Aus dem Institut für Medizinische Biometrie
Universitätsklinikum Heidelberg
Geschäftsführender Direktor: Prof. Dr. sc. hum. Meinhard Kieser

Comparison of different additional benefit assessment methods for oncology treatments

Inauguraldissertation
zur Erlangung des
Doctor scientiarum humanarum (Dr. sc. hum.)

an der Medizinischen Fakultät Heidelberg
der Ruprecht-Karls-Universität

vorgelegt von
Christopher Alexander Büsch
aus
Neuwied am Rhein

2024 (Jahr der Einreichung)

Dekan: Prof. Dr. rer. nat. Michael Boutros

Doktorvater: Prof. Dr. sc. hum. Meinhard Kieser

Contents

Abbreviations and Symbols

ABAM	additional benefit assessment method
ADEMP	structure of simulation studies proposed by Morris et al. (2019) including aims, data-generation mechanism, estimands, methods and performance measures
ANV	Arzneimittel-Nutzenbewertungsverordnung, engl. regulation for early benefit assessment of new pharmaceuticals
ASCO	American Society of Clinical Oncology
AUC	area under the curve
BP	bonus points
CBS	clinical benefit score
CI	confidence interval
CR	complete response
ctl	control group
designHR	design hazard ratio used for sample size calculation
ESMO	European Society for Medical Oncology
FN	false negative
FP	false positive
FPR	false positive rate
gain	absolute difference in median survival times
GBA	Gemeinsamer Bundesaussusch, engl. Federal Joint Committee
GK-SV	Spitzenverband Bund der Krankenkassen, engl. National Association of Statutory Health Insurance Funds
HR	hazard ratio
HR-CI	hazard ratio confidence interval
HR-PE	HR point estimate
HR^{-}	lower 95% confidence interval limit of the HR-PE
HR^{+}	upper 95% confidence interval limit of the HR-PE
HR_{var}	factor for deviance between designHR and trueHR
IQWiG	Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen, engl. Institute for Quality and Efficiency in Health Care

IQWiG_{RR}	original IQWiG method
med_{ctl}	median survival time of control group
med_{trt}	median survival time of treatment group
$\text{Mod-IQWiG}_{\text{HR}}$	modified IQWiG method
NHB	net health benefit
OS	overall survival
p_C	overall censoring rate
p_C^{ctl}	Censoring rate of control group
p_C^{trt}	Censoring rate of treatment group
PE	point estimate
PR	partial response
QoL	quality of life
r	allocation ratio
ROC	receiver operating characteristic
ROC01	point on the ROC curve which is closest to the point (0,1)
RR	relative risk
TN	true negative
TP	true positive
TPR	true positive rate
trt	treatment group
trueHR	true underlying HR of data generation
TS	toxicity score

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Background

During the development of a new treatment many requirements for the market authorization must be met through various stages, as illustrated in Figure 1. After in vitro research, the effectiveness is tested with pre-clinical treatment development investigations. Only treatments, which satisfy the expectations of the company and the conditions of the authorities, are taken to the last and essential step for market approval: the clinical development and registration trials in humans. These trials are divided into three phases, where phase III trials are the basis of market approval. The main objective of phase III trials is to verify whether the results of the previous studies can also be achieved in a larger and broader population of patients and hence ultimately proof efficacy of the new treatment. Furthermore, interactions with other treatments and side effects of the new treatment can be spotted. Thus, a trial with a large patient size and a long follow-up is performed, where the needed sample size is usually calculated with the help of the observed effect of the prior performed studies. After positive significant results of these phase III trials, a marketing authorisation application including all trials (pre-clinical and clinical) with analyses and conclusions can be submitted to an appropriate authority. If the treatment is approved, the additional benefit of the new treatment is compared to the already established treatments with the help of the results of the phase III trial(s) (see white path of Figure 1). This assessment can decide on the amount of reimbursement of the new treatment on the market and can help to manage the uncertainty of patients regarding the treatment medical effectiveness and toxicity (Weeks et al., 2012).

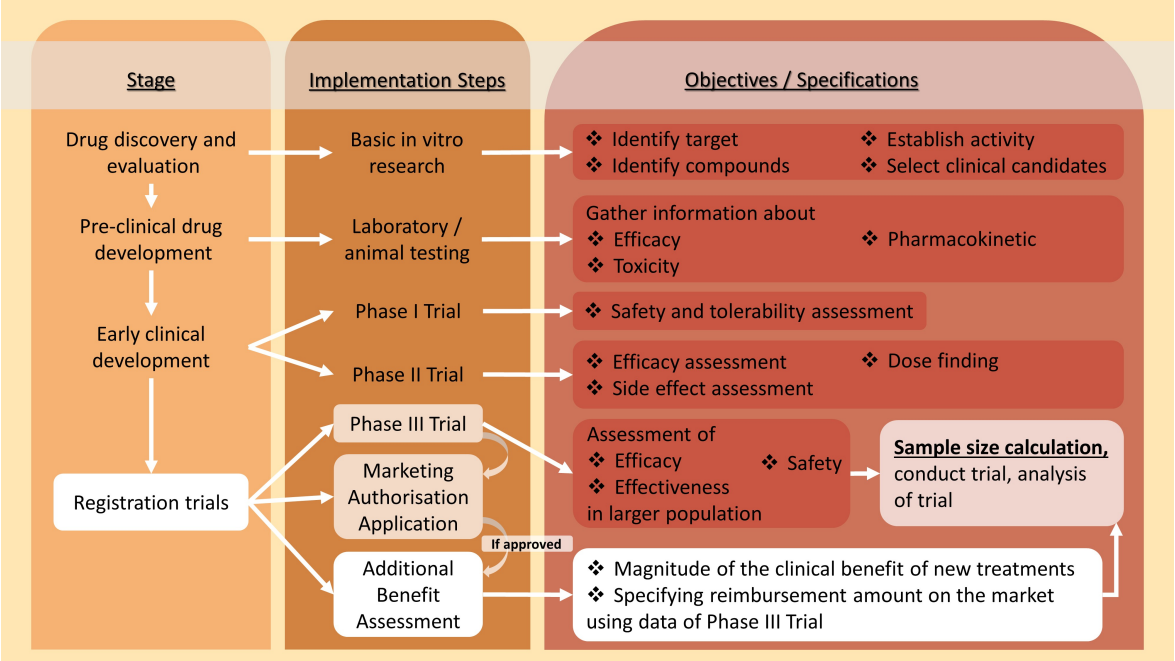

Figure 1: *Stages of treatment development*

This thesis focuses on non-curative or advanced diseases like many types of cancer. Thus, time-to-event endpoints such as overall survival (OS) are assumed to be present. For these types of endpoints, three different additional benefit assessment methods have been developed so far. Two of them have been developed in Europe (Institute for Quality and Efficiency in Health Care (IQWiG), European Society for Medical Oncology (ESMO)) and one method in the United States of America (American Society of Clinical Oncology (ASCO)). A short summary of these methods is provided in Table 1.

In Germany, IQWiG evaluates the additional benefit of new treatments, which are approved by the appropriate regulatory authorities. This evaluation is commissioned by the Federal Joint Committee (GBA, germ. Gemeinsamer Bundesaussusch) and is carried out by comparing the new treatment against established treatments on the market. For time-to-event endpoints, IQWiG determines the degree of additional benefit by comparing the upper limit of the 95% hazard ratio (HR) confidence interval (CI) against specific relative risk based thresholds. This comparison categorises the new treatment into the following three categories: major, considerable, and minor (more than marginal) added benefit (Skipka et al., 2016). For other endpoint types like continuous endpoints, the IQWiG assessment consists out of three more and hence overall six categories: major, considerable, minor (more than marginal),

nonquantifiable (potentially minor, considerable or major), no added benefit, and less benefit (than the appropriate comparator therapy). IQWiG has chosen relative effect measures like the HR to evaluate the benefit of new treatments because of the non-transferability of absolute measures, e.g. the risk difference, from clinical studies to real life practice (Skipka et al., 2016). Moreover, for classification into the above mentioned ranking IQWiG uses the two-sided 95% CI instead of the point estimate because the variance and hence the precision of the point estimate is considered in the CI. Skipka et al. (2016) also mention another advantage being that the probability of statistical errors of a CI can be controlled by the used significance level, e.g. for a 95% CI it is limited to 5%. Based on the determined category of IQWiG, the GBA decides on the additional benefit of the new treatment which can play an important role in the negotiation of the reimbursement amount on the market between the National Association of Statutory Health Insurance Funds (GKV-SV, germ. Spitzenverband Bund der Krankenkassen) and the pharmaceutical company.

Table 1: *Overview of additional benefit assessment methods*

Method	Main statistical quantity	Levels of outcome
ASCO	HR-PE	Continuous
ESMO	Lower limit of the 95% HR-CI (HR^-)	Ordinal (5 categories): • 1 to 3: low benefit • 4 to 5: substantial benefit
IQWiG	Upper limit of the 95% HR-CI (HR^+)	Ordinal (3 categories): • minor added benefit • considerable added benefit • major added benefit

Abbreviations: CI: Confidence interval, HR: Hazard ratio, HR^+: Upper 95% confidence interval limit of the HR-PE, HR^-: Lower 95% confidence interval limit of the HR-PE, PE: Point estimate

ESMO developed a method named Magnitude of Clinical Benefit Scale Version 1.1, which has two different forms. Form 1 is used for the curative setting of adjuvant and neoadjuvant therapies, where no time-to-event data like OS or progression free survival (PFS) is present. For these cases other measures like disease-free survival or recurrence-free survival can be used. Form 2 is used for non-curative settings and is divided into form 2a if the primary outcome is OS and form 2b if it is PFS. Since this thesis focuses on the non-curative setting with OS as primary outcome, form 2a is of interest, which uses a dual rule consisting out of a relative and absolute benefit component to compute a preliminary scale. Based on the

lower limit of the 95% HR-CI (relative benefit rule) and the observed absolute difference in median treatment outcomes (absolute benefit rule) the new treatment is categorised into four categories. This preliminary scale is then adjusted for toxicity, quality of life outcomes or other bonus adjustments of the new treatment so that finally, a ranking into 5 categories results. Here categories 1 to 3 represent low and categories 4 and 5 substantial benefit, respectively (Cherny et al., 2015, 2017). The classification of ESMO affects the price paid for the new medication in an indirect way. Cherny et al. (2015) mention that the ESMO method is applied to every new anti-cancer treatment which has been approved by the European Medicine Agency. Furthermore, ESMO emphasizes every treatment with substantial benefit (category 5 or 4) in their guideline with the intent to stimulate prompt usage in Europe with the help of other health authorities. Hence, ESMO's classification has a rather indirect influence on the treatment's price.

Besides these two additional benefit assessment methods developed in Europe, an additional method is used in the United States of America, which has been developed by ASCO and is named Value Framework Net Health Benefit (NHB) Score. Since the United States has a different health care system compared to Europe, the high costs of new treatments are often not covered by statutory health insurance and have to be payed by the patients. This leads to difficulties for many patients because they either do not have private health insurance or cannot pay for the treatment on their own. This is especially the case for cancer treatments as these treatments are relatively expensive (Danzon and Taylor, 2010). Thus, the main aim of the society was to allow physicians and patients to assess the current possible treatment options and make a shared decision between the different treatments and their price. Hence, ASCO juxtaposes the cost of the treatment against the NHB score without influencing the pricing of the treatment (Schnipper et al., 2015, 2016), which is the only method to do so. The NHB score is defined differently for advanced diseases and for potentially curable diseases (adjuvant therapy). As this thesis is focused on advanced diseases, the NHB score consists of the clinical benefit, toxicity, and bonus points (quality of life, treatment-free interval, cost and other characteristics of the new treatment). The main element of the NHB score is the assessment of the clinical benefit, which is calculated by subtracting the HR point estimate (PE) from 1 and then multiplying the result by 100. This score can then be adjusted by

the other components of the method, for example percentage of toxicity grades adjustments, leading to the final NHB score.

The main difference of the three above described methods for the assessment of the additional benefit of new treatments is the evaluation of the clinical benefit. IQWiG uses the upper limit of the 95% HR, ESMO uses the lower limit of the 95% HR-CI, and ASCO uses the HR-PE. At first glance, there are several points of criticism to each of the assessment methods. For example, the use of the HR-PE for the assessment of the clinical benefit could penalize studies of substantial benefit by ignoring the precision of the estimate. In contrast, the upper or lower limit of the HR-CI take into account the variability of the estimate and hence should provide more information. Nevertheless, the use of the lower confidence interval limit could lead to a higher probability of a better grading because it may credit studies with a smaller sample size and hence with a wider confidence interval.
Moreover, the statistical quantities used for the three methods can but do not have to be estimated using several studies, for example by using a meta analysis of two phase III studies. This thesis focuses on single phase III studies with OS as primary endpoint and thus cases, where two or more studies are combined for the methods application are not considered. Since this simplification influences all benefit assessment methods in the same way, this does not create any impact on the interpretation of the performed research. Furthermore, this thesis' focus is on non-curable diseases and hence the advanced disease framework of ASCO and non-curative settings of ESMO are implemented, i.e. overall survival and no progression free survival efficacy endpoint is used. In addition, to achieve a fair comparison of the statistical approaches between these methods, the statistical components of the methods are evaluated and compared. Consequently, adjustments regarding non-statistical components like cost, toxicity, quality of life or bonus point adjustments were not implemented. As a result ESMO consists only out of the preliminary scale ranging from 1 to 4 and thus the maximal categories of IQWiG and ESMO are comparable (major added benefit is considered as the equivalent of substantial improvement).

1.2 Previous work

Despite the publications for the additional benefit assessment methods itself (Schnipper et al., 2015, 2016; Cherny et al., 2015, 2017; Skipka et al., 2016), the following research aspects have already been investigated empirically.

Cherny et al. (2019) calculated ASCO cutoff values corresponding to categories of ESMO. In their empirical investigation 102 randomized controlled trials were used for calculation of the cutoff values, resulting in an ASCO score of 46 or greater and 41 or less to define substantial benefit (category 4) and low benefit (category 1–3), respectively. In addition, the relationship between ESMO and ASCO has been examined on the same sample of 102 trials using the Spearman correlation between both methods. Other publications also using real studies for the ASCO and ESMO application show different inconsistent negligible to low correlation results of 0.17 (Cheng et al., 2017), 0.397 (Del Paggio et al., 2018), and 0.40 (Becker et al., 2017) between ASCO and ESMO.

The aspect whether HR^- or HR-PE might be better for the assessment of additional benefit was investigated by Dafni et al. (2017). They stated that HR^- should be preferred over HR-PE as well as the fact that ESMO does not show discriminatory behavior in over-/underpowered trials. In the same spirit the inventors of IQWiGs method (Skipka et al., 2016) mentioned that HR-CI provides more information than HR-PE through the included variability in these estimates and that HR-PE might introduce potential bias to the additional benefit assessment. In addition, two Letters to the Editors (Muhonen et al., 2015; Wild et al., 2016) raised the concern that HR^- might lead to higher grades, especially in studies with smaller sample sizes which could lead to deliberately overpowered studies.

1.3 Aim of this thesis

Besides the different health care systems in the world and the fact that the methods were developed for different purposes, they share the objective to provide an assessment of the clinical additional benefit of new treatments. As described above, the main difference between the three methods is the use of different statistical quantities for the clinical additional benefit assessment: HR^-, HR-PE, and HR^+.

So far, the three different methods have only been compared empirically and hence this thesis

aims to close this knowledge gap. Furthermore, in previous research the complete methods were applied based on the results of clinical trials and hence did not focus on the statistical quantities. Moreover, all three methods and their approaches have never been compared collectively. Hence, the aim of this thesis is to obtain a better understanding of the differences between the additional benefit assessment methods and to answer the question which statistical quantity has the best properties to assess additional benefit.
To achieve these objectives, this thesis evaluates and compares the above described three methods for the clinical assessment of the additional benefit of new cancer treatments by means of comprehensive simulation studies. Furthermore, it is investigated which category of ESMO and IQWiG corresponds to which ASCO score in order to enable an easier comparison between all three methods. The simulation studies are constructed with the focus to implement realistic phase III trials and hence comprise different failure time distributions, treatment effects, power, allocation ratios, censoring types, and censoring rates. Furthermore, scenarios with non-proportional hazards, underpowered trials, and overpowered trials are investigated.

1.4 Structure of this thesis

This thesis is structured as follows: In Chapter 2, the methods are provided, introducing the methodological tools, the required knowledge needed for the elaboration of the results, a detailed description of the additional benefit assessment methods, and an in-depth description of the performed simulation studies. The results are presented in Chapter 3, which consists of three sections: Section 3.1 presents the results of a simulation study using a censoring mechanism, which introduces bias in the HR estimation (Simulation 1). The results investigating the robustness of Simulation 1 are shown in Section 3.2 (Simulation 2). Section 3.3 outlines the application of the additional benefit assessment methods and determined ASCO cutoff application illustrated by two clinical studies. In Chapter 4, the results are discussed together with their contribution to research as well as limitations and directions for further research. In Chapter 5 the thesis is summarized in English and Chapter 6 provides a direct translation to German. Additional results of the simulation studies are presented in Appendix A, while Appendix B contains information on R-Code structure and execution of the programs to determine the results of this thesis. The performed simulation studies are reproducible with

the provided R-Code `https://www.github.com/cbuesch/SumulationStudyABAM`. Moreover, Appendix C contains information of simulation parameters of Simulation 2.

Chapter 2

Methodology

2.1 Background of time-to-event analysis

In the following, the theoretical background in time-to-event analysis will be given. The fundamental terms and formulas that form the basis of the simulation studies will be introduced. One of the main advantages of time-to-event analysis is that it can handle data of patients with unknown event time, e.g. due to a patient being event-free at end of study, patient cannot be followed up (e.g. moved away) or other reasons. Hence, it can handle incomplete data. This occurrence of incomplete data is called censoring. Censored data should in any case not be excluded from the analysis because important information would be lost and thereby the average time-to-event and probability of an event would be incorrectly estimated. For instance, lets assume a patient shows no event at the end of study due to the positive effect of the new treatment. If one would leave this kind of patients out of the analysis and ignore this patient is event-free throughout the study, the effect of the new drug would be underestimated. In addition, one has to distinguish between three different censoring types:

- Right censoring: the event occurs later than the observed time.
- Left censoring: the event has occurred already before the study began.
- Interval censoring: the exact event time is unknown but it is known that the event occurred between two time points.

In this thesis, left censoring has no relevance because all patients are assumed to be alive at the beginning of the trial. Moreover, as a result of the simulation studies and hence of the generated event times it can be assumed that one knows the exact event times of each patient and thus no interval censoring is present as well.

Basics

$T > 0$ is a continuous random variable which defines the time to an event of interest. The **distribution** is denoted by $F(t)$ and the **density** by $f(t)$. Given this setting the **survival function** $S(t)$ is defined the following way:

$$S(t) := \mathbb{P}(T > t) = 1 - \mathbb{P}(T \leqslant t) = 1 - F(t) = \int_t^\infty f(v)dv.$$

This function represents the probability that an event occurs at time point t and therefore the function must be monotonic decreasing. Moreover, the survival function is always equal to 1 at time point zero, $S(0) = 1$.

The density $f(t)$ and the **hazard** $h(t)$ can be expressed in terms of the survival function as:

$$\begin{aligned} f(t) &:= \lim_{\Delta t \searrow 0} \frac{\mathbb{P}(t \leqslant T \leqslant t + \Delta t)}{\Delta t} = F'(t) = -S'(t), \\ h(t) &:= \lim_{\Delta t \searrow 0} \frac{\mathbb{P}(t \leqslant T \leqslant t + \Delta t | T \geqslant t)}{\Delta t} = \frac{f(t)}{\mathbb{P}(T \geqslant t)} = \frac{f(t)}{S(t)} = -\frac{S'(t)}{S(t)}, \end{aligned}$$

where $F'(t)$ and $S'(t)$ represent the first derivative of $F(t)$ and $S(t)$, respectively. The hazard function is the ratio of the density function $f(t)$ and the survival function $S(t)$ and describes the failure rate or force of mortality.

Moreover, $H(t)$ denotes the **cumulative hazard** and hence is defined as follows

$$H(t) := \int_0^t h(x)dx = \int_0^t -\frac{d \ln S(x)}{dx}dx = -\ln S(t).$$

Cox proportional hazard model

With the help of the proportional hazard model proposed by Cox (1972) the influence of covariates (e.g., age, treatment, etc.) on the survival time can be analysed. Let $\mathbf{b} = (b_1, ..., b_p)$ be the p-dimensional covariate vector. The model is defined as follows:

$$h(t, \mathbf{b}) := h_0(t) \cdot \exp(\mathbf{b}^\star \cdot \boldsymbol{\beta}),$$

where $h_0(t)$ is an arbitrary baseline-hazard, $\mathbf{b}^\star$ is the transpose of $\mathbf{b}$, and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)$ is the parameter vector.

Based on this model, the **hazard ratio (HR)**, which is the ratio of the hazard rates between two patients i and j $(i \neq j)$ with covariate vectors $\mathbf{b}^i = (b_1^i, \ldots, b_p^i)$ and $\mathbf{b}^j = (b_1^j, \ldots, b_p^j)$, can be calculated as:

$$\text{HR} = \frac{h(t, \mathbf{b}^i)}{h(t, \mathbf{b}^j)} = \exp((\mathbf{b}^i - \mathbf{b}^j)^\star \cdot \boldsymbol{\beta}).$$

Since the baseline hazards of both patients are the same, they cancel each other out and hence, the HR is constant over time. "Proportional" refers to the fact that the HR is the constant of proportionality and describes the factor by which these two hazards differ at every time point.

Estimation of the parameter vector $\boldsymbol{\beta}$

Let $\mathbf{b}^i = (b_1^i, \ldots, b_p^i)$ be the p dimensional covariate vector for patient i at the observed event time t_i. In addition, a censoring time or event time is handled by a binary variable d_i, which is equal to 1 when an event occurred and equal to 0 when the event time is not known, hence the patient is censored. For the estimation of the parameter vector $\boldsymbol{\beta}$ Cox (Cox, 1972) estimated the conditional probability that patient i has an event at time point t_i given the previous observations. This function is called partial likelihood function L_i and can be written as:

$$L_i(\boldsymbol{\beta}) = \frac{h(t_i, \mathbf{b}^i)}{\sum\limits_{\mathrm{j} \in \{j | \mathrm{t}_j \geqslant \mathrm{t}_i\}} h(t_i, \mathbf{b}^j)} = \frac{h_0(t_i) \cdot \exp(\mathbf{b}^i \cdot \boldsymbol{\beta})}{\sum\limits_{j \in \{j | \mathrm{t}_j \geqslant \mathrm{t}_i\}} h_0(t_i) \cdot \exp(\mathbf{b}^j \cdot \boldsymbol{\beta})} = \frac{\exp(\mathbf{b}^i \cdot \boldsymbol{\beta})}{\sum\limits_{j \in \{j | \mathrm{t}_j \geqslant \mathrm{t}_i\}} \exp(\mathbf{b}^j \cdot \boldsymbol{\beta})},$$

where the summation is only over the set of patients j where no event or censoring has occurred before time point i.

With the help of the partial likelihood function and the assumption that the patients are independent from each other, the joint partial likelihood function is defined as

$$L(\boldsymbol{\beta}) = \prod_{i \in \{1, \ldots, n | d_i = 1\}} L_i(\boldsymbol{\beta}).$$

This function is used for the estimation of the effect of the covariates without modeling the change of the hazard over time. Hence, one has to maximize this function with respect to $\boldsymbol{\beta}$ to obtain a maximum likelihood estimate of the model parameters. An easier way to perform

this is to maximize the log joint partial likelihood function:

$$l(\boldsymbol{\beta}) = \sum_{i \in \{1,\ldots,n | d_i = 1\}} \left(\mathbf{b}^i \cdot \boldsymbol{\beta} - \log \sum_{j \in \{j : \mathrm{t}_j \geqslant \mathrm{t}_i\}} \exp(\mathbf{b}^j \cdot \boldsymbol{\beta}) \right).$$

Thus, the parameter vector can be estimated by solving the equations

$$\frac{\partial l(\boldsymbol{\beta})}{\partial \beta_k} = 0,$$

where $k \in \{1, \ldots, p\}$.

Interpretation of the HR

If the parameter vector only contains one binary covariate like for example the treatment assignment of a study, the HR of patient i receiving the active treatment ($b_1^i = 1$) and patient j receiving the control treatment ($b_1^j = 0$) reduces to

$$\mathrm{HR} = \frac{h(t, b_1^i)}{h(t, b_1^j)} = \exp\left((b_1^i - b_1^j) \cdot \beta_1\right) = \exp(\beta_1).$$

There are three possible situations:

1. HR < 1: The risk of an event for a patient treated with the active treatment ($b_1^i = 1$) is less than that of a patient treated with the control treatment ($b_1^j = 0$).

2. HR $= 1$: The risk of an event is the same in both treatment groups at any given time point during the study.

3. HR > 1: The risk of an event for a patient treated with the active treatment ($b_1^i = 1$) is higher than that of a patient treated with the control treatment ($b_1^j = 0$).

2.1.1 Distributions of time-to-event data

For the performed simulation studies in this thesis, exponential, Weibull, and Gompertz distributions are of importance for the generation of time-to-event data. Therefore, the probability density function, cumulative distribution function, survival function, hazard function, cumulative hazard function, median, expected value, and variance of the three distributions

are illustrated below in Table 2. Furthermore, conditions for fulfilling proportional hazards for the distributions are illustrated below.

Table 2: *Distributions of time-to-event data*

Distribution	Exponential	Weibull	Gompertz
Probability density function $f(t)$	$\lambda \cdot \exp(-\lambda \cdot t)$, where $t \geqslant 0, \lambda > 0$	$k \cdot \lambda \cdot \left(t \cdot \lambda\right)^{k-1} \cdot \exp\left(-\left(t \cdot \lambda\right)^{k}\right)$, where $t \geqslant 0$, $k \in (0, \infty)$ and $\lambda \in (0, \infty)$	$b \cdot \exp(a \cdot t) \cdot \exp\left(-\frac{b}{a} \cdot (\exp(a \cdot t) - 1)\right)$, where $t \geqslant 0$, $a \in [0, \infty)$ and $b \in (0, \infty)$
Cumulative distribution function $F(t)$	$1 - \exp(-\lambda \cdot t)$	$1 - \exp\left(-\left(t \cdot \lambda\right)^{k}\right)$	$1 - \exp\left(-\frac{b}{a} \cdot (\exp(a \cdot t) - 1)\right)$
Survival function $S(t)$	$\exp(-\lambda \cdot t)$	$\exp\left(-\left(t \cdot \lambda\right)^{k}\right)$	$\exp\left(-\frac{b}{a} \cdot (\exp(a \cdot t) - 1)\right)$
Hazard function $h(t)$	λ	$k \cdot \lambda \cdot \left(t \cdot \lambda\right)^{k-1}$	$b \cdot \exp(a \cdot t)$
Cumulative hazard function $H(t)$	$\lambda \cdot t$	$-\left(t \cdot \lambda\right)^{k}$	$-\frac{b}{a} \cdot (\exp(a \cdot t) - 1)$
Median	$\frac{\ln(2)}{\lambda}$	$\frac{(\ln(2))^{\frac{1}{k}}}{\lambda}$	$\frac{1}{a} \cdot \ln\left(1 + \frac{a}{b} \cdot \ln(2)\right)$
Expected value	$\frac{1}{\lambda}$	$\frac{1}{\lambda} \cdot \Gamma(\frac{1}{k} + 1)$	$\frac{1}{a} \cdot \exp\left(\frac{b}{a}\right) \cdot \mathrm{E}_1\left(\frac{b}{a}\right)$
Variance	$\frac{1}{\lambda^2}$	$\frac{1}{\lambda^2} \cdot \left[\Gamma\left(1 + \frac{2}{k}\right) - \left(\Gamma\left(1 + \frac{1}{k}\right)\right)^2\right]$	$\frac{2}{a^2} \exp\left(\frac{b}{a}\right) \cdot \left\{ -\frac{b}{a}\, {}_3F_3\left[\begin{matrix}1, & 1, & 1; \\ 2, & 2, & 2;\end{matrix} \frac{b}{a}\right] + \frac{1}{2}\left[\frac{\pi^2}{6} + \left(\gamma + \ln\left(\frac{b}{a}\right)\right)^2\right]\right\} - \left[\frac{1}{b} \exp\left(\frac{b}{a}\right) \mathrm{E}_1\left(\frac{b}{a}\right)\right]^2$

Notes: $\Gamma(n) = \int_0^\infty \exp(-t) \cdot t^{n-1} dt$, $\mathrm{E}_n(z) = \int_1^\infty \frac{\exp(-z \cdot t)}{t^n} dt, n > 0, z > 0$ *is defined by Abramowitz and Stegun (1965),* $\gamma \approx 0.57722$ *is the Euler-Mascheroni constant and* ${}_pF_q\left[\begin{matrix}a_1, & \ldots, & a_p \\ b_1, & \ldots, & b_p;\end{matrix} z\right] = \sum_{k=0}^{\infty} \frac{(a_1)_k \cdots (a_p)_k}{(b_1)_k \cdots (b_p)_k} \frac{z^k}{k!}$ *denotes the generalized hypergeometric function (Askey and Daalhuis, 2010).*

In case of Weibull distribution, $k \in (0, \infty)$ is called shape and $\lambda \in (0, \infty)$ scale parameter. The shape parameter can be interpreted in the following way:

- $k < 1$ indicates that the failure rate decreases over time leading to a monotonically decreasing hazard function.
- $k = 1$ indicates that the failure rate is constant over time leading to a constant hazard function.
- $k > 1$ indicates that the failure rate increases over time leading to a monotonically increasing hazard function.

If the shape parameter is set to 1, the distribution simplifies to an exponential distribution where the scale parameter of the Weibull distribution is equal to the parameter of the exponential distribution $\left(\lambda_{\text{exp}} = \lambda_{\text{weibull}}\right)$.
In case of Gompertz distribution, $a \in [0, \infty)$ is the shape parameter and $b \in (0, \infty)$ is the

rate parameter of the distribution. The shape parameter can be interpreted in the following way:

- $a < 0$ indicates that the failure rate decreases over time leading to a monotonically decreasing hazard function.
- $a = 0$ indicates that the failure rate is constant over time leading to a constant hazard function.
- $a > 0$ indicates that the failure rate increases over time leading to a monotonically increasing hazard function.

2.1.1.1 Condition for proportional hazards

The proportional hazard assumption between two distributions is fulfilled, if the ratio of the two hazard functions ($\mathrm{HR}(t) = \frac{h_1(t)}{h_2(t)}$) is constant over time t.
In case of two exponential distributions, the HR reduces to $\mathrm{HR}(t) = \frac{\lambda_1}{\lambda_2}$, where λ_1 and λ_2 are the parameters of two exponential distributions. Since time t does not influence the HR, proportional hazards are always fulfilled in case of two exponential distributions.
In case of two Weibull distributions, the HR reduces to

$$\mathrm{HR}(t) = \frac{k_1 \cdot \lambda_1 \cdot \Big(t \cdot \lambda_1\Big)^{k_1-1}}{k_2 \cdot \lambda_2 \cdot \Big(t \cdot \lambda_2\Big)^{k_2-1}} = \frac{\lambda_1^{k_1} \cdot k_1 \cdot t^{k_1-1}}{\lambda_2^{k_2} \cdot k_2 \cdot t^{k_2-1}},$$

where k_1, k_2, λ_1, and λ_2 are the shape and scale parameters of two Weibull distributions. To achieve a constant HR over time t, the shape parameter k has to be chosen to be equal for both Weibull distributions ($k = k_1 = k_2$) leading to $\mathrm{HR}(t) = \frac{\lambda_1^k}{\lambda_2^k}$.
In case of two Gompertz distributions, the HR reduces to

$$\mathrm{HR}(t) = \frac{b_1 \cdot \exp(a_1 \cdot t)}{b_2 \cdot \exp(a_2 \cdot t)},$$

where a_1, a_2, b_1, and b_2 are the shape and rate parameters of two Gompertz distributions. To achieve a constant HR over time t, the shape parameter a has to be chosen to be equal for both Gompertz distributions ($a = a_1 = a_2$) leading to $\mathrm{HR}(t) = \frac{b_1}{b_2}$.

2.2 Additional benefit assessment methods (ABAMs)

In the following, the additional benefit assessment methods are described, which are investigated in this thesis. As outlined in Section 1.3 this thesis is focused on the statistical aspects of the methods for advanced or non-curable diseases with OS outcome. Thus, only these parts of the methods are described.

Parts of this Section 2.2 are already published in the articles *A Comprehensive Comparison of Additional Benefit Assessment Methods Applied by Institute for Quality and Efficiency in Health Care and European Society for Medical Oncology for Time-to-Event Endpoints After Significant Phase III Trials — a Simulation Study* by Büsch et al. (2022) and *A Comparison of Additional Benefit Assessment Methods for Time-to-Event Endpoints Using Hazard Ratio Point Estimates or Confidence Interval Limits by Means of a Simulation Study* by Büsch et al. (2024). The manuscripts have been written by the lead author but may contain comments and corrections from the co-authors and the reviewers.

2.2.1 Institute for Quality and Efficiency in Health Care (IQWiG)

In case of time-to-event endpoints, IQWiG uses relative risk (RR) based thresholds derived for a binary outcome for the hazard ratio as effect measurement. Hence, in this thesis, IQWiG_{RR} is defined as the method used for survival endpoints. For the main classification, the upper limit of the 95% HR-CI (HR^+) is compared to RR based thresholds of 0.85 and 0.95. Thus, $\text{HR}^+<0.85$ is considered as major, $0.85 \leqslant \text{HR}^+ < 0.95$ as considerable, and $\text{HR}^+ \geqslant 0.95$ as minor added benefit. An overview of the used thresholds by IQWiG is given in Table 3, where the categories "less benefit", "no added benefit", and "nonquantifiable added benefit" cannot be assigned because a statistically significant increase in the survival time ($\text{HR} < 1$) has to be already shown by one or multiple phase III studies. Otherwise, the additional benefit assessment method would not be applied. Thus, only the three remaining categories "minor added benefit", "considerable added benefit", and "major added benefit" can be assigned.

