learning, and fine-tuning strategies for treatment effect estimation models.

Model Reliability and Baseline Comparison

To assess the reliability of the image-based CATE estimation models in clinical imaging use cases, and to explore possible reasons for limitations in their performance, multiple

analyses were performed on the semi-synthetic NSCLC-Radiomics dataset and the EORTC RCT dataset to address RQ2.3 from different angles.

Comparison to regression and alternative baselines. In the comparison of the image-based CATE estimation models against simple tabular-only regression baselines (Table 5.4,Table 5.8), the clinical tabular-only regression baselines generally outperformed the image-based deep learning models on both datasets in terms of either factual outcome prediction metrics on the hold-out test set or treatment-related metrics (decision accuracy, PEHE, policy value, policy risk). This held true even when the tabular data included only information that was presumed to be prognostic, and, in the case of the semi-synthetic NSCLC-Radiomics dataset, when all the information about predictive effects was only present in the images. This suggests that the image-based models failed to reliably extract an additional predictive signal from the imaging inputs. Nevertheless, the better decision accuracy of the survival-outcome CATE estimation models on the NSCLC-Radiomics indicated that the models were at least partially able to do so.

Another possible implication is that the tabular-only regression baselines were more stable and less likely to overfit, for example, by learning spurious correlations, due to having a higher bias and lower variance compared to more flexible deep learning models. This factor is especially important in smaller datasets, which is often the case for clinical imaging datasets such as the EORTC and NSCLC-Radiomics datasets. The poorer performance of deep learning models for tabular data compared to classical tree-based methods has also been discussed by Shwartz-Ziv et al. (2022), who mention the need for more extensive hyperparameter optimization as one of the reasons.

As noted earlier, the results underscore that the clinical tabular data from both datasets already provide valuable prognostic information, which is not only important for predicting factual outcomes but also for reliably estimating heterogeneous effects as a consequence.

The aforementioned considerations about bias-variance trade-off and the influence of the dataset size are also in line with the better performance of the survival outcome prediction model ("Surv-CNN") compared to the CATE estimation version ("Surv-TARNet") in terms of C-index on the EORTC dataset. The better performance could be explained by the fact that the former one-headed model was trained using the full training dataset. In contrast, the two-headed model could only update each head using data from the respective treatment arms, which reduces the effective sample size for each head. While this offers more freedom, it comes at the cost of being able to generalize robustly. Additionally, the treatment effects of the EORTC study were much weaker than the prognostic effects, as supported by the study of Wick et al. (2017), which found no significant average treatment effect with respect to overall survival between the treatment and control groups.

While the experiments in this thesis using the one-headed S-Learner versions of the Surv-TARNet models attempted to take advantage of the shared similarities between

treatment and control groups, as well as the negligible average treatment effect, they did not consistently outperform the two-headed Surv-TARNet models. This suggests that more research is needed to identify alternative metalearners or CATE estimation architectures that optimally introduce dataset-specific inductive biases, as suggested by Curth and Schaar (2021).

Reliability of survival analysis and treatment recommendations. The treatment recommendations made by both the binary-outcome and survival-outcome image-based CATE estimation models generally led to Kaplan-Meier curves that did not show a meaningful benefit of following the recommendations by the ensembled models compared to the actual randomly assigned treatment. As seen in Figure 5.7, Figure 5.10, and Figure 5.12 (a), there were no significant differences between the patients who received the recommended treatment and those who received the anti-recommended treatment. The Kaplan-Meier curves of the individual subgroups that were stratified according to the treatment recommendations mostly did not show a meaningful separation either (Figure 5.8, Figure 5.11, and Figure 5.12).

The only exception was the binary-outcome model on the EORTC dataset, which showed a slightly larger difference between the "recommendation followed" patients and the "anti-recommendation followed" patients and also a slight separation between control and treated patients within the recommended subgroups. However, in that case, the disagreement between the predictions of the cross-validation folds suggested that the models were unstable towards variations in the training data. While comparing the results across folds was useful for assessing robustness, the ensemble-level treatment recommendations used for stratification of patients for the Kaplan-Meier curves may partly mask this instability, as averaging across folds can yield seemingly better performance by chance.

These results were another indication that the models likely did not learn meaningful treatment policies from the data, but instead often overfitted or sometimes produced trivial recommendation policies (e.g. by recommending T=1 to everyone), which could be due to limitations of the model but also due to the size of the dataset and the underlying signal itself. They again highlight the broader issue of model reliability and the importance of validating treatment recommendations beyond selecting the best models based on a single metric such as policy value or policy risk.

Predictive Signal of the EORTC Dataset

One major aim of this part of the thesis was to assess whether there are heterogeneous treatment effect signals in the EORTC dataset using the proposed image-based CATE estimation models, which, to the best of my knowledge, have not been investigated previously for this dataset. Such a signal would point to predictive imaging biomarker

candidates for the tested experimental treatment using bevacizumab, which also ties to RQ2.3.

The analysis of the semi-synthetic NSCLC-Radiomics dataset provides a useful reference point for the predictive imaging biomarker analysis. Even though the nature of the simulated outcomes ensured that a predictive biomarker signal was present in the dataset, the model was not able to recover the predictive imaging biomarker accurately, as shown by the direct comparison of the estimated CATE with the value of the ground truth predictive imaging biomarker used in the data simulations, i.e. the radiomics feature "flatness", which displayed no significant correlation (Figure 5.9).

This highlights that the weak treatment recommendation performance observed in both datasets could not be attributed solely to the absence of a predictive biomarker signal, but most likely also reflects common methodological limitations affecting both the NSCLC-Radiomics and EORTC dataset, such as limited sample size, noise, and unstable training.

To analyze the results for the EORTC dataset without access to a confirmed ground truth predictive imaging biomarker, the analysis relied on the same quantitative evaluation as proposed in the evaluation protocol of the first part of the thesis (Section 4.1.3). Only the ensembled binary-outcome CATE estimation model showed a statistically significant biomarker-by-treatment interaction. Its relative predictive strength, with values greater than 1, was much stronger compared to the results of the survival-outcome model, which failed to show evidence for the presence of a predictive imaging biomarker.

The observed survival treatment effect stratified by tertiles, which were determined according to the estimated CATE of the models (uplift bins), also supported this finding. The binary-outcome model showed a stronger trend towards a larger treatment effect in the higher CATE tertiles compared to the survival-outcome model, for which this trend was not apparent.

However, the evidence for the presence of a predictive imaging biomarker in the EORTC dataset was not robust or conclusive for RQ2.3, which was evidenced by the large confidence intervals of the tertile plot combined with the high variability of the results across cross-validation folds, and the lack of a meaningful separation in the Kaplan-Meier curves stratified by treatment recommendations suggest that.

To further investigate the information content of the imaging data in the EORTC dataset, the capacity of the models' image encoder and the potential presence of confounders, an additional set of experiments evaluated whether the available tabular covariates (patient age, sex, WHO performance status, and the tumor volume) could be predicted directly from the MRI scans. These analyses indicated that the imaging data carried very limited signal for predicting the covariates, aside from the tumor volume, with model performance close to chance for most classification tasks. This provides further evidence that the clinical tabular data contain information complementary to the imaging data. The imaging data, in turn, likely do not contain information about potential confounders,

as suggested by the performance close to chance in classifying the applied treatment. Even tumor volume regression yielded only modest performance, and the inclusion of ground truth tumor segmentation masks yielded little additional benefit.

These findings reflect limitations in the model capacity, but also the dataset size, both of which likely contributed to overfitting. Learning to directly regress tumor volume from image data alone is a challenging task without additional model supervision, and the models were likely unable to reliably disentangle subtle imaging features associated with measuring the tumor volume. The wider implication of these results is that the imaging modality of the EORTC dataset, at least in this setup, may not only lack predictive but also prognostic information, which could help to explain the limited success of the image-based CATE estimation models.

While the results do not provide robust evidence for the presence of a predictive imaging biomarker for bevacizumab in the EORTC dataset, they nevertheless demonstrate how the proposed image-based CATE estimation model for survival outcomes can be applied to explore treatment effect heterogeneity in clinical imaging datasets and gain insights into the robustness and limitations of such as models. Additionally, the results also show that the quantitative evaluation methods from the first part of the thesis (described in Section 4.1.3) can support the systematic analysis of predictive imaging biomarker discovery.

General Strengths and Limitations

This part of the thesis, to the best of my knowledge, introduced the first study to extend image-based CATE estimation models with multimodal inputs to survival outcomes and to comprehensively evaluate their feasibility for making treatment recommendations. It presented the first application of such a methodology to a clinical imaging MRI dataset from an RCT in glioblastoma patients, which allowed a post-hoc exploratory investigation of the possible presence of a predictive imaging biomarker without handcrafted features or a separate feature extraction step as proposed in the first part of the thesis. In this context, this thesis is also the first to demonstrate the value of integrating pre-trained image encoders trained on MRI data for treatment effect estimation.

A comprehensive evaluation was enabled by evaluating the model in diverse settings. These included two very distinct medical imaging datasets of two anatomical regions and imaging methods, the lung CT dataset NSCLC-Radiomics and the brain MRI dataset from the EORTC trial, which represent a semi-synthetic dataset with outcome simulated using a known predictive imaging biomarker signal and a real RCT dataset, respectively. Additionally, the comparisons were performed across two types of outcomes (binary and survival), multiple combinations of multimodal inputs (imaging data, segmentation masks and clinical tabular data), different training strategies (multitask and transfer learning) and architectures (S-learner versus T-learner, ResNet versus ResEnc-L encoders trained from scratch or pre-trained).

Comparing the results across this diverse experimental and evaluation setup revealed several limitations, which may have affected the factual outcome prediction performance, as well as the treatment effect estimation and treatment recommendation performance. At the same time, this setup allowed clarifying where potential failures of the models stemmed from.

One of the main dataset-related limitations that affected both datasets was the small sample sizes ($n = 415$ for NSCLC-Radiomics and $n = 427$ for EORTC) compared to related heterogeneous treatment effect studies, which could be counteracted by augmentation and regularization strategies only to a limited extent. For instance, Durso-Finley et al. (2022) had access to a cohort of $n = 1817$ patients and Ma et al. (2023) to a dataset of $n = 656$ cases for their image-based CATE estimation study for regression or classification outcomes, while Schrod et al. (2022) used a dataset with $n = 1545$ for their method for tabular inputs survival outcomes, which allowed them to make reliable statements about their results.

In the setting of this study, the impact of the dataset size is amplified by different factors. First, the effective number of survival outcomes is reduced due to censoring, and modeling the stochastic time-to-event processes is inherently more complex than standard regression. Second, learning meaningful representations from often noisy and heterogeneous imaging data additionally increases the difficulty compared to using tabular inputs. Unlike segmentation tasks, where each image pixel or voxel has a label and provides a signal, treatment effect estimation for survival outcomes relies on a single patient-level outcome per sample. Additionally, the overall survival time might also not be directly tied to distinct visual features, as is often the case in classification tasks. Third, each treatment group head of the TARNet-like architecture is only trained on the respective subgroup of patients, reducing the effective training size for each head.

The ability of the CATE estimation models to make optimal treatment recommendations for the patients was likely not only affected by the small sample size of the datasets themselves, but also by the weak treatment effects present in the datasets. Both datasets had an average treatment effect close to zero, indicating that neither treatment was extremely beneficial or harmful, which made it inherently more challenging for the models to make a certain treatment decision. For NSCLC-Radiomics dataset, this was a deliberate simulation choice when generating the outcomes to obtain survival times in a realistic range. In this context, the coefficient of the predictive imaging biomarker (a z-standardized feature with a mean of zero) was set to a value that ensured a plausible rather than an artificially strong heterogeneous treatment effect size. Due to the lower signal-to-noise ratio, it is therefore difficult to make statements about whether a predictive imaging biomarker is truly present in the EORTC dataset, as the signals may be too weak for the current model to capture.

In addition to the limited dataset size, a higher model complexity of the CNN-based models (compared to regression models) could have contributed to overfitting and variations

across folds when training on the two clinical imaging datasets, despite measures to counteract it, such as dropout, weight decay, extensive image augmentation, label-smoothing, and the IPM balancing term of the BITES loss function to compensate for the imbalanced treatment arms.

The hyperparameter tuning process was also impacted by these limitations in the training, but more importantly, depended heavily on the choice of metrics used for model selection. Different metrics often led to different rankings of the model performance, as they prioritize different aspects. For example, factual outcome prediction metrics such as the C-index were used for preliminary hyperparameter tuning in the experiments as they are more stable and more discriminative compared to the observed policy value, but can lead to different model selections than metrics more relevant to the task of treatment effect estimation. The latter, computed only on factual outcomes, can have the same value even with different treatment effect estimations as long as the recommended treatment is identical, which makes it less informative as a model optimization metric.

This stresses the importance of a comprehensive evaluation rather than relying on a single metric, which was demonstrated in this thesis through multiple analysis approaches, which has also been raised by Lillelund et al. (2025).

Broader Implications

This study's primary clinical application was to evaluate whether the proposed image-based CATE estimation models for survival outcomes could stratify patients provide evidence for a predictive imaging biomarker in pre-treatment MRI scans associated with a treatment benefit of bevacizumab in glioblastoma patients using data from the EORTC-26101 trial. The methodological contributions of this study thereby provided a new causal inference perspective to the post-hoc exploratory analysis of this dataset, which is fundamentally different from prior related work focusing on features derived from diffusion MRI (Schell et al. 2020), perfusion MRI (Kickingereder et al. 2015), or radiomics features (Kickingereder et al. 2016; Grossmann et al. 2017; Ammari et al. 2021) as potential pre-defined biomarker candidates without directly modeling the heterogeneous treatment effects.

Despite methodological advancements that modeled survival outcomes, combined multimodal inputs and leveraged pre-trained image encoders, the experiments in this thesis did not indicate conclusive evidence for a predictive imaging biomarker or strong heterogeneous treatment effects, which is consistent with the earlier mentioned prior works in glioblastoma. While some of them showed evidence for a predictive biomarker (Kickingereder et al. 2015), none of them could confirm it, but often reported evidence for a prognostic biomarker and the challenge of disentangling predictive and prognostic signals (Schell et al. 2020). This challenge specifically what CATE estimation is designed to address, as discussed in Section 6.1. The results from the experiments of this thesis altogether represent valuable findings demonstrating the inherent challenges of deep-

learning-based CATE estimation models when dealing with clinical imaging data under real-world constraints.

These challenges highlight the requirements for performing image-based CATE estimation on survival outcomes, especially regarding data, offer several lessons for future research directions, which are discussed in Section 6.4. Reliable treatment recommendation and predictive biomarker discovery fundamentally depends on the accurately estimating heterogeneous treatment effect estimation, which was difficult to achieve in both studied datasets due to small sample sizes, weak treatment effects, a high variability and imbalances.

The proposed image-based CATE estimation model remains a valuable tool for future clinical research. It offers an additional way of retrospectively analyzing RCT data with survival outcomes and deriving data-driven insights directly from clinical imaging data without requiring image feature candidates, for example to generate new data-driven hypotheses about patient subgroups and to inform the design of future, more targeted clinical trials.

6.3 General Discussion

This thesis addressed challenges related to supporting image-based treatment decision-making, but from two different perspectives. The first part introduced the task of discovering predictive imaging biomarkers, which are useful for guiding treatment decisions, without requiring a separate image feature extraction step. It established the methodological foundations for evaluating the discovery of predictive imaging biomarkers using experiments on datasets with simulated continuous outcomes and known ground truth imaging biomarkers. While the concepts of this part were not limited to biomedical data and applicable to any scenario where images can be used to predict treatment effects, as illustrated in the toy experiments using datasets such as CMNIST and CUB-200-2011, the second part of this thesis focused on clinical imaging datasets and survival outcomes from RCTs. It addressed the challenge of applying image-based CATE estimation models to datasets with more complex real-world constraints than those in the first part to make individualized treatment recommendations based on images and provided a rigorous evaluation of survival treatment effects.

Despite tackling distinct research questions, both parts of this thesis are connected through the use of image-based treatment effect estimation models to analyze RCT data, with the shared goal of advancing personalized medicine. The evaluation protocol established in the first part of the thesis can be directly applied to assess the performance of the extended CATE estimation model proposed in the second part. Vice versa, the extensions such as integrating multimodal inputs, survival outcomes and pre-trained encoders can be employed to potentially enhance predictive biomarker discovery. As

demonstrated in Section 5.2.2 of the experiments, the same evaluation methods are also applicable for assessing whether such a model is able to identify predictive imaging biomarkers in the clinical imaging datasets with survival outcomes from the second part. Altogether, the work bridges the gap between causal inference, and medical imaging for predictive imaging biomarker discovery and image-based treatment decision-making.

The experiments from both studies leverage semi-synthetic data for validation, where outcomes are simulated using pre-defined image features as imaging biomarkers. Notably, the NSCLC-Radiomics dataset was used in both parts. The first part simulated continuous outcomes based on a linear biomarker–outcome model using different combinations of predictive and prognostic biomarker strength parameters, whereas the second part simulated semi-synthetic time-to-event outcomes by scaling the real-world outcomes with a single pre-defined parameter for the predictive biomarker strength. A similar series of experiments as in the first part could be repeated in future work using multiple parameter settings for the treatment effect to determine how strong the predictive imaging biomarker must be for it to be reliably detected by the model under clinically realistic conditions with survival outcomes.

Even though semi-synthetic data is inherently less realistic, it was crucial for the experimental validation setup to include ground-truth treatment effects, as discussed previously. Its utility for benchmarking the discovery of predictive biomarkers has also been stressed by (Curth, Svensson, et al. 2021) and (Crabbé et al. 2022). Publicly accessible RCT or even observational image datasets with a verified predictive imaging biomarker that would be suitable for benchmarking treatment effect estimation methods remain extremely limited. Therefore, the semi-synthetic setup based on the NSCLC-Radiomics dataset remains a viable alternative for the validation of image-based treatment effect estimation methods, particularly given the current lack of any widely accepted benchmark for imaging data. As argued by Brouillard et al. (2024), the availability of higher-quality datasets, including realistic synthetic datasets, and benchmarks is important for the development of causal models, both of which are addressed by the contributions of this thesis.

Beyond the availability of benchmarks, the two parts of this thesis also revealed other data-related limitations that impacted them both. One of the most critical factors was the limited dataset size, which had a stronger impact on the clinical imaging datasets (NSCLC-Radiomics and EORTC), and to some extent also on ISIC 2018, while being less pronounced on the larger CMNIST and CUB-200-2011 datasets. This, in turn, also affected the model performance and led to overfitting on the smaller datasets, which was partly mitigated by data augmentation. However, its effectiveness heavily depends on the dataset and the underlying treatment effect estimation task itself, as a heavy image augmentation scheme can negatively impact the prognostic or predictive signals. In this context, other design choices can influence the performance. For example, choosing an image cropped to the pathology or tumor bounding box (as done for the NSCLC-Radiomics dataset) or the whole image (as done for the EORTC dataset) as input can either exclude potentially

relevant information or introduce redundant and noisy information, respectively. This highlights the need for data-centric adaptations to strike a balance between over- and underfitting when applying image-based CATE estimation models.

The underlying motivation throughout the thesis was to explore whether end-to-end learning approaches can be used to discover predictive imaging biomarkers and estimate treatment effects from images without requiring handcrafted features or a separate feature extraction step, as is commonly done in radiomics pipelines. This was successfully demonstrated, in particular, in the first part of the thesis using simple synthetic linear outcomes. A key advantage of the proposed approach is that relevant information can be learned directly from the raw images without imposing prior assumptions that may bias the discovery process, such as defining a region of interest for feature extraction. This makes the approach potentially more general and flexible and well-suited for exploratory biomarker discovery, especially when larger and more diverse datasets become available. Although no direct comparisons to radiomics were conducted in this work, the methods explored here should be viewed as complementary, rather than competitive with existing radiomics pipelines.

Radiomics features have the advantage of being more explicitly interpretable and easier to integrate into regression-based or causal inference models for tabular data. The experiments from the second part of the thesis, where simpler clinical tabular regression models often outperformed deep image-based models, imply that tabular models with radiomics features may also be more stable in such a low-sample setting. Combining the proposed image-based CATE estimation approaches with radiomics may therefore be a promising future direction in small or noisy clinical datasets, which are discussed further in Section 6.4.

6.4 Future Research Directions

The challenges and limitations discussed in earlier sections of this chapter outline several open problems that need to be addressed to improve the robustness, scalability, and clinical applicability of the proposed image-based CATE estimation methods and their downstream tasks. They motivate several directions for future research, spanning data-centric solutions, model development, evaluation methodology, and clinical applications.

Data Scale and Availability

While the experiments in this thesis demonstrated the feasibility of estimating treatment effects from medical imaging data, they also highlighted that one of the most critical current limitations is the availability of suitable data. As previously emphasized by Curth et al. (2024), the performance of treatment effect estimation models fundamentally depends on the quality and quantity of data available. For example, the available image signal

could be enhanced by employing more advanced imaging modalities such as different MRI techniques (e.g. diffusion or perfusion imaging) or contrast agents. However, simply acquiring more imaging data is often infeasible, particularly in clinical studies, due to high cost, time, and logistical complexity.

For this reason, causal machine learning methods for CATE estimation have often focused on more abundant observational datasets, where the treatment is not assigned randomly and can depend on patient covariates. Leveraging such real-world data is promising for future work including predictive biomarker discovery, as noted by (Weberpals et al. 2025), as it is often more representative of the real patient populations. Since the approaches used in this thesis are based on models, such as TARNet (Shalit et al. 2017) and BITES (Schrod et al. 2022), which were originally developed for observational settings, they could similarly be applied to data with non-randomized treatments in future work, as long as assumption that there are no unmeasured confounders holds. In this setting it needs to be considered that common evaluation metrics such as risk need to be adjusted using inverse probability of treatment weighting (Austin et al. 2015).

Another possible solution is to leverage multiple studies through either pooling multiple datasets or federated learning, which aims to train a model across multiple sites while preserving privacy. Works from Makhija et al. (2024), L. Han et al. (2025), and Ogier du Terrail et al. (2025) employ federated learning for treatment effect estimation primarily on tabular data or electronic health records and could be extended to imaging data.

While this thesis addressed the lack of standardized benchmarks through semi-synthetic datasets and a tailored evaluation setup for image-based treatment effect estimation, further work can expand on this foundation to establish more generalizable benchmarks across different disease areas and downstream tasks. Such a benchmark would further support method development, enable reproducible comparisons of different image-based CATE estimation models, and facilitate broader clinical translation. Although Cadei et al. (2024) has introduced a visual causal inference benchmark (ISTAnt) using an RCT dataset with ant videos, similar standardized benchmarks in the medical imaging domain, especially in combination with survival outcomes, remain relevant for future work. To that end, future studies could apply extended image-based CATE estimation model developed in this thesis to other publicly available large-scale clinical imaging datasets, such as the observational RADCURE dataset (Welch et al. 2023; Welch et al. 2024), which contains pre-treatment CT scans of 3,346 head and neck cancer patients along with radiotherapy treatment data and survival outcomes, to help establish realistic benchmarks for other use cases further.

Multimodal Integration

This thesis explored multimodal integration as an avenue for improving deep-learning-based CATE estimation performance. To leverage the often superior performance of the clinical tabular-only regression baselines, the integration of Cox regression models with

the image-based deep-learning models could be explored further to allow the network to prioritize the prognostic tabular information during training. For example, possible strategies to achieve this could include late fusion approaches combing the final outputs of the models and other information fusion strategies beyond simple concatenation and a DAFT block. To account for dataset imbalances across treatment arms, sampling strategies during training and balancing approaches, similar to the IPM term of the BITES loss function, could be combined with the fusion strategies of representations to handle modality-specific imbalances, such as the representation balancing approach used by Ma et al. (2024). Future work could also explore the incorporating of additional data modalities, such as genomics or electronic health records data, which may offer complementary information to improve the treatment recommendation performance, but may also require specialized architectures to process modalities such as text or structured inputs.

Pre-training, Transfer Learning and Foundation Models

Using pre-trained image encoders offered some modest benefits in the experiments presented in this thesis, suggesting that pre-training and transfer learning for image-based treatment effect estimation are promising areas for future work, especially when the dataset size is limited or noisy. An immediate next step would be to explore a broader range of pre-trained image encoders trained using different self-supervised learning strategies. Well-suited examples include the ones pre-trained on a large-scale brain MRI dataset by Wald et al. (2025) such as VoCo (Wu et al. 2024) or SimCLR (Chen et al. 2020). Another step would be to compare different fine-tuning strategies, such as using frozen image encoders, different multi-stage warm-up schedules or parameter-efficient fine-tuning (Z. Han et al. 2024).

An important extension of this work would be to systematically evaluate these pre-trained image encoders within CATE estimation models for survival outcomes. Motivated by the time-to-event pre-training framework proposed by Huo et al. (2025) and the radiology foundation model proposed by Dancette et al. (2025), where pre-training on a large-scale observational medical imaging dataset led to improvements in image-based survival prediction, similar techniques could be adapted for treatment effect estimation and predictive imaging discovery.

Furthermore, recent progress in pre-training causal models on large-scale observational datasets for treatment effect estimation on tabular or other structured data, such as the CURE framework (R. Liu et al. 2024), the methods proposed by Zhou et al. (2025) or Zhang et al. (2024), could inspire similar strategies for image-based models.

Together with the approaches presented in this thesis, these advancements also motivate the development of a foundation model for image-based treatment effect estimation, either specialized for specific disease areas (e.g. glioblastoma) or designed for predictive imaging biomarker discovery from clinical imaging datasets in general. Combining

image-based encoders with large vision–language models could potentially further guide the interpretation of predictive imaging biomarkers and leverage multimodal clinical data with text-based clinical information (Ligero et al. 2025; Xiang et al. 2025).

Importantly, the evaluation protocol and the experimental setup with (semi-)synthetic outcomes based on pre-defined imaging biomarkers, as proposed in the first part of this thesis, can be readily used to benchmark future models with respect to predictive biomarker discovery, a task that remains largely overlooked in the development of image-based foundation models.

Evaluation and Interpretability

Especially in low-data regimes, evaluating predictive imaging biomarker discovery and image-based treatment effect estimation remains challenging, as shown in the second part of the thesis. The evaluation strategies presented in this work could be further enhanced in the future by integrating uncertainty quantification. This may help characterize the confidence level of the employed models and increase trust in the resulting treatment recommendations (Durso-Finley et al. 2023), especially when data limitations lead to overfitting or a high variance across cross-validation folds, and support the quantitative evaluation results of the predictive imaging biomarker discovery task.

The qualitative evaluation in this thesis, using attribution maps to analyze potential imaging biomarker candidates, showed that the interpretation becomes difficult when the attribution maps focus on the same region. Such results could indicate that the model identifies potentially overlapping predictive and prognostic imaging biomarkers with different meanings or an imaging biomarker that is both predictive and prognostic. This motivates the development of more advanced XAI methods to support treatment effect estimation in the image domain, such as using counterfactual explanations (Goyal et al. 2019) or methods that combine this with learning disentangled representations to separate predictive from prognostic contributions (Chu et al. 2021; Martínez 2021).

Applications

The treatment recommendations produced by the CATE estimation models in this thesis followed a simple policy, where patients were assigned to treatment if the estimated treatment effect was positive. For the translation to clinical practice, future work could explore risk-aware treatment recommendation policies that also account for factors such as cost or potential negative side effects and adjust the treatment decision threshold accordingly.

While risk-aware treatment recommendations have been investigated in the context of categorical or continuous outcomes by Durso-Finley et al. (2022), such future work could explore how such an approach can be extended to image-based survival-outcome models from this thesis. Further work could also consider taking the full estimated survival curves into account for modeling time-varying treatment, instead of computing the CATE from median survival times alone.

The clinical translation of the models developed in this thesis would also require addressing additional challenges, such as managing distribution shifts across sites, detecting unfavorable or erroneous treatment recommendations via failure detection. Ultimately, prospective studies will be needed to validate both predictive imaging biomarker and treatment effects in real-world settings.

7 Conclusion

This thesis investigated how image-based treatment effect estimation models can contribute to discovering predictive imaging biomarkers for predicting future treatment benefits and making image-based treatment recommendations using data from randomized studies. The main motivation behind this work was its potential in helping to gain scientific insights into what kind of information clinical imaging data can provide regarding patient outcomes and to ultimately contribute to advancing image-based personalized medicine and improving patient outcomes.

In the following, the main findings and contributions are summarized and related to the research questions posed in Section 1.2.

Summary of Contributions

Part 1: Evaluating heterogeneous treatment effect estimation models for predictive imaging biomarker discovery. The first part of this thesis introduced an approach for discovering predictive imaging biomarkers from pre-treatment imaging data. This was defined as a novel machine learning task, where predictive imaging biomarkers are identified in a data-driven manner using causal inference models—specifically, deep-learning-based models for estimating heterogeneous treatment effects from image inputs, also known as CATE estimation models. This approach is complementary to the common approach of using image features from a separate (deep-learning) feature extraction step or handcrafted features, such as radiomics features, but which might lead to potential biases. One advantage of using CATE estimation models is that by explicitly modeling heterogeneous treatment effects using these models, the risk of potentially conflating prognostic with predictive imaging biomarkers is reduced.

One of the main contributions of this first part is an evaluation protocol that was developed for this previously defined predictive imaging biomarker discovery task, which has two purposes: The first purpose is to evaluate a predictive imaging biomarker candidate identified by the model itself quantitatively and qualitatively, independent of whether

a ground truth exists. Its second purpose is validating and benchmarking the performance of a treatment effect estimation model at identifying a known predictive imaging biomarker feature.

The quantitative part of the proposed evaluation protocol supports the evaluation of the estimated predictive biomarker strength through statistical tests of the biomarker-by-treatment interaction and directly comparing the predictive effects to prognostic effects. The qualitative evaluation part employs XAI methods, specifically attribution maps, to support the visual interpretation of the identified predictive imaging biomarker candidate.

The first part of the thesis introduced benchmarking experiments for the task of identifying predictive imaging biomarkers, where outcomes are simulated using real image features by altering the strength of predictive and prognostic biomarkers. The experiments on a variety of image datasets and imaging biomarker features demonstrated that the proposed approach using image-based CATE estimation models can indeed successfully identify these imaging biomarkers (RQ1.1). Using these reproducible experiments, it could also be shown that the evaluation protocol can successfully be used to measure model performance (RQ1.2), therefore offering valuable insights for improving and further developing image-based CATE estimation models.

Part 2: Image-based heterogeneous treatment effect estimation in clinical imaging studies. Building on these insights gained from the benchmarking experiments on semi-synthetic image datasets in a controlled setting regarding the value of using image-based CATE estimation, the second part of this thesis developed approaches for translating these models to real-world clinical imaging datasets. This second part of the thesis developed and evaluated a multimodal CATE estimation model for survival outcomes in an RCT setting, combining imaging and tabular data as an input. Motivated by accounting for censoring and more accurately capturing the overall survival, which is often the main outcome of interest in clinical trials, it thereby extended previous image-based approaches limited to continuous or categorical outcomes only.

The experiments presented the first evaluation of such an image-based CATE estimation model in glioblastoma. Specifically, data from a real clinical trial (EORTC) was used to assess heterogeneous treatment effects of using bevacizumab, as well as the resulting patient treatment recommendations and the potential presence of predictive imaging biomarkers for this experimental treatment (RQ2.3).

Further, a lung cancer CT dataset (NSCLC-Radiomics) was used to simulate semi-synthetic RCT survival outcomes with a controlled treatment effect based on real observational outcomes, which was leveraged for accurately validating the treatment recommendation performances of the proposed models.

With these datasets, this work assessed whether modeling survival outcomes instead of binarized survival times could improve treatment recommendations (RQ2.1) and whether

integrating clinical tabular data, segmentation masks, and performing multitask learning has an impact on model performance and robustness (RQ2.2). This work was also the first to investigate the impact of integrating pre-trained image encoders on estimating treatment effects and making treatment recommendations from MRI data.

Using experiments and an extensive evaluation setup, including an analysis of stratified subgroups and baseline comparisons, critical dataset-related and methodological limitations in applying the image-based CATE estimation models to clinical imaging data could be revealed. These limitations include the risk of overfitting due to small sample sizes, limited treatment effect signals, and the underutilization of the clinical tabular data. While modeling survival outcomes and using pre-treatment image encoders led to modest improvements, the impact of multimodal integration and multitask learning was mostly inconsistent. Additionally, results did not indicate evidence for a strong predictive imaging biomarker in the EORTC dataset.

Even though the results suggest that methods developed for large imaging datasets with strong treatment effect signals may not directly translate to clinical imaging studies with limited sample sizes, they identified important practical and methodological challenges that future work must overcome to make accurate image-based treatment recommendations and for its downstream tasks.

Outlook

The findings of this thesis demonstrate both the opportunities offered by image-based treatment effect estimation models for discovering predictive imaging biomarkers and making image-based treatment recommendations, as well as the current limitations of applying such an approach to currently available clinical imaging data. They highlight the need for more robust and generalizable methods combined with more realistic evaluation benchmarks.

Despite the challenges, this thesis also offers important insights and outlines future research directions for the development of deep learning methods at the intersection of causal inference, computer vision, and medical research. These include leveraging observational data and employing domain-specific pre-training or the use of foundation models. The models and evaluation methods developed here may serve as a valuable foundation for advancing the field of image-based decision-making and for inspiring applications of image-based treatment effect estimation in the medical domain and beyond.

List of Own Contributions and Publications

A

This thesis was written at the Division of Medical Image Computing (MIC) at the German Cancer Research Center (DKFZ), Heidelberg, under the supervision of its head of division, Prof. Dr. Klaus Maier-Hein, in collaboration with my interdisciplinary secondary supervisor, Prof. Dr. Philipp Vollmuth, at the University Hospital Bonn and University of Bonn. I closely collaborated with Dr. Paul Jäger, a member of my Thesis Advisory Committee, and received additional input from members of MIC, including Dr. Jens Petersen, Lukas Klein, and Jonas Bohn.

All research presented in this thesis was primarily conceived and implemented by me, unless stated otherwise. I developed the methods, designed and conducted the experiments and performed the data analysis and evaluation. Prof. Dr. Philipp Vollmuth was involved in the initial conception of the clinical research direction and provided access to the EORTC data used in Part 2 of this thesis: *Image-Based Heterogeneous Treatment Effect Estimation in Clinical Imaging Studies*, including annotations and clinical information.

I have contributed to following peer-reviewed publications or works currently under review:

First Author Publication

1. **Shuhan Xiao**, Lukas Klein, Jens Petersen, Philipp Vollmuth, Paul F Jaeger[1], and Klaus H Maier-Hein[1] (2025). "Enhancing predictive imaging biomarker discovery through treatment effect analysis". In: *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. IEEE, pp. 4512–4522. 10.1109/WACV61041.2025.00443.

This publication is related to Part 1 of this thesis: *Evaluating Heterogeneous Treatment Effect Estimation Models for Predictive Imaging Biomarker Discovery*. I was responsible for

[1]Shared last authorship.

the project design, conducted the experiments and performed the analyses and wrote the initial manuscript draft. The XAI analysis was primarily carried out by Lukas Klein in close collaboration with me, and he also contributed to the related discussions in the manuscript.