In the following the derivation of the RR based thresholds by Skipka et al. (2016) are outlined: At first, a (fictional) study is planned to assess the test problem

$$H_0 : RR \geqslant RR_0 \qquad vs. \qquad H_1 : RR < RR_0,$$

where $RR_0 = 1$ and $RR = {}^{p_{trt}}/_{p_{ctl}}$ with p_{ctl} and p_{trt} represents the risk rate of the control (ctl) and treatment (trt) group, respectively. Assuming further, a significance level of α, power of $1-\beta$, true risk rates of $p_{ctl,1}$ and $p_{trt,1}$ (and thus $RR_1 = {}^{p_{trt,1}}/_{p_{ctl,1}}$), the overall sample size $N = n_{ctl} + n_{trt}$ with equal sample size in both groups can be calculated by

$$N = 4 \cdot \frac{[z_{1-\alpha}\varphi_{ctl}(p_{ctl,1}, p_{trt,1}) + z_{1-\beta}\varphi_{trt}(p_{ctl,1}, p_{trt,1})]^2}{[p_{trt,1} - RR_0 \cdot p_{ctl,1}]}, \tag{2.1}$$

where z_k is the k-quantile of $N(0,1)$ and $\varphi_{ctl}(p_{ctl,1}, p_{trt,1})$ as well as $\varphi_{trt}(p_{ctl,1}, p_{trt,1})$ are defined as follows:

$$\varphi_{ctl}(p_{ctl,1}, p_{trt,1}) := \sqrt{\Big(\frac{p_{ctl,1}}{2} + \frac{p_{trt,1}}{2}\Big)\Big(1 - \frac{p_{ctl,1}}{2} - \frac{p_{trt,1}}{2}\Big)},$$

$$\varphi_{trt}(p_{ctl,1}, p_{trt,1}) := \sqrt{\frac{p_{ctl,1}(1-p_{ctl,1})}{2} + \frac{p_{trt,1}(1-p_{trt,1})}{2}}.$$

Solving now the sample size formula (formula (2.1)) for RR_0 and assuming that the alternative hypothesis is true ($p_{trt,1} < RR_0 \cdot p_{ctl,1}$) results in

$$RR_0 = \frac{1}{p_{ctl,1}}\Big[p_{trt,1} + \frac{2}{\sqrt{N}}(z_{1-\alpha}\varphi_{ctl}(p_{ctl,1}, p_{trt,1}) + z_{1-\beta}\varphi_{trt}(p_{ctl,1}, p_{trt,1}))\Big]. \tag{2.2}$$

For one study assuming a power of $1-\beta_1$ and the null hypothesis boundary $RR_0 = 1$ (as mentioned above), the sample size formula (formula (2.1)) reduces to

$$N_1 = 4 \cdot \frac{[z_{1-\alpha}\varphi_{ctl}(p_{ctl,1}, p_{trt,1}) + z_{1-\beta}\varphi_{trt}(p_{ctl,1}, p_{trt,1})]^2}{[p_{trt,1} - p_{ctl,1}]}.$$

Since multiple studies can be combined for the assessment of new treatments, the overall sample size can increase by the factor $c > 1$, $N = cN_1$, and the overall power changes to $1-\beta_2$, which results in changes to the hypothesis boundary of formula (2.2) (still assuming that the alternative hypothesis is true ($p_{trt,1} < RR_0 \cdot p_{ctl,1}$)):

$$RR_0 = \frac{1}{p_{ctl,1}} \left[p_{trt,1} + \frac{2(z_{1-\alpha}\varphi_{ctl}(p_{ctl,1}, p_{trt,1}) + z_{1-\beta_2}\varphi_{trt}(p_{ctl,1}, p_{trt,1}))}{\sqrt{c \cdot 4\frac{[z_{1-\alpha}\varphi_{ctl}(p_{ctl,1},p_{trt,1})+z_{1-\beta_1}\varphi_{trt}(p_{ctl,1},p_{trt,1})]^2}{[p_{trt,1}-p_{ctl,1}]^2}}} \right]$$
$$= \frac{1}{p_{ctl,1}} \left[p_{trt,1} - \frac{p_{trt,1} - p_{ctl,1}}{\sqrt{c}} \cdot \frac{z_{1-\alpha}\varphi_{ctl}(p_{ctl,1}, p_{trt,1}) + z_{1-\beta_2}\varphi_{trt}(p_{ctl,1}, p_{trt,1})}{z_{1-\alpha}\varphi_{ctl}(p_{ctl,1}, p_{trt,1}) + z_{1-\beta_1}\varphi_{trt}(p_{ctl,1}, p_{trt,1})} \right].$$

In the special case when the power of the combined studies is equal ($\beta = \beta_1 = \beta_2$), the null hypothesis boundary is independent of the choice of β and reduces to

$$RR_0 = \frac{1}{p_{ctl,1}} \left[p_{trt,1} - \frac{p_{trt,1} - p_{ctl,1}}{\sqrt{c}} \right] = \left(1 - \frac{1}{\sqrt{c}}\right) RR + \frac{1}{\sqrt{c}}.$$

This hypothesis boundary, RR_0, can then be used as threshold CI_S to compare the upper limit of the 95% CI for the relative risk for the ranking of the new treatment. Assuming further a sample size twice as large due to two studies ($c = 2$), the thresholds only depend on the assumed true effect RR_1:

$$CI_S = RR_1 \left(1 - \frac{1}{\sqrt{2}}\right) + \frac{1}{\sqrt{2}}. \tag{2.3}$$

IQWiG further defines an RR_1 of 0.50 and 0.83 as an effect of major and considerable added benefit, respectively (Skipka et al., 2016). Hence, plugging these values into formula (2.3) results into the rounded thresholds of 0.85 and 0.95 shown in Table 3.

The comparison of the upper limit of the CI to specific thresholds is only a hypothesis shift where a value different from 1 is chosen for RR_0, which results in reduced power. However, since data of multiple studies can be combined and thus pooled 95% CI limits can be obtained, the power can be increased and thus the power reduction can be compensated. Although the derivation of this threshold is based on two studies, IQWiG still uses the same thresholds even if different amounts of studies are combined for a pooled 95% CI.

Table 3: *Categories of IQWiG$_{RR}$ and Mod-IQWiG$_{HR}$ for overall survival*

Categories	IQWiG$_{\mathrm{RR}}$	Mod-IQWiG$_{\mathrm{HR}}$
less benefit	Any statistically significant increase in survival time is at least classified as "minor added benefit", since for all-cause mortality the ANV's requirement that an effect should be "more than marginal" is regarded to be fulfilled by the outcome itself (Skipka et al. (2016))	
no added benefit		
nonquantifiable added benefit		
minor added benefit	$\mathrm{HR}^{+} \in [0.95, 1)$	$\mathrm{HR}^{+} \in [0.93, 1)$
considerable added benefit	$\mathrm{HR}^{+} \in [0.85, 0.95)$	$\mathrm{HR}^{+} \in [0.79, 0.93)$
major added benefit	$\mathrm{HR}^{+} < 0.85$	$\mathrm{HR}^{+} < 0.79$

Abbreviations: ANV: Arzneimittel-Nutzenbewertungsverordnung (engl. Regulation for Early Benefit Assessment), CI: Confidence interval, HR: Hazard ratio, HR^{+}: Upper 95% confidence interval limit of the HR-PE, PE: Point estimate

2.2.1.1 Mod-IQWiG$_{\mathrm{HR}}$: Modifying thresholds of IQWiG

As mentioned above, IQWiG$_{\mathrm{RR}}$ uses RR based thresholds for the comparison to upper 95% HR-CI limit for the additional benefit assessment of new treatments in case of time-to-event endpoints. To investigate the influence of these wrongly-scaled thresholds, the RR based thresholds were transformed to HR based ones using the conversion formula proposed by VanderWeele (2020):

$$RR = \frac{1 - 0.5^{\sqrt{\mathrm{HR}}}}{1 - 0.5^{\sqrt{\frac{1}{\mathrm{HR}}}}}.$$

This formula, however, is not analytical solvable. Thus, a numerical approach (optimization) using minimization without derivatives introduced by Brent and Brent (1974) was used to calculate the HR based thresholds, which is implemented in the uniroot function of the stats package in R. The solutions for the transformed thresholds are 0.79 and 0.93. This version of IQWiG$_{\mathrm{RR}}$ using HR based thresholds instead of RR based ones was defined as Mod-IQWiG$_{\mathrm{HR}}$. Table 3 gives an an overview of IQWiG$_{\mathrm{RR}}$ and Mod-IQWiG$_{\mathrm{HR}}$.

2.2.2 European Society for Medical Oncology (ESMO)

The main aim of the society was to develop a clear and unbiased statement representing the magnitude of the clinical benefit from new treatments regarding oncology treatments. Thus, in 2015 Cherny et al. (2015) published their first version of a validated and reproducible tool called the ESMO Magnitude of Clinical Benefit Scale. After the first implementation, a revised version was published with modifications of the shortcomings of its first version (Cherny et al., 2017). ESMO defined a dual rule consisting out of a relative and absolute rule. The relative and absolute benefit is assessed using the lower limit of the 95% HR-CI (HR^{-}) and the observed gain, defined as difference between observed median survival in the treatment (med_{trt}) and control group (med_{ctl}), respectively. These estimates are compared to specific thresholds leading to a preliminary ordinal rating for the classification consisting of categories 1 to 4. In addition, category 4 can already be achieved without fulfilling the relative and absolute rule, if the survival rate increased by $\geqslant 10\%$ at key milestones. Table 4 gives an overview of specific thresholds and the key milestones stratified by different observed med_{ctl} (Cherny et al., 2015, 2017).

If only the relative benefit rule is applied, the method compares only HR^{-} against the thresholds 0.7 and 0.65 in case of $med_{ctl} \leqslant 12$ months and $med_{ctl} > 12$ months, respectively (see Table 4). Thus, only the categories "1" or "greater than 1" can be achieved using only the relative benefit rule instead of the dual rule.

Table 4: *ESMO categories: Non-curative setting for overall survival as efficacy endpoint (form 2a)*

ESMO categories for overall survival			
	$\text{med}_{\text{ctl}} \leqslant 12$ mon.	$\text{med}_{\text{ctl}} \in (12, 24]$ mon.	$\text{med}_{\text{ctl}} > 24$ mon.
1	$\text{HR}^- > 0.7$ <u>OR</u> gain < 1.5 mon.	$\text{HR}^- > 0.75$ <u>OR</u> gain < 1.5 mon.	$\text{HR}^- > 0.75$ <u>OR</u> gain < 4 mon.
2	$\left[\begin{array}{c}\text{HR}^- \leqslant 0.65 \text{ } \underline{\text{AND}} \\ \text{gain} \in [1.5, 2) \text{ mon.}\end{array}\right]$ <u>OR</u> $\left[\begin{array}{c}\text{HR}^- \in (0.65, 0.7] \text{ } \underline{\text{AND}} \\ \text{gain} \geqslant 1.5 \text{ mon.}\end{array}\right]$	$\left[\begin{array}{c}\text{HR}^- \leqslant 0.7 \text{ } \underline{\text{AND}} \\ \text{gain} \in [1.5, 3) \text{ mon.}\end{array}\right]$ <u>OR</u> $\left[\begin{array}{c}\text{HR}^- \in (0.7, 0.75] \text{ } \underline{\text{AND}} \\ \text{gain} \geqslant 1.5 \text{ mon.}\end{array}\right]$	$\left[\begin{array}{c}\text{HR}^- \leqslant 0.7 \text{ } \underline{\text{AND}} \\ \text{gain} \in [4, 6) \text{ mon.}\end{array}\right]$ <u>OR</u> $\left[\begin{array}{c}\text{HR}^- \in (0.7, 0.75] \text{ } \underline{\text{AND}} \\ \text{gain} \geqslant 4 \text{ mon.}\end{array}\right]$
3	$\text{HR}^- \leqslant 0.65$ <u>AND</u> gain $\in [2, 3)$ mon.	$\text{HR}^- \leqslant 0.7$ <u>AND</u> gain $\in [3, 5)$ mon.	$\text{HR}^- \leqslant 0.7$ <u>AND</u> gain $\in [6, 9)$ mon.
4	$\left[\begin{array}{c}\text{HR}^- \leqslant 0.65 \text{ } \underline{\text{AND}} \\ \text{gain} \geqslant 3 \text{ mon.}\end{array}\right]$ <u>OR</u> $\left[\begin{array}{c}\text{Increase in 2 year} \\ \text{survival} \geqslant 10\%\end{array}\right]$	$\left[\begin{array}{c}\text{HR}^- \leqslant 0.7 \text{ } \underline{\text{AND}} \\ \text{gain} \geqslant 5 \text{ mon.}\end{array}\right]$ <u>OR</u> $\left[\begin{array}{c}\text{Increase in 3 year} \\ \text{survival} \geqslant 10\%\end{array}\right]$	$\left[\begin{array}{c}\text{HR}^- \leqslant 0.7 \text{ } \underline{\text{AND}} \\ \text{gain} \geqslant 9 \text{ mon.}\end{array}\right]$ <u>OR</u> $\left[\begin{array}{c}\text{Increase in 5 year} \\ \text{survival} \geqslant 10\%\end{array}\right]$
5	Only achievable with toxicity, QoL or other bonus point adjustments		

<u>Notes:</u> Category 1-3 is defined as low benefit and 5-4 as substantial improvement. <u>Abbreviations:</u> CI: Confidence interval, gain: Absolute difference in median survival times, HR: Hazard ratio, HR^-: Lower 95% confidence interval limit of the HR-PE, med_{ctl}: Median survival time in the control group, mon.: Month(s), PE: Point estimate

2.2.3 American Society of Clinical Oncology (ASCO)

Schnipper et al. (2015, 2016) developed the American Society of Clinical Oncology Value Framework, which is defined as the sum of a clinical benefit score (CBS) and bonus points (BP) to calculate the continuous Net Health Benefit (NHB) score. The NHB reflects the clinical point of view of the treatment. Furthermore, information about the costs of the treatment is also given besides the NHB score allowing the consideration of the financial impact of the treatment for the patient. The main component CBS is defined as

$$\text{CBS} = 100 \cdot (1 - \text{HR-PE}),$$

where HR-PE is the HR point estimate of the significant phase III trial (or a combined effect of multiple studies). The BP consists out of 4 components. This thesis focus is on the statistical elements of the methods, hence only the statistical BP component "Tail of the survival curve" is described: The time point on the survival curve that is two times the median OS of the control group needs to be identified. If the proportion of patients alive in the treatment compared to the control group improved by 50% or more (assuming $> 20\%$ surviving in control group), 20 points are rewarded. If less than 20% survived in the control group, no points are rewarded.

Table 5: *Composition of ASCO for overall survival*

ASCO	
clinical benefit score (CBS)	$\text{CBS} = 100 \cdot (1 - \text{HR-PE})$
bonus points (BP)	Tail of the survival curve: The time point on the survival curve that is $2 \cdot \text{med}_{\text{ctl}}$, is identified. If the proportion of patients alive in the treatment compared to the control arm improved by 50% or greater (assuming $> 20\%$ surviving in control arm), 20 points are rewarded.
net health benefit score (NHB)	NHB = CBS+BP

Abbreviations: HR: Hazard ratio, med_{ctl}: Median survival time in the control group, PE: Point estimate

2.3 Simulation studies

In the following two simulation studies are described. Firstly, in Section 2.3.1 a simulation study using an approach where the censoring times are dependent on the event times (Simulation 1) and hence might introduce bias into the HR and HR-CI estimation is outlined. If this suspicion turns out to be true, it affects all additional benefit assessment method equally and thus should not bias the comparison in a major way. Nevertheless, to rule out any uncertainty in Section 2.3.2 another simulation study is performed using an unbiased approach for the censoring mechanism (Simulation 2). Both simulation studies are described in detail below using the ADEMP structure proposed by Morris et al. (2019) for planning simulation studies with the goal of improving the design, analysis, and report of simulations. The abbreviation "ADEMP" stands for Aims, Data-generating mechanisms, Estimands, Methods, and Performance measures. The performed simulation studies are reproducible with the provided R-Code at `https://www.github.com/cbuesch/SumulationStudyABAM`. Additional information on R-Code structure and execution of the programs to determine the results of this thesis can be found in Appendix B.

2.3.1 Simulation 1 (censoring times dependent on event times)

Parts of this Section 2.3.1 are already published in the article *A Comprehensive Comparison of Additional Benefit Assessment Methods Applied by Institute for Quality and Efficiency in Health Care and European Society for Medical Oncology for Time-to-Event Endpoints After Significant Phase III Trials - a Simulation Study* by Büsch et al. (2022). The manuscript has been written by the lead author but may contain comments and corrections from the co-authors and the reviewers.

2.3.1.1 Aim

The aim of this simulation study is the comparison of the statistical aspects of the additional benefit assessment methods ASCO, IQWiG_{RR}, Mod-IQWiG_{HR}, and ESMO in non-curative setting with overall survival as efficacy endpoint.

The three used quantities on which the methods are based (lower 95% HR-CI limit, HR-

PE, and upper 95% HR-CI limit) are assessed using sensitivity and specificity. Hence, the question which statistical approach might be better for the assessment of additional benefit will be examined. Further goals are the assessment of the relationship degree between all methods as well as to determine which ASCO cutoff values correspond to the categories of ESMO, IQWiG_{RR}, and Mod-IQWiG_{HR}. This enables an estimate of how the other methods would assess a new treatment without the need to apply them.

2.3.1.2 Data-generating mechanisms

The methods are applied after a statistically significant phase III trial based on the log-rank test. To perform phase III trials the failure time, censoring time, and sample size calculation needed to be determined. Subsequently, the precise definition of used data generation including required parameters and choice of distribution is outlined:
Let T be the event time, C the censoring time, A the accrual time, and FU the follow-up time. Moreover, dur denotes the duration of the study and is defined as $dur := A + FU$. The density, distribution, and survival functions are denoted with f, F, and S, respectively. A distinguishment between the true treatment effect (trueHR), which is used for the data generation, and the design treatment effect (designHR), which is assumed for sample size calculation, is made. HR_{var} is defined to measure the deviance between designHR and trueHR: trueHR=designHR$\cdot\text{HR}_{\text{var}}$.

- Data-generating algorithm: Each simulated randomized clinical phase III trial with time-to-event outcome is defined to compare a treatment against a control group with an allocation ration r and sample size of n_{trt} and n_{ctl} for treatment and control group, respectively. The alogorithm consist out of two steps:

  1. Set a seed to create reproducible results.

  2. Generate independent failure times T with the failure times f_{trt} and f_{ctl} for the treatment and control group, n_{trt} and n_{ctl} times, respectively. In addition, for each patient generate independent right-censoring time C and independent administrative censoring time A, which takes the accrual time into account. Then select the minimum out of C, A, and T, which represents the final observed time-to-event data of the trial. Therefore, the final data tuple is of the form

$(\min(T, C, A), \mathbb{1}(T \leqslant \min(C, A)))$, where the entries represent the event time and cause of event (failure or censoring), respectively.

- Scenarios / specification of parameters: An extensive simulation study was performed by generating different scenarios of phase III trials making a comprehensive comparison between all additional benefit assessment methods (IQWiG_{RR}, $\text{Mod-IQWiG}_{\text{HR}}$, ESMO, and ASCO) possible. Furthermore, the influence of incorrect assumed designHR (over- / underpowered trials), various failure time distribution, non-proportional hazards (late treatment effects), various allocation ratios, various censoring mechanism, and informative censoring on the different additional benefit assessment methods can be examined. Each scenario consists out of multiple parameter combinations, i.e. sub-scenarios. In the following an overview of all scenarios and sub-scenarios is given and in Table 6 the differences between the performed scenarios are highlighted in bold.

  1. *Standard Scenario (Scenario 1)*: Failure time distribution following an exponential distribution and a combination of administrative and exponential censoring for the generation of censoring times. The following parameter combinations with assumptions are used, which leads to $(5 \cdot 31 \cdot 2 \cdot 3 =)$ 930 sub-scenarios:

     - $\text{med}_{ctl} \in \{6, 12, 18, 24, 30\}$
     - $\text{designHR} \in \{0.3, 0.32, 0.34, ..., 0.86, 0.88, 0.9\}$
     - $\text{HR}_{\text{var}} = 1$
     - type-II-error rate $\beta \in \{0.1, 0.2\}$, hence power of 90% and 80%
     - type-I-error rate $\alpha = 0.05$ (two-sided)
     - allocation ratio $r = 1$
     - combination out of administrative censoring (accrual time of 2 years and a follow-up time of $2 \cdot \text{med}_{ctl}$) and exponential censoring so that an overall censoring rate of $p_C \in \{20\%, 40\%, 60\%\}$ equal in both treatment groups was achieved.

Table 6: *Overview of simulation scenarios of simulation study with censoring times dependent on event times*

Scenario	Parameters differences between the scenarios			
	HR_{var}	Failure time distribution	r	Censoring distribution
Standard Scenario	1	exponential	1	$p_C \in \{20\%, 40\%, 60\%\}$ equal in both treatment groups with administrative (accrual: 2 years, follow-up: $2 \cdot med_{ctl}$) and exponential censoring
Scenario 2 (**HR_{var}**)	**Overpowered trials:{0.8,0.9} Underpowered trials:{1.1,1.2}**	exponential	1	$p_C \in \{20\%, 40\%, 60\%\}$ equal in both treatment groups with administrative (accrual: 2 years, follow-up: $2 \cdot med_{ctl}$) and exponential censoring
Scenario 3 (**failure time**)	1	**Weibull and Gompertz**	1	$p_C \in \{20\%, 40\%, 60\%\}$ equal in both treatment groups with administrative (accrual: 2 years, follow-up: $2 \cdot med_{ctl}$) and exponential censoring
Scenario 4 (**non-prop. hazards**)	1	exponential **with delayed treatment effect** ($start_{\mathbf{trt}} = \frac{1}{3} \cdot med_{\mathbf{ctl}}$)	1	$p_C \in \{20\%, 40\%, 60\%\}$ equal in both treatment groups with administrative (accrual: 2 years, follow-up: $2 \cdot med_{ctl}$) and exponential censoring
Scenario 5 (**r**)	1	exponential	$\{\frac{1}{2}, \frac{2}{1}\}$	$p_C \in \{20\%, 40\%, 60\%\}$ equal in both treatment groups with administrative (accrual: 2 years, follow-up: $2 \cdot med_{ctl}$) and exponential censoring
Scenario 6 (**censoring**)	1	exponential	1	$p_C \in \{20\%, 40\%, 60\%\}$ equal in both treatment groups **with only exponential censoring**
Scenario 7 (**informative censoring**)	1	exponential	1	$p_C \in \{20\%, 40\%, 60\%\}$ **unequal** in both treatment groups **($p_C^{ctl} = 0.2$ & $p_C^{trt} = 0.4$ or $p_C^{ctl} = 0.4$ & $p_C^{trt} = 0.2$)** with administrative (accrual: 2 years, follow-up: $2 \cdot med_{ctl}$) and exponential censoring

Notes: Differences to the standard scenario regarding the parameter choice were highlighted in bold. The following parameters were chosen to be identical in each scenario and hence are not shown in the table: type-I-error rate of 0.05, power of 90% and 80%, $med_{ctl} \in \{6, 12, 18, 24, 30\}$ and designHR $\in \{0.3, 0.32, ..., 0.88, 0.9\}$. Abbreviations: designHR: Design hazard ratio used for sample size calculation, HR: Hazard ratio, HR_{var}: Factor for deviance between designHR and trueHR, p_C: Overall censoring rate, p_C^{ctl} and p_C^{trt}: Censoring rate of control and treatment group, med_{ctl}: Median survival time of control group, r: Allocation ratio, $start_{trt}$: Treatment starting time point, trueHR: True underlying HR of data generation

2. *Influence of incorrectly assumed designHR for sample size calculation (Scenario 2):* Incorrect assumed treatment effects lead to over- and underpowered trials. Hence, for the simulation a designHR unequal to the trueHR was choosen: $HR_{var} \in \{0.8, 0.9, 1.1, 1.2\}$. Hence, sub-scenarios with $HR_{var}<1$ are overpowered and sub-scenarios with $HR_{var}>1$ are underpowered trials. Otherwise, the same parameters as in the Standard Scenario were used. This leads to $(5 \cdot 31 \cdot 4 \cdot 2 \cdot 3 =)$ 3,720 sub-scenarios.

3. *Influence of different underlying failure time distributions (Scenario 3):* Two different failure time distributions (Weibull and Gompertz) were used instead of the

exponential distribution in the Standard Scenario. Otherwise, the same parameters as in the Standard Scenario were used.

In general, non-proportional hazards have to be assumed as the hazards of Weibull and Gompertz distribution are time-dependent. Since the Cox regression, which is performed to apply the additional benefit assessment methods, assumes proportional hazards (constant hazard ratio over time between the two treatment groups), the shape parameter of the Weibull and Gompertz distribution was fixed to specific values in both treatment groups causing again proportional hazards (see Section 2.1.1 for more detail). An example using a designHR of 0.9 (designHR=trueHR) and med_{ctl} of 6 months can be found in Figure 2.

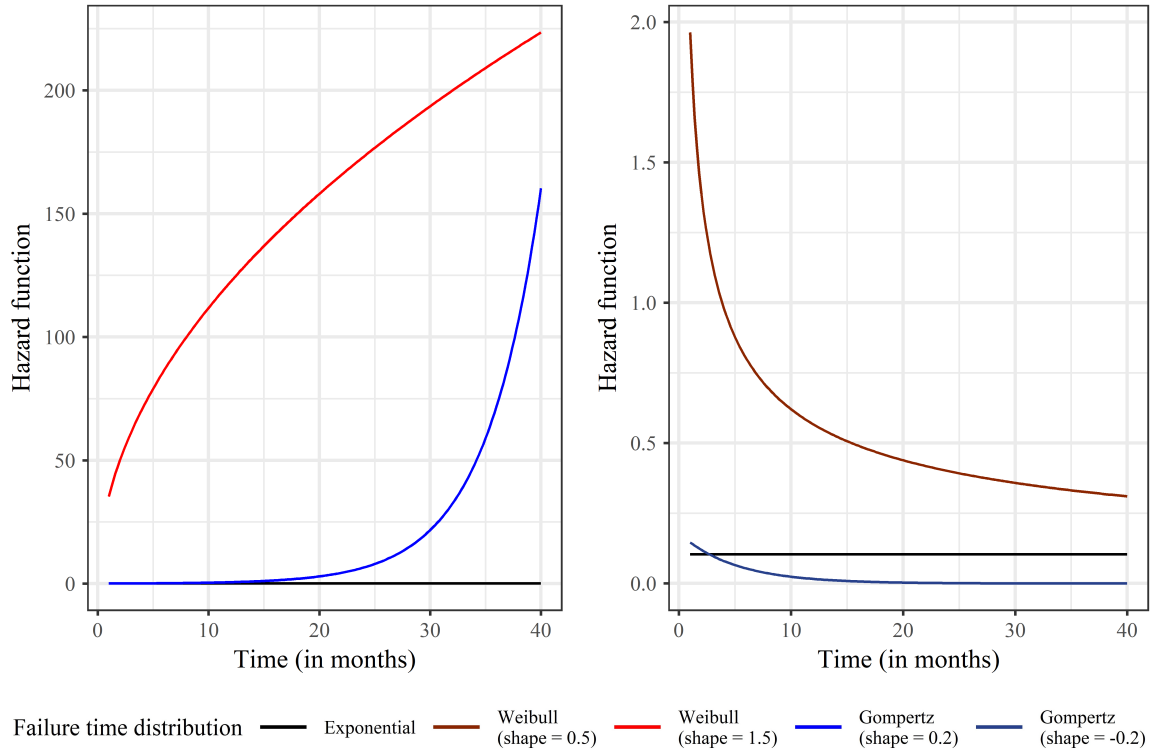

Figure 2: *Hazard functions of exponential, Weibull, and Gompertz distribution with assumed parameters for a designHR = 0.9, designHR=trueHR, and* $med_{ctl} = 6$ *months.*
Abbrevations: designHR: Design hazard ratio used for sample size calculation, HR: Hazard ratio, med_{ctl}*: Median survival time of control group, trueHR: True underlying HR of data generation*

- Weibull distribution: $k \in \{0.5, 1.5\}$

  These two parameter values for k were chosen to include sub-scenarios where the failure rate decreases ($k < 1$, failures occur earlier in time) and increases over time ($k > 1$, failures occur later in time). This leads to $(5 \cdot 31 \cdot 2 \cdot \cdot 3 \cdot 2 =)$ 1,860 sub-scenarios.

- Gompertz distribution: $a \in \{-0.2, 0.2\}$

  These two parameter values for a were chosen to include sub-scenarios where the failure rate decreases ($a < 0$, failures occur earlier in time) and increases over time ($a > 0$, failures occur later in time). This leads to $(5 \cdot 31 \cdot 2 \cdot \cdot 3 \cdot 2 =)$ 1,860 sub-scenarios.

4. *Influence of non-proportional hazards using late treatment effects for the treatment group (Scenario 4):* For this objective, the same parameters as in the Standard Scenario were used. This leads to $(5 \cdot 31 \cdot 2 \cdot 3 =)$ 930 sub-scenarios. The underlying failure time distribution of both treatment groups, however, was chosen to be exponential with a delayed treatment effect for the treatment group using a piece-wise exponential failure time distribution. Hence, the HR is not constant over time leading to one kind of non-proportional hazards. The underlying distribution of the treatment group, $F_{trt}(x)$, was chosen to be exponential using the distribution parameter λ_{ctl} of the control group until the delayed treatment start at time point start_{trt} and λ_{trt} after start_{trt}:

$$F_{ctl}(x) = 1 - \exp(-\lambda_{ctl} \cdot x),$$

$$F_{trt}(x) = \begin{cases} 1 - \exp(-\lambda_{ctl} \cdot x) & , x \in [0, \text{start}_{trt}] \\ 1 - \exp(-\lambda_{ctl} \cdot \text{start}_{trt}) \cdot \exp(-\lambda_{trt} \cdot (x - \text{start}_{trt})) & , \text{otherwise,} \end{cases}$$

   where F_{ctl} and F_{trt} are the cumulative distribution functions of the treatment and control group, $\lambda_{ctl} > 0$ and $\lambda_{trt} > 0$ are the parameters of the corresponding exponential distributions, and start_{trt} $(= \frac{1}{3} \cdot \text{med}_{ctl})$ is the time point of treatment effect start for the treatment group.

   ESMO uses the gain of the new treatment (absolute difference of median treatment outcomes) to establish the different categories. Hence, if $\text{med}_{trt} \approx \text{med}_{ctl}$, the

method would only assign the lowest category to a new treatment. To not penalize ESMO for its design, start_{trt} was set to $\frac{1}{3}$ of the assumed median survival time of the control group (med_{ctl}) for the simulations ($\text{start}_{trt} \ll \text{med}_{ctl}$). An example using a designHR of 0.7 (designHR=trueHR), med_{ctl} of 12 months, and start_{trt} of 4 months can be found in Figure 3.

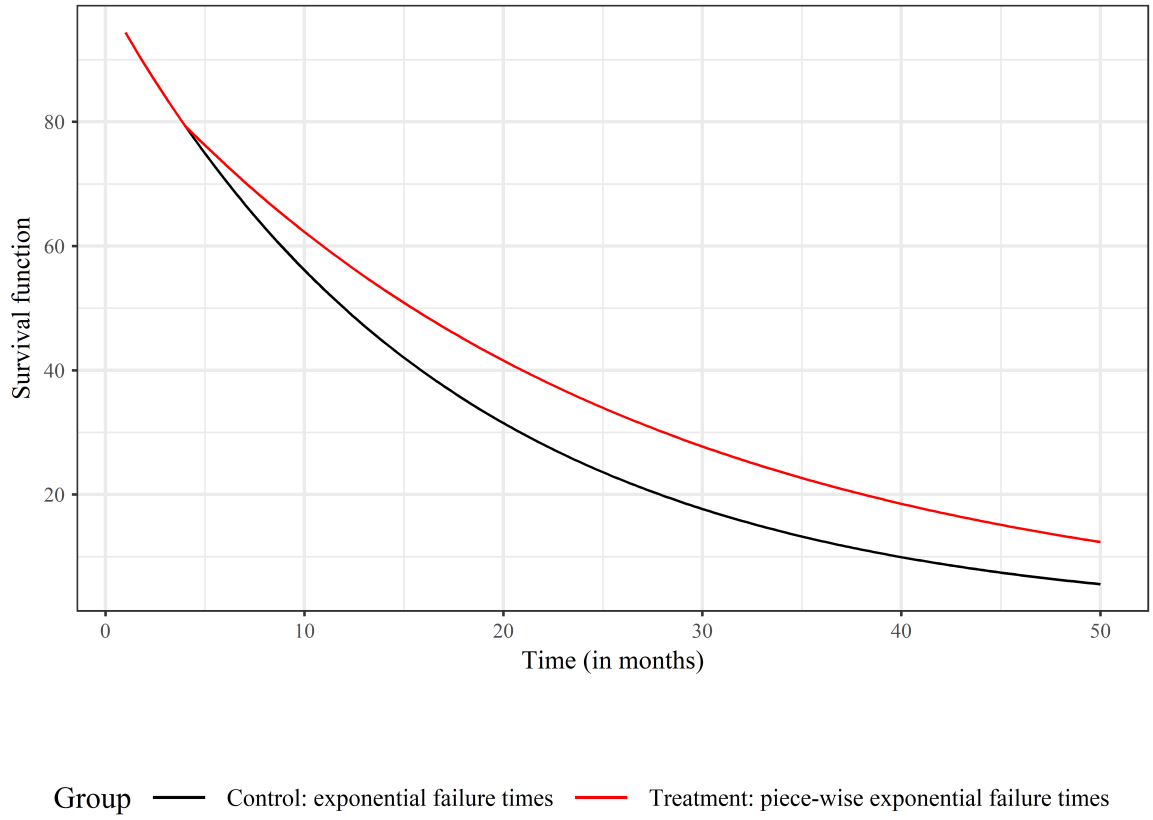

Figure 3: *Survival functions of piece-wise exponential distribution with late treatment effect, assuming a designHR of 0.7, designHR=trueHR, med_{ctl} = 12 months, and med_{trt} = 4 months ($start_{trt} = 1/3 \cdot med_{ctl}$).*
Abbrevations: designHR: Design hazard ratio used for sample size calculation, HR: Hazard ratio, med_{ctl}: Median survival time of control group, med_{trt}: Median survival time of treatment group, trueHR: True underlying HR of data generation

5. *Influence of unequal sample sizes (Scenario 5)*: To simulate unequal sample sizes between the treatment groups, an allocation ratio unequal 1 was used: $r \in \{\frac{1}{2}, \frac{2}{1}\}$. Otherwise, the same parameters as in the Standard Scenario were used. This leads to $(5 \cdot 31 \cdot 2 \cdot 2 \cdot 3 =)$ 1,860 sub-scenarios.

6. *Influence of using only exponential distributed censoring distribution without administrative censoring (Scenario 6)*: The same parameters as in the Standard Scenario were used. This leads to $(5 \cdot 31 \cdot 2 \cdot 3 =)$ 930 sub-scenarios. Only the generated censoring time is defined as exponential censoring distribution without administrative censoring ($C \sim \exp(\lambda_C)$) so that an overall censoring rate of $p_C \in \{20\%, 40\%, 60\%\}$ equal in both treatment groups was achieved.

7. *Influence of informative censoring due to treatment (Scenario 7):* To simulate informative censoring due to the new treatment, unequal censoring rates p_C^{ctl} and p_C^{trt} were defined in the treatment and control group, respectively:

   - $p_C^{ctl} = 0.2$ & $p_C^{trt} = 0.4$
   - $p_C^{ctl} = 0.4$ & $p_C^{trt} = 0.2$

   Due to the fact that one does not assume informative censoring at the planning stage of a trial, the formula of the probability of an event for the sample size calculation is computed with the censoring rate of the control group:

   $$\mathbb{P}(D) = 1 - p_C^{ctl}.$$

   More information regarding the performed sample size calculation follows. in this section. In addition, a combination of both censoring methods (administrative censoring and exponential censoring) was used for this scenario using a fixed accrual time (a) of 2 years and a follow-up time (FU) of $2 \cdot \text{med}_{\text{ctl}}$ (as in the Standard Scenario). Otherwise, the same parameters as in the Standard Scenario were used. This leads to $(5 \cdot 31 \cdot 2 \cdot 2 =)$ 620 sub-scenarios.

- Failure time distribution of control and treatment group (f_{ctl} and f_{trt}):
  - Assuming failure times follow an exponential distribution, the median overall survival time of the control group (med_{ctl}), designHR, and trueHR ($=\text{HR} \cdot \text{HR}_{\text{var}}$) needed to be fixed to a specific value for calculation of the required parameters:

    $$f_C \sim \exp(\lambda_{ctl}), \qquad f_{trt} \sim \exp(\lambda_{trt}),$$

where λ_{ctl} was calculated using the assumed med_{ctl}:

$$\text{med}_{ctl} = \frac{\ln(2)}{\lambda_{ctl}} \Rightarrow \lambda_{ctl} = \frac{\ln(2)}{\text{med}_{ctl}} \tag{2.4}$$

and λ_{trt} was calculated using the trueHR, the proportional hazards assumption (ratio of the hazards is constant over time) as well as the conversion of λ_{ctl} (formula (2.4)):

$$\text{trueHR} = \frac{h_{trt}(t)}{h_{ctl}(t)} = \frac{\lambda_{trt}}{\lambda_{ctl}} \overset{(2.4)}{\Rightarrow} \lambda_{trt} = \text{trueHR} \cdot \frac{\ln(2)}{\text{med}_{ctl}}.$$

– Assuming failure times follow a Weibull distribution, med_{ctl}, designHR, and trueHR needed to be fixed to a specific value to calculate the required parameters:

$$f_{ctl} \sim \text{Weibull}(\lambda_{ctl}, k_{ctl}),$$

$$f_{trt} \sim \text{Weibull}(\lambda_{trt}, k_{trt}),$$

where λ_{ctl} was calculated using the assumed med_{ctl}:

$$\text{med}_{ctl} = \frac{(\ln(2))^{1/k_{ctl}}}{\lambda_{ctl}} \Rightarrow \lambda_{ctl} = \frac{(\ln(2))^{1/k_{ctl}}}{\text{med}_{ctl}} \tag{2.5}$$

and λ_{trt} was calculated using the trueHR, the proportional hazards assumption (ratio of the hazards is constant over time) as well as the conversion of λ_{ctl} (formula (2.5)):

$$\begin{aligned} \text{trueHR} = \frac{h_{trt}(t)}{h_{ctl}(t)} = \frac{\lambda_{trt}^{k_{trt}} \cdot k_{trt} \cdot t^{k_{trt}-1}}{\lambda_{ctl}^{k_{ctl}} \cdot k_{ctl} \cdot t^{k_{ctl}-1}} &\overset{(*)}{=} \frac{\lambda_{trt}^{k}}{\lambda_{ctl}^{k}} \\ \overset{(2.5)}{\Rightarrow} \lambda_{trt} = \left(\frac{\text{trueHR} \cdot \ln(2)}{\text{med}_{ctl}^{k}} \right)^{\frac{1}{k}}&, \end{aligned}$$

where at $(*)$ the shape parameter k was chosen to be identical for both treatment groups to achieve a constant hazard ratio over time ($k_{ctl} = k_{trt}$).

– Assuming failure times follow a Gompertz distribution, med_{ctl}, designHR, and trueHR needed to be fixed to a specific value to calculate the required parameters:

$$f_{ctl} \sim \text{gompertz}(a_{ctl}, b_{ctl}),$$

$$f_{trt} \sim \text{gompertz}(a_{trt}, b_{trt}),$$

where b_{ctl} was calculated using med_{ctl}:

$$\text{med}_{ctl} = \frac{1}{a_{ctl}} \cdot \ln\left(1 + \frac{a_{ctl}}{b_{ctl}} \cdot \ln(2)\right) \Rightarrow b_{ctl} = \frac{a_{ctl} \cdot \ln(2)}{\exp(\text{med}_{ctl} \cdot a_{ctl}) - 1} \tag{2.6}$$

and b_{trt} was calculated using the trueHR, the proportional hazards assumption (ratio of the hazards is constant over time) as well as the conversion of b_{ctl} (formula (2.6):

$$\begin{aligned} \text{trueHR} = \frac{h_{trt}(t)}{h_{ctl}(t)} = \frac{b_{trt} \cdot \exp(a_{trt} \cdot x)}{b_{ctl} \cdot \exp(a_{ctl} \cdot x)} &\overset{(*)}{=} \frac{b_{trt}}{b_{ctl}} \\ \overset{(2.6)}{\Rightarrow} b_{trt} = \frac{\text{trueHR} \cdot a \cdot \ln(2)}{\exp(\text{med}_{ctl} \cdot a) - 1}&, \end{aligned}$$

where at $(*)$ the shape parameter a was chosen to be same for both treatment groups to achieve a constant hazard ratio over time ($a_{ctl} = a_{trt}$).

– Assuming failure times follow a piece-wise exponential distribution with an additional late treatment effect for the treatment group, med_{ctl}, designHR, and trueHR needs to be fixed to a specific value. To achieve a late treatment effect for the treatment group, a piece-wise exponential distribution was chosen in the following way:

$$F_{ctl}(x) = 1 - \exp(-\lambda_{ctl} \cdot x),$$

$$F_{trt}(x) = \begin{cases} 1 - \exp(-\lambda_{ctl} \cdot x) & , x \in [0, \text{start}_{trt}] \\ 1 - \exp(-\lambda_{ctl} \cdot \text{start}_{trt}) \cdot \exp(-\lambda_{trt} \cdot (x - \text{start}_{trt})) & , \text{otherwise,} \end{cases}$$

where $F_{ctl}(x)$ and $F_{trt}(x)$ are the cumulative distribution functions of the treatment and control group, $\lambda_{ctl} > 0$ and $\lambda_{trt} > 0$ are the parameters of the corresponding

exponential distributions, and start_{trt} $(= \frac{1}{3} \cdot \text{med}_{ctl})$ is the time-point where the treatment effect sets in. The failure times of the treatment groups were generated using the inversion method by Kolonko (2008, Chapter 8, pg. 85-95): Assuming a uniform random variable U on the interval [0,1], $X := F_{trt}^{-1}(U)$ is F_{trt} is distributed, meaning $\mathbb{P}(X \leqslant t) = F_{trt}, t \in \mathbb{R}$. Therefore, the inversion of the cumulative distribution $F_{trt}(x)$ is given by

$$F_{trt}^{-1}(y) = \begin{cases} \frac{\ln(1-y)}{-\lambda_{ctl}} & , y \in [0, 1 - \exp(-\lambda_{ctl} \cdot \text{start}_{trt})] \\ \frac{\ln(1-y)+\lambda_{ctl} \cdot \text{start}_{trt}}{-\lambda_{trt}} + \text{start}_{trt} & , \text{otherwise.} \end{cases}$$

Additionally, λ_{ctl} and λ_{trt} were defined in the same way as in the exponential case. Hence,

$$\begin{aligned} \text{med}_{ctl} &\overset{!}{=} \frac{\ln(2)}{\lambda_{ctl}} \Rightarrow \lambda_{ctl} = \frac{\ln(2)}{\text{med}_{ctl}}, \\ \text{HR} \cdot \text{HR}_{\text{var}} &\overset{!}{=} \frac{h_{trt}(t)}{h_{ctl}(t)} = \frac{\lambda_{trt}}{\lambda_{ctl}} \overset{(2.4)}{\Rightarrow} \lambda_{trt} = (\text{HR} \cdot \text{HR}_{\text{var}}) \cdot \frac{\ln(2)}{\text{med}_{ctl}}. \end{aligned}$$

- Censoring time distribution: Depending on the simulation scenario three different types of censoring time generation were implemented. In the following these three types are described:

  1. Assuming only *administrative censoring*, the censoring time is set to be uniformly distributed:

     $$\text{A} \sim \mathcal{U}(a) + \text{FU},$$

     where A is the administrative censoring from tuple mentioned in the data generaion algorithm. In addition, a represents the accrual time and FU the follow-up time.

  2. Assuming a *specific given censoring rate*, p_C, the simulated censoring time is set to be *exponentially distributed*:

     $$\text{C} \sim \exp(\lambda_C),$$

where λ_C is calculated for every failure time t separately so that the specific censoring rate $\mathrm{p_C}$ is met:

$$\mathbb{P}(T_C \leqslant t) = 1 - \exp(-\lambda_C \cdot t) \stackrel{\text{def}}{=} \mathrm{p_C}$$

$$\Rightarrow \lambda_C = -\frac{\ln(1 - \mathrm{p_C})}{t}.$$

Consequently the censoring is dependent on the failure time.

3. Assuming a *combination of administrative and exponential censoring*, the simulated censoring time was generated in two steps:

   (a) Generate administrative censoring times A as explained in 1.

   (b) Generate exponential censoring time given a specific censoring rate for the remaining simulated events, where the administrative censoring times were not smaller than simulated event times, so that an overall censoring proportion of $\mathrm{p_C}$ is achieved:

   $$\mathrm{C} \sim \exp(\lambda_C),$$

   where λ_C is calculated for every failure time t separately so that the specific censoring rate $\mathrm{p_C}$ is met:

   $$\mathbb{P}(T_C \leqslant t) = 1 - \exp(-\lambda_C \cdot t) \stackrel{\text{def}}{=} \mathrm{p_C^{needed}}$$

   $$\Rightarrow \lambda_C = -\frac{\ln(1 - \mathrm{p_C^{needed}})}{t},$$

   where $\mathrm{p_C^{needed}}$ is the specific censoring rate still needed to achieve an overall censoring proportion of $\mathrm{p_C}$. Hence,

   $$\mathrm{p_C^{needed}} = \frac{\mathrm{p_C} \cdot (n_{AC} + n_{SC}) - n_{AC} \cdot \mathrm{p_C^{AC}}}{n_{SC}},$$

   where n_{AC} and $\mathrm{p_C^{AC}}$ are the sample size and rate of censored patients due to administrative censoring (step (a)), respectively. In addition, n_{SC} is the remaining sample size of patients which can be censored by the specific censoring rate in step (b).

- Sample size calculations were performed for each sub-scenario with the approach of Schoenfeld (1981, 1983):

  1. Calculate the required number of events:

  $$d = \frac{(1+r)^2}{r} \cdot \frac{(z_{1-\frac{\alpha}{2}} + z_{1-\beta})^2}{(\ln(\text{designHR}))^2},$$

  where α is the type-I-error rate, β is the type-II-error rate, r is the sample size ratio between the treatment and the control group ($r = n_{trt}/n_{ctl}$), designHR is the assumed hazard ratio / treatment effect, and z_k is the k-quantile of $N(0,1)$.

  2. Calculate the probability of an event $\mathbb{P}(D)$ and divide the number of required events d by this probability to get the required sample size N. Hence, $N = \frac{d}{\mathbb{P}(D)} = \frac{d}{p_C}$.

- Software: The simulation was performed using the software R version 4.2.1 (R Core Team, 2021) in combination with RStudio version 2022.07.2 (RStudio Team, 2021) and packages "tidyverse" (Wickham et al., 2019), "survival" (Therneau, 2023; Therneau and Grambsch, 2000), "flexsurv" (Jackson, 2016), "cutpointr" (Thiele and Hirschfeld, 2021), "vcd" (Meyer et al., 2022), and "pcaPP" (Filzmoser et al., 2022) for data generation and analysis. Additionally, the package "ggpubr" (Kassambara, 2023) and "ggplot2" (Wickham, 2016) was to produce the graphics.

2.3.1.3 Estimands

In the following the estimands for comparing the additional benefit assessment methods are described.

1. Estimation of median, θ_1^{ASCO}, for the method with continuous outcome (ASCO) and category rates for methods with ordinal outcome, $\theta_1^{i,j}$, where i and j represent category i of method j={IQWiG$_{\text{RR}}$, Mod-IQWiG$_{\text{HR}}$, and ESMO}, respectively. This includes the estimation of the maximal category rates $\theta_1^{\max,j}$.

2. Estimation of the relationship between the methods using pairwise correlation, $\theta_2(x, y)$, where x and y are two of the four methods (ASCO, IQWiG$_{\text{RR}}$, Mod-IQWiG$_{\text{HR}}$, and ESMO).

3. Selection of best statistical quantity for additional benefit assessment (lower 95% HR-CI limit, HR-PE, and upper 95% HR-CI limit) using sensitivity and specificity. More detail is provided in the Section 2.3.1.5.

4. Estimation of ASCO cutoff values, which are consistent with categories of ESMO, IQWiG$_{\text{RR}}$, and Mod-IQWiG$_{\text{HR}}$.

2.3.1.4 Methods

As there are two versions of IQWiGs method considered, four methods are compared in total denoted by ASCO, IQWiG$_{\text{RR}}$, Mod-IQWiG$_{\text{HR}}$, and ESMO. A detailed overview of all methods is provided in Section 2.2. The methods are applied to a statistically significant phase III trial based on the log-rank test. This thesis focuses on the application of single phase III trials and does not consider cases where two or more phase III trials are used for more precise parameter estimations.

For the application of the methods the HR-PE with corresponding 95% Wald-CI, and the 2-, 3- and 5-year survival increase were required. Additionally, for ASCO bonus point adjustment "tail of the curve" and ESMO absolute benefit rule med$_{\text{ctl}}$ or med$_{\text{trt}}$ had to be calculated. However, if the survival curve does not fall below 50%, e.g. due to large treatment effects, the median survival time cannot be observed. In this case, a conservative approach was implemented, using the last observed censoring or event time point of the survival curve instead.

2.3.1.5 Performance measures, number of iterations (n_{sim}) and seeds of the simulation

Performance measures:

1. The category rate of each method (with an ordinal outcome) in each sub-scenario is estimated by its proportion:

$$\widehat{\theta_1^{i,j}} = \frac{M_i^j}{\text{Number of significant trials}},$$

where M_i is the number of trials where method $j = \{\text{ESMO}, \text{IQWiG}_{\text{RR}}, \text{Mod-IQWiG}_{\text{HR}}\}$ assigned the treatment an additional benefit category i. The amount of categories differ for each method:

- IQWiG_{RR} and $\text{Mod-IQWiG}_{\text{HR}}$: "minor added benefit", "considerable added benefit", and "major added benefit"
- ESMO: 1, 2, 3, 4

This also includes the proportion of the maximal category in each sub-scenario and each method, where the number of trials with a maximal added benefit treatment assignment is divided by the number of significant trials:

$$\widehat{\theta_1^{\text{max, j}}} = \frac{\text{Number of maximal categories}}{\text{Number of significant trials}},$$

where a maximal score is defined differently for each method:

- IQWiG_{RR} and $\text{Mod-IQWiG}_{\text{HR}}$: "major added benefit"
- ESMO: 4

Since ASCO has a continuous outcome, median estimates of the NHB score (θ_1^{ASCO}) are reported.

2. To estimate the relationship between the methods, pairwise Spearman correlation was calculated using the interpretation provided by Mukaka (2012) and examining the com-

plete range of categories for the two methods:

$$\widehat{\theta_2}(x,y) = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{(n-1)s_x s_y},$$

where n is the sample size, x_i, $\bar{x}$, and s_x are the individual scores/categories, sample means, and sample standard deviations of the rank-converted scores/categories of the additional benefit assessment method x. Analogously, the same applies for y_i, $\bar{y}$, and s_y for additional benefit assessment method y.

As a sensitivity analysis for the relationship assessment Kendall-τ_{b} was calculated.

3. To evaluate which statistical quantity (lower 95% HR-CI limit, HR-PE, and upper 95% HR-CI limit) is better suited for the additional benefit assessment of new treatments, sensitivity, and specificity were estimated. In the following true positives rate (TPR) instead of sensitivity and false positive rates (FPR) instead of 1 - specificity will be used. FPR and TPR were estimated for thresholds ranging from 0.2 to 1 which are used for defining a maximal additional benefit classification using HR-PE, HR^{-}, and HR^{+}. In this context, a true positive and false positive event is defined as deserved classification of a maximal category, or, respectively, not deserved classification of a maximal category. Furthermore, a ground truth was needed for the estimation of TPR and FPR but since no gold standard for additional benefit assessment method exists, a maximal category was assumed to be justified if trueHR $< \delta_{deserved}$ was met for different cut-offs values of $\delta_{deserved}$ (0.7, 0.75, and 0.8).

   For illustration purposes, ROC curves using the TPR and FPR estimations were used.

   Example for TPR and FPR calculation:

   Lets assume a sub-scenario with a specific designHR, power, censoring rate, and $\mathrm{med}_{\mathrm{ctl}}$. If further a specific $\delta_{deserved}$ is assumed, the simulation provides estimates for the number of false positives (FP), true negatives (TN), true positives (TP), and false negatives (FN) like shown in Table 7. Hence, FPR and TPR can be estimated in the following way:

$$\widehat{\mathrm{FPR}} = \frac{\mathrm{FP}}{\mathrm{TN} + \mathrm{FP}},$$

$$\widehat{\mathrm{TPR}} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}.$$

Table 7: *Example cross table for FPR and TPR calculation.*

		Deserved maximal score (trueHR $< \delta_{deserved}$)	
		yes	no
Achieved maximal score by ESMO, IQWiG$_{RR}$ or Mod-IQWiG$_{HR}$	yes	TP	FP
	no	FN	TN

Abbreviations: FPR: False positive rate, FN: False negative, FP: False positive, HR: Hazard ratio, TP: True positive, TN: True negative, TPR: True positive rate, trueHR: True underlying HR of data generation, $\delta_{deserved}$: Ground truth of deserved maximal category for TPR and FPR calculation (justified if trueHR $< \delta$)

It is obvious that FPR and TPR can only be estimated when TN + FP $\neq 0$ and TP + FN $\neq 0$, respectively. However, in sub-scenarios with trueHR $< \delta_{deserved}$ the sum TN + FP can be zero and vice versa in sub-scenarios with trueHR $\geq \delta_{deserved}$ the sum TP + FN can be zero. For example, lets assume $\delta_{deserved} = 0.7$ and the sub-scenario with designHR = 0.8 and $HR_{var} = 1$ (trueHR = 0.8) of the Standard Scenario. Hence, a maximal category is not deserved (trueHR$>\delta_{deserved}$). Thereby, a FP can only occur if an additional benefit assessment method awards a maximal category and a TN can only occur if an additional benefit assessment method does not award a maximal category. In other words, in this case TP and FN cannot occur and thus TPR cannot be estimated.

Hence, to still be able to calculate FPR, TPR and thus display ROC curves, the complete simulated treatment effect range was used. Thereby, in each simulation scenario all sub-scenario with designHRs ranging from 0.3 until 0.9 were combined while fixing the other parameters of the simulation scenario (e.g.: censoring rate, power, allocation ratio, HR_{var}); for example in the Standard Scenario remained $(5 \cdot 2 \cdot 3 =)$ 30 instead of $(5 \cdot 31 \cdot 2 \cdot 3 =)$ 930 sub-scenarios.

Since the additional benefit assessment methods are applied for all new treatments including ones with large and small treatment effects, the approach of combining the complete range of treatment effects for an overall valuation of the methods can be seen as realistic representation.

4. To investigate which ESMO, IQWiG$_{RR}$, and Mod-IQWiG$_{HR}$ category corresponds to which ASCO score, the maximizing weighted Cohens kappa approach for cutoff value

determination was used (Cohen, 1960; Chang et al., 2015):

$$\kappa = 1 - \frac{\sum_{i=1}^{k}\sum_{j=1}^{k} w_{ij} \cdot x_{ij}}{\sum_{i=1}^{k}\sum_{j=1}^{k} w_{ij} \cdot m_{ij}},$$

where $i = 1, \ldots, k$ and $j = 1, \ldots, k$ are the categories of both methods, x is the observed probability matrix, w the quadratic weights matrix, and m the expected probability matrix. Since ESMO, IQWiG$_{\text{RR}}$, and Mod-IQWiG$_{\text{HR}}$ have an ordinal outcome, disagreements close to the diagonal implies a smaller disagreement than far from the diagonal. Thus, Cohen kappa with "Fleiss-Cohen" weights were used.

To determine how different methods for cutoff determination impact the results and to achieve a fair comparison between the research of this thesis and Cherny et al. (2019), who calculated ASCO cutoff values corresponding to categories of ESMO using 102 real studies, the same methods for cutoff determination were used. Hence, beside the maximizing Cohens kappa approach, sensitive analysis using receiver operating characteristic (ROC) and Svenssons method (Svensson, 2000a,b) were performed. For the former, the categories were separated pairwise and considered optimal when the point on the ROC curve is closest to the point (0,1) (ROC01). For the latter, the cutoffs were defined so that same marginal distribution of ordinal method and continuous ASCO were present.

Number of iterations for each sub-scenario ($\mathbf{n_{sim}}$):

As one of the main estimands of this simulation study is the proportion of maximal category ($\theta_1^{\max}$) and the most additional benefit assessment methods have an ordinal rating scale, the number of iterations for each sub-scenario is based on this estimand:

An indicator variable, Y_i, was defined for iteration i to be 1 if an additional benefit assessment method awarded a maximal category; otherwise Y_i was set to be 0:

$$Y_i := \begin{cases} 1, & \text{if the maximal category is awarded,} \\ 0, & \text{another category is awarded.} \end{cases}$$

Therefore, Y_i is by definition a Bernoulli variable with a coverage probability of p, which can be estimated by the sample proportion. The variance of the simulation-based estimate of p

is given by

$$\hat{p} = \frac{p \cdot (1-p)}{\mathrm{n_{sim}}},$$

where $\mathrm{n_{sim}}$ is the number of iterations. This estimation is derived by the variance of a Bernoulli variable, which is given by $p \cdot (1-p)$, and the fact that the simulation generated independent and identically distributed Bernoulli variables. Furthermore, it can be shown that

$$\frac{p \cdot (1-p)}{\mathrm{n_{sim}}} \leqslant \frac{1}{4 \cdot \mathrm{n_{sim}}}.$$

The final number of iterations $\mathrm{n_{sim}}$ can then be calculated by assuming a variance of p less than some pre-specified threshold δ

$$\mathrm{n_{sim}} \geqslant \frac{1}{(4 \cdot \delta)}.$$

In each simulated sub-scenario of this simulation study, the requirement of a standard deviation of 0.25% for the coverage probability of a maximal score was implemented, assuming a constant error variance. Thus, δ was set to $2.5 \cdot 10^{-5}$, which corresponds to a standard deviation of 0.5%, resulting in a number of $\mathrm{n_{sim}} = 10,000$ iterations (phase III trials) for each sub-scenario.

Since the additional benefit assessment methods are applied after a significant phase III trial, only $\mathrm{n_{sim}} - \mathrm{n_{ns}}$ trials were used for the method comparison, where $\mathrm{n_{ns}}$ is the number of non-significant trials. As a sample size calculation is performed, the resulting number of significant iterations should be around 8,000 or 9,000, depending on the power of 90% or 80%. In case of underpowered or overpowered trials like in Scenario 2, the remaining number of significant iterations might be less or more, respectively. In this scenario, the difference between actual and the planned number of significant iterations of 8,000 or 9,000 was not adjusted by increasing or decreasing $\mathrm{n_{sim}}$.

Seeds of the simulation:

To create reproducible results and to achieve comparability between the different scenarios, the same integer numbers were used as seeds at the beginning of the 10,000 iterations of each sub-scenario. Hence, 10,000 integer numbers were once randomly drawn out of a sample ranging from 1 to 1 billion without replacement.

2.3.2 Simulation 2 (censoring times independent of event times)

Parts of this Section 2.3.2 are already published in the article *A Comparison of Additional Benefit Assessment Methods for Time-to-Event Endpoints Using Hazard Ratio Point Estimates or Confidence Interval Limits by Means of a Simulation Study* by Büsch et al. (2024). The manuscript has been written by the lead author but may contain comments and corrections from the co-authors and the reviewers.

2.3.2.1 Aim

The possible introduced bias by the data generation of Simulation 1 (described in Section 2.3.1) is not affecting the method comparison to a substantial degree because it is affecting all compared methods equally. The aim of this additional simulation is to investigate the robustness of the results obtained from Simulation 1 (details in Section 2.3.1). The amount of different scenarios, however, is reduced to the ones which showed the main differences in the already performed simulation study. The ADEMP structures of both simulation studies are in many aspects similar and hence, only the changed aspects of the ADEMP structure are mentioned below.

2.3.2.2 Data-generating mechanisms

- Scenarios / specification of parameters

  As this simulation study was performed to investigate the robustness of the results of the simulation study described in Section 2.3.1, the amount of different scenarios is reduced to the first four scenarios. Furthermore, in sub-scenarios with a monotonically decreasing hazard rate like in scenarios with Weibull and Gompertz failure time distribution, the hazard of an event is getting smaller over time. Thus, patients with no event until a specific time t, almost cannot have an event anymore and must be censored at the end of the trial (administrative censoring). Unfortunately, in some sub-scenarios these patients are often present leading to censoring rate larger than the intended p_C. To still achieve the targeted censoring rate without biasing the hazard ratio estimation, only the sub-scenarios with p_C=60% were used.

  An overview of all scenarios and sub-scenarios is given in Table 8, where the differences

between the performed scenarios are highlighted in bold.

Table 8: *Overview of simulation scenarios of simulation study with censoring times independent on event times*

Scenario	Parameters differences between the scenarios			
	HR_{var}	Failure time distribution	r	Censoring distribution
Standard Scenario	1	exponential	1	$p_C = 60\%$ equal in both treatment groups with administrative (accrual: 2 years, follow-up: $2 \cdot med_{ctl}$) and exponential censoring
Scenario 2 (**HR_{var}**)	**Overpowered trials:{0.8,0.9} Underpowered trials:{1.1,1.2}**	exponential	1	$p_C = 60\%$ equal in both treatment groups with administrative (accrual: 2 years, follow-up: $2 \cdot med_{ctl}$) and exponential censoring
Scenario 3 (**failure time**)	1	**Weibull and Gompertz**	1	$p_C = 60\%$ equal in both treatment groups with administrative (accrual: 2 years, follow-up: $2 \cdot med_{ctl}$) and exponential censoring
Scenario 4 (**non-prop. hazards**)	1	exponential **with delayed treatment effect** ($start_{trt} = \frac{1}{3} \cdot med_{ctl}$)	1	$p_C = 60\%$ equal in both treatment groups with administrative (accrual: 2 years, follow-up: $2 \cdot med_{ctl}$) and exponential censoring

Notes: Differences to the standard scenario regarding the parameter choice were highlighted in bold. The following parameters were chosen to be identical in each scenario and hence are not shown in the table: significance level α of 0.05, power of 90% and 80%, $med_{ctl} \in \{6, 12, 18, 24, 30\}$ and designHR $\in \{0.3, 0.32, ..., 0.88, 0.9\}$. Abbreviations: α: Significance level, designHR: Design hazard ratio used for sample size calculation, HR: Hazard ratio, HR_{var}: Factor for deviance between designHR and trueHR, p_C: Overall censoring rate, med_{ctl}: Median survival time of control group, $start_{trt}$: Treatment starting time point, trueHR: True underlying HR of data generation

- Failure time distribution of control and treatment group (f_{ctl} and f_{trt}) were defined as in Simulation 1. Details can be found in Section 2.3.1.2.

- Censoring time distribution:

  To achieve a realistic phase III trial, the censoring was defined as a combination of independent administrative censoring and independent right-censoring with an overall targeted censoring proportion of p_C for all scenarios and without introducing bias to the HR estimation. Hence, administrative censoring, independent right-censoring and failure times needed to be independent from each other. To achieve this, the following considerations were made:

  1. *Censoring proportion for administrative censoring with an accrual period:*

     A patient with an event time $T = t$, who enters the trial at time point $Acc = a$ after study initiation, will be censored if the event would happen after end of the study, i.e. if $t + a > dur$. Taking this into account, the censoring proportion can

be calculated in the following way:

$$\begin{aligned} p_C &= \mathbb{P}[T + Acc > dur] \\ &= \int \mathbb{P}[T + Acc > dur \mid A = a] f_{Acc}(a) da \\ &= \int \mathbb{P}[T > dur - a] f_{Acc}(a) da \\ &= \int S_T(dur - a) f_{Acc}(a) da. \end{aligned}$$

2. *Censoring proportion for administrative and independent censoring:*
   In this case, it is assumed that all patients are recruited at the same time leading to two reasons a patient can be censored: Firstly, a patient did not have an event over the period of the trial and was not lost to follow-up. Hence, this patient has a censoring time $C = c$ and event time $T = t$ later than the trial duration ($dur < \min(c, t)$), which is defined as administrative censoring ($A = dur$). Secondly, a patient did have a censoring event $C = c$ occurring before the event $T = t$ and the study duration ($c < \min(t, dur)$); e.g. a patient moves away and is lost to follow-up.
   Furthermore, in case a patient with an event time after the end of the study ($T > dur$) must be censored administratively (at the last follow-up visit) or censored because of a censoring event during the course of the study. In addition, a patient with an event during the trial, $T \leqslant dur$, can only be censored if the censoring occurs before the event. Taking this into account, the censoring proportion can be calculated solving the following integral:

$$\begin{aligned} p_C &= \int_0^{dur} \int_0^t f_{T,C}(t, c) dc dt + \int_{dur}^{\infty} \underbrace{\int_0^{\infty} f_{T,C}(t, c) dc}_{f_T(t)} dt \\ &= \int_0^{dur} f_T(t) F_C(t) dt + S_T(dur). \end{aligned}$$

3. *Censoring proportion for administrative (with accrual period) and independent censoring:*

   The two above explained considerations need to be combined. This means that the time under observation for each patient reduces by the time $Acc = a$ that the patient enters the study after its initiation. Hence, integrating the expression from the last step over the accrual distribution yields:

$$\begin{aligned} p_C &= \int \left(\int_0^{dur-a} f_T(t) F_C(t) dt + S_T(dur - a) \right) f_{Acc}(a) da \\ &= \int \int_0^{dur-a} f_T(t) F_C(t) f_{Acc}(a) dt da + \int S_T(dur - a) f_{Acc}(a) da. \end{aligned} \tag{2.7}$$

In the performed simulation study the censoring time distribution C is assumed to be exponential distributed ($C \sim \text{Exp}(\lambda_C)$). Furthermore, three different failure time distributions are assumed, leading to different p_C equations. Each failure time distribution has a different effect on the censoring probability p_C and the censoring parameter λ_C. More information regarding the different distributions and the obtaining of λ_C can be found in Appendix C.

- Sample size calculations were performed as in Simulation 1. Details can be found in Section 2.3.1.2.

- Software: The same software as in Simulation 1 was used. Details can be found in Section 2.3.1.2.

2.3.2.3 Estimands

The first three estimands as in Simulation 1 were used. Details can be found in Section 2.3.1.3.

2.3.2.4 Methods

The same methods as in Simulation 1 was used. Details can be found in Section 2.3.1.4.