Co-Authorships (not directly referenced in this thesis)

1. Maximilian Fischer, Peter Neher, Michael Götz, **Shuhan Xiao**, Silvia Dias Almeida, Peter Schüffler, Alexander Muckenhuber, Rickmer Braren, Jens Kleesiek, Marco Nolden, and Klaus Maier-Hein (2022). 'Deep Learning on Lossily Compressed Pathology Images: Adverse Effects for ImageNet Pre-trained Models'. In: *International Workshop on Medical Optical Imaging and Virtual Microscopy Image Analysis*. Springer, pp. 73–83

2. Maximilian Fischer, Peter Neher, Peter Schüffler, **Shuhan Xiao**, Silvia Dias Almeida, Constantin Ulrich, Alexander Muckenhuber, Rickmer Braren, Michael Götz, Jens Kleesiek, and Klaus Maier-Hein (2023). 'Enhanced diagnostic fidelity in pathology whole slide image compression via deep learning'. In: *International Workshop on Machine Learning in Medical Imaging*. Springer, pp. 427–436

3. Maximilian Fischer, Peter Neher, Tassilo Wald, Silvia Dias Almeida, **Shuhan Xiao**, Peter Schüffler, Rickmer Braren, Michael Götz, Alexander Muckenhuber, Jens Kleesiek, and Klaus Maier-Hein (2024). 'Learned image compression for he-stained histopathological images via stain deconvolution'. In: *International Workshop on Medical Optical Imaging and Virtual Microscopy Image Analysis*. Springer, pp. 97–107

4. Maximilian Fischer, Peter Neher, Peter Schüffler, Sebastian Ziegler, **Shuhan Xiao**, Robin Peretzke, David Clunie, Constantin Ulrich, Michael Baumgartner, Alexander Muckenhuber, Silvia Dias Almeida, Michael Gőtz, Jens Kleesiek, Marco Nolden, Rickmer Braren, and Klaus Maier-Hein (2025). 'Unlocking the potential of digital pathology: Novel baselines for compression'. In: *Journal of Pathology Informatics* 17, p. 100421. ISSN: 2153-3539. DOI: https://doi.org/10.1016/j.jpi.2025.100421

5. Arvin von Salomon, Jessica Kächele, Thomas Nonnenmacher, Markus Bujotzek, **Shuhan Xiao**, Andres Martinez Mora, Marina Hajiyianni, Ekaterina Menis, Martin Grözinger, Fabian Bauer, Veronika Riebl, Thomas Hielscher, Britta Besemer, Saif Afat, Ullrich Graeven, Adrian Ringelstein, Mathias Hänel, Dieter Fedders, Alexandra Ljimani, Gerald Antoch, Andreas H Mahnken, Elias Mai, Marc S Raab, Hartmut Goldschmidt, Tim F Weber, Heinz-Peter Schlemmer, Stefan Delorme, Klaus Maier-Hein, Peter Neher, and Markus Wennmann (2025). 'Automatische Detektion von fokalen Läsionen im MRT bei Patienten mit Multiplem Myelom–eine multizentrische Machbarkeitsstudie'. In: *RöFo-Fortschritte auf dem Gebiet der Röntgenstrahlen und der bildgebenden Verfahren*. Vol. 197. S 01. Georg Thieme Verlag KG, ab157

6. Stefan Denner, David Zimmerer, Dimitrios Bounias, Markus Bujotzek, **Shuhan Xiao**, Raphael Stock, Lisa Kausch, Philipp Schader, Tobias Penzkofer, Paul F. Jäger, and Klaus Maier-Hein (2025). In: *Computers in Biology and Medicine* 196, p. 110640. ISSN: 0010-4825. DOI: https://doi.org/10.1016/j.compbiomed.2025.110640. URL: https://www.sciencedirect.com/science/article/pii/S0010482525009916

7. Alexandra Ertl, **Shuhan Xiao**, Stefan Denner, Robin Peretzke, David Zimmerer, Peter Neher, Fabian Isensee, and Klaus Maier-Hein (2025). *nnLandmark: A Self-Configuring Method for 3D Medical Landmark Detection*. Pre-published. arXiv: 2504.06742 [cs.CV]. URL: https://arxiv.org/abs/2504.06742

8. Maximilian Fischer, Alexander Muckenhuber, Robin Peretzke, Luay Farah, Constantin Ulrich, Sebastian Ziegler, Philipp Schader, Lorenz Feineis, Hanno Gao, **Shuhan Xiao**, Marco Nolden, Katja Steiger, Jens Sieveke, Rickmer Braren, Jens Kleesiek, Peter Schüffler, Peter Neher, and Klaus Maier-Hein (2025). 'Contrastive virtual staining enhances deep learning-based PDAC subtyping from H&E-stained tissue cores'. In: *The Journal of Pathology*. DOI: 10.1002/path.6491

9. Markus Wennmann, Jessica Kächele, Arvin von Salomon, Tobias Nonnenmacher, Markus Bujotzek, **Shuhan Xiao**, Andres Martinez Mora, Thomas Hielscher, Marina Hajiyianni, Ekaterina Menis, Martin Grözinger, Fabian Bauer, Veronika Riebl, Lukas Rotkopf, Kevin Sun Zhang, Saif Afat, Britta Besemer, Martin Hoffmann, Adrian Ringelstein, Ullrich Graeven, Dieter Fedders, Mathias Hänel, Gerald Antoch, Roland Fenk, Andreas H. Mahnken, Christoph Mann, Theresa Mokry, Marc-Steffen Raab, Niels Weinhold, Elias Karl Mai, Hartmut Goldschmidt, Tim Frederik Weber, Stefan Delorme, Peter Neher, Heinz-Peter Schlemmer, and Klaus Maier-Hein (2025). 'Automated Detection of Focal Bone Marrow Lesions From MRI: A Multi-center Feasibility Study in Patients with Monoclonal Plasma Cell Disorders'. In: *Academic Radiology*. DOI: https://doi.org/10.1016/j.acra.2025.06.034

Additionally, I co-supervised the Master's thesis by Niklas Woebs with the title "Predicting overall survival time of glioblastoma patients using deep learning methods trained on irregular time series".

Appendix

B

B.1 Additional Results

Table B.1: Additional ablation results for experiments on the NSCLC-Radiomics dataset. The results are shown for three variations of the survival-outcome model for image and tabular inputs: the model with concatenated representation and no regularization term, the model with a DAFT module (Wolf et al. 2022) for multimodal integration instead, and the model trained using the BITES loss function (Schrod et al. 2022) with the hyperparameter $\alpha = 0.01$ for the IPM regularization term. Reported are the fraction of correctly assigned treatments (Decision Accuracy), root PEHE ($\sqrt{\epsilon_{PEHE}}$), observed policy value $\hat{V}_{Pol}$, and Antolini's C-index, with mean $\pm$ SD across folds. All metrics are IPCW-adjusted, except for the C-index.

Split	Method	Decision Acc	$\sqrt{\epsilon_{PEHE}} \downarrow [10^3 \text{d}]$	$\hat{V}_{Pol} \uparrow [10^3 \text{d}]$	C-Index
Validation	Survival	**0.54 ± 0.09**	**4.0 ± 1.8**	0.46 ± 0.14	**0.57 ± 0.03**
	Survival + DAFT	0.51 ± 0.08	4.5 ± 2.3	0.39 ± 0.12	0.55 ± 0.04
	Survival + α = 0.01	0.52 ± 0.03	4.7 ± 2.5	**0.84 ± 0.60**	0.53 ± 0.04

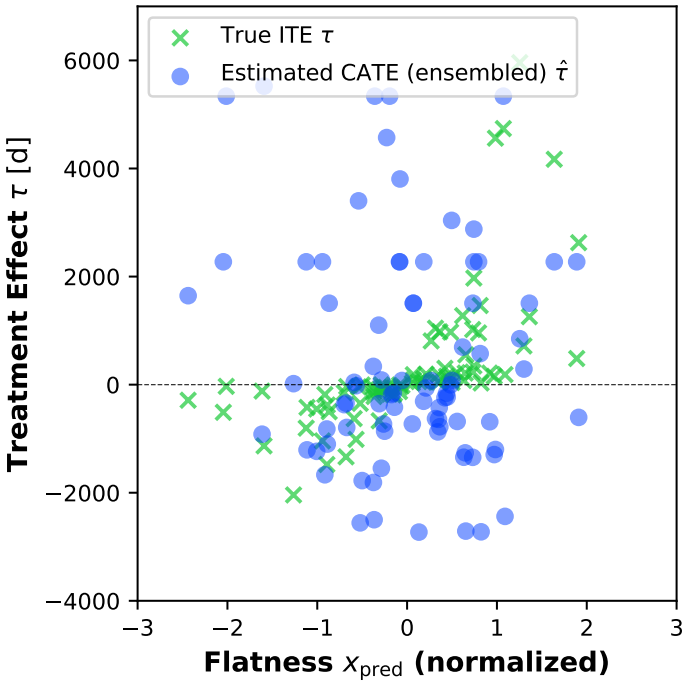


Figure B.1: ensembled result, -0.1195

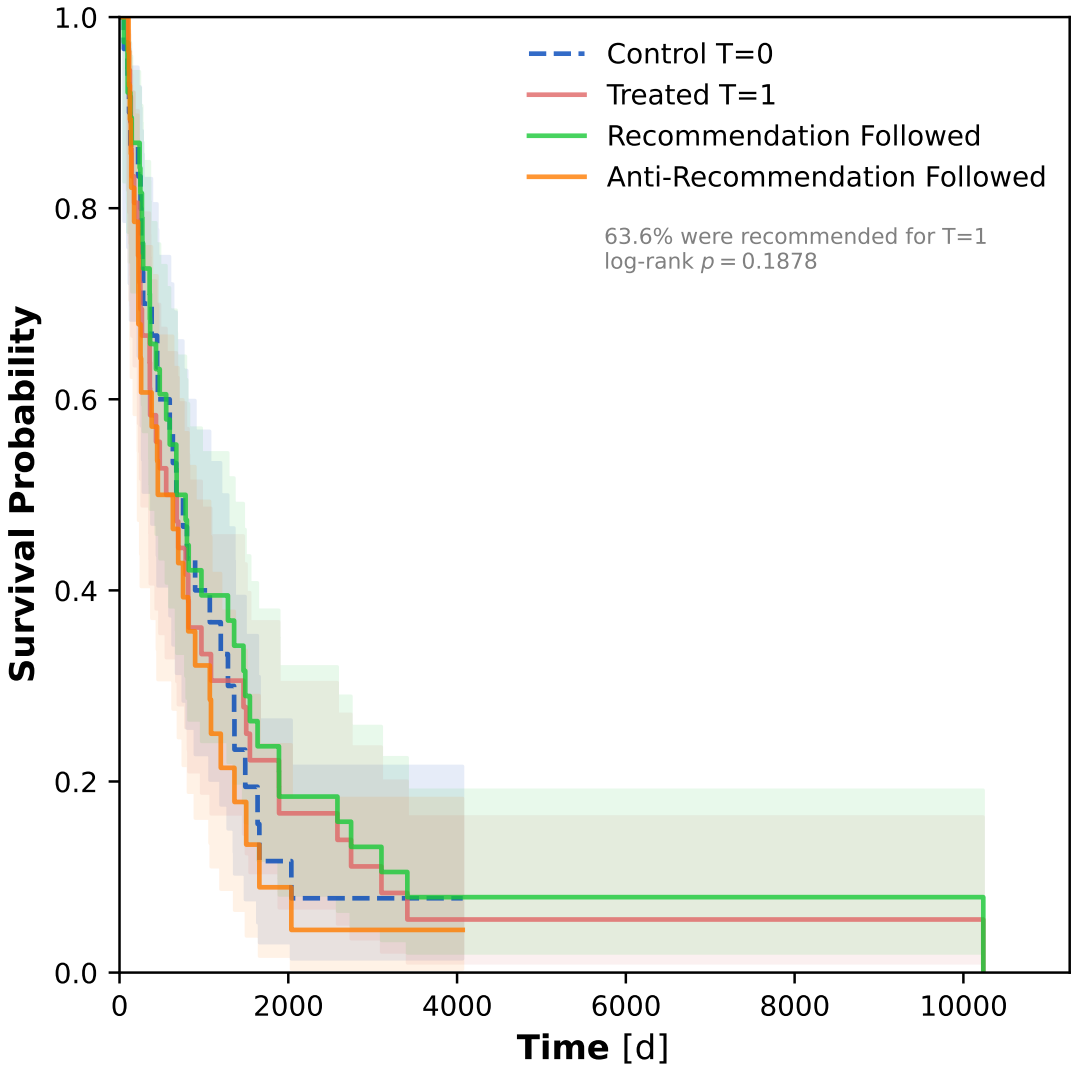
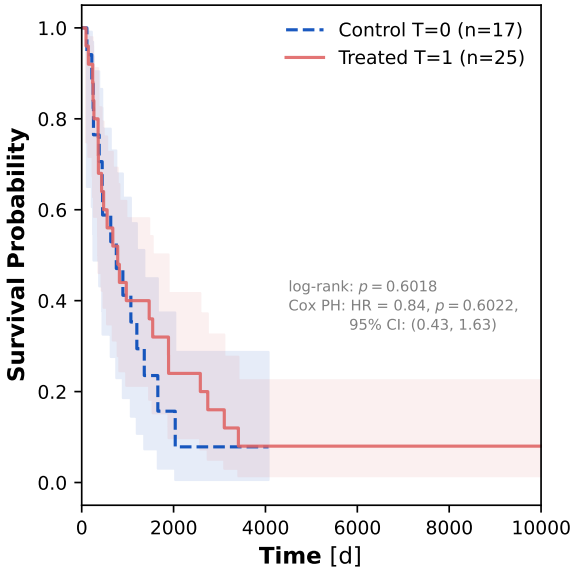
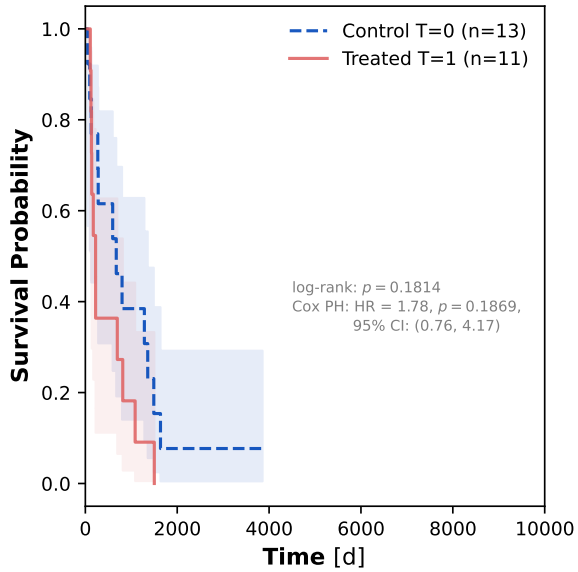


Figure B.2: Kaplan-Meier curves on the NSCLC-Radiomics dataset comparing the survival probability for patients who received the treatment recommended by the estimated CATE (green) versus those who did not (orange). For reference, the curves for the treated group ($T = 1$, red) and control group ($T = 0$, blue) are also shown. Results are based on the validation set of fold 2, using recommendations from the best-performing survival-outcome model for that fold (image-only input). Log-rank p-values are reported for reference.



(a) $\hat{\tau}_i > 0$ subgroup



(b) $\hat{\tau}_i \leq 0$ subgroup

Figure B.3: Kaplan–Meier curves on the NSCLC-Radiomics dataset for observed patient subgroups stratified by the sign of the estimated CATE ($\hat{\tau}_i > 0$ vs. $\hat{\tau}_i \leq 0$), where a positive CATE indicates that the patient is predicted to benefit from treatment $T = 1$. Within each subgroup, curves compare the survival probability for patients who were actually treated ($T = 1$, red) versus the control group ($T = 0$, blue). Results are shown on the validation set of fold 2 using the image-only model. Log-rank p-values and Cox proportional hazards results are reported for reference.

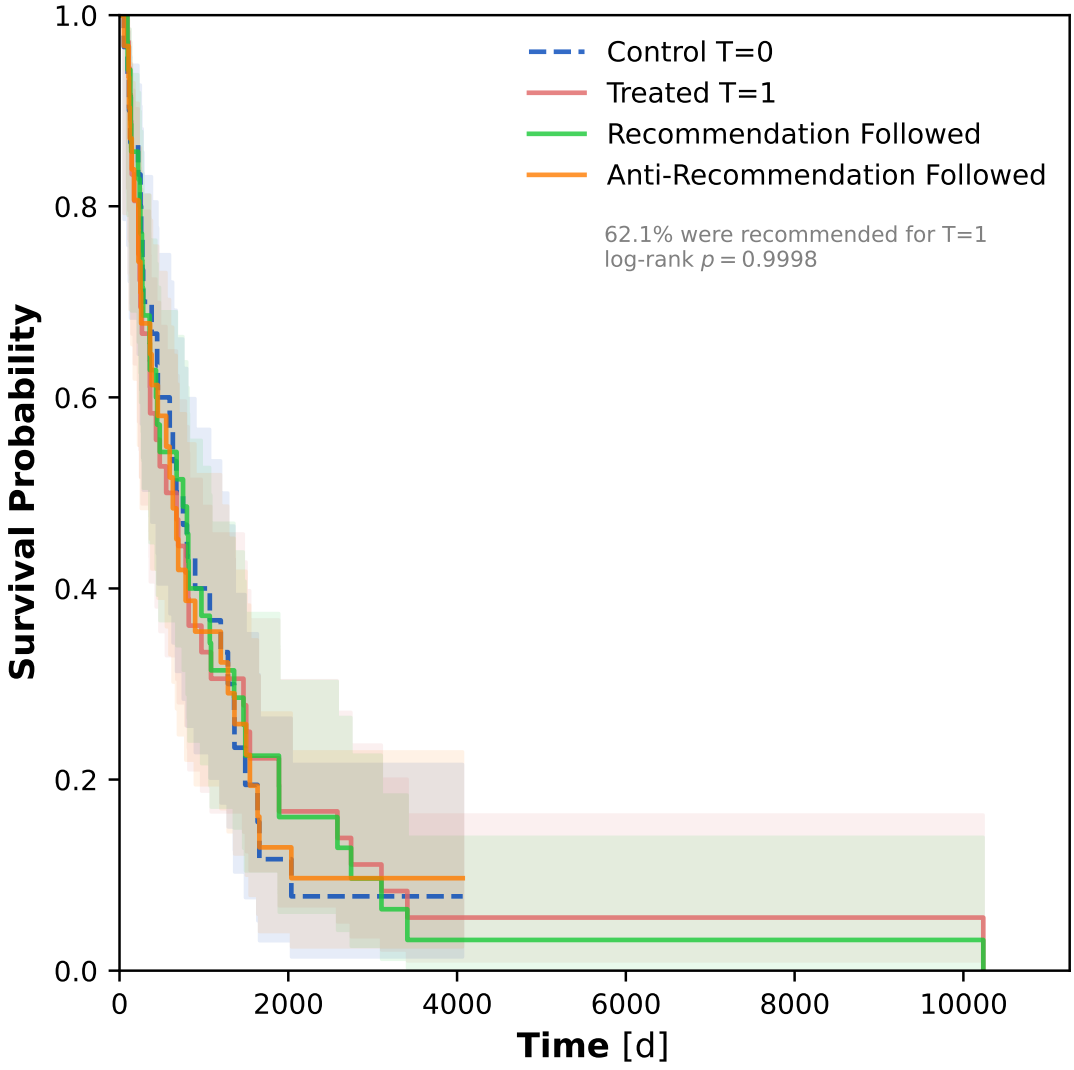
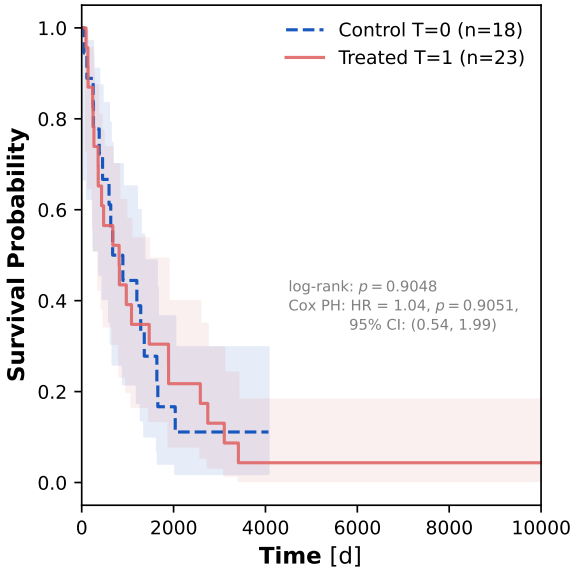
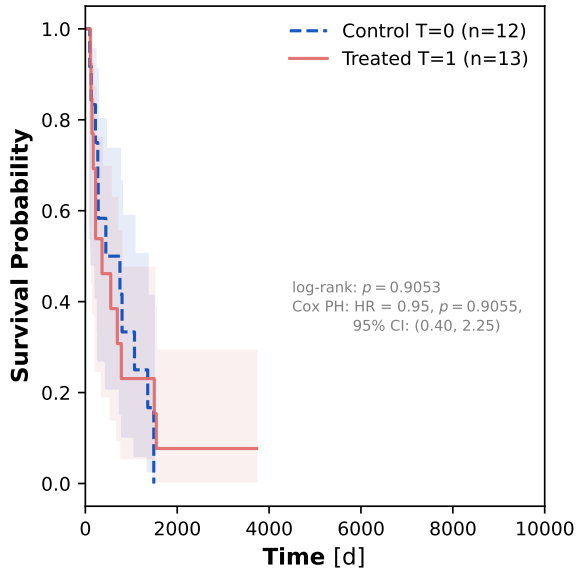