2.3.2.5 Performance measures, number of iterations (n_{sim}) and seeds of the simulation

The same performance measures, n_{sim} and seeds as in Simulation 1 was used. Details can be found in Section 2.3.1.5.

Chapter 3

Results

In this chapter, the results are outlined. Section 3.1 and Section 3.2 show the results of Simulation 1 and Simulation 2, respectively. Overall in both simulation studies 15,810 subscenarios were investigated. Section 3.3 depicts the application of the additional benefit assessment methods and the results of the determined ASCO cutoff application on two study examples.
Additional results of the performed simulation studies can be found in Appendix A.

3.1 Simulation 1

Parts of this Section 3.1 are already published in the article *A Comprehensive Comparison of Additional Benefit Assessment Methods Applied by Institute for Quality and Efficiency in Health Care and European Society for Medical Oncology for Time-to-Event Endpoints After Significant Phase III Trials — a Simulation Study* by Büsch et al. (2022). The manuscript has been written the lead author but may contain comments and corrections from the co-authors and the reviewers.
In the following, the results of Simulation 1 are shown, where the generated censoring times are dependent on the event times and hence might introduce bias into the HR and HR-CI estimation. Further information of the ADEMP structure of the simulation study is outlined in Section 2.3.1.
Firstly, the description of the additional benefit assessment methods is shown for each Scenario, where the ASCO score distribution is illustrated using boxplots separated into the categories of ESMO, IQWiG$_{\text{RR}}$, Mod-IQWiG$_{\text{HR}}$, and overall, respectively (y-axis). Secondly, the relationship between the methods is shown by displaying pairwise Spearman correlations between the additional benefit assessment methods. Thirdly, ROC curves are displayed to investigate the best statistical quantity for additional benefit assessment. Fourthly, the ASCO cutoff values for corresponding ESMO, IQWiG$_{\text{RR}}$, and Mod-IQWiG$_{\text{HR}}$ are laid out. Lastly, the possible introduced bias due to the data generation is investigated by displaying HR

estimation bias using line figures.

Important to mention is that in case of underpowered studies, the planned power of 80% or 90% was not achieved due to overoptimistic treatment effect assumptions (Scenario 2 with $HR_{var}>1$). For example, over all sub-scenarios for $HR_{var}=1.1$ and $HR_{var}=1.2$ a mean power of 69.84% and 50.00% was present, respectively. This power reduction occurs especially for small treatment effects as a larger sample size is needed. As the additional benefit assessment methods are only applied after a significant study, the results of this scenario are based on a very limited amount of observations. This has to be kept in mind for interpretation.

3.1.1 Comparison of additional benefit assessment methods

In the following, the ASCO score distribution (x-axis) is illustrated for each Scenario using boxplots separated into the categories of ESMO, $IQWiG_{RR}$, Mod-$IQWiG_{HR}$, and overall (abbreviated "all"), respectively (y-axis). Furthermore, the relative frequencies, i.e. proportions, of each category of the ordinal methods (ESMO, $IQWiG_{RR}$, and Mod-$IQWiG_{HR}$) are shown on the y-axis label. For these comparisons the results of all sub-scenarios of each simulation scenario are combined and shown in separate panels. For Scenario 3a and 3b (different failure time distributions) the results (ASCO distribution and category proportion of ordinal methods) for the sub-scenarios with increasing and decreasing hazards over time are shown combined and additionally separately color coded in the respective panel. The reason for this illustration is that the additional benefit assessment methods respond differently on increasing and decreasing hazards.

The figures for the pairwise comparisons of ASCO and each of the ordinal additional benefit assessment methods are shown in the following subsections: Firstly, ESMO vs. ASCO, secondly $IQWiG_{RR}$ vs. ASCO, and lastly Mod-$IQWiG_{HR}$ vs. ASCO.

3.1.1.1 ESMO vs. ASCO

In the Standard Scenario (Figure 4 and 5, upper left panel) ESMO shows a very high rate of 80.57% for the maximum category. Other categories mainly do hardly exist. For example category 2 and 3 have only a rate of 5.29% and 2.53%, respectively. Category 1 has again an increased rate of 11.62% compared to category 2 and 3. This tendency or even an higher rate of the maximal category is present in almost all scenarios: For the scenarios

with Weibull distributed failure times (Scenario 3), non-proportional hazards (Scenario 4), different treatment allocation ratios (Scenario 5, Figure 5), and only exponential censoring without administrative censoring (Scenario 6) the distribution is similar to the Standard Scenario. The maximal category rate is even larger in overpowered and underpowered studies (Scenario 2, Figure 4) compared to the Standard Scenario, e.g. the maximal category rate of 94.39% and 90.08% is present, respectively. In cases of informative censoring (Scenario 7, Figure 5) the maximal category rate is also larger than in the Standard Scenario, e.g. 92.52% and 95.79%, where the censoring rate of the control group is smaller than in the treatment group and vice versa, respectively. An exception of this tendency is present with Gompertz distributed failure time (Scenario 3) where the categories are more evenly distributed but the maximal category has still the highest rate, i.e. the maximal category rate is at 44.25%.
ASCO has an overall median score of 42.41 in the Standard Scenario (Figure 4, upper left panel). In case of other simulated scenarios, the overall median ASCO score stays very similar. Only in case of wrongly assumed treatment effects leading to over- and underpowered studies (Scenario 2; Figure 4, upper middle and right panel) as well as informative censoring with a censoring rate larger in the control than in the treatment group (Scenario 7; $p_C^{ctl} > p_C^{trt}$, Figure 5, lower right panel), the median ASCO score is increased to 49.60, 46.99, and 49.36, respectively, compared to the Standard Scenario. Furthermore, non-proportional hazards (Scenario 4; Figure 4, lower right panel) decrease the median score to 36.95.

As the majority of sub-scenarios are categorized as a maximal category by ESMO, the overall median ASCO score and the median ASCO score of ESMOs maximal category are very similar. Furthermore, the median values of the ASCO score of ESMO categories do increase with increasing ESMO category and are quite similar in all scenarios, e.g. in the Standard Scenario the median ASCO score is 13.44, 18.90, 23.00, and 48.72 for ESMO category 1, 2, 3, and 4, respectively. Only with Gompertz distributed failure times this is not the case, meaning that category 3 has a higher median ASCO score than category 4, i.e. 56.39 and 54.68, respectively. This can be explained by the different category rate distributions of increasing and decreasing hazards: With the former, category 4 has a rate of 15.45% and with the latter a rate of 78.50% is present. Hence, for the combined results of increasing and decreasing hazards, the sub-scenarios with decreasing hazards influence the combined median ASCO score (Figure 4, Scenario 3b: Gompertz, blue boxplots) in a stronger way.

For category 3 it is the other way round, meaning that the rate is higher in sub-scenarios with increasing hazards, i.e. 16.12% vs. 0.01%. Thus, the combined category 4 rate is mainly explained by sub-scenarios with decreasing hazards and category 3 by sub-scenarios with increasing hazards. Moreover, as the median ASCO score of category 3 for increasing hazards is larger than the median ASCO score of category 4 for decreasing hazards (56.41 vs. 49.75), the combined results for Gompertz distributed failure times have a smaller median ASCO score for category 4 than category 3.

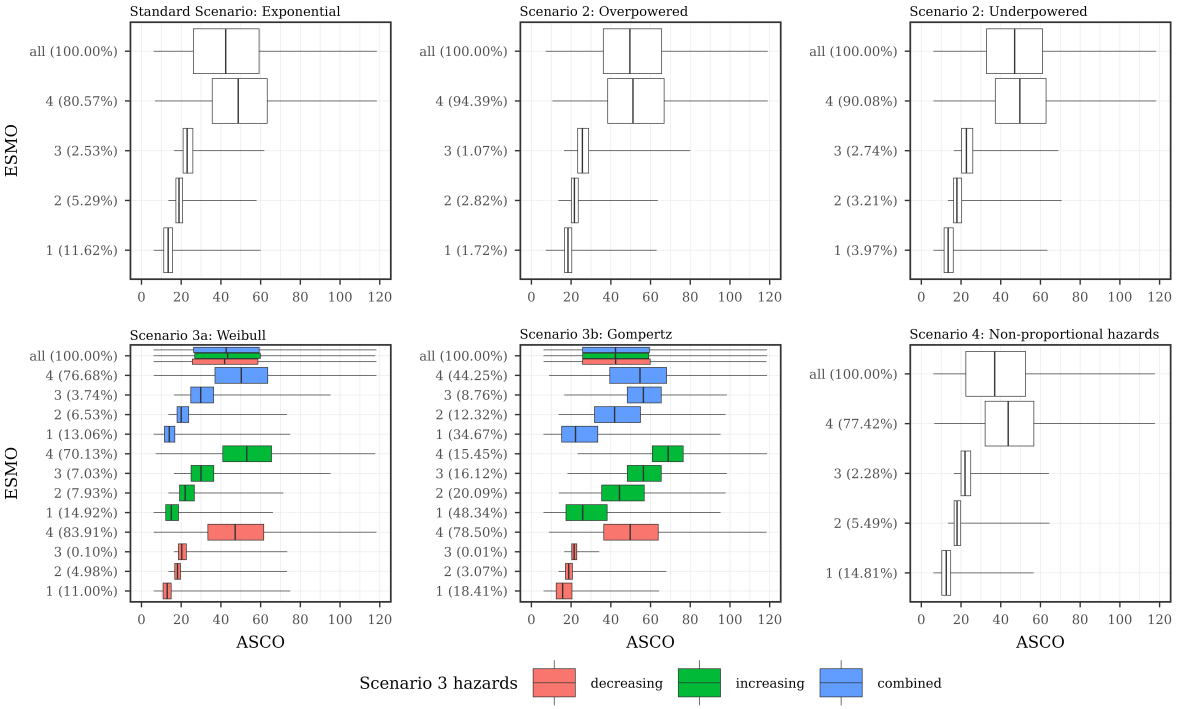

Figure 4: *Description of ASCO score distribution (x-axis) separated into the categories of ESMO and overall (y-axis) using boxplots for Scenarios 1 to 4.*

The results of all sub-scenarios of each scenario are combined and shown in separate panels. The overall ASCO score is abbreviated to "all". The proportions of each ESMO category are shown on the y-axis label. For Scenario 3a and 3b the ASCO distribution and category proportion of ESMO are shown combined (blue) and separately (red and green) for the sub-scenarios with increasing and decreasing hazards over time.

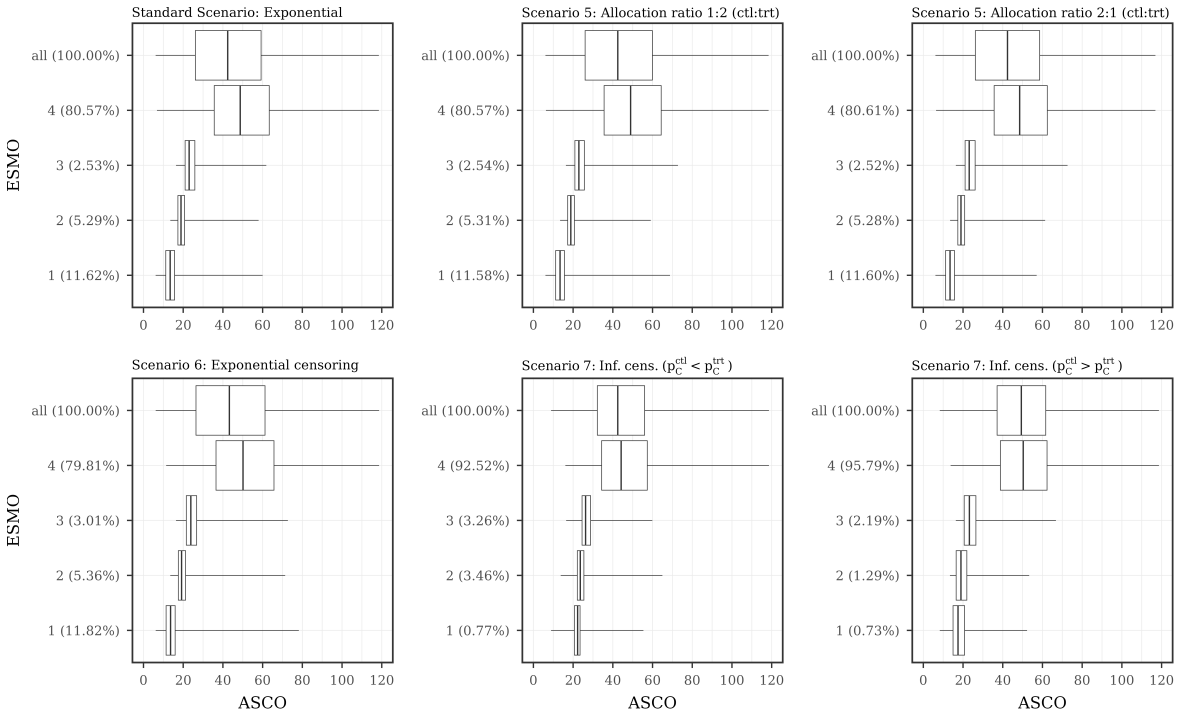

Figure 5: *Description of ASCO score distribution (x-axis) separated into the categories of ESMO and overall (y-axis) using boxplots for Scenarios 1 and 5 to 7.*

The results of all sub-scenarios of each scenario are combined and shown in separate panels. The overall ASCO score is abbreviated to "all". The proportions of each ESMO category are shown on the y-axis label. Abbreviations: p_C^{ctl} and p_C^{trt}: Censoring rate of control and treatment group

3.1.1.2 IQWiG$_{RR}$ vs. ASCO

IQWiG$_{RR}$ has an overall maximal category rate of 56.70% in the Standard Scenario (Figure 6 and 7, upper left panel), which is less than ESMOs maximal category rate. Other underlying failure time distributions like Weibull and Gompertz (Scenario 3, Figure 6), different allocation ratios (Scenario 5, Figure 7) or without administrative censoring (Scenario 6, Figure 7) do impact the rate distributions of IQWiG$_{RR}$s categories only marginally. However, using wrongly assumed treatment effects for the sample size calculation and hence leading to overpowered or underpowered studies do influence IQWiG$_{RR}$ compared to the Standard Scenario (Figure 6). In case of overpowered studies, the maximal category rate increases by 30.76% to 87.46%. This increase is more than twice as large than the increase of ESMO maximal category (94.39%-80.57%=13.82%) because ESMO does already have a very large rate in the Standard Scenario and hence cannot increase much further. Underpowered studies do not influence IQWiG$_{RR}$ category distribution compared to the Standard Scenario. ESMO, however, still shows (as described above) a similar increase in the maximal category rate as for overpowered studies. If non-proportional hazards are present (Scenario 4) the rate of the maximal category is reduced to 38.13% and the minimal category increased to 24.88% compared to 56.70% and 12.50% in the Standard Scenario, respectively. Similar results with category rates reduced to 46.40% or increased to 21.23% for the maximal and minimal category can be seen in case of informative censoring with a censoring rate larger in the control than in the treatment group (Scenario 7 ($p_C^{ctl} > p_C^{trt}$), Figure 7 lower right panel). If the treatment group has the larger censoring rate ($p_C^{ctl} < p_C^{trt}$) an increase of the maximal category rate is present, similar to overpowered studies (Scenario 2).

As for the ASCO and ESMO comparison, the median values of the ASCO score of IQWiG$_{RR}$ categories do increase with increasing IQWiG$_{RR}$ category and are quite similar in all scenarios, e.g. in the Standard Scenario the median ASCO score is 55.94, 26.07, and 16.52 for IQWiG$_{RR}$ category major, considerable, and minor, respectively. Only with overpowered studies (Scenario 2) and present informative censoring (Scenario 7, $p_C^{ctl} < p_C^{trt}$) this behavior does change, meaning that the category "considerable" has a lower median ASCO score than category minor. Furthermore, the overall median ASCO score is very similar to the maximal category of IQWiG$_{RR}$ (Scenario 2, overpowered: 49.60 vs. 52.39; Scenario 7, $p_C^{ctl} < p_C^{trt}$: 42.48 vs. 45.07) because the rate of maximal category of IQWiG$_{RR}$ is very large.

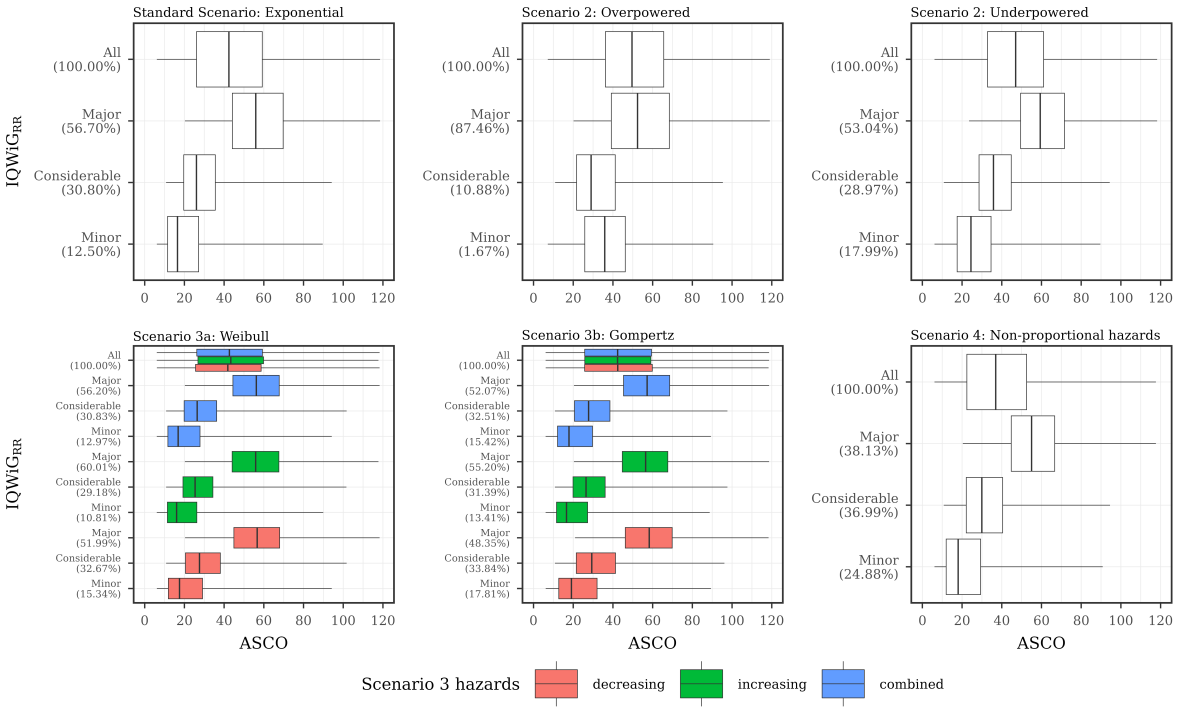

Figure 6: *Description of ASCO score distribution (x-axis) separated into the categories of $IQWiG_{RR}$ and overall (y-axis) using boxplots for Scenarios 1 to 4.*

The results of all sub-scenarios of each scenario are combined and shown in separate panels. The overall ASCO score is abbreviated to "all". The proportions of each $IQWiG_{RR}$ category are shown on the y-axis label. For Scenario 3a and 3b the ASCO distribution and category proportion of $IQWiG_{RR}$ are shown combined (blue) and separately (red and green) for the sub-scenarios with increasing and decreasing hazards over time.

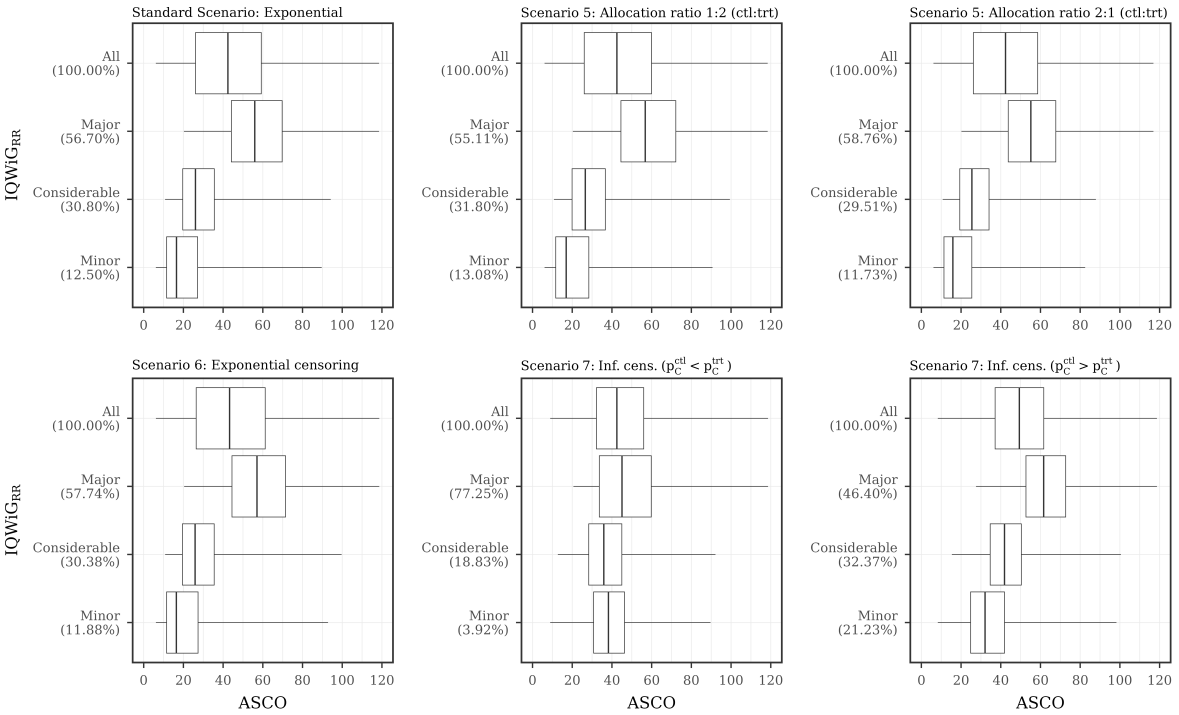

Figure 7: *Description of ASCO score distribution (x-axis) separated into the categories of $IQWiG_{RR}$ and overall (y-axis) using boxplots for Scenarios 1 and 5 to 7.*

The results of all sub-scenarios of each scenario are combined and shown in separate panels. The overall ASCO score is abbreviated to "all". The proportions of each $IQWiG_{RR}$ category are shown on the y-axis label. Abbreviations: p_C^{ctl} and p_C^{trt}: Censoring rate of control and treatment group

3.1.1.3 Mod-IQWiG$_{\text{HR}}$ vs. ASCO

Mod-IQWiG$_{\text{HR}}$ has the lowest rate of the maximal category of all three ordinal additional benefit methods. For example in the Standard Scenario, Mod-IQWiG$_{\text{HR}}$ has a rate of 42.62% compared to 56.70% and 80.57% of IQWiG$_{\text{RR}}$ and ESMO, respectively. Otherwise, Mod-IQWiG$_{\text{HR}}$ behaves very similar to IQWiG$_{\text{RR}}$ over the range of scenarios: The most scenarios (Scenario 2 (with underpowered studies), 3, 5, and 6) do not affect IQWiG$_{\text{RR}}$ meaning that the category rate distribution stays very similar to the Standard Scenario. Furthermore, in these scenarios the median ASCO score of Mod-IQWiG$_{\text{HR}}$ categories does increase with increasing Mod-IQWiG$_{\text{HR}}$ category and is quite similar (as for the comparison of ASCO/ESMO and ASCO/IQWiG$_{\text{RR}}$), e.g. in the Standard Scenario the median ASCO score is 61.41, 29.00, and 17.59 for Mod-IQWiG$_{\text{HR}}$ category major, considerable, and minor, respectively. In case of overpowered studies (Scenario 2) and present informative censoring (Scenario 7, $p_C^{ctl} < p_C^{trt}$), this behavior does change (as for the comparison of IQWiG$_{\text{RR}}$ and ASCO), meaning that category considerable has a lower median ASCO score than category minor. Furthermore, the overall median ASCO score is closest to the maximal category compared to the other categories of Mod-IQWiG$_{\text{HR}}$ (Scenario 2, overpowered: 49.60 vs. 56.00; Scenario 7, $p_C^{ctl} < p_C^{trt}$: 42.48 vs. 52.98). This, however, is not as similar as for IQWiG$_{\text{RR}}$ because the maximal category rate of Mod-IQWiG$_{\text{RR}}$ is not as large as for IQWiG$_{\text{RR}}$.

Furthermore, the same scenarios as for IQWiG$_{\text{RR}}$ influence Mod-IQWiG$_{\text{HR}}$ compared to the Standard Scenario: Overpowered studies lead to a maximal category rate increase of 31.36%(=73.98%-42.62%), which is similar to IQWiG$_{\text{RR}}$'s increase of 30.76%. The rate of the maximal category, however, is still less than for IQWiG$_{\text{RR}}$ (and ESMO) because in the Standard Scenario the rate is already smaller. Non-proportional hazards (Scenario 4) lead to a reduction in almost half of the maximal category rate for Mod-IQWiG$_{\text{HR}}$ compared to the Standard Scenario, which is again very similar to IQWiG$_{\text{RR}}$. Furthermore, informative censoring (Scenario 7) with the control group having a higher censoring rate lead to a decrease of the maximal category and to very similar rates for all three categories (32.49%, 38.33%, and 29.18%). If the treatment group has a higher censoring rate, the highest two categories have very similar rates (49.69% and 44.02%), where both category rates are slightly increased compared to the Standard Scenario.

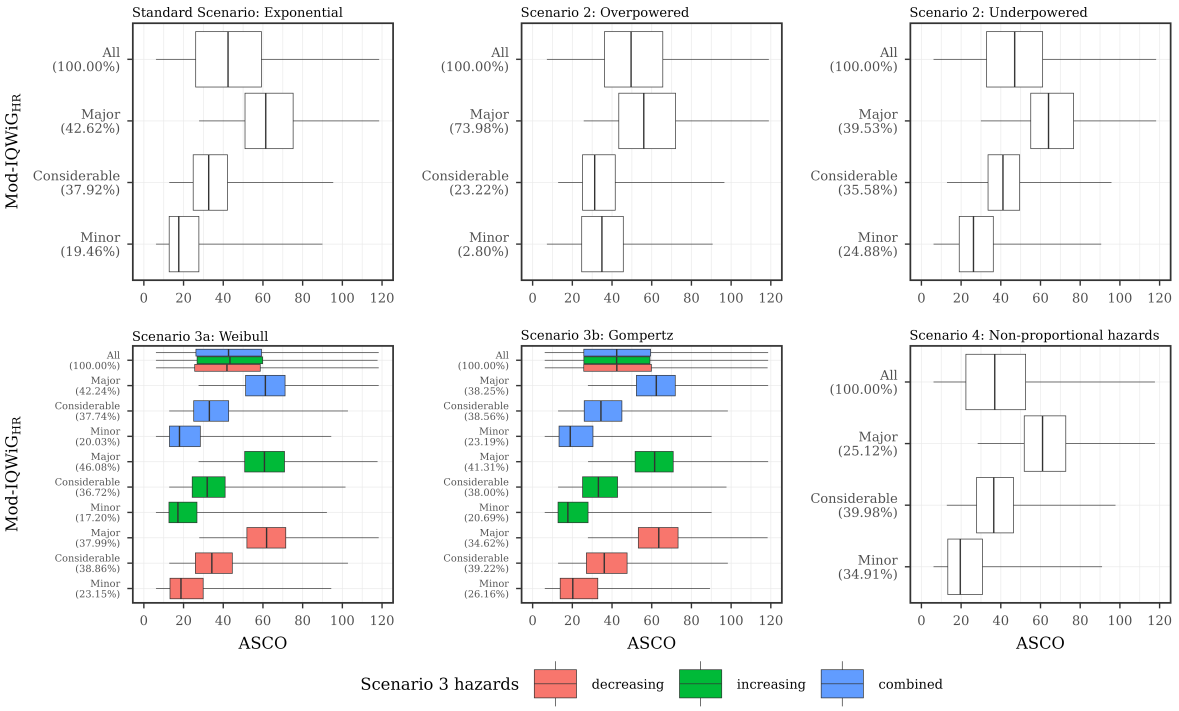

Figure 8: *Description of ASCO score distribution (x-axis) separated into the categories of Mod-IQWiG$_{HR}$ and overall (y-axis) using boxplots for Scenarios 1 to 4.*

The results of all sub-scenarios of each scenario are combined and shown in separate panels. The overall ASCO score is abbreviated to "all". The proportions of each Mod-IQWiG$_{HR}$ category are shown on the y-axis label. For Scenario 3a and 3b the ASCO distribution and category proportion of Mod-IQWiG$_{HR}$ are shown combined (blue) and separately (red and green) for the sub-scenarios with increasing and decreasing hazards over time.

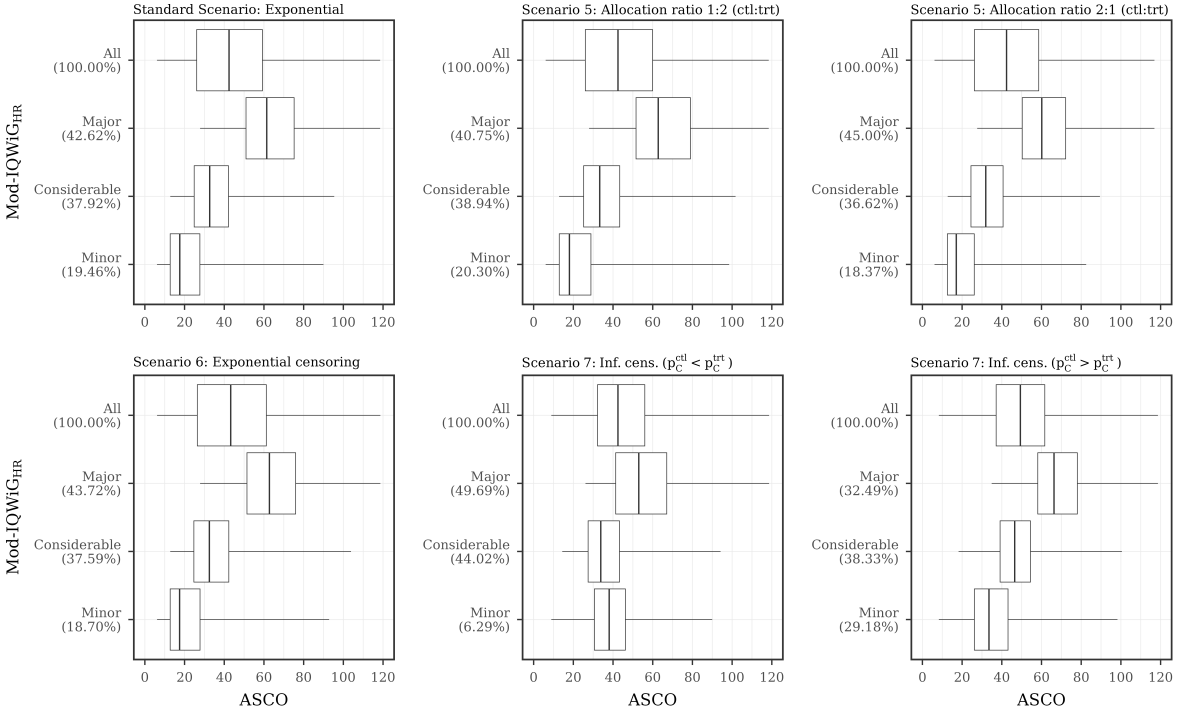

Figure 9: *Description of ASCO score distribution (x-axis) separated into the categories of Mod-IQWiG$_{HR}$ and overall (y-axis) using boxplots for Scenarios 1 and 5 to 7.*

The results of all sub-scenarios of each scenario are combined and shown in separate panels. The overall ASCO score is abbreviated to "all". The proportions of each Mod-IQWiG$_{HR}$ category are shown on the y-axis label. Abbreviations: p_C^{ctl} and p_C^{trt}: Censoring rate of control and treatment group

3.1.2 Relationship between additional benefit assessment methods

In this Section, the results of the pairwise Spearman correlation between all four additional benefit assessment methods are illustrated using heatmaps and line figures. For the former, all sub-scenarios - e.g. different treatment effects (designHR), different censoring rates, etc. - were combined before calculating pairwise Spearman correlations for each scenario. For the latter, the figures are separated by trueHR, med_{ctl}, power, p_{C}, and a scenario specific parameter, which is unique for each scenario; e.g. for Scenario 2 the parameter HR_{var} for wrongly assumed treatment effects is used. In sub-scenarios with very small trueHR (corresponding to large treatment effects), only the same category is assigned by the ordinal additional benefit assessment methods (IQWiG_{RR}, Mod-IQWiG_{HR}, and ESMO) leading to non-computable correlations. Hence, in the line figures of sub-scenarios with large treatment effects, missing values are present. In case of the heatmap, this issue did not occur because all sub-scenarios including small and large treatment effects were combined and hence the whole range of the

ordinal additional benefit assessment methods were assigned. Thus, Spearman correlation could be calculated. Furthermore, the heatmaps and line figures use the same colour codes for the pairwise method correlation results.
Since the results of the line figures of sub-scenarios with different underlying censoring rates and power are very similar, the following results are mostly focused on sub-scenarios with a censoring rate of 60% (p_C=60%) and a power of 90%. The results for other assumed censoring rates, i.e. 20% and 40%, and a power of 80% can be found in Appendix A.1 (only for the Standard Scenario both power sub-scenarios are shown in the Results section).

Figure 10 illustrates the pairwise Spearman correlation (x-axis) using the above mentioned heatmap, where all sub-scenarios of each scenario (y-axis) are combined. Some scenarios, however, were again separated by specific parameters as different results are present. For example, Scenario 2 is separated into underpowered ($HR_{var} > 1$) and overpowered ($HR_{var} < 1$) studies. The specific deviation is described on the y-axis label.
In the Standard Scenario all pairwise comparisons with ESMO (ASCO/ESMO, IQWiG$_{RR}$/ESMO, and Mod-IQWiG$_{HR}$/ESMO) show a moderate positive correlation of 0.54, 0.58, and 0.68, while all other pairwise comparisons have a high positive correlation of 0.74 for ASCO/IQWiG$_{RR}$, 0.79 for ASCO/Mod-IQWiG$_{HR}$, and even 0.84 for Mod-IQWiG$_{HR}$/IQWiG$_{RR}$. Other scenarios lead to different correlation results compared to the Standard Scenario. For example, in case of over- and underpowered studies, the ASCO/ESMO correlation is strongly reduced to a low positive correlation of 0.39 and 0.5, respectively. This behavior is again present for all comparisons with ESMO while the other pairwise comparisons such as ASCO/IQWiG$_{RR}$, ASCO/Mod-IQWiG$_{HR}$, and Mod-IQWiG$_{HR}$/IQWiG$_{RR}$ are only influenced by overpowered studies. The correlation in underpowered studies stays similar to the Standard Scenario.
Different failure time distributions (Scenario 3) do generally not change the correlation between the methods compared to the Standard Scenario. The only exceptions are the ESMO comparisons where a reduced correlation is present for the Gompertz distribution. For the ASCO/ESMO comparison this can be explained by the not increasing median ASCO score with increasing ESMO categories. For the sub-scenarios with increasing and decreasing hazards the median ASCO score does increase with increasing ESMO category as described in detail in Section 3.1.1.1 and hence the correlation between ASCO/ESMO is similar to the

Standard Scenario. Pairwise Spearman correlation behavior of IQWiG$_{\text{RR}}$/ESMO and Mod-IQWiG$_{\text{HR}}$/ESMO can be interpreted similarly. In case of Weibull failure time distribution, the correlation values stay similar for all pairwise comparisons compared to the Standard Scenario. Splitting this scenario into sub-scenarios with increasing and decreasing hazards, however, shows that the ESMO comparisons have an increased correlation for sub-scenarios with increasing hazards and a decreased correlation for sub-scenarios decreasing hazards. For example, the Spearman correlation of ESMO/ASCO is increased to 0.75 and decreased to 0.63 compared to the Standard Scenario with 0.68 for increasing and decreasing hazards, respectively. All other pairwise comparisons without ESMO are not influenced by increasing or decreasing hazards.

Even though non-proportional hazards (Scenario 4) influenced the descriptive measures of the additional benefit methods, the pairwise Spearman correlation is not affected as much compared to the Standard Scenario because all methods are influenced by non-proportional hazard to some degree leading to only small changes in their relationship to one another. Different allocation ratios (Scenario 5) and using only exponential censoring (Scenario 6) do not influence all pairwise comparisons compared to the Standard Scenario. Informative censoring (Scenario 7) with $p_C^{ctl} < p_C^{trt}$, however, does again impact all comparisons leading to a heavily reduced correlation value, where even negligible correlation values of lower than 0.30 are present. For $p_C^{ctl} > p_C^{trt}$ only comparisons with ESMO are impacted leading also to a heavily reduced correlation compared to the Standard Scenario. The other pairwise correlations stay similar to the Standard Scenario.

Similar results can be seen using Kendall-τ_{b} instead of Spearman correlation (see Appendix A.1). The only difference is that Kendall-τ_{b} results show overall smaller values than Spearman correlation.

In the following sub-sections the results of the pairwise Spearman correlation between all four additional benefit assessment methods are shown using the described line figures from above for simulated scenarios without combining sub-scenarios.

Figure 10: *Pairwise Spearman correlation of the additional benefit assessment methods (x-axis) for the different scenarios (y-axis) where all sub-scenarios were combined for correlation calculation.*

Abbreviations: designHR: Design Hazard Ratio used for sample size calculation, HR_{var}: Factor for deviance between designHR and trueHR, n_{ctl} and n_{trt}: Sample size of control and treatment group, p_C^{ctl} and p_C^{trt}: Censoring rate of control and treatment group, r: Allocation ratio (n_{trt}/n_{ctl}), trueHR: True underlying Hazard Ratio of data generation

3.1.2.1 Standard Scenario

Focusing on each sub-scenario of the Standard Scenario with a censoring rate of 60% (p_C=0.6, Figure 11), all pairwise method comparisons with ESMO have similar Spearman correlation curves (black, turquoise, and blue lines) ranging from low to high positive correlation. In particular, decreasing underlying treatment effect (increasing trueHR) lead to stronger correlation between the methods, e.g. in sub-scenarios with med_{ctl}=18 months and power of 90% correlation of 0.0165 and 0.6128 is present for ASCO/ESMO with a trueHR of 0.50 and 0.78, respectively. The correlation curve, however, has its maximum at "moderate" trueHR of around 0.84 with a correlation value of approximately 0.75 to 0.90 for all sub-scenarios, i.e. all panels of Figure 11. For example, in the sub-scenario with med_{ctl}=12 months and power of 90%, the comparisons ASCO/ESMO, IQWiG$_{RR}$/ESMO, and Mod-IQWiG$_{HR}$/ESMO have their maximum correlation of 0.9048, 0.7081, and 0.8312 at trueHR of 0.84, 0.82, and 0.84, respectively. Nevertheless, only at moderate treatment effects these high correlation values are present. All other sub-scenarios with different trueHR have negligible to low correlation values.

The other three pairwise comparisons have more similar correlation values over all treatment effects. For example, Mod-IQWiG$_{HR}$/IQWiG$_{RR}$ (purple line) has a correlation value of 0.7685 at med_{ctl}=12 months, power of 90%, and trueHR=0.3. With larger trueHR (smaller treatment effect) the correlation decreases, e.g. at trueHR=0.9 a correlation value of 0.5611 is present for the same sub-scenario. Furthermore, ASCO/IQWiG$_{RR}$ (red line) as well as ASCO/Mod-IQWiG$_{HR}$ (green line) show a moderate to very high positive correlation over the complete treatment effect range with a maximal correlation value of approximately 0.90 as for the pairwise comparisons of ESMO. For example, at med_{ctl}=12 months and power of 90% the maximal correlation value is 0.8991 and 0.9010 for ASCO/IQWiG$_{RR}$ and ASCO/Mod-IQWiG$_{HR}$, respectively.

These observations are in line with the heatmap (Figure 10), where all sub-scenarios were combined resulting in a Spearman correlation value of 0.68, 0.74, and 0.79 for ASCO/ESMO, ASCO/IQWiG$_{RR}$, and ASCO/Mod-IQWiG$_{HR}$, respectively.

Different simulated power and med_{ctl} marginally influence the above described correlation pattern, e.g. at med_{ctl}=18 months the peak of ESMO comparisons is a bit smaller at around 0.75.

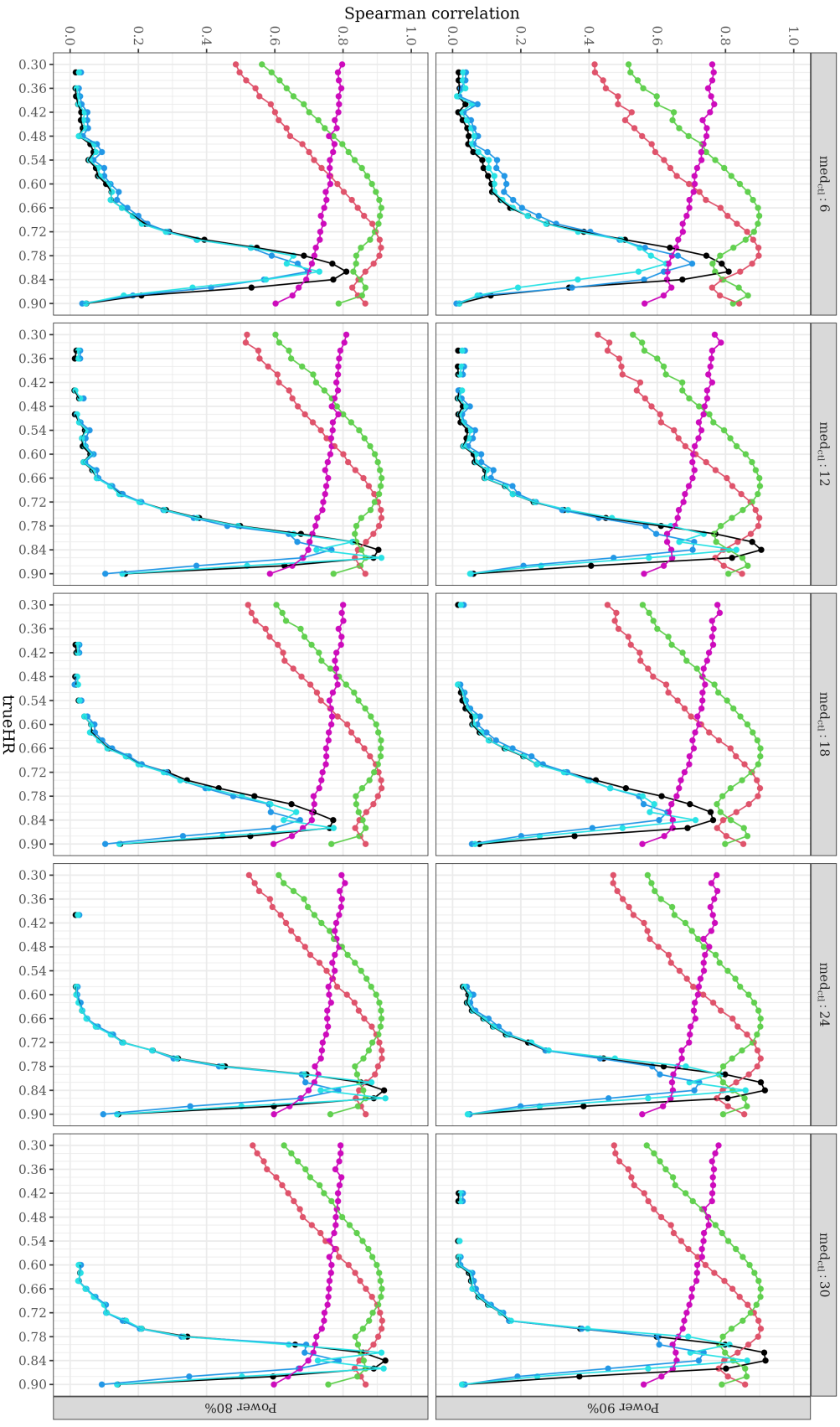

Figure 11: *Pairwise Spearman correlation (y-axis) separated by trueHR (x-axis) of Standard Scenario with p_C=60%.*

Each row of panel stands for different power and each coloumn of panel stands for different med_{ctl}. Abbreviations: med_{ctl}: Median survival time in the control group, p_C: Censoring rate, trueHR: True underlying Hazard Ratio of data generation

3.1.2.2 Influence of incorrectly assumed designHR for sample size calculation (Scenario 2)

Figure 12 illustrates the results of pairwise Spearman correlation between all methods for each sub-scenario of Scenario 2 and is constructed in the same way as Figure 11. It consists, however, out of five instead of two rows, where each row of panel stands for a different HR_{var} and each coloumn of panel stands for a different med_{ctl}. The first two rows are the sub-scenarios with overpowered studies and the last two rows are the sub-scenarios with underpowered studies. Furthermore, the middle row is the same as the top row of Figure 11. The range of trueHR dependents on HR_{var} and hence is different for each row, e.g. in the first row trueHR ranges from 0.24 ($= 0.3 \cdot 0.8$) to 0.72 ($= 0.9 \cdot 0.8$).

The bottom two rows (underpowered studies) can only be compared to the Standard Scenario when focusing on trueHRs ranging from 0.36 to 0.90 because in this range all rows have simulated trueHR sub-scenarios and hence Spearman correlation results are present. The behaviour of all pairwise Spearman correlation comparisons are similar in the underpowered studies compared to the Standard Scenario in the above mentioned range of the trueHR. The correlation maxima, however, of all pairwise ESMO comparisons (black, blue, and turquoise line) are smaller with increasing HR_{var} meaning that in the lowest row of panels this extreme is more visible as in the panel row second to last. This behavior leads to a reduced Spearman correlation when combining all sub-scenarios as shown before in Figure 11. Nevertheless, in general with a smaller treatment effect, the correlation is still increasing.

Furthermore, the pairwise comparison between ASCO and both IQWiG methods (green and red line) flattens out more in underpowered studies with decreasing treatment effect compared to the Standard Scenario. In the Standard Scenario these comparisons have a local maximum at moderate treatment effect before decreasing and increasing again with smaller treatment effects. In sub-scenarios with underpowered studies and trueHR>0.9, the Spearman correlation value is reduced with decreasing treatment effect for all pairwise comparisons.

Similar as in underpowered studies, the top two rows (overpowered studies) can only be compared to the Standard Scenario when focusing on trueHRs ranging from 0.30 to 0.72 because in this range all rows have simulated trueHR sub-scenarios and hence Spearman correlation results are present. In this trueHR range of the overpowered studies, all pairwise Spearman correlations are reduced compared to the Standard Scenario, especially with smaller treat-

ment effects. For example, Mod-IQWiG$_{\mathrm{HR}}$/IQWiG$_{\mathrm{RR}}$ (violet line) has the largest reduction. It is also reflected in all pairwise method correlations when combing all sub-scenarios as shown above (Figure 10).

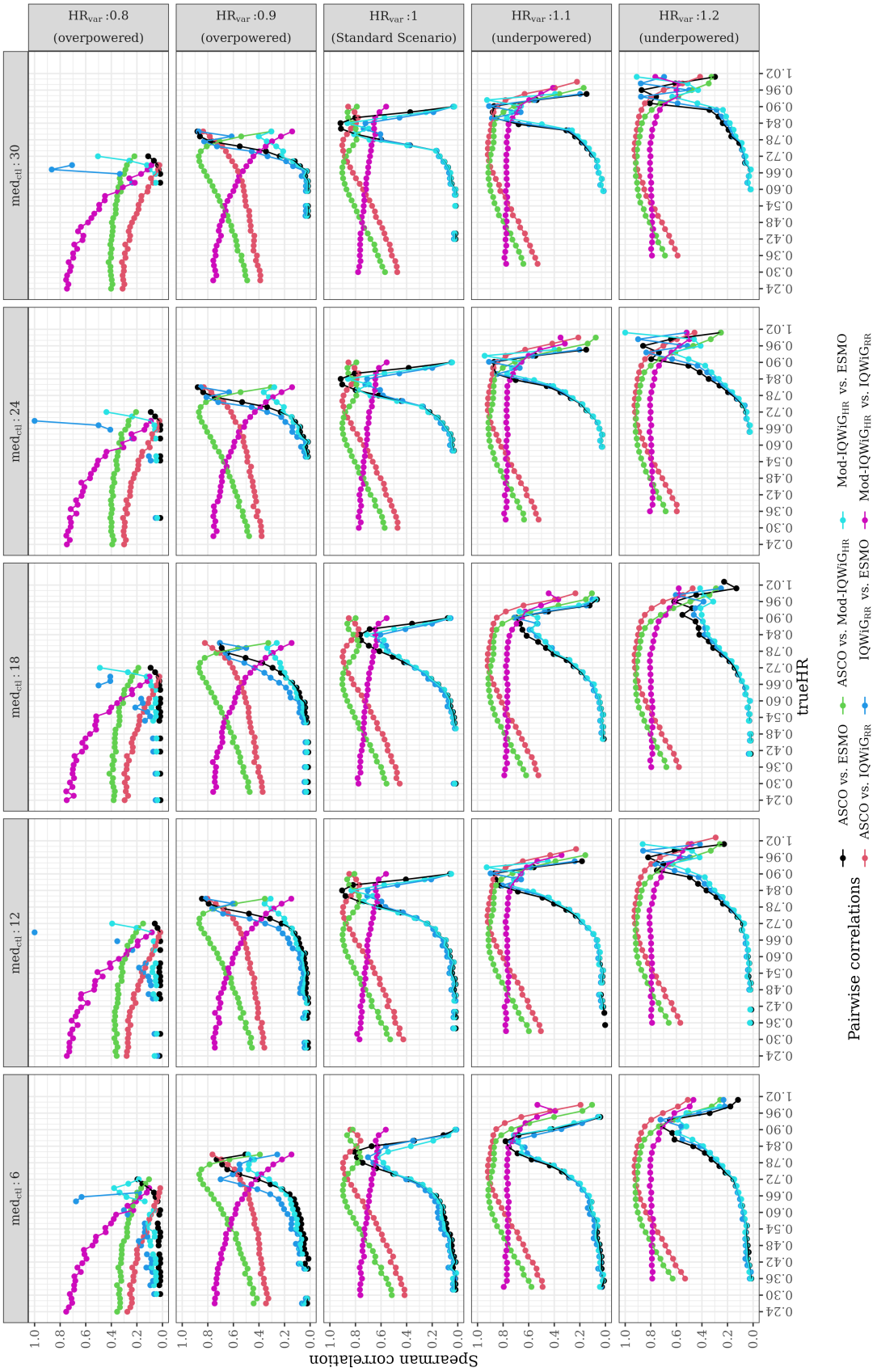

Figure 12: *Pairwise Spearman correlation (y-axis) separated by trueHR (x-axis) of Scenario 2 with p_C=60% and 90% power.*
Each row of panel stands for different HR_{var} and each coloumn of panel stands for different med_{ctl}. The middle row is the same as the top row of Figure 11. Abbreviations: designHR: Design Hazard Ratio used for sample size calculation, HR_{var}: Factor for deviance between designHR and trueHR, med_{ctl}: Median survival time in the control group, p_C: Censoring rate, trueHR: True underlying Hazard Ratio of data generation

3.1.2.3 Influence of different underlying failure time distributions (Scenario 3)

Figures 13 and 14 describe the pairwise Spearman correlations for each sub-scenario with Gompertz and Weibull distributed failure times, respectively. Each row of panel stands for different used shape parameters resulting in increasing, decreasing or constant hazards (Standard Scenario) over time.

Underlying Gompertz distributed failure time distribution heavily influences the pairwise correlations compared to the Standard Scenario (last row). For example, over time increasing hazards in sub-scenarios with very large treatment effects results in an increase from no correlation to a moderate positive correlation of around 0.25 to 0.5 for all pairwise ESMO comparisons compared to the Standard Scenario. In sub-scenarios with moderate to small treatment effect, the pairwise correlation drops down to almost no correlation. This behavior is precise the other way round in the Standard Scenario. The other pairwise comparisons ASCO/Mod-IQWiG$_{\text{HR}}$, ASCO/IQWiG$_{\text{RR}}$, and Mod-IQWiG$_{\text{HR}}$/IQWiG$_{\text{RR}}$ (green, red, and violet line) are not influenced and hence are very similar to the Standard Scenario (bottom row of panels). Decreasing hazards (upper row of panels), however, do only influence all pairwise method comparisons marginally compared to the Standard Scenario.

Weibull distributed failure times (see Figure 14) influence the pairwise Spearman correlation only marginally and hence is not as strong as for underlying Gompertz distribution. In case of over time decreasing hazards (upper row of panels), all pairwise comparisons including ESMO (black, blue, and turquoise line) only have high positive correlations with moderate treatment effect and hence the correlation does not increase over the range of trueHR as for the Standard Scenario. This behavior is exactly the opposite in case of increasing hazards (middle row of panels), where the pairwise Spearman correlation is increasing with larger treatment effect even stronger than in the Standard Scenario and hence these pairwise comparisons have an overall Spearman correlation which is slightly larger than for the Standard Scenario, e.g. 0.75 vs. 0.68 (heatmap, Figure 10). The other pairwise comparisons ASCO/Mod-IQWiG$_{\text{HR}}$, ASCO/IQWiG$_{\text{RR}}$, and Mod-IQWiG$_{\text{HR}}$/IQWiG$_{\text{RR}}$ (green, red, and violet line) are not influenced by increasing and decreasing hazards and hence are very similar to the Standard Scenario (bottom row of panels).

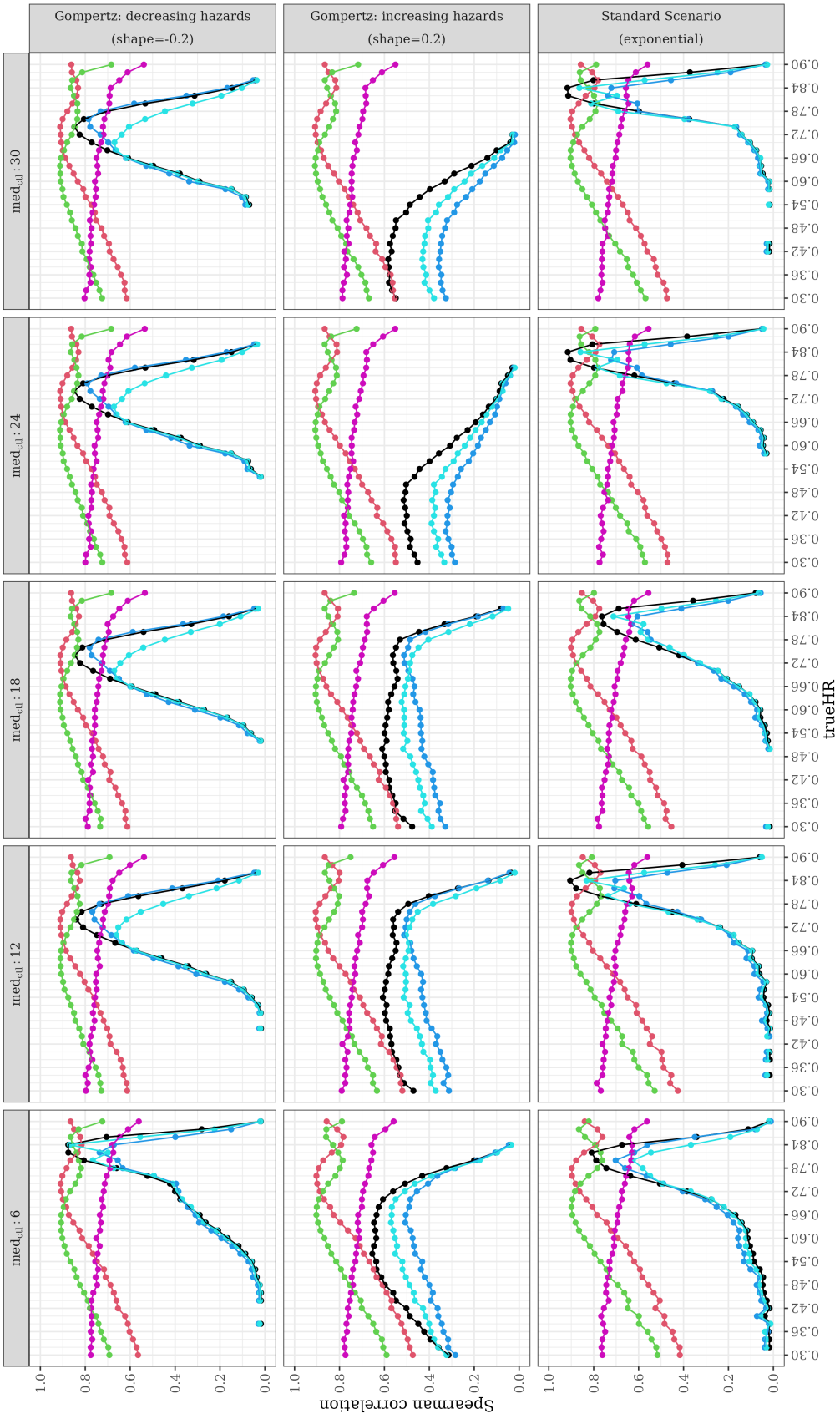

Figure 13: *Pairwise Spearman correlation (y-axis) separated by trueHR (x-axis) of Scenario 3 with Gompertz failure time distribution, p_C=60% and 90% power.*

Each row of panel stands for different shape parameter of the Gompertz distribution and each coloumn of panel stands for different med_{ctl}. The bottom row is the same as the top row of Figure 11. Abbreviations: med_{ctl}: Median survival time in the control group, p_C: Censoring rate, trueHR: True underlying Hazard Ratio of data generation

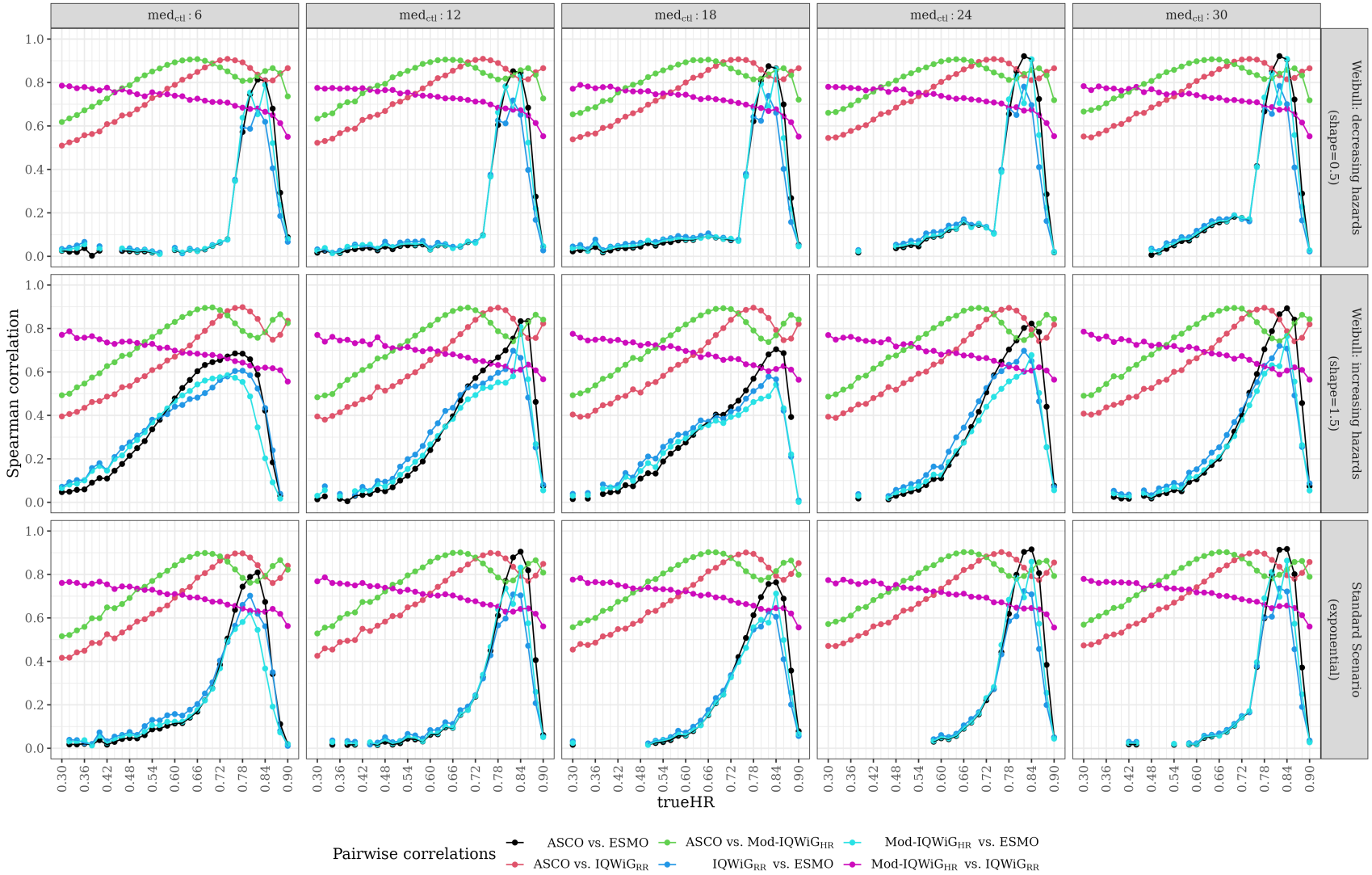

Figure 14: *Pairwise Spearman correlation (y-axis) separated by trueHR (x-axis) of Scenario 3 with Weibull failure time distribution, p_C=60% and 90% power.*

Each row of panel stands for different shape parameter of the Weibull distribution and each coloumn of panel stands for different med$_{ctl}$. The bottom row is the same as the top row of Figure 11. Abbreviations: med$_{ctl}$: Median survival time in the control group, p_C: Censoring rate, trueHR: True underlying Hazard Ratio of data generation

3.1.2.4 Influence of informative censoring due to treatment (Scenario 7)

Figure 15 shows the results of informative censoring on the pairwise Spearman correlation between the methods for each sub-scenario. The panels in the middle row show the sub-scenarios with a larger censoring rate in the treatment group and the bottom row of panels the sub-scenarios with a larger censoring rate in the control group. The top row illustrates again the results of the Standard Scenario, i.e. without informative censoring and hence equal censoring rates in both treatment groups.

In case of a larger censoring rate in the treatment group, all pairwise Spearman correlations are over the complete range of trueHR reduced compared to the Standard Scenario. For example, the Spearman correlation increases with smaller treatment effect for all pairwise ESMO comparisons as for the Standard Scenario. The reached maximum is, however, reduced. Moreover, with decreasing treatment effect the Mod-IQWiG$_{\text{HR}}$/IQWiG$_{\text{RR}}$ correlation is reduced as in the Standard Scenario but the reduction with present informative censoring is stronger. ASCO/IQWiG$_{\text{RR}}$ does increase with decreasing treatment effect, but from a trueHR approximately greater than 0.6 the correlation is reduced strongly, which is different to the Standard Scenario.

In case of a larger censoring rate in the control group, all pairwise ESMO comparisons do not reach the maxima of the Standard Scenario. Hence, the overall Spearman correlation is reduced. For the other three pairwise comparisons the overall Spearman correlation stays similar to the Standard Scenario as already shown in Figure 10. The behavior over the range of the treatment effects, however, is different. For ASCO/Mod-IQWiG$_{\text{HR}}$ and ASCO/IQWiG$_{\text{RR}}$, the correlation with a large treatment effect is stronger than in the Standard Scenario. With small treatment effects the correlation is smaller than in the Standard Scenario. Hence, these divergence behaviors cancel each other out and lead to similar Spearman correlation when combining all sub-scenarios.

Important to mention is, that with a larger censoring rate in the control group and with small treatment effects, i.e. trueHR$>$0.82, all simulated studies did not show a statistically significant log rank test between the treatment groups. Hence, the additional benefit assessment methods and Spearman correlations between them could not be calculated and are missing in the corresponding panels.

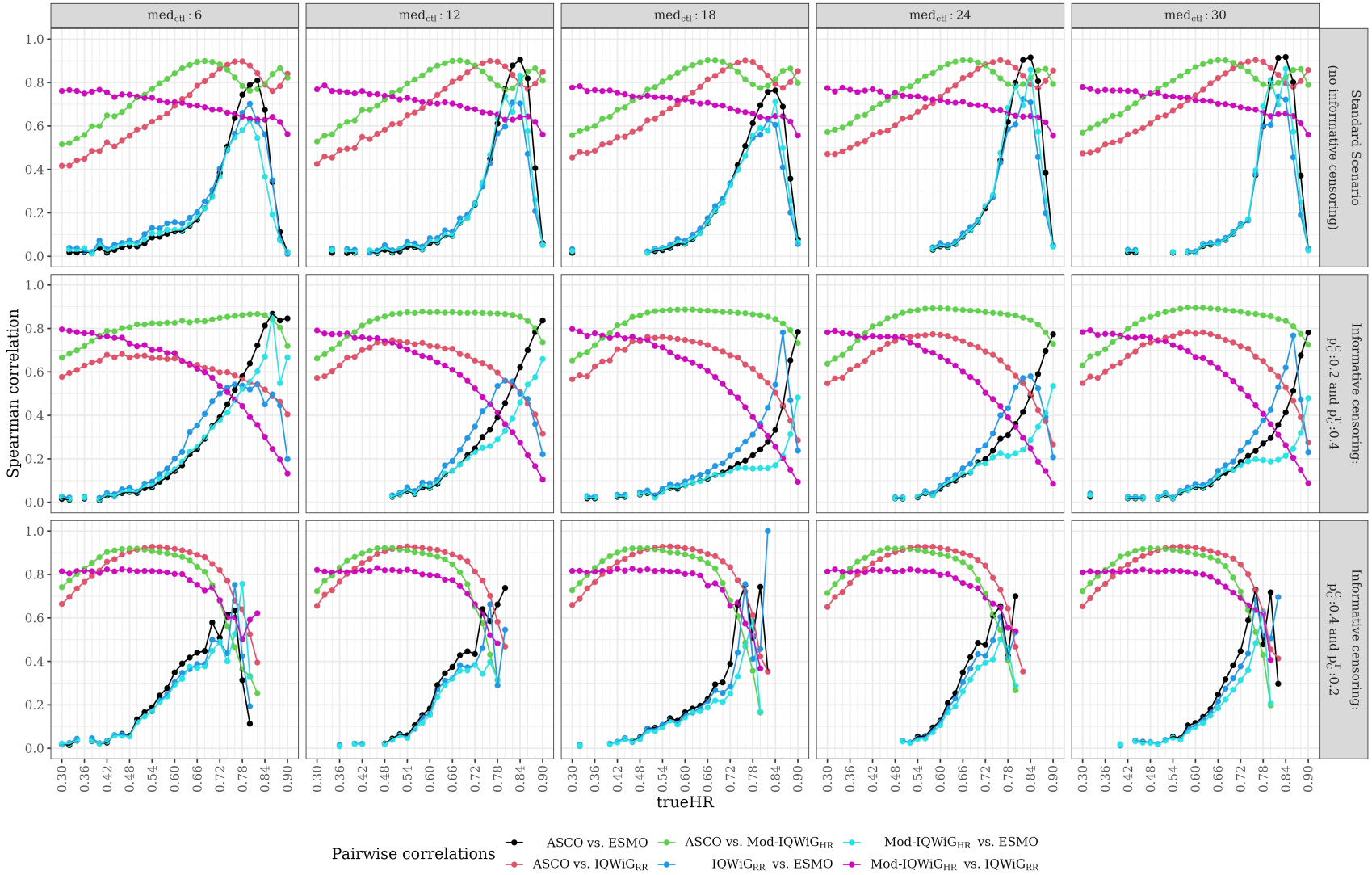

Figure 15: *Pairwise Spearman correlation (y-axis) separated by trueHR (x-axis) of Scenario 7 with 90% power.*

Each row of panel stands for different informative censoring sub-scenariosand each coloumn of panel stands for different med_{ctl}. The top row is the same as the top row of Figure 11, where p_C=60%. Abbreviations: med_{ctl}: Median survival time in the control group, p_C: Censoring rate, trueHR: True underlying Hazard Ratio of data generation, p_C^{ctl} and p_C^{trt}: Censoring rate of control and treatment group

3.1.2.5 Remaining scenarios

All other remaining scenarios like non-proportional hazards using late treatment effects for the treatment group (Scenario 4), unequal sample sizes (Scenario 5), and using exponential censoring distribution (Scenario 6) have very similar pairwise Spearman correlation results as in the Standard Scenario. Corresponding figures can be found in Appendix A.1.