Figure B.4: Kaplan-Meier curves on the NSCLC-Radiomics dataset comparing the survival probability for patients who received the treatment recommended by the ground truth ITE (green) versus those who did not (orange). For reference, the curves for the treated group ($T = 1$, red) and control group ($T = 0$, blue) are also shown. Results are based on the validation set of fold 2. Log-rank p-values are reported for reference.

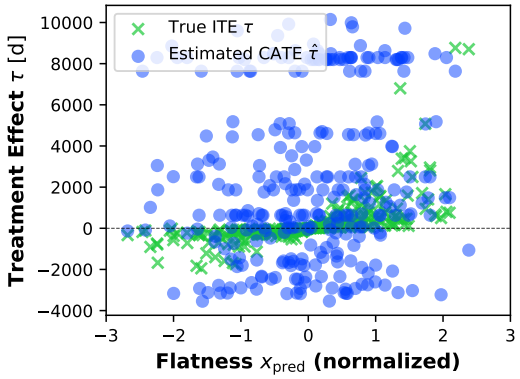


(a) $\tau_i > 0$ subgroup

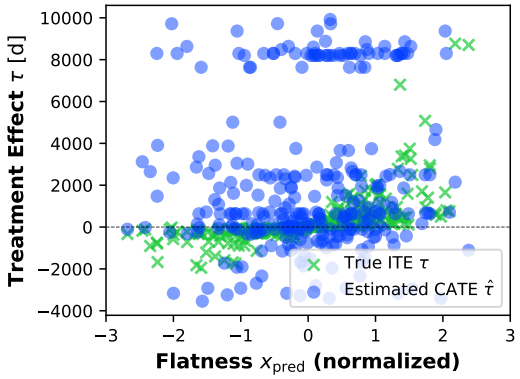


(b) $\tau_i \leq 0$ subgroup

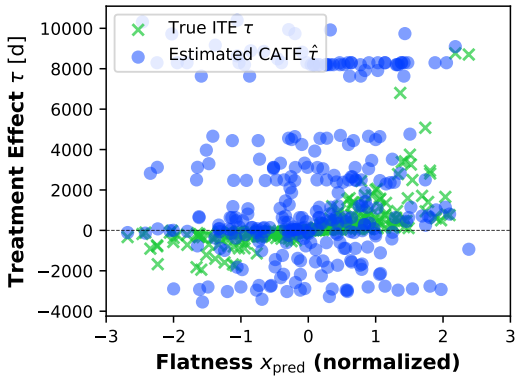
Figure B.5: Kaplan–Meier curves on the NSCLC-Radiomics dataset for observed patient subgroups stratified by the sign of the ground-truth ITE ($\tau_i > 0$ vs. $\tau_i \leq 0$), where a positive ITE indicates that the patient truly benefits from treatment $T = 1$. Within each subgroup, curves compare the survival probability for patients who were actually treated (T=1T=1, red) versus the control group ($T = 0$, blue). Results are shown on the validation set of fold 2. Log-rank p-values and Cox proportional hazards results are reported for reference.



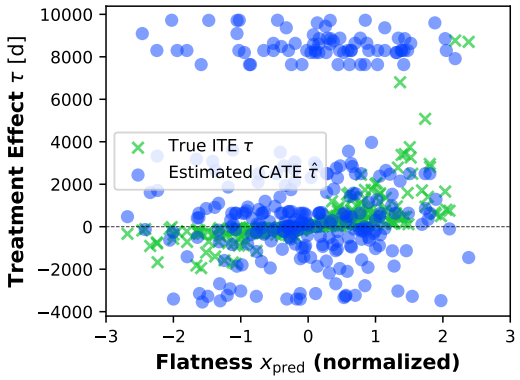
(a) Input modality: image



(b) Input modalities: image + segmentation



(c) Input modalities: image + tabular



(d) Input modalities: Image + Tabular + Segmentation

Figure B.6: Scatter plot of the z-score-normalized radiomics feature flatness vs. estimated ITE computed on each validation fold of the 5-fold cross validation models trained on different input modalities on the NSCLC-Radiomics dataset. For comparison, the relationship between flatness, which was used to simulate the semi-synthetic treatment effect, and the true ITE is shown.

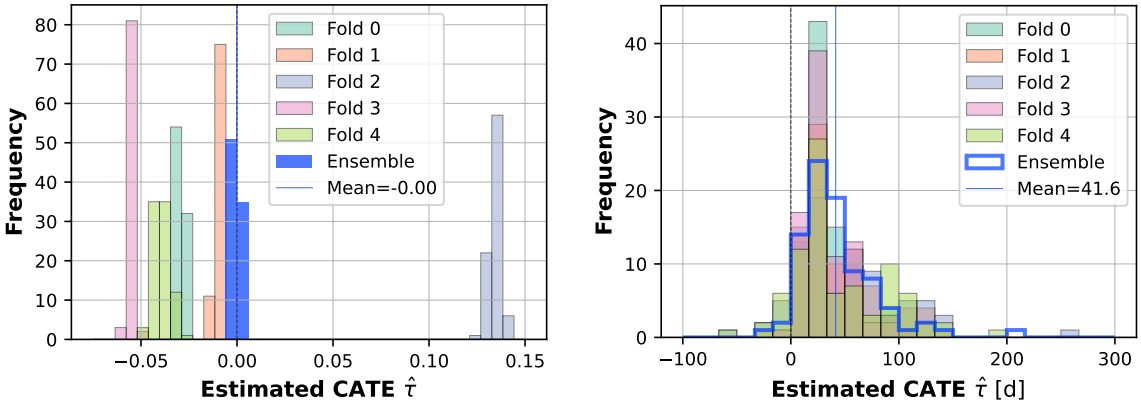
Table B.2: Full comparison of the proposed two-headed TARNet-like CATE estimation model with a single-headed S-Learner architecture sharing the same backbone and configuration. Reported are the observed policy risk $\widehat{R}_{Pol}$ or policy value $\widehat{V}_{Pol}$, as well as the Balanced Accuracy, AUROC and Antolini's C-index, with mean $\pm$ SD across folds. All metrics are IPCW-adjusted, except for the C-index.

a Performance of binary-outcome CATE estimation models.

Split	Model	Modalities		Multitask	$\widehat{R}_{Pol} \downarrow$	Balanced Acc ↑	AUROC ↑
		Image	Tabular				
Val.	S-Learner	✓	-	-	0.858 ± 0.038	0.520 ± 0.045	0.52 ± 0.06
		✓	✓	-	0.858 ± 0.038	0.506 ± 0.014	0.54 ± 0.06
		✓	✓	✓	**0.810 ± 0.053**	0.500 ± 0.000	0.58 ± 0.06
	TARNet	✓	-	-	0.858 ± 0.038	**0.541 ± 0.092**	0.50 ± 0.14
		✓	✓	-	0.858 ± 0.038	**0.541 ± 0.092**	0.51 ± 0.14
		✓	✓	✓	0.831 ± 0.047	0.518 ± 0.041	**0.61 ± 0.05**
Test	S-Learner	✓	-	-	0.821 ± 0.000	0.499 ± 0.002	**0.55 ± 0.01**
		✓	✓	-	0.821 ± 0.000	0.500 ± 0.000	0.53 ± 0.01
		✓	✓	✓	**0.805 ± 0.023**	0.500 ± 0.000	0.46 ± 0.07
	TARNet	✓	-	-	0.821 ± 0.000	**0.501 ± 0.000**	0.50 ± 0.00
		✓	✓	-	0.821 ± 0.000	**0.501 ± 0.000**	0.51 ± 0.02
		✓	✓	✓	0.811 ± 0.022	0.500 ± 0.000	0.49 ± 0.02

b Performance of survival-outcome CATE estimation models.

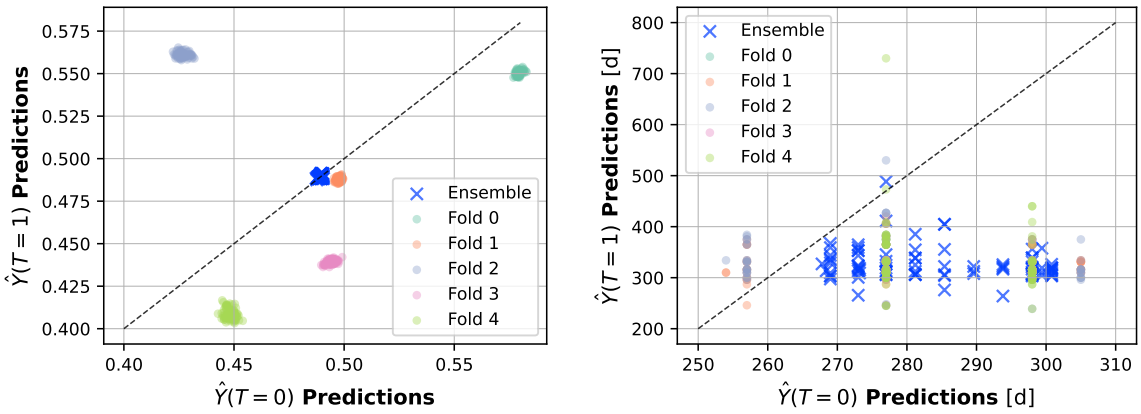
Split	Model	Modalities		Multitask	$\widehat{V}_{Pol} \uparrow [10^3 \text{d}]$	C-Index ↑
		Image	Tabular			
Val.	S-Learner	✓	-	-	0.187 ± 0.080	0.492 ± 0.027
		✓	✓	-	**0.200 ± 0.073**	0.506 ± 0.030
		✓	✓	✓	0.107 ± 0.009	0.504 ± 0.019
	TARNet	✓	-	-	0.137 ± 0.041	0.497 ± 0.027
		✓	✓	-	0.147 ± 0.046	0.501 ± 0.026
		✓	✓	✓	0.109 ± 0.007	**0.527 ± 0.035**
Test	S-Learner	✓	-	-	0.178 ± 0.015	0.520 ± 0.030
		✓	✓	-	**0.184 ± 0.013**	0.527 ± 0.020
		✓	✓	✓	0.161 ± 0.000	0.522 ± 0.028
	TARNet	✓	-	-	0.165 ± 0.015	0.546 ± 0.004
		✓	✓	-	0.165 ± 0.014	0.556 ± 0.017
		✓	✓	✓	0.161 ± 0.000	**0.576 ± 0.011**



(a) Histogram for the binary-outcome model.

(b) Histogram for the survival-outcome model.

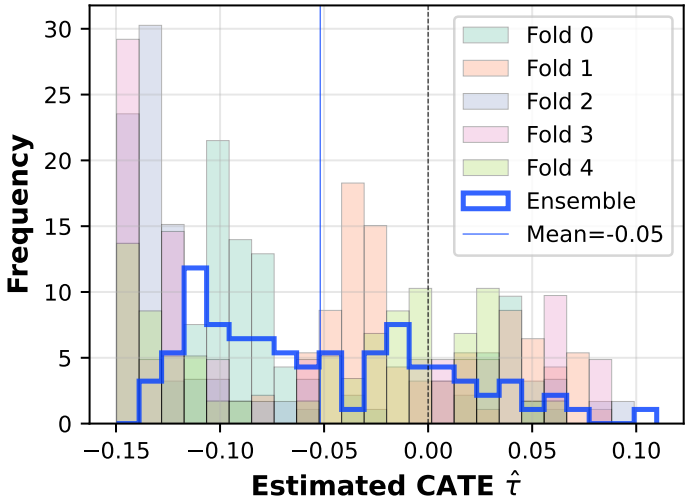
Figure B.7: Histograms showing the distribution of the estimated CATEs $\hat{\tau}$ on EORTC dataset for the individual cross-validation folds and for the resulting ensemble. Results are shown on the hold-out test set using cross-validation models selected based on validation performance.



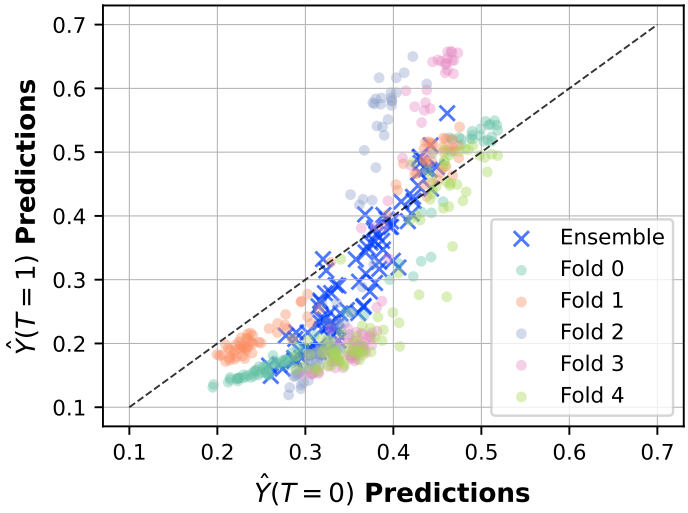
(a) Binary-outcome model predictions (probabilities).

(b) Survival-outcome model predictions (median survival time).

Figure B.8: Scatter plots showing the predicted control group outcome $\hat{Y}(T=0)$ versus the predicted treatment group outcome $\hat{Y}(T=1)$ on EORTC dataset for the individual cross-validation folds and for the resulting ensemble. Results are shown on the hold-out test set using cross-validation models selected based on validation performance.



(a) Histogram of estimated CATEs.



(b) Scatter plot of the predicted probabilities.

Figure B.9: Results for the estimated CATEs and predicted outcome probabilities of the control group $\hat{Y}(T=0)$ versus the treatment group outcome $\hat{Y}(T=1)$ predicted outcome probabilities of a model fine-tuned a pre-trained ResEnc-L encoder for different folds and ensemble on the EORTC dataset. Results are shown on the hold-out test set using cross-validation models selected based on validation performance (MAE encoder released by (Wald et al. 2025)).

Table B.3: Evaluation results of the predictive strength of the estimated CATE from binary-outcome CATE estimation models using the ResEnc-L backbone, trained from scratch or fine-tuned from a pre-trained MAE-encoder (Wald et al. 2025) on the EORTC dataset. The policy value ($\widehat{V}_{Pol}$) is computed using survival outcomes. Reported are the same Cox regression statistics as in Table 5.10: the p -value from the Wald test for biomarker-by-treatment interaction term, the ratio of absolute Wald z -statistics $|z_{pred}/z_{prog}|$ for the predictive vs. prognostic term, and the p -value of the likelihood ratio test.