3.1.3 ROC

In this Section, the results for ROC curves and the AUCs of the ROC curves of all ordinal additional benefit assessment methods and all related statistical quantities are shown to evaluate which statistical quantity is better suited for the additional benefit assessment of new treatments. Furthermore, TPR and FPR results of statistical quantities of specific sub-scenarios are shown as additional information. FPR and TPR were estimated for thresholds ranging from 0.2 to 1 which are used for defining a maximal additional benefit classification using HR-PE, HR^{-}, and HR^{+}. In this context, a true positive and false positive event is defined as deserved classification of a maximal category, or, respectively, not deserved classification of a maximal category. Furthermore, a ground truth was needed for the estimation of TPR and FPR but since no gold standard for additional benefit assessment method exists, a maximal category was assumed to be justified if trueHR $< \delta_{deserved}$ was met for different cut-offs values of $\delta_{deserved}$ (0.7, 0.75, and 0.8). For the TPR and FPR calculation all sub-scenarios - e.g. different treatment effects (designHR), different censoring rates, etc. - were combined (see Section 2.3.1.4 for more details).
Since the results of sub-scenarios with different underlying censoring rates are very similar, the following results are focused on sub-scenarios with a censoring rate of 60% (p_C=60%) and a power of 90%. The results for other assumed censoring rates, i.e. 20% and 40%, and a power of 80% can be found in Appendix A.1.

3.1.3.1 Standard Scenario

Figure 16 shows the ROC curves with estimated FPR and TPR of the Standard Scenario with a power of 90%, med_{ctl} of 6 months, and p_C of 60%. The different ROC lines represents the three statistical quantities (HR-PE, HR^{-}, and HR^{+}) used for additional benefit. Furthermore, the TPR and FPR of each of the additional benefit assessment methods with ordinal

outcome is shown as well (ESMO, Mod-IQWiG$_{\text{HR}}$, IQWiG$_{\text{RR}}$).

HR$^-$ (blue line) shows over the complete range of thresholds for a theoretical maximal category the best ROC curve meaning that it is closest to the perfect classifier with 0% FPR and 100% TPR (top left corner of each panel). The next best statistical quantity is HR-PE (yellow line) followed by HR$^+$ (black line). This described behavior of the three statistical quantities can be seen for all δ_{deserved} values, i.e. for all panels.

Important to note is that the choice of the threshold used to define a maximal category is more important than the statistical quantity itself. For example, in Figure 16 and defining HR$^+$<0.75 as maximal category would lead to a small FPR of 0.0357, 0.0124 or 0.0006 for δ_{deserved} equal to 0.7, 0.75 or 0.8, respectively. On the other hand defining HR$^-$<0.75 as maximal category would lead to a large FPR of 0.6956, 0.6195 or 0.3967 for δ_{deserved} equal to 0.7, 0.75 or 0.8, respectively. Hence, even though HR$^-$ might be in general the better statistical quantity for additional benefit assessment regarding ROC curves, the threshold used for the definition of a maximal category is equally important.

ESMO has the largest TPR of the three additional benefit assessment methods with a categorical outcome. FPR, however, is also large compared to IQWiG$_{\text{RR}}$ and Mod-IQWiG$_{\text{HR}}$, except for δ_{deserved}=0.8 (right panel) where the FPR is quite similar for all three methods (IQWiG$_{\text{RR}}$: 0.0805, Mod-IQWiG$_{\text{HR}}$: 0.0060, and ESMO: 0.0242). Thus, ESMO is the most liberal method, while Mod-IQWiG$_{\text{HR}}$ has the lowest FPR and TPR for all δ_{deserved} values and hence can be interpreted as the most conservative method.

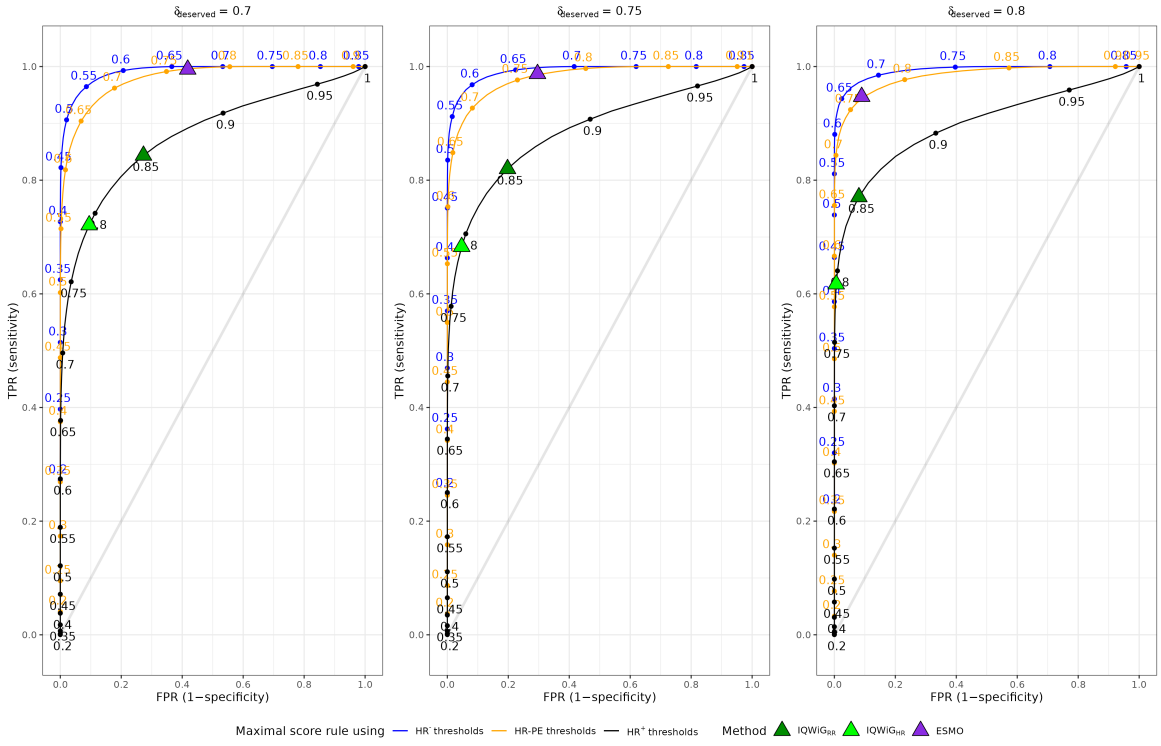

Figure 16: *ROC curves of Standard Scenario with p_C=60%, med_{ctl}=6, and power of 90%.*
For each panel the sub-scenarios with designHRs ranging from 0.3 until 0.9 were used for FPR and TPR estimation, where thresholds ranging from 0.2 to 1 were used for defining a maximal additional benefit classification using HR-PE, HR^-, and HR^+. For TRP and FPR calculation a ground truth was needed: A maximal category was assumed to be justified if trueHR < $\delta_{deserved}$ was met for different cut-offs values of $\delta_{deserved}$ (0.7, 0.75, and 0.8). In addition, TPR and FPR of all ordinal additional benefit assessment methods were calculated. Each panel stands for different $\delta_{deserved}$ values. Abbreviations: FPR: False Positive Rate, HR-PE: Hazard Ratio Point Estimate, HR^-: Lower 95% confidence interval limit of the HR-PE, HR^+: Upper 95% confidence interval limit of the HR-PE, med_{ctl}: Median survival time in the control group, p_C: Censoring rate, trueHR: True underlying Hazard Ratio of data generation,ROC: Receiver Operating Characteristic, TPR: True Positive Rate, trueHR: True underlying Hazard Ratio of data generation, $\delta_{deserved}$: Ground truth of deserved maximal category for TPR and FPR calculation (justified if trueHR < δ)

If one would define that the FPR should not be larger than 5% or 10% - similar to the used significance level of 5% in clinical studies for controlling the type-I-error rate - HR^- still obtains higher TPR values for different δ_{deserved} compared to HR^+ and HR-PE. This is illustrated in detail in Table 9, where the TPR results of Figure 16 for all three statistical quantities under the condition that FPR is smaller than 5% or 10%, are shown. Furthermore, it can be seen that the threshold defining a maximal category to achieve a FPR of less than 5% or 10% and a δ_{deserved} of 0.8 leads to very similar threshold values as used in the additional benefit assessment methods of IQWiG_{RR} (δ_{deserved} = 0.8: 0.83 or 0.85; used threshold in method:

0.85) and ESMO ($\delta_{\text{deserved}} = 0.8$: 0.66 or 0.68; used threshold in method: 0.65 and 0.7). Nevertheless, the larger TPR values for HR^- and hence ESMO, once again illustrates that $IQWiG_{RR}$ is the more conservative method and ESMO the more liberal one.

In addition, Table 10 illustrates a similar consideration as Table 9, where the minimal needed thresholds for all three statistical quantities under the condition that FPR is smaller than 5% or 10% in all sub-scenarios of the Standard Scenario are shown. Using these resulting thresholds the corresponding TPR average and range (min, max) are shown as well. The results are very similar to Table 9 and reinforce the described behavior from above: Using δ_{deserved} of 0.8 leads to a very similar threshold value as used in the additional benefit assessment method of $IQWiG_{RR}$. The needed threshold for HR^-, however, is slightly reduced but still similar to the threshold used in ESMO. Moreover, HR^- has still larger TPR values than HR^+.

Table 9: *Threshold and corresponding TPR allowing only a FPR of 5% or 10% of the Standard Scenario with power of 90%, med_{ctl} of 6 months, and p_C of 60%*

δ_{deserved}	Method	FPR $\leqslant$ 5%		FPR $\leqslant$ 10%	
		Threshold	TPR	Threshold	TPR
0.7	HR^-	0.52	0.93	0.55	0.96
0.7	HR-PE	0.63	0.87	0.66	0.92
0.7	HR^+	0.76	0.65	0.79	0.72
0.75	HR^-	0.58	0.95	0.60	0.97
0.75	HR-PE	0.68	0.90	0.70	0.93
0.75	HR^+	0.79	0.68	0.81	0.73
0.8	HR^-	0.66	0.95	0.68	0.97
0.8	HR-PE	0.74	0.91	0.76	0.94
0.8	HR^+	0.83	0.72	0.85	0.77

Abbreviations: FPR: False positive rate, HR: Hazard ratio, HR^-: Lower 95% confidence interval limit of the HR-PE, HR^+: Upper 95% confidence interval limit of the HR-PE, med_{ctl}: Median survival time of control group, p_C: Overall censoring rate, PE: Point estimate, TPR: True positive rate, trueHR: True underlying HR of data generation, $\delta_{deserved}$: Ground truth of deserved maximal category for TPR and FPR calculation (justified if trueHR < δ)

Table 10: *Minimal threshold and corresponding TPR given as mean and range (min, max) allowing only a FPR of 5% or 10% for the complete Standard Scenario*

δ_{deserved}	Method	FPR $\leqslant$ 5%		FPR $\leqslant$ 10%	
		Threshold	TPR	Threshold	TPR
0.7	HR^-	0.50	0.90 (0.87, 0.94)	0.53	0.94 (0.91, 0.97)
0.7	HR-PE	0.62	0.83 (0.79, 0.87)	0.65	0.88 (0.85, 0.92)
0.7	HR^+	0.76	0.52 (0.45, 0.65)	0.79	0.60 (0.52, 0.72)
0.75	HR^-	0.55	0.91 (0.88, 0.94)	0.58	0.95 (0.92, 0.97)
0.75	HR-PE	0.66	0.84 (0.81, 0.88)	0.69	0.90 (0.87, 0.93)
0.75	HR^+	0.79	0.56 (0.49, 0.68)	0.81	0.61 (0.54, 0.73)
0.8	HR^-	0.64	0.93 (0.90, 0.95)	0.66	0.95 (0.93, 0.97)
0.8	HR-PE	0.73	0.88 (0.85, 0.91)	0.75	0.91 (0.88, 0.93)
0.8	HR^+	0.83	0.61 (0.54, 0.72)	0.85	0.67 (0.60, 0.77)

Abbreviations: FPR: False positive rate, HR: Hazard ratio, HR^-: Lower 95% confidence interval limit of the HR-PE, HR^+: Upper 95% confidence interval limit of the HR-PE, PE: Point estimate, TPR: True positive rate, trueHR: True underlying HR of data generation, $\delta_{deserved}$: Ground truth of deserved maximal category for TPR and FPR calculation (justified if trueHR $< \delta$)

The AUC values of the ROC curves for all sub-scenarios of the Standard Scenario including the one shown in Figure 16 are shown in a nested loop-plot (see Figure 17). This nested loop-plot is split into two panels, one for sub-scenarios with 90% and one for 80% power. In each panel a nested loop plot is shown, meaning that the simulation results are reordered into a lexicographical order. Hence, the results (AUC values) are arranged consecutively on the horizontal axis and the criterion for the different sub-scenarios is presented on the vertical axis (p_C and med_{ctl}). Furthermore, the AUC results for each statistical quantity used for the additional benefit assessment are highlighted in different colors and the AUC results for the three δ_{deserved} values are connected with each other for each of the sub-scenarios.

As already described above this figure underlines that HR^- has the largest AUC value with a mean of 0.9925 and hence ROC curve closest to 0% FPR and 100% TPR over all sub-scenarios. HR-PE has consistent smaller AUC values with a mean of 0.9809, which is still quite similar to HR^-. HR^+, however, shows the lowest AUC values with a mean of 0.8553 (slightly increasing with larger δ_{deserved}). Overall, other sub-scenarios with different med_{ctl}, p_C, and power of the Standard Scenario show no differences in AUC values for the statistical quantities and hence, the above described pattern between the statistical quantities stays the same (see Figure 17).

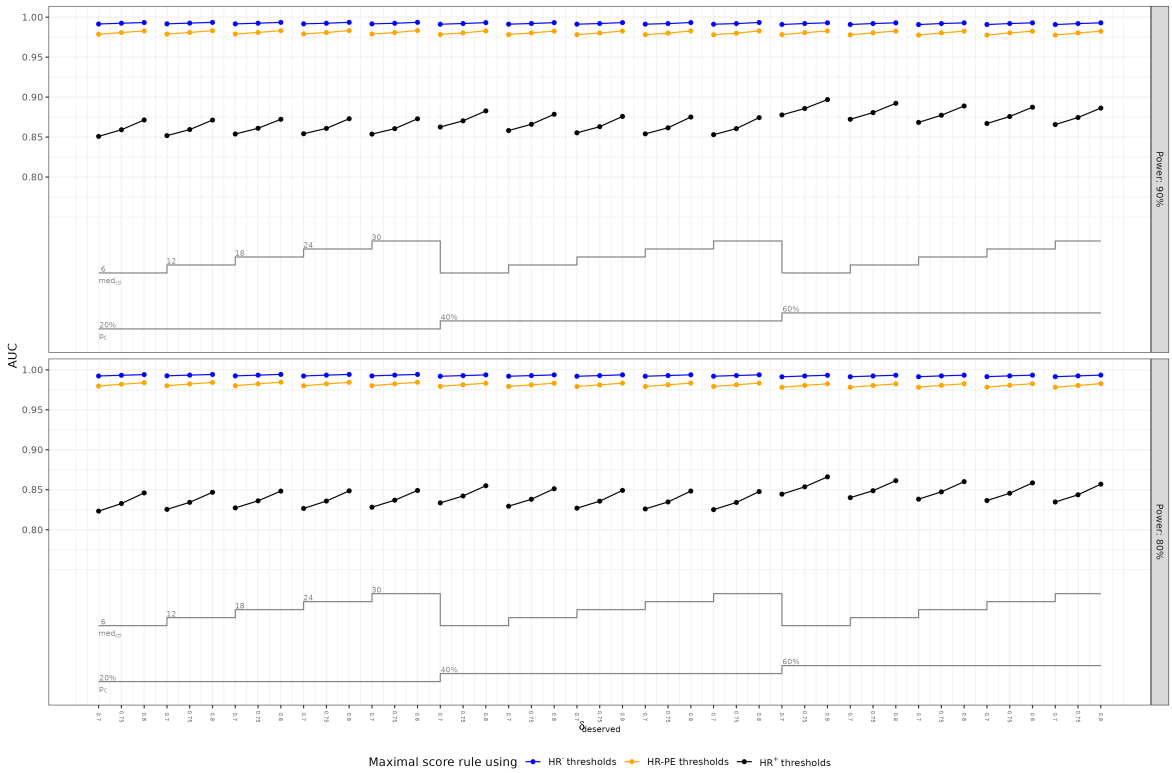

Figure 17: *AUC of ROC curves (y-axis) separated by $\delta_{deserved}$ (x-axis) of Standard Scenario for each sub-scenario.*

Each panel stands for a different power. Abbreviations: AUC: Area Under the Curve, HR-PE: Hazard Ratio Point Estimate, HR^{-}: Lower 95% confidence interval limit of the HR-PE, HR^{+}: Upper 95% confidence interval limit of the HR-PE, med_{ctl}: Median survival time in the control group, p_C: Censoring rate, trueHR: True underlying Hazard Ratio of data generation, ROC: Receiver Operating Characteristic, $\delta_{deserved}$: Ground truth of deserved maximal category for TPR and FPR calculation (justified if trueHR < δ)

Using constant sample size:

Figure 18 shows the ROC with estimated FPR and TPR of the same sub-scenario of Standard Scenario as above (power: 90%, med_{ctl}=6, and p_C=0.6). However, to examine, whether the sample size calculation leads to the above described differences of the statistical quantities, a simulation with constant sample size was performed. A sample size of 500 per group, independent of the actual treatment effect, was used for each sub-scenario. This resulted in exactly the same ROC curves for all three quantities, i.e. all three lines overlap (blue, yellow, and black). To reach the same FPR and TPR for the different statistical quantities different thresholds defining the maximal category have to be chosen. All statistical quantities show the same AUC values within each sub-scenario. The corresponding AUC results of all sub-scenarios of the Standard Scenario, i.e. the nested loop plot, can be found in Appendix A.1 (Figure 33).

ESMO is again the most liberal method with the largest TPR and FPR. IQWiG$_{\text{RR}}$ and Mod-IQWiG$_{\text{HR}}$ have lower FPR with slightly reduced TPR values. For example, with a δ_{deserved} of 0.8, ESMO has TPR and FPR of 0.97 and 0.60, while IQWiG$_{\text{RR}}$ has 0.87 and 0.11, respectively.

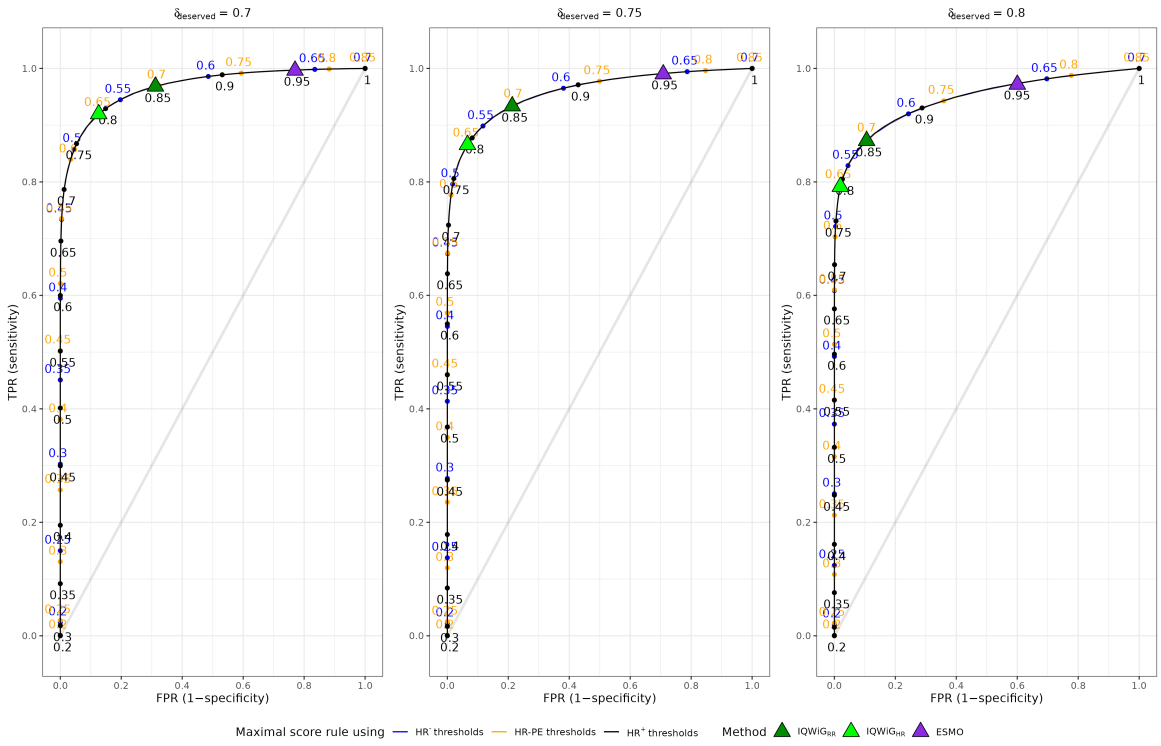

Figure 18: *ROC curves of the Standard Scenario with p_C=60%, med_{ctl}=6, and power of 90% with constant sample size.*

No sample size calculation for data generation was performed. Instead a sample size of 500 per group was used for each sub-scenario. For each panel the sub-scenarios with designHRs ranging from 0.3 until 0.9 were used for FPR and TPR estimation, where thresholds ranging from 0.2 to 1 were used for defining a maximal additional benefit classification using HR-PE, HR$^-$, and HR$^+$. For TRP and FPR calculation a ground truth was needed: A maximal category was assumed to be justified if trueHR $< \delta_{deserved}$ was met for different cut-offs values of $\delta_{deserved}$ (0.7, 0.75, and 0.8). In addition, TPR and FPR of all ordinal additional benefit assessment methods were calculated. Each panel stands for different $\delta_{deserved}$ values. Abbreviations: ESMO: European Society for Medical Oncology, FPR: False Positive Rate, HR-PE: Hazard Ratio Point Estimate, HR$^-$: Lower 95% confidence interval limit of the HR-PE, HR$^+$: Upper 95% confidence interval limit of the HR-PE, IQWiG: Institute for quality and efficiency in health care, IQWiG$_{RR}$: Original IQWiG method, med_{ctl}: Median survival time in the control group, Mod-IQWiG$_{HR}$: Modified IQWiG method using upper confidence interval limit based on IQWiG thresholds (transformation into HR based thresholds using the conversion formula proposed by VanderWeele (2020), p_C: Censoring rate, ROC: Receiver Operating Characteristic, TPR: True Positive Rate, trueHR: True underlying Hazard Ratio of data generation, $\delta_{deserved}$: Ground truth of deserved maximal category for TPR and FPR calculation (justified if trueHR $< \delta$)

Reason for this difference in TPR, FPR, and corresponding AUC values between with sample size calculation and with constant sample size can be attributed to the difference in precise estimation of HR, HR^-, and HR^+. Therefore, Figure 19 shows these estimations regarding the simulation with sample size calculation and with constant sample size for the same sub-scenario as above (power: 90%, med_{ctl}=6, and p_C=0.6) using only significant studies. Hence, especially in sub-scenarios with small treatment effects and with constant sample size (right panel of Figure 19) the remaining significant studies for the estimations is reduced, leading to not as precise estimations. For example, with a trueHR of 0.9, the estimated median of HR is 0.78.

Furthermore, in the case of constant sample size (right panel), the slope of the estimations of all three statistical quantities stays very similar over the range of trueHRs. In case of performed sample size calculation (left panel), the slope is different between the statistical quantities. Hence, combining all these sub-scenarios of the right panel for FPR and TPR calculation lead to similar results. For example, defining a HR^-<0.53 as maximal category leads to a TPR of 0.9159 and FPR 0.1196 (for $\delta_{deserverd}$=0.7). For HR-PE and HR^+ different thresholds can be found that lead to similar TPR and FPR values: For HR-PE<0.65 and HR^+<0.79 a TPR of 0.9213 and 0.9187 as well as a FPR of 0.1302 and 0.1238 are present, respectively. The threshold values of 0.53, 0.65, and 0.79 can also be seen to be very close in the ROC curve of this sub-scenario (Figure 18, left panel). In the performed simulation the used thresholds for defining a maximal category ranged from 0.2 to 1 in 0.01 equidistant steps. If the step sequence would be smaller, the remaining difference in TPR and FPR would also be smaller.

In case a sample size calculation is performed (left panel), the range (i.e. size of the boxplots) of the three statistical quantity estimates are larger for large treatment effects, as the sample size is smaller, and vice versa in case of small treatment effects. Hence, over the range of trueHR the estimates are not following a similar slope as for the simulation with constant sample size (right panel). HR^- and HR-PE estimations have the steepest increase with decreasing treatment effect (increasing trueHR). HR^+ estimation does not increase as steep as the other two quantities. Hence, combining all these sub-scenarios of the left panel for FPR and TPR calculation lead to different results, as with steeper increase of the estimates, it is easier to separate the sub-scenarios in a more distinct way.

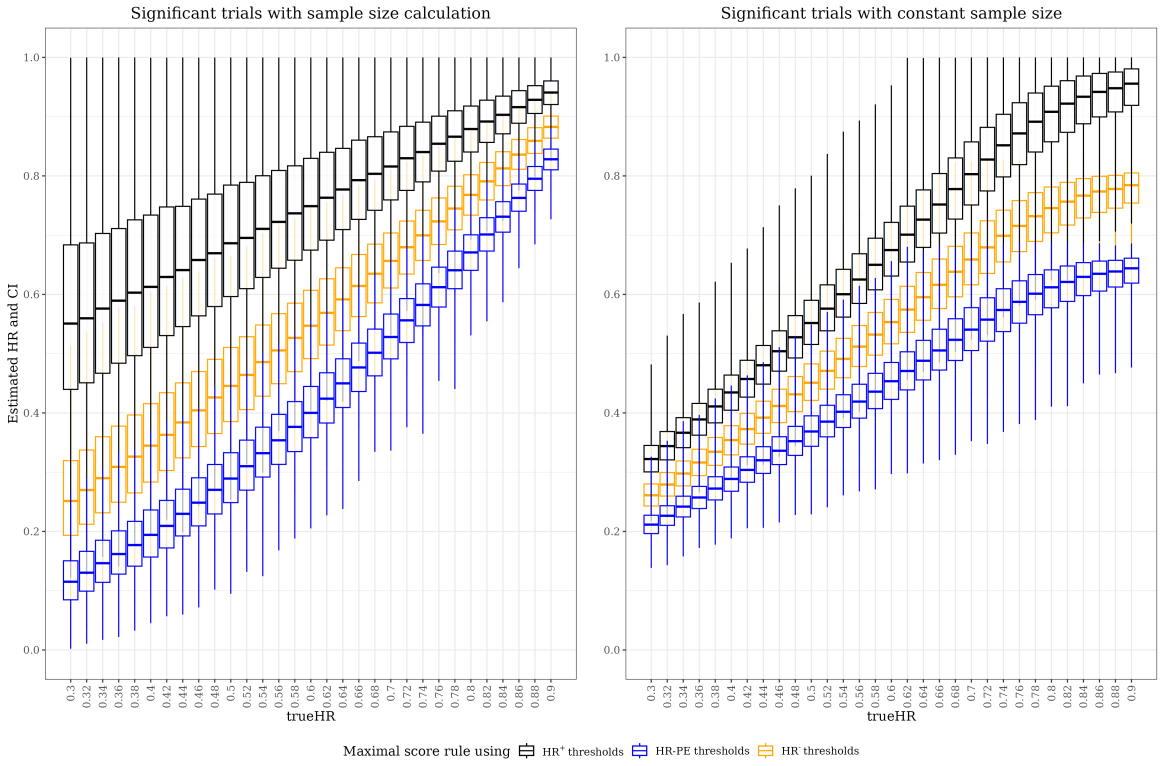

Figure 19: *Description of HR-PE, HR^-, and HR^+ estimation distribution (y-axis) separated by trueHR (x-axis) using boxplots with sample size calculation (left panel) and with constant sample size (right panel) of the Standard Scenario with power of 90%, med_{ctl} = 6, and p_C = 60%*

In case of no performed sample size calculation for the data generation, a sample size of 500 per group was used for each sub-scenario. Only significant studies were used for HR-PE, HR^-, and HR^+ estimation. Abbreviations: CI: Confidence Interval, HR-PE: Hazard Ratio Point Estimate, HR^-: Lower 95% confidence interval limit of the HR-PE, HR^+: Upper 95% confidence interval limit of the HR-PE, med_{ctl}: Median survival time in the control group, p_C: Censoring rate, trueHR: True underlying Hazard Ratio of data generation

3.1.3.2 Remaining scenarios

All other remaining scenarios have generally no influence on TPR, FPR, and corresponding ROC as well as AUC values compared to the Standard Scenario. Corresponding figures can be found in Appendix A.1.

One exception is present for Scenario 2 with overpowered studies, i.e. $HR_{var}<1$. Here, the FPR of $IQWiG_{RR}$ is strongly increased compared to the Standard Scenario (Figure 20), e.g. for $HR_{var}=0.9$ and $\delta_{deserved} = 0.8$ from 0.0805 to 0.7357. ESMO and Mod-$IQWiG_{HR}$ have increased FPR as well, e.g. for $HR_{var}=0.9$ and $\delta_{deserved}=0.8$ from 0.0896 to 0.1537 and from 0.0060 to 0.0539, respectively. These increases, however, are not as strong. Overall,

Mod-IQWiG$_{\text{HR}}$ is still the most conservative method. The general comparison between the statistical quantities is not affected, meaning that HR^{-} is still closest to the perfect classifier with 0% FPR and 100% TPR (top left corner of each panel).

Similar different behavior compared to the Standard Scenario can also be seen in cases with informative censoring (Scenario 7), where the treatment group has a larger censoring rate than the control group. The results are shown in Figure 21. All methods have again an increased FPR compared to the Standard Scenario, while IQWiG$_{\text{RR}}$ is influenced the strongest, e.g. 0.0805 vs. 0.9134 for $\delta_{\text{deserved}} = 0.8$, and Mod-IQWiG$_{\text{HR}}$ is still the most conservative method. The general comparison between the statistical quantities is again not affected.

Furthermore, different underlying failure time distributions (Scenario 3) with increasing hazards over time reduce TPR of ESMO drastically. For example, in the same sub-scenario as shown in Figure 16 of the Standard Scenario and with $\delta_{\text{deserved}} = 0.8$ the TPR is reduced by 0.2451 (0.9470 vs. 0.7019) for Weibull distributed failure times (shape=1.5). IQWiG$_{\text{RR}}$ and Mod-IQWiG$_{\text{HR}}$ are hardly influenced by these distributions leading to similar TPR and FPR values, e.g. in the same sub-scenario: 0.7704 vs. 0.7823 (TPR) and 0.0805 vs. 0.0826 (FPR) for IQWiG$_{\text{RR}}$, respectively, and 0.6170 vs. 0.6333 (TPR) and 0.0060 vs. 0.0056 (FPR) for Mod-IQWiG$_{\text{HR}}$. The general comparison between the statistical quantities is again not affected. In addition, in case of decreasing hazards over time, similar results as for the Standard Scenario are present.

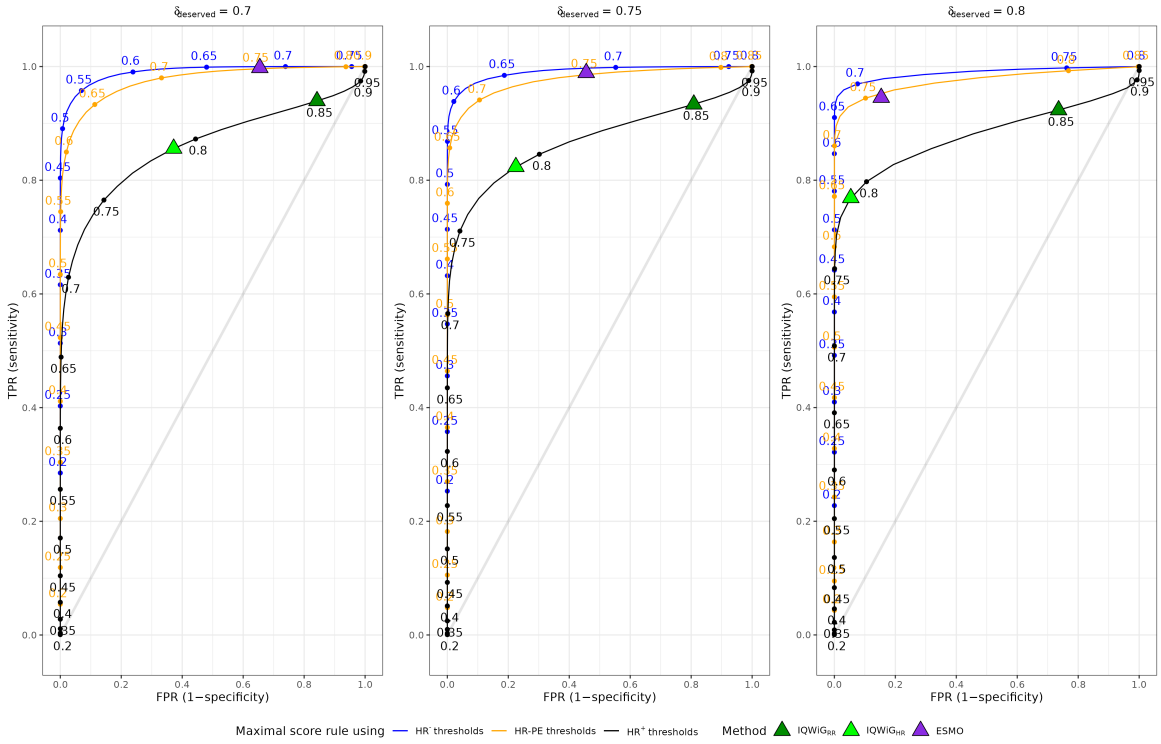

Figure 20: *ROC curves of Scenario 2 with p_C=60% med_{ctl}=6, HR_{var}=0.9 (overpowered studies), and power of 90%.*

For each panel the sub-scenarios with designHRs ranging from 0.3 until 0.9 were used for FPR and TPR estimation, where thresholds ranging from 0.2 to 1 were used for defining a maximal additional benefit classification using HR-PE, HR^-, and HR^+. For TRP and FPR calculation a ground truth was needed: A maximal category was assumed to be justified if trueHR < $\delta_{deserved}$ was met for different cut-offs values of $\delta_{deserved}$ (0.7, 0.75, and 0.8). In addition, TPR and FPR of all ordinal additional benefit assessment methods were calculated. Each panel stands for different $\delta_{deserved}$ values. Abbreviations: designHR: Design Hazard Ratio used for sample size calculation, FPR: False Positive Rate, HR-PE: Hazard Ratio Point Estimate, HR^-: Lower 95% confidence interval limit of the HR-PE, HR^+: Upper 95% confidence interval limit of the HR-PE, HR_{var}: Factor for deviance between designHR and trueHR, med_{ctl}: Median survival time in the control group, p_C: Censoring rate, ROC: Receiver Operating Characteristic, TPR: True Positive Rate, trueHR: True underlying Hazard Ratio of data generation, $\delta_{deserved}$: Ground truth of deserved maximal category for TPR and FPR calculation (justified if trueHR < δ)

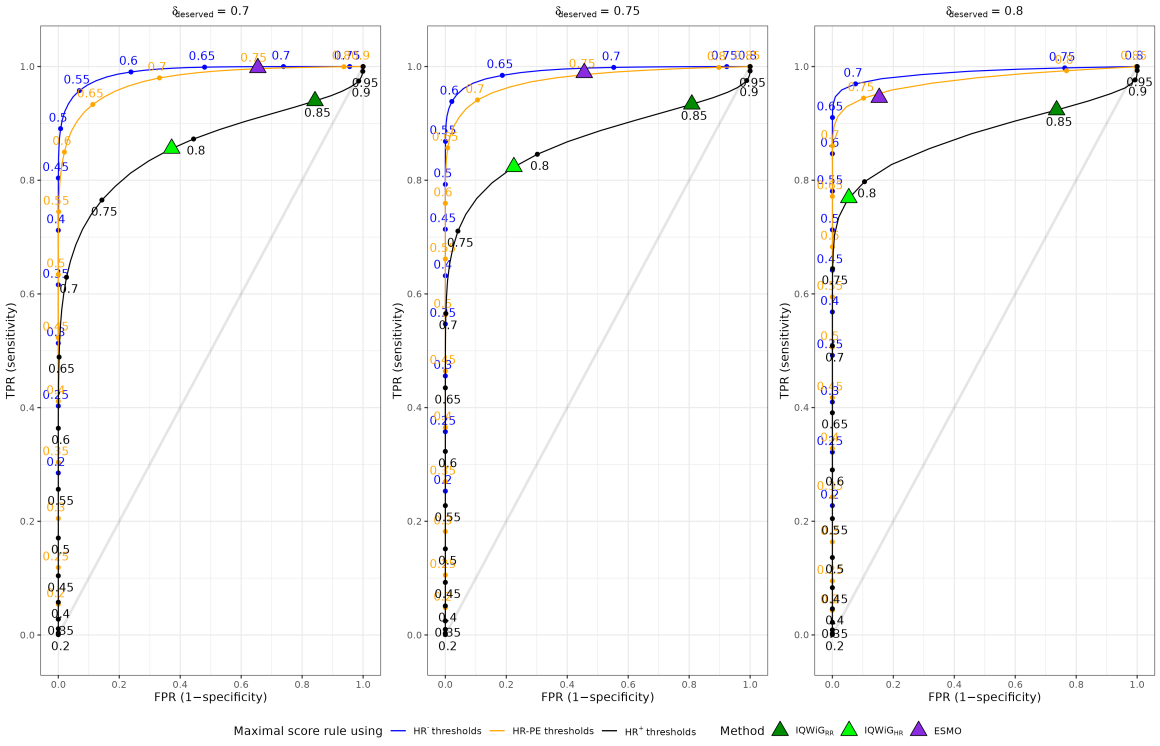

Figure 21: *ROC curves of Scenario 7 with* med_{ctl}*=6,* $p_C^{ctl} < p_C^{trt}$ *(*p_C^{ctl}*=20%,* p_C^{trt}*=40%), and power of 90%.*

For each panel the sub-scenarios with designHRs ranging from 0.3 until 0.9 were used for FPR and TPR estimation, where thresholds ranging from 0.2 to 1 were used for defining a maximal additional benefit classification using HR-PE, HR^-*, and* HR^+*. For TRP and FPR calculation a ground truth was needed: A maximal category was assumed to be justified if trueHR <* $\delta_{deserved}$ *was met for different cut-offs values of* $\delta_{deserved}$ *(0.7, 0.75, and 0.8). In addition, TPR and FPR of all ordinal additional benefit assessment methods were calculated. Each panel stands for different* $\delta_{deserved}$ *values. Abbreviations: FPR: False Positive Rate, HR-PE: Hazard Ratio Point Estimate,* HR^-*: Lower 95% confidence interval limit of the HR-PE,* HR^+*: Upper 95% confidence interval limit of the HR-PE,* med_{ctl}*: Median survival time in the control group,* p_C^{ctl} *and* p_C^{trt}*: Censoring rate of control and treatment group, ROC: Receiver Operating Characteristic, TPR: True Positive Rate, trueHR: True underlying Hazard Ratio of data generation,* $\delta_{deserved}$*: Ground truth of deserved maximal category for TPR and FPR calculation (justified if trueHR <* δ*)*

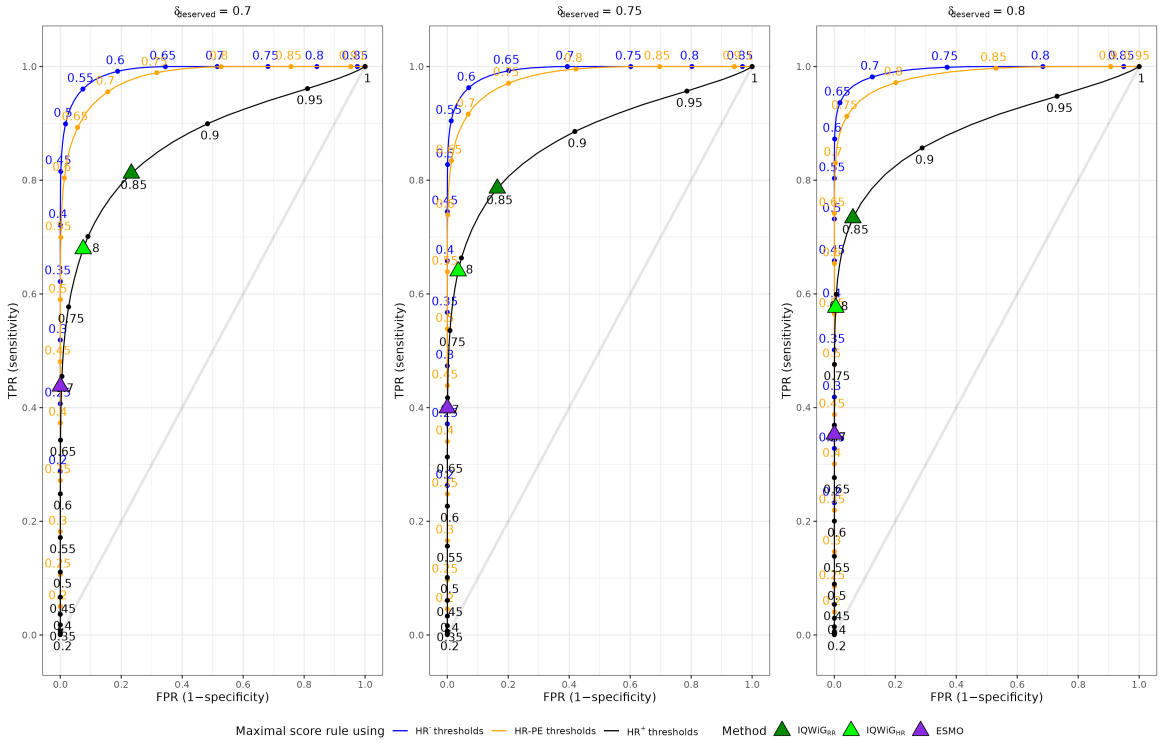

Figure 22: *ROC curves of Scenario 3 with Gompertz failure time distribution (shape of 0.2: increasing hazards over time), med_{ctl}=6, p_C=60%, and power of 90%.*

For each panel the sub-scenarios with designHRs ranging from 0.3 until 0.9 were used for FPR and TPR estimation, where thresholds ranging from 0.2 to 1 were used for defining a maximal additional benefit classification using HR-PE, HR^-, and HR^+. For TRP and FPR calculation a ground truth was needed: A maximal category was assumed to be justified if trueHR < $\delta_{deserved}$ was met for different cut-offs values of $\delta_{deserved}$ (0.7, 0.75, and 0.8). In addition, TPR and FPR of all ordinal additional benefit assessment methods were calculated. Each panel stands for different $\delta_{deserved}$ values. Abbreviations: FPR: False Positive Rate, HR-PE: Hazard Ratio Point Estimate, HR^-: Lower 95% confidence interval limit of the HR-PE, HR^+: Upper 95% confidence interval limit of the HR-PE, med_{ctl}: Median survival time in the control group, p_C: Censoring rate, ROC: Receiver Operating Characteristic, TPR: True Positive Rate, trueHR: True underlying Hazard Ratio of data generation, $\delta_{deserved}$: Ground truth of deserved maximal category for TPR and FPR calculation (justified if trueHR < δ)

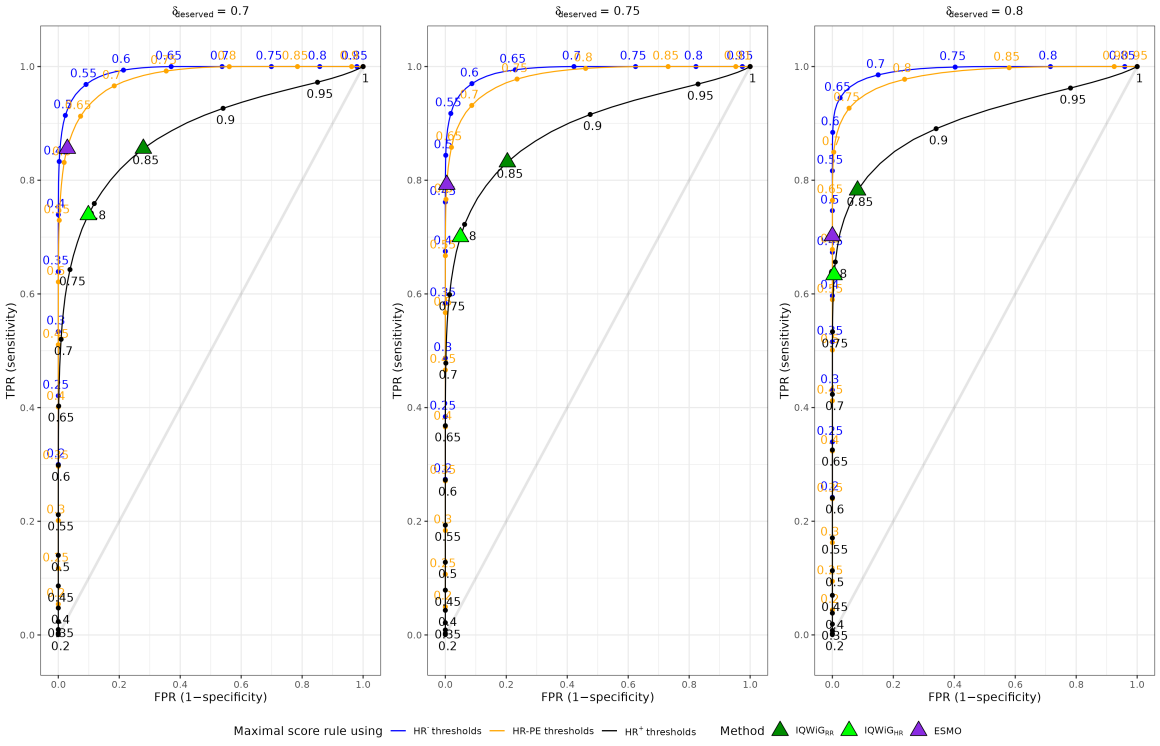

Figure 23: *ROC curves of Scenario 3 with Weibull failure time distribution (shape of 1.5: increasing hazards over time), med_{ctl}=6, p_C=60%, and power of 90%.*

For each panel the sub-scenarios with designHRs ranging from 0.3 until 0.9 were used for FPR and TPR estimation, where thresholds ranging from 0.2 to 1 were used for defining a maximal additional benefit classification using HR-PE, HR^-, and HR^+. For TRP and FPR calculation a ground truth was needed: A maximal category was assumed to be justified if trueHR < $\delta_{deserved}$ was met for different cut-offs values of $\delta_{deserved}$ (0.7, 0.75, and 0.8). In addition, TPR and FPR of all ordinal additional benefit assessment methods were calculated. Each panel stands for different $\delta_{deserved}$ values. Abbreviations: FPR: False Positive Rate, HR-PE: Hazard Ratio Point Estimate, HR^-: Lower 95% confidence interval limit of the HR-PE, HR^+: Upper 95% confidence interval limit of the HR-PE, med_{ctl}: Median survival time in the control group, p_C: Censoring rate, ROC: Receiver Operating Characteristic, TPR: True Positive Rate, trueHR: True underlying Hazard Ratio of data generation, $\delta_{deserved}$: Ground truth of deserved maximal category for TPR and FPR calculation (justified if trueHR < δ)

3.1.4 Optimal cutoff determination

The calculated ASCO cutoff values, which correspond to categories of ESMO, IQWiG$_{RR}$, and Mod-IQWiG$_{HR}$, using the maximizing weighted Cohens kappa approach, ROC01, and Svensson method are shown in Figure 24.

It can be seen that for ESMO all three ASCO cutoff values separating the ASCO scale in the four categories of ESMO are very similar over all simulated scenarios (top three panels), i.e. the cutoff values for the Standard Scenario are 17.06, 20.07, and 23.30 using the maximizing weighted Cohens kappa approach. This reflects the high rate of maximal categories and hence the liberal behavior of ESMO as the cutoff value for the maximal ESMO category is quite low and hence almost the complete range of ASCO values are categoriesed into the maximal ESMO category. Furthermore, with underlying Gompertz distributed failure times with increasing hazards, the cutoff values are increased to 37.72, 52.90, and 67.36, which logically increases the cutoff values of the overall Gompertz scenario to 25.79, 30.89, and 54.83. The sub-scenario with decreasing hazards is not affected and hence similar cutoff values as for the Standard Scenario are present. This reflects the unusual behavior of ESMO to the Gompertz distribution with increasing hazards as already mentioned in the sections above.

For IQWiG$_{RR}$ (middle two panels) and Mod-IQWiG$_{HR}$ (bottom two panels) the two cutoff values separating ASCO into three categories of IQWiG$_{RR}$ and Mod-IQWiG$_{HR}$ are more distinguished, e.g. 20.06 and 36.48 for IQWiG$_{RR}$ as well as 22.56 and 45.34 for Mod-IQWiG$_{HR}$ of the Standard Scenario using the maximizing weighted Cohens kappa approach. These values stay very similar over all scenarios.

In general, all methods used for the calculation of the ASCO cutoff values (maximizing weighted Cohens kappa, ROC01, and Svensson method) lead to very similar results. Only ROC01 results in larger ASCO cutoff values, especially for the cutoff value separating the lowest two categories of ESMO, IQWiG$_{RR}$, and Mod-IQWiG$_{HR}$ (left three panels). Cohens kappa approach and Svensson cutoff values are very similar for all additional benefit assessment methods and simulated scenarios.

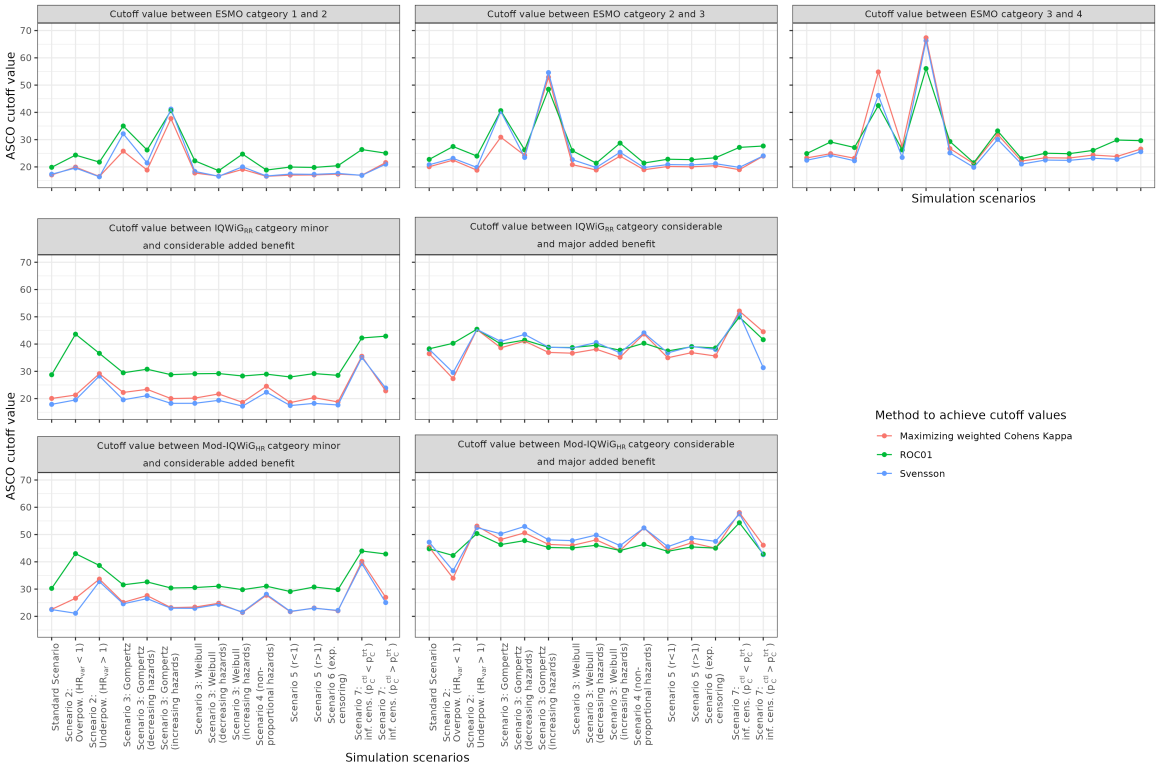

Figure 24: *Optimal ASCO cutoffs (y-axis) for the different scenarios (x-axis) where all sub-scenarios were combined for cutoff determination.*

Each row of panel stands for ASCO cutoff values between different categories of ordinal additional benefit assessment methods and each coloumn of panel stands for different methods. Abbreviations: designHR: Design Hazard Ratio used for sample size calculation, HR_{var}: Factor for deviance between designHR and trueHR, n_{ctl} and n_{trt}: Sample size of control and treatment group, p_C^{ctl} and p_C^{trt}: Censoring rate of control and treatment group, r: Allocation ratio (n_{trt}/n_{ctl}), trueHR: True underlying Hazard Ratio of data generation

3.1.5 Bias evaluation

As already mentioned in the ADEMP structure of this simulation (see Section 2.3.1.2), the data generation mechanism was defined with the goal to achieve a specific censoring rate using a combination of administrative and exponential censoring mechanisms. To achieve this, the censoring times were defined dependent on the generated event times, which could introduce bias into the simulation study. This possible bias was investigated with an additionally performed simulation. For this data of the Standard Scenario with n_{sim}= 10,000 and a constant sample size of 1,000 (500 per treatment group) for each sub-scenario was generated. The results of this assessment is shown in Table 11 and Figure 25. As anticipated the censoring mechanism for exponentially distributed censoring achieving a specific censoring rate introduced bias to the HR estimation. With larger simulated censoring rate (p_C) and smaller med_{ctl}, the HR bias was increased, i.e. sub-scenarios with $p_C = 0.6$ (blue line) and $med_{ctl} = 6$ months (left panel) have the largest HR bias. Furthermore, treatment effects of trueHR of around 0.5 lead to larger bias in the HR estimation. For example, with $med_{ctl} = 6$ months, $p_C = 0.6$, and power of 90%, the HR bias is -0.0378, -0.0471, and -0.0118 for trueHR of 0.3, 0.5, 0.9, respectively. Similar results can also be seen in Table 11, where all different trueHR of each sub-scenario are combined for the calculation of the mean HR bias and Monte Carlo SE of the HR bias. Here, the same results can be seen: With larger p_C and smaller med_{ctl}, the HR bias was increased.

Overall, the HR was slightly underestimated, i.e. the overall HR bias of the Standard Scenario, where all sub-scenarios were combined, was -0.0126 (Monte Carlo SE of HR bias: 0.000505).

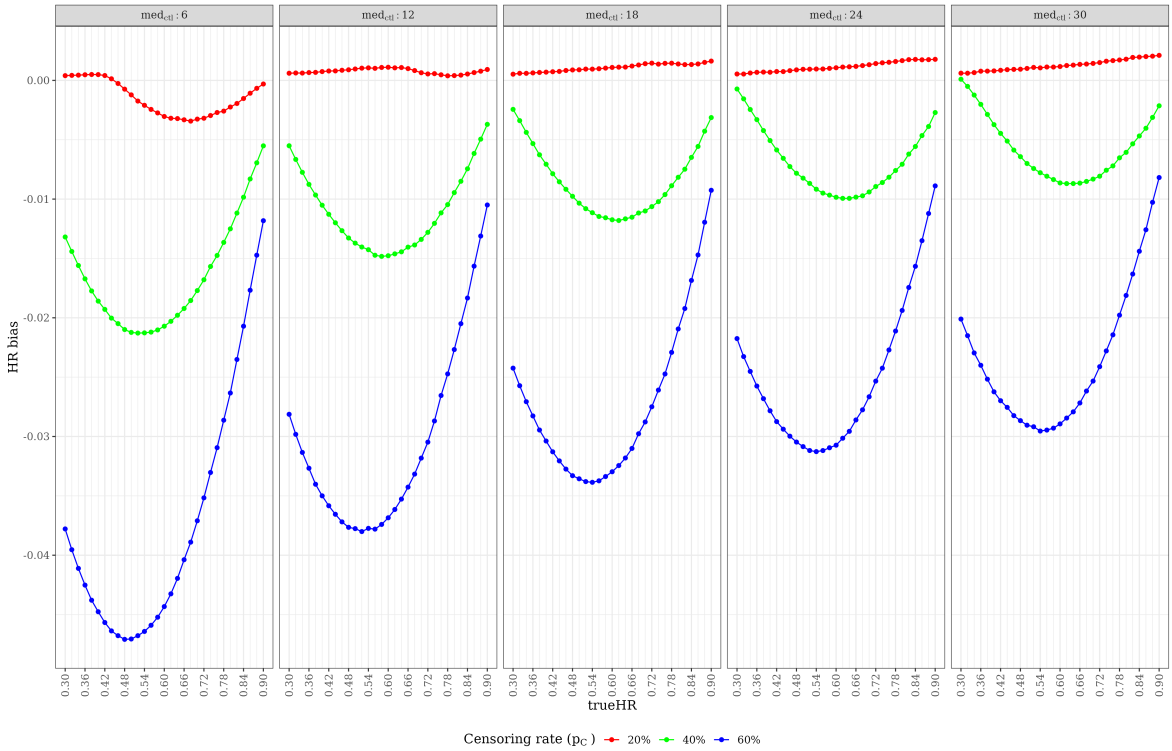

Figure 25: *Biased introduced to HR estimation (y-axis) separated by trueHR (x-axis) of Standard Scenario of Simulation 1 (outlined in Section 2.3.1) with power of 90%, n_{sim}= 10,000, and constant sample size (N=1,000; 500 per group).*

Each column of panel stands for different med_{ctl}. The simulated censoring rate is colour coded. Non-significant studies were included in this analysis. Abbreviations: med_{ctl}: Median survival time in the control group, p_C: Censoring rate, trueHR: True underlying Hazard Ratio of data generation

Table 11: *Mean HR bias (rounded to the 4^{th} decimal place) with Monte Carlo SE (rounded to the 6^{th} decimal place) of Standard Scenario of Simulation 1 (outlined in Section 2.3.1) with power of 90%, n_{sim}= 10,000, and constant sample size (N=1,000; 500 per group).*

Scenario		Introduced bias	
Censoring rate (p_C)	med_{ctl}	HR Bias	Monte Carlo SE of HR Bias
0.2	6	-0.0015	0.000432
0.2	12	0.0008	0.000438
0.2	18	0.0011	0.000445
0.2	24	0.0011	0.000449
0.2	30	0.0013	0.000453
0.4	6	-0.0166	0.000484
0.4	12	-0.0110	0.000488
0.4	18	-0.0085	0.000490
0.4	24	-0.0068	0.000491
0.4	30	-0.0057	0.000492
0.6	6	-0.0373	0.000573
0.6	12	-0.0305	0.000580
0.6	18	-0.0272	0.000583
0.6	24	-0.0251	0.000585
0.6	30	-0.0235	0.000587

Notes: Non-significant studies were included in this analysis. Abbreviations: HR: Hazard ratio, med_{ctl}: Median survival time of control group, p_C: Overall censoring rate, SE: Standard error

3.2 Simulation 2

Parts of this Section 3.2 are already published in the article *A Comparison of Additional Benefit Assessment Methods for Time-to-Event Endpoints Using Hazard Ratio Point Estimates or Confidence Interval Limits by Means of a Simulation Study* by Büsch et al. (2024). The manuscript has been written by the lead author but may contain comments and corrections from the co-authors and the reviewers.

To investigate the robustness of the results of Simulation 1, another simulation study was performed using an unbiased approach for the censoring mechanism. Further information of the ADEMP structure of the simulation study is outlined in Section 2.3.2. In the following, the results of Simulation 2 are shown.

This Section focuses only on important findings, where differences to the results of Simulation 1 are highlighted. Corresponding figures of the results not shown in this section can be found in Appendix A.2.

Overall, the results of Simulation 2 are very similar to the first simulation. One difference is present with small median survival times in the control group, i.e. $\text{med}_{\text{ctl}} \in \{6, 12\}$. Here, pairwise Spearman correlations between ESMO and the other additional benefit assessment methods have a reduced maximal value (black, blue, and turquise lines, Figure 26 first two panel-coloumns of middle row) compared to Simulation 1 (black, blue, and turquise lines, Figure 11 first two panel-coloumns of top row). For example, for med_{ctl}=6 and power 90% (left panel of middle row of Figure 12 and left panel of top row of Figure 26 for Simulation 1 and 2, respectively) the maximum Spearman correlation value between ASCO/ESMO is 0.8094 and 0.5240 for Simulation 1 and Simulation 2, respectively. The trueHR, however, where the maxima occurs, did not change. Other pairwise comparisons (ASCO/Mod-IQWiG$_{\text{HR}}$, ASCO/IQWiG$_{\text{RR}}$, and IQWiG$_{\text{RR}}$/Mod-IQWiG$_{\text{HR}}$) are very similar in all sub-scenarios as for Simulation 1.

In case of overpowered studies (Scenario 2, HR_{var}=0.8), the comparison ASCO/Mod-IQWiG$_{\text{HR}}$ has overall larger correlation values and the correlation increases with decreasing treatment effect, while in Simulation 1 this was the other way round (see Figure 26, second row from the top). For other HR_{var} values, this different behavior to Simulation 1 is not present.

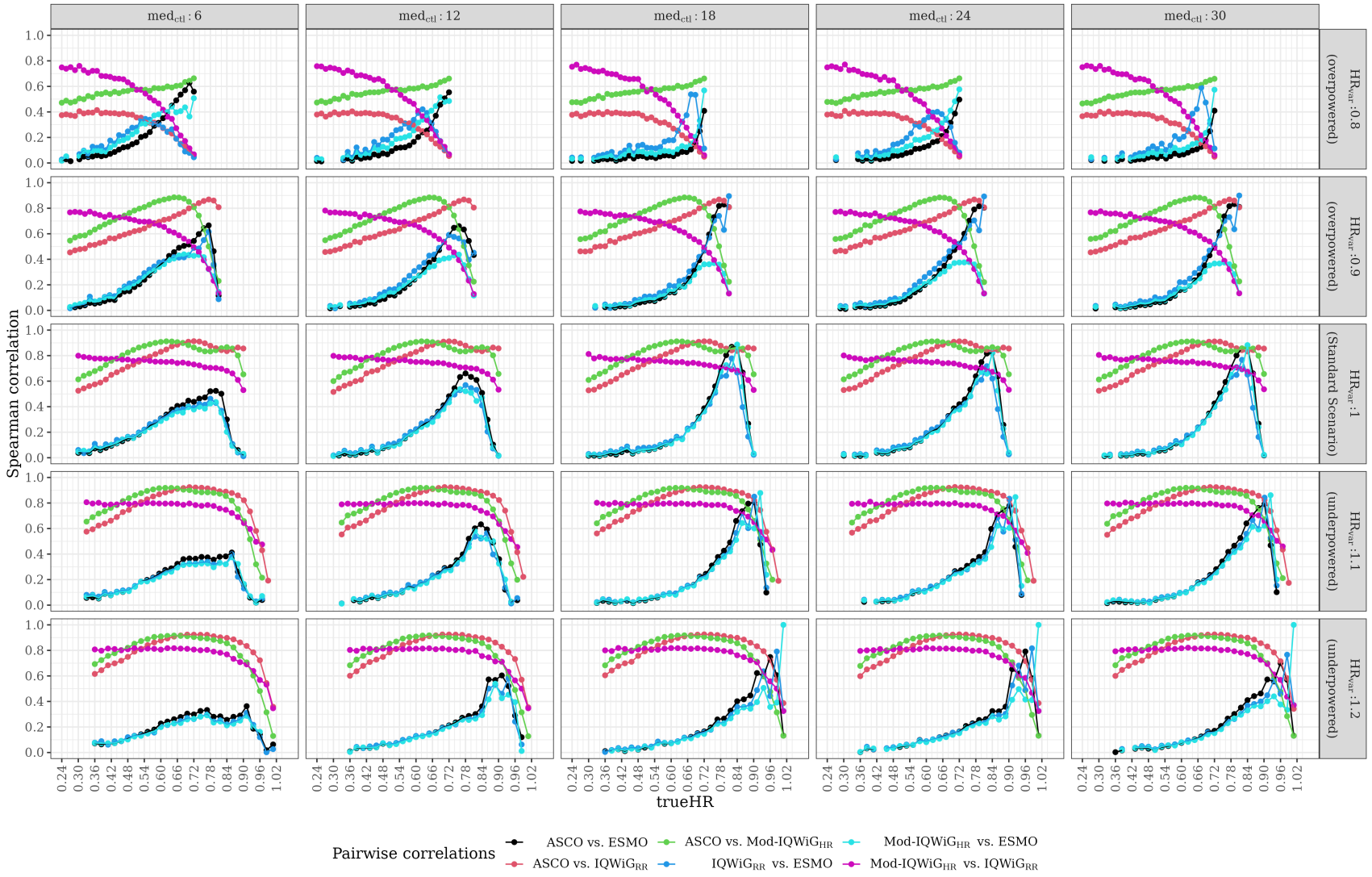

Figure 26: *Pairwise Spearman correlation (y-axis) separated by trueHR (x-axis) of Scenario 2 (Simulation 2) with p_C=60% and 90% power.*

Each row of panel stands for different HR_{var} and each column of panel stands for different med_{ctl}. The middle row is the same as the top row of Figure 11. Abbreviations: designHR: Design Hazard Ratio used for sample size calculation, HR_{var}: Factor for deviance between designHR and trueHR, med_{ctl}: Median survival time in the control group, p_C: Censoring rate, trueHR: True underlying Hazard Ratio of data generation

Moreover, the AUC values of the ROC curves of Simulation 1 and 2 of the Standard Scenario are shown in Figure 27, where the former is shown in transparent colours. Furthermore, the mean AUC values for all three statistical quantities of Standard Scenario 1 are shown in Table 12. It can be seen that for the statistical quantity HR^+ (black line) the AUC values are slightly reduced in Simulation 2, e.g. on average by 0.0180, 0.0185, 0.0185 for δ_{deserved} 0.7, 0.75, and 0.8, respectively. The results of Simulation 1 and Simulation 2 cannot be distinguished from the other two statistical quantities (HR^- and HR-PE) and hence have very similar results compared to Simulation 1. The overall comparison between the three statistical quantities, however, stays the same: HR^- and HR-PE have very large AUC values, where HR^- performs slightly better, and HR^+ has the lowest AUC results.