Split	Model	$\widehat{V}_{Pol} \uparrow [10^3\text{d}]$	**Wald** p	$\lvert z_{pred}/z_{prog}\rvert$	**LR** p
Test	ResEnc-L Bin-TARNet (From Scratch)	0.173 ± 0.010	0.731	0.35	0.731
	ResEnc-L Bin-TARNet (MAE)	0.185 ± 0.017	0.662	0.60	0.660

List of Acronyms

AI artificial intelligence

AP average precision

ATE average treatment effect

AUROC area under the receiver operating characteristic curve

BEV bevacizumab

BITES Balanced Individual Treatment Effect for Survival Data

C-index concordance index

CATE conditional average treatment effect

CE cross-entropy

CI confidence interval

CNN convolutional neural network

CT computed tomography

CUB Caltech-UCSD Birds

DKFZ German Cancer Research Center

EG Expected Gradients

EORTC European Organisation for Research and Treatment of Cancer

FLAIR fluid-attenuated inversion recovery

GPU graphics processing unit

Grad-CAM Gradient-weighted Class Activation Mapping

HR hazard ratio

HTE heterogeneous treatment effect

HU Hounsfield unit

IPCW inverse probability of censoring weighting

IPM integral probability metric

ISIC International Skin Imaging Collaboration

ITE individual treatment effect

MAE mean absolute error

MAE masked autoencoder

MIC Medical Image Computing

MITK Medical Imaging Interaction Toolkit

MNIST Modified National Institute of Standards and Technology database

MRI magnetic resonance imaging

MSE mean squared error

NSCLC non-small-cell lung cancer

OS overall survival

PEHE precision of estimating heterogeneous effects

RCT randomized controlled trial

ResEnc residual encoder

ResNet residual neural network

RQ research question

SD standard deviation

SGD stochastic gradient descent

T1-w T1-weighted

T2-w T2-weighted

TARNet Treatment-Agnostic Representation Network

WHO World Health Organization

XAI explainable artificial intelligence

List of Figures

List of Tables

Bibliography

Aerts, H. J. W. L., L. Wee, E. Rios Velazquez, R. T. H. Leijenaar, C. Parmar, P. Grossmann, S. Carvalho, J. Bussink, R. Monshouwer, B. Haibe-Kains, D. Rietveld, F. Hoebers, M. M. Rietbergen, C. R. Leemans, A. Dekker, J. Quackenbush, R. J. Gillies, and P. Lambin (2014). *Data From NSCLC-Radiomics (version 4)*. https://doi.org/10.7937/K9/TCIA.2015.PF0M9REI. Data set. The Cancer Imaging Archive (cit. on p. 38).

Aerts, Hugo J. W. L., Emmanuel Rios Velazquez, Ralph T. H. Leijenaar, Chintan Parmar, Patrick Grossmann, Sara Cavalho, Johan Bussink, Rene Monshouwer, Benjamin Haibe-Kains, Derek Rietveld, Frank Hoebers, Michelle M. Rietbergen, C. Rene Leemans, Andre Dekker, John Quackenbush, Robert J. Gillies, and Philippe Lambin (2014). 'Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach'. In: *Nature communications* 5.1, p. 4006 (cit. on pp. 10, 39, 42).

Alaa, Ahmed M, Michael Weisz, and Mihaela Van Der Schaar (2017). 'Deep counterfactual networks with propensity-dropout'. In: *arXiv preprint arXiv:1706.05966* (cit. on p. 24).

Aloui, Ahmed, Juncheng Dong, Cat P Le, and Vahid Tarokh (2023). 'Transfer learning for individual treatment effect estimation'. In: *Uncertainty in Artificial Intelligence*. PMLR, pp. 56–66 (cit. on p. 29).

Alymani, Nayef A, Murray D Smith, David J Williams, and Russell D Petty (2010). 'Predictive biomarkers for personalised anti-cancer drug use: discovery to clinical implementation'. In: *European Journal of Cancer* 46.5, pp. 869–879 (cit. on p. 3).

Ameratunga, Malaka, Nick Pavlakis, Helen Wheeler, Robin Grant, John Simes, and Mustafa Khasraw (2018). 'Anti-angiogenic therapy for high-grade glioma'. In: *Cochrane Database of Systematic Reviews* 11 (cit. on p. 4).

Ammari, Samy, Raoul Sallé de Chou, Tarek Assi, Mehdi Touat, Emilie Chouzenoux, Arnaud Quillent, Elaine Limkin, Laurent Dercle, Joya Hadchiti, Mickael Elhaik, et al. (2021). 'Machine-learning-based radiomics MRI model for survival prediction of recurrent glioblastomas treated with bevacizumab'. In: *Diagnostics* 11.7, p. 1263 (cit. on pp. 4, 121).

Amsterdam, WAC van, JJC Verhoeff, PA de Jong, T Leiner, and MJC Eijkemans (2019). 'Eliminating biasing signals in lung cancer images for prognosis predictions with deep learning'. In: *NPJ digital medicine* 2.1, pp. 1–6 (cit. on p. 24).

Antolini, Laura, Patrizia Boracchi, and Elia Biganzoli (2005). 'A time-dependent discrimination index for survival data'. In: *Statistics in medicine* 24.24, pp. 3927–3944 (cit. on pp. 18, 55).

Antonia, Scott J, Augusto Villegas, Davey Daniel, David Vicente, Shuji Murakami, Rina Hui, Takayasu Kurata, Alberto Chiappori, Ki H Lee, Maike De Wit, et al. (2018). 'Overall survival with durvalumab after chemoradiotherapy in stage III NSCLC'. In: *New England Journal of Medicine* 379.24, pp. 2342–2350 (cit. on p. 44).

Arango-Argoty, Gustavo, Damian E Bikiel, Gerald J Sun, Elly Kipkogei, Kaitlin M Smith, Sebastian Carrasco Pro, Elizabeth Y Choe, and Etai Jacob (2025). 'AI-driven predictive biomarker discovery with contrastive learning to improve clinical trial outcomes'. In: *Cancer Cell* 43.5, pp. 875–890 (cit. on pp. 22, 25).

Arjovsky, Martín, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz (2019). 'Invariant Risk Minimization'. In: *CoRR* abs/1907.02893. URL: http://arxiv.org/abs/1907.02893 (cit. on p. 38).

Ascarza, Eva (2018). 'Retention futility: Targeting high-risk customers might be ineffective'. In: *Journal of marketing Research* 55.1, pp. 80–98 (cit. on p. 101).

Austin, Peter C and Elizabeth A Stuart (2015). 'Moving towards best practice when using inverse probability of treatment weighting (IPTW) using the propensity score to estimate causal treatment effects in observational studies'. In: *Statistics in medicine* 34.28, pp. 3661–3679 (cit. on p. 125).

Bahamyirou, Asma, Mireille E Schnitzer, Edward H Kennedy, Lucie Blais, and Yi Yang (2022). 'Doubly robust adaptive LASSO for effect modifier discovery'. In: *The International Journal of Biostatistics* 18.2, pp. 307–327 (cit. on pp. 3, 22, 25).

Ballman, Karla V (2015). 'Biomarker: predictive or prognostic?' In: *Journal of clinical oncology: official journal of the American Society of Clinical Oncology* 33.33, pp. 3968–3971 (cit. on pp. 2, 9, 13, 17, 34).

Ballman, Karla V, Jan C Buckner, Paul D Brown, Caterina Giannini, Patrick J Flynn, Betsy R LaPlant, and Kurt A Jaeckle (2007). 'The relationship between six-month progression-free survival and 12-month overall survival end points for phase II trials in patients with glioblastoma multiforme'. In: *Neuro-oncology* 9.1, pp. 29–38 (cit. on pp. 10, 44).

Bender, Ralf, Thomas Augustin, and Maria Blettner (2005). 'Generating survival times to simulate Cox proportional hazards models'. en. In: *Statistics in Medicine* 24.11. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/sim.2059, pp. 1713–1723. ISSN: 1097-0258. DOI: 10.1002/sim.2059. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.2059 (visited on 01/29/2022) (cit. on p. 44).

Bo, Na, Jong-Hyeon Jeong, Erick Forno, and Ying Ding (2025). 'Evaluating Meta-Learners to Analyze Treatment Heterogeneity in Survival Data: Application to Electronic Health Records of Pediatric Asthma Care in COVID-19 Pandemic'. In: *Statistics in medicine* 44.3-4, e10333 (cit. on p. 22).

Boileau, Philippe, Nina Ting Qi, Mark J van der Laan, Sandrine Dudoit, and Ning Leng (2023). 'A flexible approach for predictive biomarker discovery'. In: *Biostatistics* 24.4, pp. 1085–1105 (cit. on pp. 3, 22, 25).

Bortolotto, Chandra, Andrea Lancia, Chiara Stelitano, Marianna Montesano, Elisa Merizzoli, Francesco Agustoni, Giulia Stella, Lorenzo Preda, and Andrea Riccardo Filippi (2021). 'Radiomics features as predictive and prognostic biomarkers in NSCLC'. In: *Expert Review of Anticancer Therapy* 21.3, pp. 257–266 (cit. on p. 39).

Braghetto, Anna, Francesca Marturano, Marta Paiusco, Marco Baiesi, and Andrea Bettinelli (2022). 'Radiomics and deep learning methods for the prediction of 2-year overall survival in LUNG1 dataset'. In: *Scientific Reports* 12.1, p. 14132 (cit. on p. 42).

Brouillard, Philippe, Chandler Squires, Jonas Wahl, Konrad P Kording, Karen Sachs, Alexandre Drouin, and Dhanya Sridhar (2024). 'The landscape of causal discovery data: Grounding causal discovery in real-world applications'. In: *arXiv preprint arXiv:2412.01953* (cit. on p. 123).

Cadei, Riccardo, Lukas Lindorfer, Sylvia Cremer, Cordelia Schmid, and Francesco Locatello (2024). 'Smoke and mirrors in causal downstream tasks'. In: *Advances in Neural Information Processing Systems* 37, pp. 26082–26112 (cit. on pp. 22, 24, 125).

Califf, Robert M (2018). 'Biomarker definitions and their applications'. In: *Experimental biology and medicine* 243.3, pp. 213–221 (cit. on p. 9).

Caramanna, Ivan, Julie M de Kort, Alba A Brandes, Walter Taal, Michael Platten, Ahmed Idbaih, Jean Sebastien Frenel, Wolfgang Wick, Chandrakanth Jayachandran Preetha, Martin Bendszus, Vollmuth Philipp, Jaap C Reijneveld, Martin Klein, et al. (2022). 'Corticosteroids use and neurocognitive functioning in patients with recurrent glioblastoma: Evidence from European Organization for Research and Treatment of Cancer (EORTC) trial 26101'. In: *Neuro-Oncology Practice* 9.4, pp. 310–316 (cit. on p. 10).

Cardoso, M Jorge, Wenqi Li, Richard Brown, Nic Ma, Eric Kerfoot, Yiheng Wang, Benjamin Murrey, Andriy Myronenko, Can Zhao, Dong Yang, et al. (2022). 'Monai: An open-source framework for deep learning in healthcare'. In: *arXiv preprint arXiv:2211.02701* (cit. on p. 40).

Chaddad, Ahmad, Michael Jonathan Kucharczyk, Paul Daniel, Siham Sabri, Bertrand J Jean-Claude, Tamim Niazi, and Bassam Abdulkarim (2019). 'Radiomics in glioblastoma: current status and challenges facing clinical implementation'. In: *Frontiers in oncology* 9, p. 374 (cit. on p. 10).

Chapfuwa, Paidamoyo, Serge Assaad, Shuxi Zeng, Michael J Pencina, Lawrence Carin, and Ricardo Henao (2021). 'Enabling counterfactual survival analysis with balanced representations'. In: *Proceedings of the Conference on Health, Inference, and Learning*, pp. 133–145 (cit. on pp. 23, 27).

Chen, Tianqi and Carlos Guestrin (2016). 'Xgboost: A scalable tree boosting system'. In: *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pp. 785–794 (cit. on p. 27).

Chen, Ting, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton (2020). 'A simple framework for contrastive learning of visual representations'. In: *International conference on machine learning*. PmLR, pp. 1597–1607 (cit. on p. 126).

Chinot, Olivier L, Wolfgang Wick, Warren Mason, Roger Henriksson, Frank Saran, Ryo Nishikawa, Antoine F Carpentier, Khe Hoang-Xuan, Petr Kavan, Dana Cernea, et al. (2014). 'Bevacizumab plus radiotherapy–temozolomide for newly diagnosed glioblastoma'. In: *New England Journal of Medicine* 370.8, pp. 709–722 (cit. on p. 4).

Chiu, Fang-Ying and Yun Yen (2023). 'Imaging biomarkers for clinical applications in neuro-oncology: current status and future perspectives'. In: *Biomarker Research* 11.1, p. 35 (cit. on p. 4).

Chu, Jiebin, Zhoujian Sun, Wei Dong, Jinlong Shi, and Zhengxing Huang (2021). 'On learning disentangled representations for individual treatment effect estimation'. In: *Journal of Biomedical Informatics* 124, p. 103940 (cit. on p. 127).

Codella, Noel, Veronica Rotemberg, Philipp Tschandl, M. Emre Celebi, Stephen Dusza, David Gutman, Brian Helba, Aadi Kalloo, Konstantinos Liopyris, Michael Marchetti, Harald Kittler, and Allan Halpern (2019). 'Skin Lesion Analysis Toward Melanoma Detection 2018: A Challenge Hosted by the International Skin Imaging Collaboration (ISIC)'. In: *arXiv preprint arXiv:1902.03368*. URL: http://arxiv.org/abs/1902.03368 (cit. on p. 38).

Cox, David R (1972). 'Regression models and life-tables'. In: *Journal of the Royal Statistical Society: Series B (Methodological)* 34.2, pp. 187–202 (cit. on pp. 16, 26, 27).

Crabbé, Jonathan, Alicia Curth, Ioana Bica, and Mihaela van der Schaar (2022). 'Benchmarking Heterogeneous Treatment Effect Models through the Lens of Interpretability'. In: *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*. URL: https://openreview.net/forum?id=ddPXQt-gM-- (cit. on pp. 3, 22, 25, 35, 65, 123).

Cui, Can, Haichun Yang, Yaohong Wang, Shilin Zhao, Zuhayr Asad, Lori A Coburn, Keith T Wilson, Bennett A Landman, and Yuankai Huo (2023). 'Deep multimodal fusion of image and non-image data in disease diagnosis and prognosis: a review'. In: *Progress in Biomedical Engineering* 5.2, p. 022001 (cit. on p. 114).

Curth, Alicia, Changhee Lee, and Mihaela van der Schaar (2021). 'Survite: Learning heterogeneous treatment effects from time-to-event data'. In: *Advances in Neural Information Processing Systems* 34, pp. 26740–26753 (cit. on pp. 23, 27, 49).

Curth, Alicia, Richard W Peck, Eoin McKinney, James Weatherall, and Mihaela van Der Schaar (2024). 'Using machine learning to individualize treatment effect estimation: challenges and opportunities'. In: *Clinical Pharmacology & Therapeutics* 115.4, pp. 710–719 (cit. on pp. 3, 124).

Curth, Alicia and Mihaela van der Schaar (2021). 'On inductive biases for heterogeneous treatment effect estimation'. In: *Advances in Neural Information Processing Systems* 34, pp. 15883–15894 (cit. on pp. 33, 117).

Curth, Alicia, David Svensson, Jim Weatherall, and Mihaela van der Schaar (2021). 'Really Doing Great at Estimating CATE? A Critical Look at ML Benchmarking Practices

in Treatment Effect Estimation'. In: *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*. URL: https://openreview.net/forum?id=FQLzQqGEAH (cit. on pp. 25, 32, 43, 123).

Dancette, Corentin, Julien Khlaut, Antoine Saporta, Helene Philippe, Elodie Ferreres, Baptiste Callard, Théo Danielou, Léo Alberge, Léo Machado, Daniel Tordjman, Julie Dupuis, Korentin Le Floch, Jean Du Terrail, Mariam Moshiri, Laurent Dercle, Tom Boeken, Jules Gregory, Maxime Ronot, François Legou, Pascal Roux, Marc Sapoval, Pierre Manceron, and Paul Hérent (2025). *Curia: A Multi-Modal Foundation Model for Radiology*. arXiv: 2509.06830 [cs.CV]. URL: https://arxiv.org/abs/2509.06830 (cit. on pp. 28, 126).

Davidson-Pilon, Cameron (2019). 'lifelines: survival analysis in Python'. In: *Journal of Open Source Software* 4.40, p. 1317. DOI: 10.21105/joss.01317. URL: https://doi.org/10.21105/joss.01317 (cit. on pp. 52, 56).

Delgado, Amanda and Achuta Kumar Guddati (2021). 'Clinical endpoints in oncology-a primer'. In: *American journal of cancer research* 11.4, p. 1121 (cit. on pp. 6, 26).

Deng, Li (2012). 'The MNIST Database of Handwritten Digit Images for Machine Learning Research [Best of the Web]'. In: *IEEE Signal Processing Magazine* 29.6, pp. 141–142. DOI: 10.1109/MSP.2012.2211477 (cit. on p. 38).

Deshpande, Shachi, Kaiwen Wang, Dhruv Sreenivas, Zheng Li, and Volodymyr Kuleshov (2022). 'Deep multi-modal structural equations for causal effect estimation with unstructured proxies'. In: *Advances in Neural Information Processing Systems* 35, pp. 10931–10944 (cit. on p. 22).

Durso-Finley, Joshua, Jean-Pierre Falet, Raghav Mehta, Douglas L Arnold, Nick Pawlowski, and Tal Arbel (2023). 'Improving image-based precision medicine with uncertainty-aware causal models'. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, pp. 472–481 (cit. on pp. 3, 22, 24, 127).

Durso-Finley, Joshua, Jean-Pierre Falet, Brennan Nichyporuk, Arnold Douglas, and Tal Arbel (2022). 'Personalized prediction of future lesion activity and treatment effect in multiple sclerosis from baseline MRI'. In: *International Conference on Medical Imaging with Deep Learning*. PMLR, pp. 387–406 (cit. on pp. 3, 22, 24, 33, 49, 56, 101, 120, 127).

Efthimiou, Orestis, Jeroen Hoogland, Thomas PA Debray, Valerie Aponte Ribero, Wilma Knol, Huiberdina L Koek, Matthias Schwenkglenks, Séverine Henrard, Matthias Egger, Nicolas Rodondi, et al. (2025). 'Measuring the Performance of Survival Models to Personalize Treatment Choices'. In: *Statistics in Medicine* 44.7, e70050 (cit. on p. 27).

Efthimiou, Orestis, Jeroen Hoogland, Thomas PA Debray, Michael Seo, Toshiaki A Furukawa, Matthias Egger, and Ian R White (2023). 'Measuring the performance of prediction models to personalize treatment choice'. In: *Statistics in medicine* 42.8, pp. 1188–1206 (cit. on pp. 15, 55).

Eisenhauer, Elizabeth A, Patrick Therasse, Jan Bogaerts, Lawrence H Schwartz, Danielle Sargent, Robert Ford, Janet Dancey, Stephen Arbuck, Steve Gwyther, Margaret Mooney, et al. (2009). 'New response evaluation criteria in solid tumours: revised

RECIST guideline (version 1.1)'. In: *European journal of cancer* 45.2, pp. 228–247 (cit. on p. 1).

Elazab, Ahmed, Changmiao Wang, Syed Jamal Safdar Gardezi, Hongmin Bai, Qingmao Hu, Tianfu Wang, Chunqi Chang, and Baiying Lei (2020). 'GP-GAN: Brain tumor growth prediction using stacked 3D generative adversarial networks from longitudinal MR Images'. In: *Neural Networks* 132, pp. 321–332 (cit. on p. 10).

Erion, Gabriel, Joseph D Janizek, Pascal Sturmfels, Scott M Lundberg, and Su-In Lee (2021). 'Improving performance of deep learning models with axiomatic attribution priors and expected gradients'. In: *Nature machine intelligence* 3.7, pp. 620–631 (cit. on pp. 36, 40, 65).

Falcon, William and The PyTorch Lightning team (Mar. 2019). *PyTorch Lightning*. Version 1.4. DOI: 10.5281/zenodo.3828935. URL: https://github.com/Lightning-AI/lightning (cit. on p. 50).

Faraggi, David and Richard Simon (1995). 'A neural network model for survival data'. In: *Statistics in medicine* 14.1, pp. 73–82 (cit. on pp. 16, 48).

Ferrara, Napoleone, Kenneth J Hillan, and William Novotny (2005). 'Bevacizumab (Avastin), a humanized anti-VEGF monoclonal antibody for cancer therapy'. In: *Biochemical and biophysical research communications* 333.2, pp. 328–335 (cit. on p. 11).

Foster, Jared C, Jeremy MG Taylor, and Stephen J Ruberg (2011). 'Subgroup identification from randomized clinical trial data'. In: *Statistics in medicine* 30.24, pp. 2867–2880 (cit. on p. 26).

Frauen, Dennis, Valentyn Melnychuk, Jonas Schweisthal, and Stefan Feuerriegel (2025). 'Treatment effect estimation for optimal decision-making'. In: *arXiv preprint arXiv:2505.13092* (cit. on p. 23).

Gareau, Daniel S, James Browning, Joel Correa Da Rosa, Mayte Suarez-Farinas, Samantha Lish, Amanda M Zong, Benjamin Firester, Charles Vrattos, Yael Renert-Yuval, Mauricio Gamboa, et al. (2020). 'Deep learning-level melanoma detection by interpretable machine learning and imaging biomarker cues'. In: *Journal of Biomedical Optics* 25.11, pp. 112906–112906 (cit. on p. 39).

Gareau, Daniel S, Joel Correa da Rosa, Sarah Yagerman, John A Carucci, Nicholas Gulati, Ferran Hueto, Jennifer L DeFazio, Mayte Suárez-Fariñas, Ashfaq Marghoob, and James G Krueger (2017). 'Digital imaging biomarkers feed machine learning for melanoma screening'. In: *Experimental dermatology* 26.7, pp. 615–618 (cit. on p. 39).

Gerds, Thomas A and Martin Schumacher (2006). 'Consistent estimation of the expected Brier score in general survival models with right-censored event times'. In: *Biometrical Journal* 48.6, pp. 1029–1040 (cit. on p. 19).

Gilbert, Mark R, James J Dignam, Terri S Armstrong, Jeffrey S Wefel, Deborah T Blumenthal, Michael A Vogelbaum, Howard Colman, Arnab Chakravarti, Stephanie Pugh, Minhee Won, et al. (2014). 'A randomized trial of bevacizumab for newly diagnosed glioblastoma'. In: *New England Journal of Medicine* 370.8, pp. 699–708 (cit. on pp. 4, 11).

Goyal, Yash, Ziyan Wu, Jan Ernst, Dhruv Batra, Devi Parikh, and Stefan Lee (2019). 'Counterfactual visual explanations'. In: *International Conference on Machine Learning*. PMLR, pp. 2376–2384 (cit. on p. 127).

Grochans, Szymon, Anna Maria Cybulska, Donata Simińska, Jan Korbecki, Klaudyna Kojder, Dariusz Chlubek, and Irena Baranowska-Bosiacka (2022). 'Epidemiology of glioblastoma multiforme–literature review'. In: *Cancers* 14.10, p. 2412 (cit. on pp. 4, 10).

Grossmann, Patrick, Vivek Narayan, Ken Chang, Rifaquat Rahman, Lauren Abrey, David A Reardon, Lawrence H Schwartz, Patrick Y Wen, Brian M Alexander, Raymond Huang, et al. (2017). 'Quantitative imaging biomarkers for risk stratification of patients with recurrent glioblastoma treated with bevacizumab'. In: *Neuro-oncology* 19.12, pp. 1688–1697 (cit. on pp. 4, 121).

Gupta, Tarun, Wenbo Gong, Chao Ma, Nick Pawlowski, Agrin Hilmkil, Meyer Scetbon, Marc Rigter, Ade Famoti, Ashley Juan Llorens, Jianfeng Gao, et al. (2024). 'The essential role of causality in foundation world models for embodied AI'. In: *arXiv preprint arXiv:2402.06665* (cit. on p. 1).

Haarburger, Christoph, Philippe Weitz, Oliver Rippel, and Dorit Merhof (2019). 'Image-based survival prediction for lung cancer patients using CNNS'. In: *2019 IEEE 16th international symposium on biomedical imaging (ISBI 2019)*. IEEE, pp. 1197–1201 (cit. on pp. 23, 28, 112).

Han, Larry, Jue Hou, Kelly Cho, Rui Duan, and Tianxi Cai (2025). 'Federated adaptive causal estimation (face) of target treatment effects'. In: *Journal of the American Statistical Association*, pp. 1–14 (cit. on p. 125).

Han, Zeyu, Chao Gao, Jinyang Liu, Jeff Zhang, and Sai Qian Zhang (2024). 'Parameter-efficient fine-tuning for large models: A comprehensive survey'. In: *arXiv preprint arXiv:2403.14608* (cit. on p. 126).

Hao, Degan, Qiong Li, Qiu-Xia Feng, Liang Qi, Xi-Sheng Liu, Dooman Arefan, Yu-Dong Zhang, and Shandong Wu (2022). 'SurvivalCNN: A deep learning-based method for gastric cancer survival prediction using radiological imaging data and clinicopathological variables'. In: *Artificial intelligence in medicine* 134, p. 102424 (cit. on pp. 23, 28).

Harrell, Frank E, Robert M Califf, David B Pryor, Kerry L Lee, and Robert A Rosati (1982). 'Evaluating the yield of medical tests'. In: *Jama* 247.18, pp. 2543–2546 (cit. on pp. 18, 55).

Harrell Jr, Frank E, Kerry L Lee, and Daniel B Mark (1996). 'Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors'. In: *Statistics in medicine* 15.4, pp. 361–387 (cit. on pp. 18, 55).

He, Kaiming, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick (2022). 'Masked autoencoders are scalable vision learners'. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16000–16009 (cit. on p. 54).

He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun (2016). 'Deep residual learning for image recognition'. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778 (cit. on pp. 33, 39).

Helland, Ragnhild Holden, Alexandros Ferles, André Pedersen, Ivar Kommers, Hilko Ardon, Frederik Barkhof, Lorenzo Bello, Mitchel S Berger, Tora Dunås, Marco Conti Nibali, et al. (2023). 'Segmentation of glioblastomas in early post-operative multimodal MRI with deep neural networks'. In: *Scientific reports* 13.1, p. 18897 (cit. on p. 10).

Hermansson, Erik and David Svensson (2021). 'On Discovering Treatment-Effect Modifiers Using Virtual Twins and Causal Forest ML in the Presence of Prognostic Biomarkers'. In: *International Conference on Computational Science and Its Applications*. Springer, pp. 624–640 (cit. on pp. 22, 25, 26, 32).

Herzog, Lisa, Pascal Bühler, Ezequiel de la Rosa, Beate Sick, and Susanne Wegener (2025). 'Outcome prediction and individualized treatment effect estimation in patients with large vessel occlusion stroke'. In: *Stroke Workshop on Imaging and Treatment Challenges*. Springer, pp. 73–82 (cit. on pp. 22, 24, 29).

Hill, Jennifer L (2011). 'Bayesian nonparametric modeling for causal inference'. In: *Journal of Computational and Graphical Statistics* 20.1, pp. 217–240 (cit. on p. 14).

Hitsch, Günter J, Sanjog Misra, and Walter W Zhang (2024). 'Heterogeneous treatment effects and optimal targeting policy evaluation'. In: *Quantitative Marketing and Economics* 22.2, pp. 115–168 (cit. on p. 15).

Ho, Seng-Beng, Mark Edmonds, and Song-Chun Zhu (2020). 'Actional-perceptual causality: Concepts and inductive learning for AI and robotics'. In: *2020 IEEE Symposium Series on Computational Intelligence (SSCI)*. IEEE, pp. 442–448 (cit. on p. 1).

Holland, Paul W (1986). 'Statistics and causal inference'. In: *Journal of the American statistical Association* 81.396, pp. 945–960 (cit. on pp. 3, 12).

Hosny, Ahmed, Hugo J Aerts, and Raymond H Mak (2019). 'Handcrafted versus deep learning radiomics for prediction of cancer therapy response'. In: *The Lancet Digital Health* 1.3, e106–e107 (cit. on p. 4).

Huo, Zepeng Frazier, Jason Alan Fries, Alejandro Lozano, Jeya Maria Jose Valanarasu, Ethan Steinberg, Louis Blankemeier, Akshay S Chaudhari, Curtis Langlotz, and Nigam Shah (2025). 'Time-to-Event Pretraining for 3D Medical Imaging'. In: *The Thirteenth International Conference on Learning Representations*. URL: https://openreview.net/forum?id=zcTLpIfj9u (cit. on pp. 23, 28, 126).

Imbens, Guido W and Jeffrey M Wooldridge (2009). 'Recent developments in the econometrics of program evaluation'. In: *Journal of economic literature* 47.1, pp. 5–86 (cit. on p. 12).

Ioffe, Sergey and Christian Szegedy (2015). 'Batch normalization: Accelerating deep network training by reducing internal covariate shift'. In: *International conference on machine learning*. pmlr, pp. 448–456 (cit. on p. 50).

Irwin, JO (1949). 'The standard error of an estimate of expectation of life, with special reference to expectation of tumourless life in experiments with mice'. In: *Epidemiology & Infection* 47.2, pp. 188–189 (cit. on p. 16).

Isensee, Fabian, Paul F Jaeger, Simon AA Kohl, Jens Petersen, and Klaus H Maier-Hein (2021). 'nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation'. In: *Nature methods* 18.2, pp. 203–211 (cit. on pp. 40, 46, 53).

Isensee, Fabian, Tassilo Wald, Constantin Ulrich, Michael Baumgartner, Saikat Roy, Klaus Maier-Hein, and Paul F Jaeger (2024). 'nnu-net revisited: A call for rigorous validation in 3d medical image segmentation'. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, pp. 488–498 (cit. on p. 53).

Ishwaran, Hemant, Udaya B Kogalur, Eugene H Blackstone, and Michael S Lauer (2008). 'Random survival forests'. In: (cit. on p. 27).

Jerzak, Connor Thomas, Fredrik Daniel Johansson, and Adel Daoud (2023). 'Image-based Treatment Effect Heterogeneity'. In: *Proceedings of the Second Conference on Causal Learning and Reasoning*. Ed. by Mihaela van der Schaar, Cheng Zhang, and Dominik Janzing. Vol. 213. Proceedings of Machine Learning Research. PMLR, pp. 531–552. URL: https://proceedings.mlr.press/v213/jerzak23a.html (cit. on pp. 1, 22, 24).

Jesson, Andrew, Sören Mindermann, Yarin Gal, and Uri Shalit (2021). 'Quantifying ignorance in individual-level causal-effect estimates under hidden confounding'. In: *International Conference on Machine Learning*. PMLR, pp. 4829–4838 (cit. on p. 14).

Jiang, Ziyang, Zhuoran Hou, Yiling Liu, Yiman Ren, Keyu Li, and David Carlson (2023). 'Estimating Causal Effects using a Multi-task Deep Ensemble'. In: *arXiv preprint arXiv:2301.11351* (cit. on pp. 22, 24).

Jin, Cheng, Heng Yu, Jia Ke, Peirong Ding, Yongju Yi, Xiaofeng Jiang, Xin Duan, Jinghua Tang, Daniel T Chang, Xiaojian Wu, et al. (2021). 'Predicting treatment response from longitudinal images using multi-task deep learning'. In: *Nature communications* 12.1, pp. 1–11 (cit. on p. 24).

Ju, Cheng, Aurélien Bibaut, and Mark van der Laan (2018). 'The relative performance of ensemble methods with deep convolutional neural networks for image classification'. In: *Journal of applied statistics* 45.15, pp. 2800–2818 (cit. on p. 56).

Kallus, Nathan and Angela Zhou (2018). 'Confounding-robust policy improvement'. In: *Advances in neural information processing systems* 31 (cit. on p. 14).

Kaplan, Edward L and Paul Meier (1958). 'Nonparametric estimation from incomplete observations'. In: *Journal of the American statistical association* 53.282, pp. 457–481 (cit. on p. 17).

Katzman, Jared L, Uri Shaham, Alexander Cloninger, Jonathan Bates, Tingting Jiang, and Yuval Kluger (2018). 'DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network'. In: *BMC medical research methodology* 18.1, p. 24 (cit. on pp. 16, 23, 26, 27, 48, 56, 82).

Kent, David M, Jessica K Paulus, David Van Klaveren, Ralph D'Agostino, Steve Goodman, Rodney Hayward, John PA Ioannidis, Bray Patrick-Lake, Sally Morton, Michael

Pencina, et al. (2020). 'The predictive approaches to treatment effect heterogeneity (PATH) statement'. In: *Annals of internal medicine* 172.1, pp. 35–45 (cit. on p. 24).

Kessler, Tobias, Daniel Schrimpf, Laura Doerner, Ling Hai, Leon D Kaulen, Jakob Ito, Martin van den Bent, Martin Taphoorn, Alba A Brandes, Ahmed Idbaih, et al. (2023). 'Prognostic markers of DNA methylation and next-generation sequencing in progressive glioblastoma from the EORTC-26101 trial'. In: *Clinical Cancer Research* 29.19, pp. 3892–3900 (cit. on p. 11).

Kickingereder, Philipp, Gianluca Brugnara, Mikkel Bo Hansen, Martha Nowosielski, Irada Pflüger, Marianne Schell, Fabian Isensee, Martha Foltyn, Ulf Neuberger, Tobias Kessler, et al. (2020). 'Noninvasive characterization of tumor angiogenesis and oxygenation in bevacizumab-treated recurrent glioblastoma by using dynamic susceptibility MRI: secondary analysis of the European Organization for Research and Treatment of Cancer 26101 Trial'. In: *Radiology* 297.1, pp. 164–175 (cit. on pp. 11, 45).

Kickingereder, Philipp, Michael Götz, John Muschelli, Antje Wick, Ulf Neuberger, Russell T Shinohara, Martin Sill, Martha Nowosielski, Heinz-Peter Schlemmer, Alexander Radbruch, et al. (2016). 'Large-scale Radiomic Profiling of Recurrent Glioblastoma Identifies an Imaging Predictor for Stratifying Anti-Angiogenic Treatment Response Radiomic Profiling of BEV Efficacy in Glioblastoma'. In: *Clinical Cancer Research* 22.23, pp. 5765–5771 (cit. on pp. 4, 10, 121).

Kickingereder, Philipp, Fabian Isensee, Irada Tursunova, Jens Petersen, Ulf Neuberger, David Bonekamp, Gianluca Brugnara, Marianne Schell, Tobias Kessler, Martha Foltyn, Inga Harting, Felix Sahm, Marcel Prager, Martha Nowosielski, Antje Wick, Marco Nolden, Alexander Radbruch, Jürgen Debus, Heinz-Peter Schlemmer, Sabine Heiland, Michael Platten, Andreas von Deimling, Martin J van den Bent, Thierry Gorlia, Wolfgang Wick, Martin Bendszus, and Klaus H Maier-Hein (2019). 'Automated quantitative tumour response assessment of MRI in neuro-oncology with artificial neural networks: a multicentre, retrospective study'. In: *The Lancet Oncology* 20.5, pp. 728–740. DOI: https://doi.org/10.1016/S1470-2045(19)30098-1. URL: https://www.sciencedirect.com/science/article/pii/S1470204519300981 (cit. on pp. 10, 45).

Kickingereder, Philipp, Benedikt Wiestler, Sina Burth, Antje Wick, Martha Nowosielski, Sabine Heiland, Heinz-Peter Schlemmer, Wolfgang Wick, Martin Bendszus, and Alexander Radbruch (2015). 'Relative cerebral blood volume is a potential predictive imaging biomarker of bevacizumab efficacy in recurrent glioblastoma'. In: *Neuro-oncology* 17.8, pp. 1139–1147 (cit. on pp. 4, 121).

Kim, Jong-Min (2025). 'Integrating Copula-Based Random Forest and Deep Learning Approaches for Analyzing Heterogeneous Treatment Effects in Survival Analysis'. In: *Mathematics* 13.10, p. 1659 (cit. on p. 23).

Kim, Steven B, Dong Sub Kim, and Xiaoming Mo (2021). 'An image segmentation technique with statistical strategies for pesticide efficacy assessment'. In: *Plos one* 16.3, e0248592 (cit. on p. 1).

Kokhlikyan, Narine, Vivek Miglani, Miguel Martin, Edward Wang, Bilal Alsallakh, Jonathan Reynolds, Alexander Melnikov, Natalia Kliushkina, Carlos Araya, Siqi Yan, and Orion Reblitz-Richardson (2020). *Captum: A unified and generic model interpretability library for PyTorch.* arXiv: 2009.07896 [cs.LG] (cit. on p. 41).

Krzykalla, Julia, Axel Benner, and Annette Kopp-Schneider (2020). 'Exploratory identification of predictive biomarkers in randomized trials with normal endpoints'. In: *Statistics in Medicine* 39.7, pp. 923–939 (cit. on p. 36).

Künzel, Sören R, Jasjeet S Sekhon, Peter J Bickel, and Bin Yu (2019). 'Metalearners for estimating heterogeneous treatment effects using machine learning'. In: *Proceedings of the national academy of sciences* 116.10, pp. 4156–4165 (cit. on pp. 14, 24, 32, 52).

Kvamme, Håvard and Ørnulf Borgan (2021). 'Continuous and discrete-time survival prediction with neural networks'. In: *Lifetime data analysis* 27.4, pp. 710–736 (cit. on p. 55).

Kvamme, Håvard, Ørnulf Borgan, and Ida Scheel (2019). 'Time-to-event prediction with neural networks and Cox regression'. In: *Journal of machine learning research* 20.129, pp. 1–30 (cit. on pp. 50, 55).

Lambin, Philippe, Ralph TH Leijenaar, Timo M Deist, Jurgen Peerlings, Evelyn EC De Jong, Janita Van Timmeren, Sebastian Sanduleanu, Ruben THM Larue, Aniek JG Even, Arthur Jochems, et al. (2017). 'Radiomics: the bridge between medical imaging and personalized medicine'. In: *Nature reviews Clinical oncology* 14.12, pp. 749–762 (cit. on pp. 4, 10).

Lee, Changhee, William Zame, Jinsung Yoon, and Mihaela Van Der Schaar (2018). 'Deephit: A deep learning approach to survival analysis with competing risks'. In: *Proceedings of the AAAI conference on artificial intelligence.* Vol. 32. 1 (cit. on p. 26).

Li, Yangming (2023). 'Deep causal learning for robotic intelligence'. In: *Frontiers in Neurorobotics* 17, p. 1128591 (cit. on p. 1).

Li, Zhi-Cheng, Jing Yan, Shenghai Zhang, Chaofeng Liang, Xiaofei Lv, Yan Zou, Huailing Zhang, Dong Liang, Zhenyu Zhang, and Yinsheng Chen (2022). 'Glioma survival prediction from whole-brain MRI without tumor segmentation using deep attention network: a multicenter study'. In: *European radiology* 32.8, pp. 5719–5729 (cit. on pp. 10, 23, 28).

Ligero, Marta, Omar SM El Nahhas, Mihaela Aldea, and Jakob Nikolas Kather (2025). 'Artificial intelligence-based biomarkers for treatment decisions in oncology'. In: *Trends in Cancer* 11.3, pp. 232–244 (cit. on p. 127).

Lillelund, Christian Marius, Shi-ang Qi, Russell Greiner, and Christian Fischer Pedersen (2025). 'Stop Chasing the C-index: This Is How We Should Evaluate Our Survival Models'. In: *arXiv preprint arXiv:2506.02075* (cit. on p. 121).

Limkin, EJ, Roger Sun, Laurent Dercle, EI Zacharaki, Charlotte Robert, Sylvain Reuzé, Antoine Schernberg, Nikos Paragios, Eric Deutsch, and Charles Ferté (2017). 'Promises and challenges for the implementation of computational medical imaging (radiomics) in oncology'. In: *Annals of Oncology* 28.6, pp. 1191–1206 (cit. on p. 10).

Liu, Mingzhu, Chirag Nagpal, and Artur Dubrawski (2024). 'Deep Survival Models Can Improve Long-Term Mortality Risk Estimates from Chest Radiographs'. In: *Forecasting* 6.2, pp. 404–417. ISSN: 2571-9394. DOI: 10.3390/forecast6020022. URL: https://www.mdpi.com/2571-9394/6/2/22 (cit. on p. 28).

Liu, Ruoqi, Pin-Yu Chen, and Ping Zhang (2024). 'CURE: A deep learning framework pre-trained on large-scale patient data for treatment effect estimation'. In: *Patterns* 5.6 (cit. on pp. 29, 126).

Liu, Zihuan, Yihua Gu, and Xin Huang (2025). 'Deep learning-based ranking method for subgroup and predictive biomarker identification in patients'. In: *Communications Medicine* 5.1, pp. 1–15 (cit. on pp. 22, 26).

Lohr, Kathleen N (1988). 'Outcome measurement: concepts and questions'. In: *Inquiry*, pp. 37–50 (cit. on p. 2).

Loshchilov, Ilya and Frank Hutter (2017). 'Decoupled weight decay regularization'. In: *arXiv preprint arXiv:1711.05101* (cit. on p. 54).

Ma, Wenao, Cheng Chen, Jill Abrigo, Calvin Hoi-Kwan Mak, Yuqi Gong, Nga Yan Chan, Chu Han, Zaiyi Liu, and Qi Dou (2023). 'Treatment Outcome Prediction for Intracerebral Hemorrhage via Generative Prognostic Model with Imaging and Tabular Data'. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, pp. 715–725 (cit. on pp. 3, 22, 24, 120).

Ma, Wenao, Cheng Chen, Yuqi Gong, Nga Yan Chan, Meirui Jiang, Calvin Hoi-Kwan Mak, Jill M Abrigo, and Qi Dou (2024). 'Causal Effect Estimation on Imaging and Clinical Data for Treatment Decision Support of Aneurysmal Subarachnoid Hemorrhage'. In: *IEEE Transactions on Medical Imaging* 43.8, pp. 2778–2789 (cit. on pp. 22, 24, 126).

Machlanski, Damian, Spyridon Samothrakis, and Paul Clarke (2023). 'Hyperparameter tuning and model evaluation in causal effect estimation'. In: *arXiv preprint arXiv:2303.01412* (cit. on p. 51).

Makhija, Disha, Joydeep Ghosh, and Yejin Kim (2024). 'Federated Learning for Estimating Heterogeneous Treatment Effects'. In: *arXiv preprint arXiv:2402.17705* (cit. on p. 125).

Mandrekar, Sumithra J and Daniel J Sargent (2009). 'Clinical trial designs for predictive biomarker validation: theoretical considerations and practical challenges'. In: *Journal of Clinical Oncology* 27.24, pp. 4027–4034 (cit. on p. 3).

Mantel, Nathan et al. (1966). 'Evaluation of survival data and two new rank order statistics arising in its consideration'. In: *Cancer Chemother Rep* 50.3, pp. 163–170 (cit. on p. 18).

Martínez, Javier Abad (2021). *Interpretability for conditional average treatment effect estimation* (cit. on p. 127).

Meng, Mingyuan, Bingxin Gu, Lei Bi, Shaoli Song, David Dagan Feng, and Jinman Kim (2022). 'DeepMTS: Deep multi-task learning for survival prediction in patients with advanced nasopharyngeal carcinoma using pretreatment PET/CT'. In: *IEEE Journal of Biomedical and Health Informatics* 26.9, pp. 4497–4507 (cit. on pp. 23, 28).

Müller, Alfred (1997). 'Integral probability metrics and their generating classes of functions'. In: *Advances in applied probability* 29.2, pp. 429–443 (cit. on p. 48).

O'Connor, James P. B., Eric O. Aboagye, Judith E. Adams, Hugo J. W. L. Aerts, Sally F. Barrington, Ambros J. Beer, Ronald Boellaard, Sarah E. Bohndiek, Michael Brady, Gina Brown, David L. Buckley, Thomas L. Chenevert, Laurence P. Clarke, Sandra Collette, Gary J. Cook, Nandita M. deSouza, John C. Dickson, Caroline Dive, Jeffrey L. Evelhoch, Corinne Faivre-Finn, Ferdia A. Gallagher, Fiona J. Gilbert, Robert J. Gillies, Vicky Goh, John R. Griffiths, Ashley M. Groves, Steve Halligan, Adrian L. Harris, David J. Hawkes, Otto S. Hoekstra, Erich P. Huang, Brian F. Hutton, Edward F. Jackson, Gordon C. Jayson, Andrew Jones, Dow-Mu Koh, Denis Lacombe, Philippe Lambin, Nathalie Lassau, Martin O. Leach, Ting-Yim Lee, Edward L. Leen, Jason S. Lewis, Yan Liu, Mark F. Lythgoe, Prakash Manoharan, Ross J. Maxwell, Kenneth A. Miles, Bruno Morgan, Steve Morris, Tony Ng, Anwar R. Padhani, Geoff J. M. Parker, Mike Partridge, Arvind P. Pathak, Andrew C. Peet, Shonit Punwani, Andrew R. Reynolds, Simon P. Robinson, Lalitha K. Shankar, Ricky A. Sharma, Dmitry Soloviev, Sigrid Stroobants, Daniel C. Sullivan, Stuart A. Taylor, Paul S. Tofts, Gillian M. Tozer, Marcel van Herk, Simon Walker-Samuel, James Wason, Kaye J. Williams, Paul Workman, Thomas E. Yankeelov, Kevin M. Brindle, Lisa M. McShane, Alan Jackson, and John C. Waterton (Mar. 2017). 'Imaging biomarker roadmap for cancer studies'. en. In: *Nature Reviews Clinical Oncology* 14.3, pp. 169–186. ISSN: 1759-4774, 1759-4782. DOI: 10.1038/nrclinonc.2016.162. URL: http://www.nature.com/articles/nrclinonc.2016.162 (visited on 01/11/2022) (cit. on p. 2).

Ogier du Terrail, Jean, Quentin Klopfenstein, Honghao Li, Imke Mayer, Nicolas Loiseau, Mohammad Hallal, Michael Debouver, Thibault Camalon, Thibault Fouqueray, Jorge Arellano Castro, et al. (2025). 'FedECA: federated external control arms for causal inference with time-to-event data in distributed settings'. In: *Nature Communications* 16.1, p. 7496 (cit. on p. 125).

Park, Ji Eun and Ho Sung Kim (2018). 'Radiomics as a quantitative imaging biomarker: practical considerations and the current standpoint in neuro-oncologic studies'. In: *Nuclear medicine and molecular imaging* 52.2, pp. 99–108 (cit. on p. 10).

Parmar, Chintan, Patrick Grossmann, Johan Bussink, Philippe Lambin, and Hugo JWL Aerts (2015). 'Machine learning methods for quantitative radiomic biomarkers'. In: *Scientific reports* 5.1, pp. 1–11 (cit. on p. 10).

Paszke, Adam, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. (2019). 'Pytorch: An imperative style, high-performance deep learning library'. In: *Advances in neural information processing systems* 32 (cit. on p. 50).

Patel, Mitesh, J Zhan, K Natarajan, Robert Flintham, Nigel Davies, Paul Sanghera, J Grist, V Duddalwar, A Peet, and V Sawlani (2021). 'Machine learning-based radiomic evaluation of treatment response prediction in glioblastoma'. In: *Clinical radiology* 76.8, 628–e17 (cit. on p. 10).

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M.

Brucher, M. Perrot, and E. Duchesnay (2011). 'Scikit-learn: Machine Learning in Python'. In: *Journal of Machine Learning Research* 12, pp. 2825–2830 (cit. on p. 52).

Petersen, Jens, Paul F Jäger, Fabian Isensee, Simon AA Kohl, Ulf Neuberger, Wolfgang Wick, Jürgen Debus, Sabine Heiland, Martin Bendszus, Philipp Kickingereder, et al. (2019). 'Deep probabilistic modeling of glioma growth'. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, pp. 806–814 (cit. on p. 10).

Pfaehler, Elisabeth, Ivan Zhovannik, Lise Wei, Ronald Boellaard, Andre Dekker, René Monshouwer, Issam El Naqa, Jan Bussink, Robert Gillies, Leonard Wee, et al. (2021). 'A systematic review and quality of reporting checklist for repeatability and reproducibility of radiomic features'. In: *Physics and imaging in radiation oncology* 20, pp. 69–75 (cit. on p. 10).

Polley, Mei-Yin C, Boris Freidlin, Edward L Korn, Barbara A Conley, Jeffrey S Abrams, and Lisa M McShane (2013). 'Statistical and practical considerations for clinical evaluation of predictive biomarkers'. In: *Journal of the National Cancer Institute* 105.22, pp. 1677–1683 (cit. on p. 34).

Poursaeed, Roya, Mohsen Mohammadzadeh, and Ali Asghar Safaei (2024). 'Survival prediction of glioblastoma patients using machine learning and deep learning: a systematic review'. In: *BMC cancer* 24.1, p. 1581 (cit. on p. 10).

Radiology (ESR), European Society of (2015). 'Medical imaging in personalised medicine: a white paper of the research committee of the European Society of Radiology (ESR)'. In: *Insights into imaging* 6.2, pp. 141–155 (cit. on p. 1).

Renfro, Lindsay A, Himel Mallick, Ming-Wen An, Daniel J Sargent, and Sumithra J Mandrekar (2016). 'Clinical trial designs incorporating predictive biomarkers'. In: *Cancer treatment reviews* 43, pp. 74–82 (cit. on p. 9).

Robins, James M and Dianne M Finkelstein (2000). 'Correcting for noncompliance and dependent censoring in an AIDS clinical trial with inverse probability of censoring weighted (IPCW) log-rank tests'. In: *Biometrics* 56.3, pp. 779–788 (cit. on p. 18).

Robins, James M and Andrea Rotnitzky (1992). 'Recovery of information and adjustment for dependent censoring using surrogate markers'. In: *AIDS epidemiology: methodological issues*. Springer, pp. 297–331 (cit. on p. 18).

Rodríguez-Camacho, Alejandro, José Guillermo Flores-Vázquez, Júlia Moscardini-Martelli, Jorge Alejandro Torres-Ríos, Alejandro Olmos-Guzmán, Cindy Sharon Ortiz-Arce, Dharely Raquel Cid-Sánchez, Samuel Rosales Pérez, Monsserrat Del Sagrario Macías-González, Laura Crystell Hernández-Sánchez, et al. (2022). 'Glioblastoma treatment: state-of-the-art and future perspectives'. In: *International journal of molecular sciences* 23.13, p. 7207 (cit. on p. 4).

Ronneberger, Olaf, Philipp Fischer, and Thomas Brox (2015). 'U-net: Convolutional networks for biomedical image segmentation'. In: *International Conference on Medical image computing and computer-assisted intervention*. Springer, pp. 234–241 (cit. on p. 53).

Rosenbaum, Paul R and Donald B Rubin (1983). 'The central role of the propensity score in observational studies for causal effects'. In: *Biometrika* 70.1, pp. 41–55 (cit. on p. 12).

Rubin, Donald B (2005). 'Causal inference using potential outcomes: Design, modeling, decisions'. In: *Journal of the American Statistical Association* 100.469, pp. 322–331 (cit. on p. 11).

Salvatore, Barbara, Maria Grazia Caprio, Billy Samuel Hill, Annachiara Sarnella, Giovanni Nicola Roviello, and Antonella Zannetti (2019). 'Recent advances in nuclear imaging of receptor expression to guide targeted therapies in breast cancer'. In: *Cancers* 11.10, p. 1614 (cit. on p. 2).

Sandmann, Thomas, Richard Bourgon, Josep Garcia, Congfen Li, Timothy Cloughesy, Olivier L Chinot, Wolfgang Wick, Ryo Nishikawa, Warren Mason, Roger Henriksson, et al. (2015). 'Patients with proneural glioblastoma may derive overall survival benefit from the addition of bevacizumab to first-line radiotherapy and temozolomide: retrospective analysis of the AVAglio trial'. In: *Journal of clinical oncology* 33.25, pp. 2735–2744 (cit. on p. 11).

Satten, Glen A and Somnath Datta (2001). 'The Kaplan–Meier estimator as an inverse-probability-of-censoring weighted average'. In: *The American Statistician* 55.3, pp. 207–210 (cit. on p. 18).

Schell, Marianne, Irada Pflüger, Gianluca Brugnara, Fabian Isensee, Ulf Neuberger, Martha Foltyn, Tobias Kessler, Felix Sahm, Antje Wick, Martha Nowosielski, Sabine Heiland, Michael Weller, Michael Platten, Klaus H Maier-Hein, Andreas Von Deimling, Martin J Van Den Bent, Thierry Gorlia, Wolfgang Wick, Martin Bendszus, and Philipp Kickingereder (2020). 'Validation of diffusion MRI phenotypes for predicting response to bevacizumab in recurrent glioblastoma: post-hoc analysis of the EORTC-26101 trial'. In: *Neuro-oncology* 22.11, pp. 1667–1676 (cit. on pp. 4, 11, 121).

Schrod, Stefan, Andreas Schäfer, Stefan Solbrig, Robert Lohmayer, Wolfram Gronwald, Peter J Oefner, Tim Beißbarth, Rainer Spang, Helena U Zacharias, and Michael Altenbuchinger (2022). 'BITES: balanced individual treatment effect for survival data'. In: *Bioinformatics* 38.Supplement_1, pp. i60–i67 (cit. on pp. 23, 27, 47–49, 52, 55, 56, 82, 120, 125, 137).

Seabold, Skipper and Josef Perktold (2010). 'statsmodels: Econometric and statistical modeling with python'. In: *9th Python in Science Conference* (cit. on p. 40).

Sechidis, Konstantinos, Konstantinos Papangelou, Paul D Metcalfe, David Svensson, James Weatherall, and Gavin Brown (2018). 'Distinguishing prognostic and predictive biomarkers: an information theoretic approach'. In: *Bioinformatics* 34.19, pp. 3365–3376 (cit. on pp. 3, 22, 25, 32).

Selvaraju, Ramprasaath R., Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra (2017). 'Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization'. In: *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 618–626. DOI: 10.1109/ICCV.2017.74 (cit. on pp. 36, 40, 65).

Shalit, Uri, Fredrik D Johansson, and David Sontag (2017). 'Estimating individual treatment effect: generalization bounds and algorithms'. In: *International Conference on Machine Learning*. PMLR, pp. 3076–3085 (cit. on pp. 14, 24, 33, 34, 125).

Shi, Claudia, David Blei, and Victor Veitch (2019). 'Adapting neural networks for the estimation of treatment effects'. In: *Advances in neural information processing systems* 32 (cit. on p. 24).

Shwartz-Ziv, Ravid and Amitai Armon (2022). 'Tabular data: Deep learning is not all you need'. In: *Information Fusion* 81, pp. 84–90 (cit. on p. 116).

Simonyan, Karen, Andrea Vedaldi, and Andrew Zisserman (2014). 'Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps'. In: *Workshop at International Conference on Learning Representations* (cit. on p. 35).

Smilkov, Daniel, Nikhil Thorat, Been Kim, Fernanda B. Viégas, and Martin Wattenberg (2017). 'SmoothGrad: removing noise by adding noise'. In: *CoRR* abs/1706.03825. arXiv: 1706.03825. URL: http://arxiv.org/abs/1706.03825 (cit. on p. 41).

Springenberg, Jost Tobias, Alexey Dosovitskiy, Thomas Brox, and Martin Riedmiller (2015). 'Striving for Simplicity: The All Convolutional Net'. In: *ICLR (workshop track)*. URL: http://lmb.informatik.uni-freiburg.de/Publications/2015/DB15a (cit. on pp. 36, 40, 65).

Strimbu, Kyle and Jorge A Tavel (2010). 'What are biomarkers?' In: *Current Opinion in HIV and AIDS* 5.6, pp. 463–466 (cit. on pp. 2, 9).

Sundararajan, Mukund, Ankur Taly, and Qiqi Yan (June 2017). 'Axiomatic Attribution for Deep Networks'. en. In: *arXiv:1703.01365 [cs]*. arXiv: 1703.01365. URL: http://arxiv.org/abs/1703.01365 (visited on 11/24/2021) (cit. on p. 65).

Susilo, Monica E, Chi-Chung Li, Kapil Gadkar, Genevive Hernandez, Ling-Yuh Huw, Jin Y Jin, Shen Yin, Michael C Wei, Saroja Ramanujan, and Iraj Hosseini (2023). 'Systems-based digital twins to help characterize clinical dose–response and propose predictive biomarkers in a phase I study of bispecific antibody, mosunetuzumab, in NHL'. In: *Clinical and Translational Science* 16.7, pp. 1134–1148 (cit. on p. 26).

Šutić, Maja, Ana Vukić, Jurica Baranašić, Asta Försti, Feđa Džubur, Miroslav Samaržija, Marko Jakopović, Luka Brčić, and Jelena Knežević (2021). 'Diagnostic, predictive, and prognostic biomarkers in non-small cell lung cancer (NSCLC) management'. In: *Journal of personalized medicine* 11.11, p. 1102 (cit. on p. 10).

Sutton, Richard S. (2019). *The Bitter Lesson*. http://www.incompleteideas.net/IncIdeas/BitterLesson.html. Accessed: 2025-10-15 (cit. on p. 4).

Svensson, David, Erik Hermansson, Nikolaos Nikolaou, Konstantinos Sechidis, and Ilya Lipkovich (2025). 'Overview and practical recommendations on using Shapley Values for identifying predictive biomarkers via CATE modeling'. In: *arXiv preprint arXiv:2505.01145* (cit. on pp. 22, 25).

Szegedy, Christian, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna (2016). 'Rethinking the inception architecture for computer vision'. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2818–2826 (cit. on p. 54).

Takeuchi, Koh, Ryo Nishida, Hisashi Kashima, and Masaki Onishi (2021). 'Grab the reins of crowds: Estimating the effects of crowd movement guidance using causal inference'. In: *arXiv preprint arXiv:2102.03980* (cit. on pp. 22, 24).

Tan, Aaron C, David M Ashley, Giselle Y López, Michael Malinzak, Henry S Friedman, and Mustafa Khasraw (2020). 'Management of glioblastoma: State of the art and future directions'. In: *CA: a cancer journal for clinicians* 70.4, pp. 299–312 (cit. on p. 10).

Tang, Fuk-Hay, Yee-Wai Fong, Shing-Hei Yung, Chi-Kan Wong, Chak-Lap Tu, and Ming-To Chan (2023). 'Radiomics-clinical AI model with probability weighted strategy for prognosis prediction in non-small cell lung cancer'. In: *Biomedicines* 11.8, p. 2093 (cit. on p. 42).

Tang, Yucheng, Dong Yang, Wenqi Li, Holger R Roth, Bennett Landman, Daguang Xu, Vishwesh Nath, and Ali Hatamizadeh (2022). 'Self-supervised pre-training of swin transformers for 3d medical image analysis'. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 20730–20740 (cit. on p. 54).

Tantalaki, Nicoleta, Stavros Souravlas, and Manos Roumeliotis (2019). 'Data-driven decision making in precision agriculture: The rise of big data in agricultural systems'. In: *Journal of agricultural & food information* 20.4, pp. 344–380 (cit. on p. 1).

Tarighati, Elaheh, Hadi Keivan, and Hojjat Mahani (2023). 'A review of prognostic and predictive biomarkers in breast cancer'. In: *Clinical and experimental medicine* 23.1, pp. 1–16 (cit. on pp. 2, 10).

Tarkhan, Aliasghar and Noah Simon (2024). 'An online framework for survival analysis: reframing Cox proportional hazards model for large data sets and neural networks'. In: *Biostatistics* 25.1, pp. 134–153 (cit. on p. 50).

Ternes, Nils, Federico Rotolo, Georg Heinze, and Stefan Michiels (2017). 'Identification of biomarker-by-treatment interactions in randomized clinical trials with survival outcomes and high-dimensional spaces'. In: *Biometrical Journal* 59.4, pp. 685–701 (cit. on p. 27).

Tian, Lu, Lihui Zhao, and Lee-Jen Wei (2014). 'Predicting the restricted mean event time with the subject's baseline covariates in survival analysis'. In: *Biostatistics* 15.2, pp. 222–233 (cit. on p. 19).

Tschandl, Philipp, Cliff Rosendahl, and Harald Kittler (2018). 'The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions'. In: *Scientific data* 5.1, pp. 1–9. DOI: 10.1038/sdata.2018.161 (cit. on p. 38).

Ulaner, Gary A, Chris C Riedl, Maura N Dickler, Komal Jhaveri, Neeta Pandit-Taskar, and Wolfgang Weber (2016). 'Molecular imaging of biomarkers in breast cancer'. In: *Journal of Nuclear Medicine* 57.Supplement 1, 53S–59S (cit. on p. 2).

Ulyanov, Dmitry, Andrea Vedaldi, and Victor Lempitsky (2016). 'Instance normalization: The missing ingredient for fast stylization'. In: *arXiv preprint arXiv:1607.08022* (cit. on p. 50).

Uno, Hajime, Tianxi Cai, Michael J Pencina, Ralph B D'Agostino, and Lee-Jen Wei (2011). 'On the C-statistics for evaluating overall adequacy of risk prediction procedures

with censored survival data'. In: *Statistics in medicine* 30.10, pp. 1105–1117 (cit. on p. 19).

Vale-Silva, Luís A and Karl Rohr (2021). 'Long-term cancer survival prediction using multimodal deep learning'. In: *Scientific Reports* 11.1, p. 13505 (cit. on pp. 23, 28).

Van Griethuysen, Joost JM, Andriy Fedorov, Chintan Parmar, Ahmed Hosny, Nicole Aucoin, Vivek Narayan, Regina GH Beets-Tan, Jean-Christophe Fillion-Robin, Steve Pieper, and Hugo JWL Aerts (2017). 'Computational radiomics system to decode the radiographic phenotype'. In: *Cancer research* 77.21, e104–e107 (cit. on p. 39).

Verhaeghe, Jarne, Femke Ongenae, and Sofie Van Hoecke (2025). 'Causalteshap: discerning predictive from prognostic features for treatment effect analysis'. In: *International Journal of Machine Learning and Cybernetics*, pp. 1–21 (cit. on pp. 3, 22, 25, 35).

Vock, David M, Julian Wolfson, Sunayan Bandyopadhyay, Gediminas Adomavicius, Paul E Johnson, Gabriela Vazquez-Benitez, and Patrick J O'Connor (2016). 'Adapting machine learning techniques to censored time-to-event health record data: A general-purpose approach using inverse probability of censoring weighting'. In: *Journal of biomedical informatics* 61, pp. 119–131 (cit. on pp. 18, 113).

Vollenweider, Michael, Manuel Schürch, Chiara Rohrer, Gabriele Gut, Michael Krauthammer, and Andreas Wicki (2025). 'Learning Personalized Treatment Decisions in Precision Medicine: Disentangling Treatment Assignment Bias in Counterfactual Outcome Prediction and Biomarker Identification'. In: *Proceedings of the 4th Machine Learning for Health Symposium*. Ed. by Stefan Hegselmann, Helen Zhou, Elizabeth Healey, Trenton Chang, Caleb Ellington, Vishwali Mhasawade, Sana Tonekaboni, Peniel Argaw, and Haoran Zhang. Vol. 259. Proceedings of Machine Learning Research. PMLR, pp. 991–1013. URL: https://proceedings.mlr.press/v259/vollenweider25a.html (cit. on p. 22).

Wah, C., S. Branson, P. Welinder, P. Perona, and S. Belongie (2011). *The Caltech-UCSD Birds-200-2011 Dataset*. Tech. rep. CNS-TR-2011-001. California Institute of Technology (cit. on p. 38).

Wald, Tassilo, Constantin Ulrich, Jonathan Suprijadi, Sebastian Ziegler, Michal Nohel, Robin Peretzke, Gregor Kohler, and Klaus Maier-Hein (2025). 'An OpenMind for 3D Medical Vision Self-supervised Learning'. In: pp. 23839–23879 (cit. on pp. 53, 54, 91, 99, 115, 126, 146, 147).

Weaver, Olena and Jessica WT Leung (2018). 'Biomarkers and imaging of breast cancer'. In: *American Journal of Roentgenology* 210.2, pp. 271–278 (cit. on p. 2).

Weberpals, Janick, Stefan Feuerriegel, Mihaela van der Schaar, and Kenneth L Kehl (2025). *Opportunities for causal machine learning in precision oncology* (cit. on p. 125).

Welch, M. L., S. Kim, A. Hope, S. H. Huang, Z. Lu, J. Marsilla, M. Kazmierski, K. Rey-McIntyre, T. Patel, B. O'Sullivan, J. Waldron, J. Kwan, J. Su, L. Soltan Ghoraie, H. B. Chan, K. Yip, M. Giuliani, Princess Margaret Head, Neck Site Group, S. Bratman, T. Tadic, et al. (2023). *Computed Tomography Images from Large Head and Neck Cohort (RADCURE) (Version 4) [Dataset]*. Dataset. DOI: 10.7937/J47W-NM11. URL: https://doi.org/10.7937/J47W-NM11 (cit. on p. 125).

Welch, Mattea L, Sejin Kim, Andrew J Hope, Shao Hui Huang, Zhibin Lu, Joseph Marsilla, Michal Kazmierski, Katrina Rey-McIntyre, Tirth Patel, Brian O'Sullivan, et al. (2024). 'RADCURE: An open-source head and neck cancer CT dataset for clinical radiation therapy insights'. In: *Medical Physics* 51.4, pp. 3101–3109 (cit. on p. 125).

Wen, Patrick Y, David R Macdonald, David A Reardon, Timothy F Cloughesy, A Gregory Sorensen, Evanthia Galanis, John DeGroot, Wolfgang Wick, Mark R Gilbert, Andrew B Lassman, et al. (2010). 'Updated response assessment criteria for high-grade gliomas: response assessment in neuro-oncology working group'. In: *Journal of clinical oncology* 28.11, pp. 1963–1972 (cit. on p. 11).

Wen, Patrick Y, Martin Van Den Bent, Gilbert Youssef, Timothy F Cloughesy, Benjamin M Ellingson, Michael Weller, Evanthia Galanis, Daniel P Barboriak, John De Groot, Mark R Gilbert, et al. (2023). 'RANO 2.0: update to the response assessment in neuro-oncology criteria for high-and low-grade gliomas in adults'. In: *Journal of Clinical Oncology* 41.33, pp. 5187–5199 (cit. on pp. 1, 11).

Wick, Wolfgang, Thierry Gorlia, Martin Bendszus, Martin Taphoorn, Felix Sahm, Inga Harting, Alba A. Brandes, Walter Taal, Julien Domont, Ahmed Idbaih, Mario Campone, Paul M. Clement, Roger Stupp, Michel Fabbro, Emilie Le Rhun, Francois Dubois, Michael Weller, Andreas von Deimling, Vassilis Golfinopoulos, Jacoline C. Bromberg, Michael Platten, Martin Klein, and Martin J. van den Bent (2017). 'Lomustine and Bevacizumab in Progressive Glioblastoma'. In: *New England Journal of Medicine* 377.20, pp. 1954–1963. DOI: 10.1056/NEJMoa1707358. eprint: https://www.nejm.org/doi/pdf/10.1056/NEJMoa1707358. URL: https://www.nejm.org/doi/full/10.1056/NEJMoa1707358 (cit. on pp. 4, 11, 45, 116).

Wick, Wolfgang, Roger Stupp, Thierry Gorlia, Martin Bendszus, Felix Sahm, Jacoline E Bromberg, Alba Ariela Brandes, Maaike J Vos, Julien Domont, Ahmed Idbaih, Jean-Sebastien Frenel, Paul M. Clement, Michel Fabbro, Emilie Le Rhun, François Dubois, Davide Musmeci, Michael Platten, Vassilis Golfinopoulos, and Martin J. Van Den Bent (2016). 'Phase II part of EORTC study 26101: The sequence of bevacizumab and lomustine in patients with first recurrence of a glioblastoma.' In: *Journal of Clinical Oncology* 34.15_suppl. PMID: pp. 2019–2019. DOI: 10.1200/JCO.2016.34.15_suppl.2019. eprint: https://ascopubs.org/doi/pdf/10.1200/JCO.2016.34.15_suppl.2019. URL: https://ascopubs.org/doi/abs/10.1200/JCO.2016.34.15_suppl.2019 (cit. on p. 45).

Wolf, Tom Nuno, Sebastian Pölsterl, Christian Wachinger, Alzheimer's Disease Neuroimaging Initiative, et al. (2022). 'DAFT: A universal module to interweave tabular data and 3D images in CNNs'. In: *NeuroImage* 260, p. 119505 (cit. on pp. 23, 28, 49, 77, 114, 137).

Wu, Linshan, Jiaxin Zhuang, and Hao Chen (2024). 'Voco: A simple-yet-effective volume contrastive learning framework for 3d medical image analysis'. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 22873–22882 (cit. on p. 126).

Xiang, Jinxi, Xiyue Wang, Xiaoming Zhang, Yinghua Xi, Feyisope Eweje, Yijiang Chen, Yuchen Li, Colin Bergstrom, Matthew Gopaulchan, Ted Kim, et al. (2025). 'A vision–language foundation model for precision oncology'. In: *Nature* 638.8051, pp. 769–778 (cit. on p. 127).

Xiao, Shuhan, Lukas Klein, Jens Petersen, Philipp Vollmuth, Paul F Jaeger, and Klaus H Maier-Hein (2025). 'Enhancing predictive imaging biomarker discovery through treatment effect analysis'. In: *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. ©2025 IEEE. IEEE, pp. 4512–4522. DOI: 10.1109/WACV61041.2025.00443 (cit. on pp. 13, 21, 22, 31, 32, 34, 37, 59, 61, 62, 64–66, 69, 70, 107).

Zabor, Emily C, Alexander M Kaizer, and Brian P Hobbs (2020). 'Randomized controlled trials'. In: *Chest* 158.1, S79–S87 (cit. on p. 3).

Zeng, Lang, Weijing Tang, Zhao Ren, and Ying Ding (2025). *Mini-batch Estimation for Deep Cox Models: Statistical Foundations and Practical Guidance*. arXiv: 2408.02839 [stat.ML]. URL: https://arxiv.org/abs/2408.02839 (cit. on p. 50).

Zhang, Jiaqi, Joel Jennings, Agrin Hilmkil, Nick Pawlowski, Cheng Zhang, and Chao Ma (2024). 'Towards causal foundation model: on duality between optimal balancing and attention'. In: *Forty-first International Conference on Machine Learning* (cit. on pp. 29, 126).

Zhao, Zhenyu and Totte Harinen (2019). 'Uplift modeling for multiple treatments with cost optimization'. In: *2019 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*. IEEE, pp. 422–431 (cit. on p. 15).

Zhou, Chuan, Yaxuan Li, Chunyuan Zheng, Haiteng Zhang, Min Zhang, Haoxuan Li, and Mingming Gong (2025). 'A Two-Stage Pretraining-Finetuning Framework for Treatment Effect Estimation with Unmeasured Confounding'. In: *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.1*. KDD '25. Toronto ON, Canada: Association for Computing Machinery, pp. 2113–2123. ISBN: 979-8-4007-1245-6. DOI: 10.1145/3690624.3709161. URL: https://doi.org/10.1145/3690624.3709161 (cit. on p. 126).

Zhou, Jie, Hany Deirawan, Fayez Daaboul, Thazin Nwe Aung, Rafic Beydoun, Fahad Shabbir Ahmed, Jeffrey H Chuang, et al. (2023). 'Integrative deep learning analysis improves colon adenocarcinoma patient stratification at risk for mortality'. In: *EBioMedicine* 94 (cit. on p. 28).

Zhu, Fucheng Warren, Connor Thomas Jerzak, and Adel Daoud (2025). 'Optimizing Multi-Scale Representations to Detect Effect Heterogeneity Using Earth Observation and Computer Vision: Applications to Two Anti-Poverty RCTs'. In: *Proceedings of the Fourth Conference on Causal Learning and Reasoning*. Ed. by Biwei Huang and Mathias Drton. Vol. 275. Proceedings of Machine Learning Research. PMLR, pp. 894–919. URL: https://proceedings.mlr.press/v275/zhu25a.html (cit. on pp. 22, 24).

Zhu, Wencan, Céline Lévy-Leduc, and Nils Ternès (2023). 'Identification of prognostic and predictive biomarkers in high-dimensional data with PPLasso'. In: *BMC bioinformatics* 24.1, p. 25 (cit. on pp. 3, 22, 25).

Zhu, Xinliang, Jiawen Yao, and Junzhou Huang (2016). 'Deep convolutional neural network for survival analysis with pathological images'. In: *2016 IEEE international conference on bioinformatics and biomedicine (BIBM)*. IEEE, pp. 544–547 (cit. on p. 28).

Ziegler, Sebastian et al. (2024). *MIC-DKFZ Image Classification Framework (OpenMind branch)*. `https://github.com/MIC-DKFZ/image_classification/tree/OpenMind`. Accessed: 2025-08-22 (cit. on p. 53).

Zwanenburg, Alex, Martin Vallières, Mahmoud A Abdalah, Hugo JWL Aerts, Vincent Andrearczyk, Aditya Apte, Saeed Ashrafinia, Spyridon Bakas, Roelof J Beukinga, Ronald Boellaard, et al. (2020). 'The image biomarker standardization initiative: standardized quantitative radiomics for high-throughput image-based phenotyping'. In: *Radiology* 295.2, pp. 328–338 (cit. on p. 39).