In case of Gompertz distributed failure times with decreasing hazard (shape=-0.2), the AUC results are reduced with increasing med_{ctl} compared to Simulation 1 (see Figure 28). This affects all statistical quantities in some degree, but HR^+ shows the largest changes compared to Simulation 1. For example, with med_{ctl}=30, power 90%, and δ_{deserved}=0.8, the AUC difference is 0.0419, 0.0119, and 0.0047 for HR^+, HR-PE, and HR^-, respectively. In case of increasing hazards (shape=0.2), only HR^+ is changed similar as in the Standard Scenario. Furthermore, all ordinal additional benefit assessment methods have slightly reduced TPR and FPR values in all scenarios compared to Simulation 1. For example Table 13 shows the mean TPR and FPR results for all ordinal additional benefit assessment methods of the Standard Scenario separated for δ_{deserved} and both simulations. The comparison between the methods, however, stays the same as in Simulation 1: ESMO is more liberal compared to IQWiG_{RR} and Mod-IQWiG_{HR} with Mod-IQWiG_{HR} being the most conservative one.

Table 12: *Mean AUC values for different statistical quantities of Standard Scenario of Simulation 1 and Simulation 2 (rounded to the 4^{th} decimal place)*

Simulation	δ_{deserved}	HR^-	HR-PE	HR^+
1	0.7	0.9911	0.9781	0.8545
2	0.7	0.9920	0.9795	0.8365
1	0.75	0.9922	0.9805	0.8633
2	0.75	0.9929	0.9817	0.8448
1	0.8	0.9931	0.9826	0.8755
2	0.8	0.9938	0.9837	0.8570

Abbreviations: AUC: Area under the curve, FPR: False positive rate, HR: Hazard Ratio, HR^+: Upper 95% confidence interval limit of the HR-PE, HR^-: Lower 95% confidence interval limit of the HR-PE, PE: Point estimate, TPR: True positive rate, trueHR: True underlying HR of data generation, $\delta_{deserved}$: Ground truth of deserved maximal category for TPR and FPR calculation (justified if trueHR $< \delta$)

Table 13: *Mean FPR and TPR of additional benefit assessment methods of Standard Scenario 1 of Simulation 1 and Simulation 2 (rounded to the 4^{th} decimal place)*

Simulation	δ_{deserved}	ESMO		IQWiG_{RR}		$\text{Mod-IQWiG}_{\text{HR}}$	
		TPR	FPR	TPR	FPR	TPR	FPR
1	0.7	0.9980	0.5148	0.7889	0.2344	0.6536	0.0789
2	0.7	0.9793	0.3932	0.6993	0.1499	0.5462	0.0417
1	0.75	0.9946	0.4041	0.7639	0.1672	0.6170	0.0390
2	0.75	0.9641	0.2903	0.6696	0.0973	0.5102	0.0183
1	0.8	0.9717	0.1683	0.7137	0.0671	0.5560	0.0054
2	0.8	0.9199	0.1137	0.6159	0.0315	0.4563	0.0017

Abbreviations: FPR: False positive rate, HR: Hazard Ratio, TPR: True positive rate, trueHR: True underlying HR of data generation, $\delta_{deserved}$: Ground truth of deserved maximal category for TPR and FPR calculation (justified if trueHR $< \delta$)

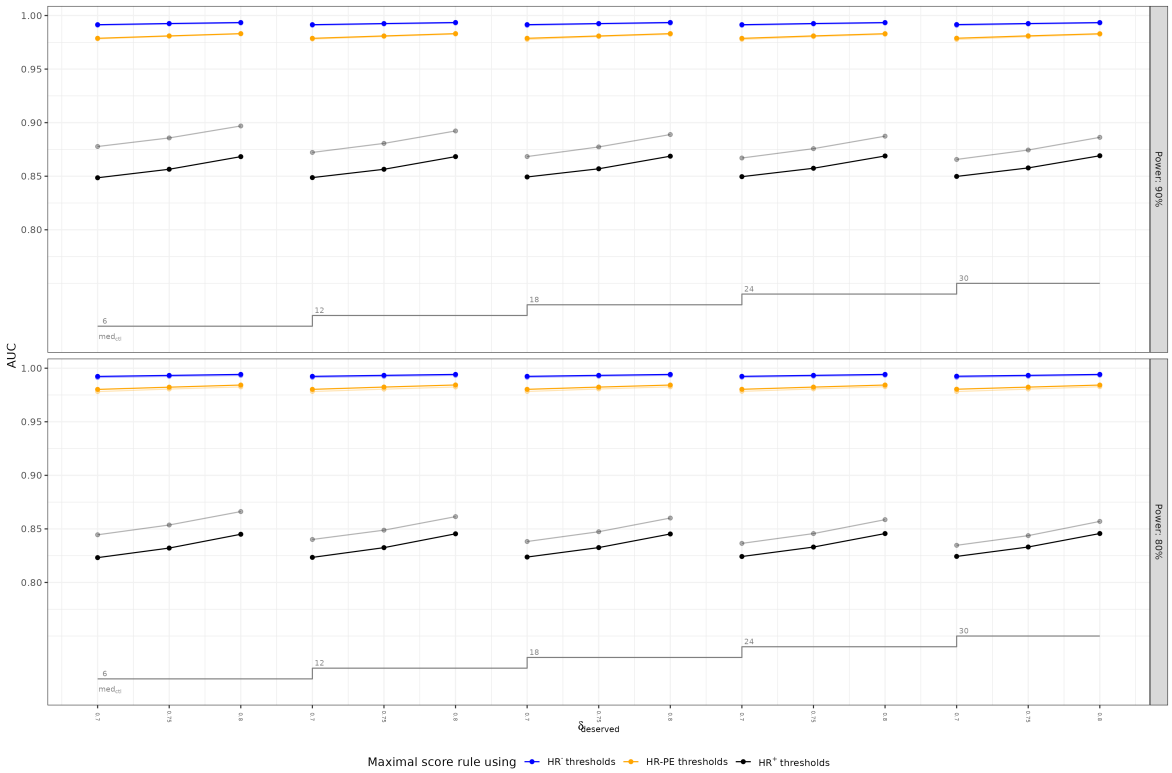

Figure 27: *AUC of ROC curves (y-axis) separated by $\delta_{deserved}$ (x-axis) of Standard Scenario with p_C=0.6 (Simulation 2).*

Each panel stands for a different power. The results of Simulation 1 are shown with transparent colours. For maximal category rule using HR^- and HR-PE thresholds, the results of Simulation 1 and 2 are almost identical. Hence, almost no difference in lines and points can be seen between both simulations. Abbreviations: AUC: Area Under the Curve, HR-PE: Hazard Ratio Point Estimate, HR^-: Lower 95% confidence interval limit of the HR-PE, HR^+: Upper 95% confidence interval limit of the HR-PE, med_{ctl}: Median survival time in the control group, p_C: Censoring rate, trueHR: True underlying Hazard Ratio of data generation, ROC: Receiver Operating Characteristic, $\delta_{deserved}$: Ground truth of deserved maximal category for TPR and FPR calculation (justified if trueHR < δ)

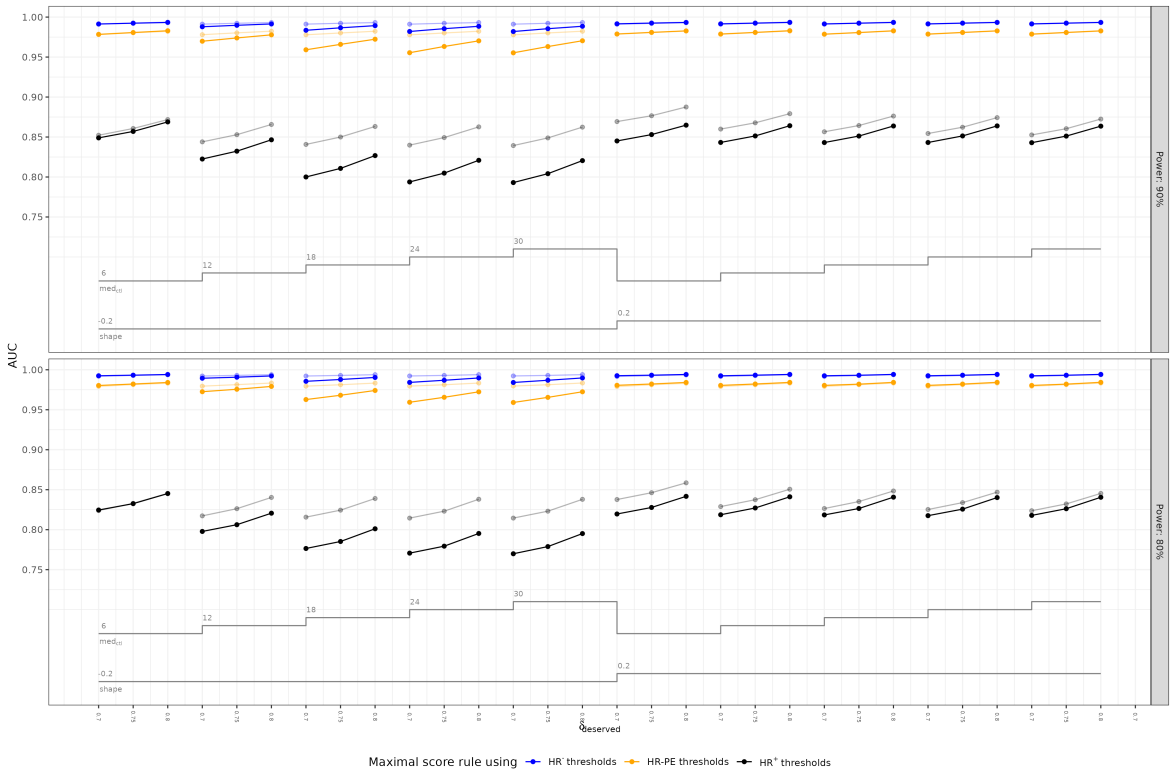

Figure 28: *AUC of ROC curves (y-axis) separated by $\delta_{deserved}$ (x-axis) of Scenario 3 with Gompertz failure time distribution and p_C=0.6 (Simulation 2).*

Each panel stands for a different power. The results of Simulation 1 are shown with transparent colours. For maximal category rule using HR^- and HR-PE thresholds, the results of Simulation 1 and 2 are almost identical. Hence, almost no difference in lines and points can be seen between both simulations. Abbreviations: AUC: Area Under the Curve, HR-PE: Hazard Ratio Point Estimate, HR^-: Lower 95% confidence interval limit of the HR-PE, HR^+: Upper 95% confidence interval limit of the HR-PE, med_{ctl}: Median survival time in the control group, p_C: Censoring rate, trueHR: True underlying Hazard Ratio of data generation, ROC: Receiver Operating Characteristic, $\delta_{deserved}$: Ground truth of deserved maximal category for TPR and FPR calculation (justified if trueHR < δ)

3.2.1 Bias evaluation

To review whether the data generation mechanism of Simulation 2 indeed leads to unbiased HR estimations, the results of the same investigation as for Simulation 1 (see Section 3.1.5) are shown in this sub-section. Hence, data of the Standard Scenario with n_{sim}= 10,000 and a constant sample size of 1,000 (500 per treatment group) for each sub-scenario was generated. The results of this assessment is shown in Table 14 and Figure 29 with the corresponding results of Simulation 1. The HR bias is always very close to zero and hence is much smaller than in Simulation 1. The simulation parameter med_{ctl} does not influence the HR estimation. Nevertheless, with decreasing treatment effect, the HR bias is slightly increased as well. For example, with med_{ctl} = 6 months, p_C = 0.6, and power of 90%, the HR bias is 0.0013 and 0.0045 for trueHR of 0.3 and 0.9, respectively. The corresponding HR bias for Simulation 1, however, are larger: -0.0378 and -0.0118, respectively.

Overall, the HR was slightly overestimated, i.e. the overall bias of the Standard Scenario was 0.00260 (Monte Carlo SE of HR bias: 0.000615) compared to the underestimation of the HR in Simulation 1 with p_C=0.6 (HR-Bias: -0.0287, Monte Carlo SE of HR bias: 0.000582).

Table 14: *Mean HR bias (rounded to the 4^{th} decimal place) with Monte Carlo SE (rounded to the 6^{th} decimal place) of Standard Scenario of Simulation 2 (outlined in Section 2.3.2) with power of 90%, p_C=0.6, n_{sim}= 10,000, and constant sample size (N=1,000; 500 per group).*

Scenario		Introduced bias	
Censoring rate (p_C)	med_{ctl}	HR Bias	Monte Carlo SE of HR Bias
0.6	6	0.0027 (-0.0373)	0.000615 (0.000573)
0.6	12	0.0026 (-0.0305)	0.000615 (0.000580)
0.6	18	0.0026 (-0.0272)	0.000615 (0.000583)
0.6	24	0.0026 (-0.0251)	0.000615 (0.000585)
0.6	30	0.0025 (-0.0235)	0.000614 (0.000587)

Abbreviations: HR: Hazard ratio, med_{ctl}: Median survival time of control group, p_C: Overall censoring rate, SE: Standard error

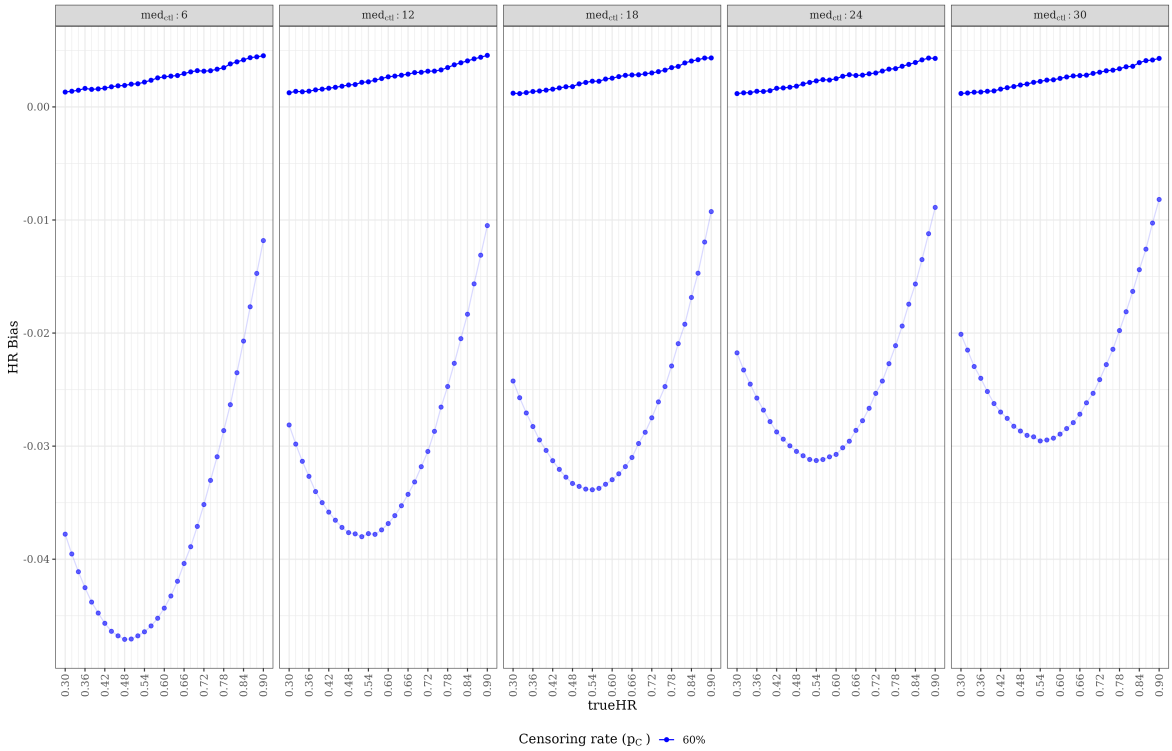

Figure 29: *HR estimation bias (y-axis) separated by trueHR (x-axis) of Standard Scenario of Simulation 2 (outlined in Section 2.3.2) with power of 90%, p_C=0.6, n_{sim}= 10,000, and constant sample size (N=1,000; 500 per group).*

Each panel stands for different med_{ctl}. The results of Simulation 1 are shown with transparent colours (see Figure 25). Non-significant studies were included in this analysis. Abbreviations: med_{ctl}: Median survival time in the control group, p_C: Censoring rate, trueHR: True underlying Hazard Ratio of data generation

3.3 Study examples

This section outlines two phase III studies in an oncology setting with OS as endpoint to exemplify the application of the additional benefit assessment methods. These studies were selected to display a large and moderate treatment effect. To reflect the setting of this thesis, the application of the additional benefit assessment methods is purely fictitious. Hence, officially the methods were not applied and for the performed application in this thesis only the statistical components of the methods are relevant as the statistical approaches between the methods are of interest. In the following, both study examples are briefly summarised:

1. Example: Al-Sarraf et al. (1998) published the results of a randomized phase III trial comparing chemoradiotherapy against radiotherapy alone in patients with nasopharyngeal cancers. The analysis of the primary outcome overall survival included 147 (69 radiotherapy and 78 chemoradiotherapy) patients and yielded an estimated hazard ratio of 0.40 (95% CI: [0.21, 0.78]).

2. Example: A study assessing the efficacy and safety of atezolizumab plus chemotherapy versus chemotherapy alone as first-line therapy for nonsquamous non-small-cell lung cancer (West et al., 2019). The analysis of the primary outcome overall survival included 679 patients in the intention-to-treat wild-type population (451 atezolizumab plus chemotherapy and 228 chemotherapy group) and yielded an estimated hazard ratio of 0.79 (95% CI: [0.64, 0.98]).

As mentioned above this thesis focuses on the statistical parts of the methods. Hence, the following statistical metrics of the two studies were required to apply these parts of the methods: HR-PE, upper (HR^+), and lower limit (HR^-) of the 95% HR-CI, median survival time of the control group (med_{ctl}), gain (=med_{trt} - med_{ctl}), increase of survival rate at a specific time point (depending on the observed med_{ctl}) and the "tail of the curve" bonus point adjustment (proportion of patients alive in the treatment compared to the control group improved by 50% or more at a specific time point). Further information about the methods application, e.g. bonus point adjustments and specific thresholds for categorization, are outlined in Section 2.2.

The application results of the additional benefit assessment methods and determined ASCO

cutoff application on the two clinical study examples are shown below in Table 15. The estimated metrics of the study examples needed for the additional benefit assessment methods application and resulting categories or scores of the two study examples are also provided in Table 15. More information on the results of the ASCO cutoff value determination is depicted in Section 3.1.4.

Table 15: *Statistical metrics (column 1), resulting categories/scores of the additional benefit assessment methods and application of ASCO cutoff values for the two study examples. For further information on the methods application see Section 2.2.*

Study	ASCO	ESMO	$IQWiG_{RR}$	Mod-$IQWiG_{HR}$
1. Al-Sarraf et al. (1998): • HR-PE: 0.40 • HR^{-}: 0.21 • HR^{+}: 0.78 • med_{ctl}: 34 months • gain: 26 months • 5 year increase: 28% • tail of the curve $\geqslant 50\%$	(1-0.40)·100 + 20 = 80	Cutoff: 4 Method: 4	Cutoff: major Method: major	Cutoff: major Method: major
2. West et al. (2019): • HR-PE: 0.79 • HR^{-}: 0.64 • HR^{+}: 0.98 • med_{ctl}: 13.9 months • gain: 4.7 months • 3 year increase: missing • tail of the curve $< 50\%$	(1-0.79)·100 + 0 = 21	Cutoff: 3 Method: 3	Cutoff: considerable* Method: minor	Cutoff: minor Method: minor

$\star$: *Applying ASCO cutoff value results in different category then method application.*
Abbreviations: CI: Confidence interval, gain: Absolute difference in median survival times, HR: Hazard ratio, HR^{+}*: Upper 95% confidence interval limit of the HR-PE,* HR^{-}*: Lower 95% confidence interval limit of the HR-PE,* med_{ctl}*: Median survival time in the control group, PE: Point estimate*

The application of the ASCO cutoff values led generally to identical categories as for the method application. The only exception is for study example 2 of West et al. (2019), where the ASCO cutoff values resulted in category "considerable added benefit" instead of "minor added benefit" for $IQWiG_{RR}$. This difference is, however, marginal because the proposed ASCO cutoff value is very close to the ASCO score of this study (cutoff value: 20.06; ASCO score: 21). Hence, the proposed ASCO cutoff values work well in real study examples.

Chapter 4

Discussion

In this Chapter, the results outlined in Chapter 3 of this thesis are discussed. The structure follows the same sequence as the different performance measures outlined in Section 2.3.1.5. In each Section, the corresponding results including the contribution to research are discussed. Additionally, the strengths and weaknesses of each additional benefit assessment method are outlined as well as limitations and directions for further research are depicted. The Discussion Chapter ends with a conclusion of the results from this thesis.

Parts of this Chapter 4 are already published in the articles *A Comprehensive Comparison of Additional Benefit Assessment Methods Applied by Institute for Quality and Efficiency in Health Care and European Society for Medical Oncology for Time-to-Event Endpoints After Significant Phase III Trials — a Simulation Study* by Büsch et al. (2022) and *A Comparison of Additional Benefit Assessment Methods for Time-to-Event Endpoints Using Hazard Ratio Point Estimates or Confidence Interval Limits by Means of a Simulation Study* by Büsch et al. (2024). The manuscripts have been written by the lead author but may contain comments and corrections from the co-authors and the reviewers.

4.1 Relationship between additional benefit assessment methods

The relationship between IQWiG_{RR}/ESMO show a low positive Spearman correlation over the range of treatment effects. The same can be said for the ASCO/ESMO relationship, only with moderate treatment effects and a median survival time in the control group larger than 12 months lead to higher correlation. Furthermore, ASCO/Mod-IQWiG_{HR} always show a stronger or at least equal high positive relationship as ASCO/IQWiG_{RR}. As IQWiG_{RR} and Mod-IQWiG_{HR} use HR^{+} for the assessment of new treatments and hence are very similar in their construction, the methods show a high positive relationship over almost every simulated scenario.

However, combining all sub-scenarios, e.g. all treatment effects, the ASCO/ESMO com-

parison shows a moderate positive relationship of 0.68, which is exactly the same as the research of Cherny et al. (2019), where 102 real studies were used to apply both methods. Other empirical investigations using also real studies for the ASCO and ESMO application show different inconsistent negligible to low correlations of 0.17 (Cheng et al., 2017), 0.397 (Del Paggio et al., 2018), and 0.40 (Becker et al., 2017) for the ASCO/ESMO relationship. This difference compared to the results of this thesis can be explained by the application of the methods on real studies where the complete methods were applied; including bonus point adjustments. To reach, however, a fair comparison of the statistical aspects of the methods, this thesis focused on the statistical quantities and hence only bonus point adjustments comprising out of statistical measures were applied.

In case of underpowered studies, all pairwise non-ESMO correlations, i.e. ASCO/IQWiG$_{\text{RR}}$, ASCO/Mod-IQWiG$_{\text{HR}}$, and IQWiG$_{\text{RR}}$/Mod-IQWiG$_{\text{HR}}$, stay similar to the Standard Scenario. All pairwise ESMO comparisons, however, show a reduced correlation, indicating that ESMO is influenced by underpowered studies. Nevertheless, with overpowered studies all methods are influenced leading to reduced pairwise correlations. Hence, all method applications and interpretations in overpowered studies should be taken with precaution. One important note is that in the scenarios with over- and underpowered studies more sub-scenarios with larger or smaller treatment effects were simulated and as the Spearman correlation is reduced in these sub-scenarios, the overall Spearman correlation, where all sub-scenarios were combined, had to be reduced.

Furthermore, informative censoring reduces the relationship between all methods. Only when the treatment group has a higher censoring rate than the control group, the non-ESMO pairwise comparisons of ASCO/IQWiG$_{\text{RR}}$, ASCO/Mod-IQWiG$_{\text{HR}}$, and IQWiG$_{\text{RR}}$/Mod-IQWiG$_{\text{HR}}$ show unchanged high positive relationships which shows again the fragile behaviour of ESMO. Important to mention is that with a larger censoring rate in the control group and small treatment effects, i.e. trueHR$>$0.82, no significant results were present in the simulation studies. Hence, the additional benefit assessment methods and Spearman correlations between them could not be calculated.

Other performed simulation scenarios such as non-proportional hazards using late treatment effects for the treatment group, unequal sample sizes, and exponential censoring distribution did not influence the described relationship behaviour between the methods.

4.2 Which of the statistical approaches provide the best properties for additional benefit assessment

The main difference between the statistical aspects of the three additional benefit assessment methods of ASCO, ESMO, and $IQWiG_{RR}$ are the used statistical quantity. ESMO uses among other aspects the lower limit of the 95% HR-CI (HR^-), IQWiG the upper limit (HR^+), and ASCO the HR-PE. This thesis evaluated which of these three statistical quantity is best for the additional benefit assessment using TPR, FPR, and corresponding AUC values of the ROC curves.

The results clearly show that HR^- has the largest TPR and smallest FPR and hence overall the largest AUC values over the complete range of simulated scenarios including different allocation ratios, censoring rates, power, failure time distributions, and more. HR-PE has very similar but nevertheless slightly smaller AUC values. On the contrary, HR^+ provides the lowest AUC values. For example, in case of the Standard Scenario, the average AUC values of HR^-, HR-PE, and HR^+ are 0.99, 0.98, and 0.86, respectively. Reason for this difference is that the slope of the estimates of HR^- and HR-PE is larger over the complete range of treatment effects. Thus, finding a cutoff value which classifies the complete range of treatment effects into deserved and not deserved maximal categories, is easier for HR^- and HR-PE than for HR^+. The complete range of treatment effects was always considered combined as the additional benefit assessment methods have to be applied for all treatments on the market and hence over the complete range of treatment effects. Interestingly, the slope difference over the range of treatment effects is only different between the statistical quantity when a sample size calculation is performed. In case of a constant sample size independent of the treatment effect, the estimates of the three statistical quantities have the same slope hence leading to the same ROC curves and the same AUC values. Important to note is that the results with a constant sample size are slightly biased because the additional benefit assessment methods can only be applied after a significant trial. Hence, only significant studies were considered leading to a small number of simulation iterations used for estimations of the statistical quantities in case of small treatment effects. Nonetheless, the results clearly depict that the sample size calculation is the reason for the differences in the statistical quantities. However, a sample size calculation is mandatory in phase III trials due to ethics, time, and costs.

Therefore, HR^- provides the best solution for additional benefit assessment.
In spite of that, the choice of the threshold used for defining a maximal category is even more important than the used statistical quantity. For example, ESMO which uses HR^-, has a very large TPR but unfortunately a large FPR as well. For example, if a deserved maximal category classification is defined as true$HR < 0.75$, an average of 0.99 and 0.40 was present for TPR and FPR in the Standard Scenario, respectively. $IQWiG_{RR}$ uses HR^+ and has a smaller FPR than ESMO, e.g. on average 0.17. Nevertheless, with HR^- it is possible to achieve a smaller FPR while maintaining a large TPR. For example, defining a maximal category with a threshold of 0.55, an average TPR of 0.91 can be achieved while only allowing a FPR of 5% and defining a deserved maximal category classification as true$HR < 0.75$. With the same constraints HR^+ achieves only an average TPR of 0.68 with a threshold of 0.79.
For the definition of deserved maximal category classification no gold standard exists and hence, no approach can be perfect. Nonetheless, in this thesis, different definitions for the deserved maximal category classification are used trying to close this knowledge gap as best as possible. The described interpretation above stays consistent over the range of different gold standard definitions: The choice of the threshold used for defining a maximal category is more important than the used statistical quantity. Nevertheless, HR^- provides the best solution for additional benefit assessment if appropriate thresholds are chosen. This substantiates the research of Dafni et al. (2017), where HR^- and HR-PE were investigated.
Hence, the concern that the use of HR^- as main statistical quantity for the additional benefit assessment would lead, especially for studies with smaller sample sizes, to higher awarded grades (Muhonen et al., 2015; Wild et al., 2016), is only partly true. ESMO, which uses HR^-, indeed awards a higher rate of maximal categories than $IQWiG_{RR}$, which uses HR^+. However, this is only due to ESMOs choice of thresholds defining the maximal category rather than the used statistical quantity. Furthermore, the assumption that HR-CI provides more information than HR-PE through the included variability in these estimates or that HR-PE might introduce possible bias (Skipka et al., 2016), is not reflected by the simulation studies. Other considered simulation scenarios such as non-proportional hazards using late treatment effects for the treatment group, unequal sample sizes, and exponential censoring distribution did not influence the described AUC, TPR, and FPR results of the statistical quantities and methods.

4.3 Corresponding ASCO cutoff values for categories of other additional benefit assessment methods

To improve practical comparison between the additional benefit assessment methods without the need to apply all methods, ASCO cutoff values were calculated, which correspond to specific categories of ESMO, IQWiG$_{\text{RR}}$, and Mod-IQWiG$_{\text{HR}}$. An ASCO score larger than 17, 20, and 23 corresponds to ESMO categories 2, 3, and 4, respectively. For IQWiG$_{\text{RR}}$ and Mod-IQWiG$_{\text{HR}}$ only two ASCO cutoffs are needed as these two methods consist out of three categories. ASCO cutoff values of 20 (23) and 36 (45) separate the score of ASCO into the three IQWiG$_{\text{RR}}$ (Mod-IQWiG$_{\text{HR}}$) categories "minor", "considerable", and "major added benefit". It is obvious that the cutoff values for IQWiG$_{\text{RR}}$ and Mod-IQWiG$_{\text{HR}}$ are wider apart from each other compared to ESMO. As ESMO consists out of four and both IQWiG methods out of three categories, the ASCO cutoff values must be closer for ESMO compared to the other two methods. Nevertheless, the ASCO cutoff values for ESMO are almost identical and hence different additional reasons must exist. One explanation is that the ASCO score can be better visually separated for IQWiG$_{\text{RR}}$ and Mod-IQWiG$_{\text{HR}}$, i.e the ASCO score distribution of the individual categories are further apart for IQWiG$_{\text{RR}}$ and Mod-IQWiG$_{\text{HR}}$ compared to ESMO. Furthermore, the maximal category of ESMO can be achieved over the complete ASCO scoring range leading to very similar ASCO cutoff values for the four ESMO categories. This also illustrates the liberal nature of ESMO.

Moreover, in this thesis, the described ASCO cutoff values were also calculated for all considered simulation scenarios such as different allocation ratios, over-/underpowered studies, different failure time distributions, and many more. Changing cutoff values between the different scenarios would suggest that the additional benefit assessment methods somehow react to the different settings. Overall, the above described cutoff values stay very similar over the range of the different scenarios. However, different failure time distributions, especially Gompertz distribution with increasing hazards over time, which still fulfill the proportional hazard assumption, resulted in inconsistent ASCO cutoff values for ESMO. As the proportional hazard assumption is still fulfilled, this difference is an unwanted behaviour.

Which ASCO score corresponds to which category of another ordinal method has also been investigated by Cherny et al. (2019). In their investigation, 102 randomized controlled trials

instead of a simulation study were used for calculation of the ASCO cutoff values for ESMO, resulting in an ASCO score of 46 or greater and 41 or less to define substantial benefit (category 4) and low benefit (category 1–3), respectively. The reason for the difference to the calculated ASCO cutoff values in this thesis is probably due to the focus on the statistical aspects of the additional benefit assessment methods. Hence, reduced versions of the methods were applied, which is the opposite of the approach followed by Cherny et al. (2019).

4.4 Strength and weaknesses of additional benefit assessment methods

Focusing on the maximal category of the ordinal additional benefit assessment methods, ESMO has the largest rate of the maximal category, which is in most scenarios over 80%. IQWiG_{RR} and $\text{Mod-IQWiG}_{\text{HR}}$ show smaller rates for their maximal category of around 55% and 40%, respectively. As the maximal category is assumed to be comparable between ESMO and IQWiG methods, it can be said that ESMO is the most liberal method while IQWiG_{RR} and $\text{Mod-IQWiG}_{\text{HR}}$ are more conservative with $\text{Mod-IQWiG}_{\text{HR}}$ being the most conservative. Even with the added absolute benefit rule added in the second ESMO version, ESMO stays the most liberal method compared to the other ones. This is verified by the maximal category rate and can also be affirmed by TPR and FPR values: ESMO has the largest FPR values of up to 0.51. IQWiG_{RR} and $\text{Mod-IQWiG}_{\text{HR}}$ have smaller FPR values of up to 0.23 and 0.08, respectively. In addition, ESMO has also the largest TPR values of up to 0.99, whilst IQWiG_{RR} and $\text{Mod-IQWiG}_{\text{HR}}$ reach only up to 0.79 and 0.65, respectively.

Moreover, all methods are affected by over- and underpowered studies, which lead to an increase of the maximal category rate or score: Both ASCO and ESMO show in both scenarios an increased score or an increased rate of the maximal category, respectively. IQWiG_{RR} and $\text{Mod-IQWiG}_{\text{HR}}$ show similar results for underpowered studies compared to the Standard Scenario and hence can be seen as conservative. In case of overpowered studies, the maximal category rate increases for IQWiG_{RR} more than twice as large compared to ESMO. As ESMO already has a large rate of over 80% in the Standard Scenario, the maximal category rate in other scenarios such as overpowered studies cannot increase much further. Hence, it is not surprising that the maximal category rate of ESMO does not increase as much as

for IQWiG_{RR} and $\text{Mod-IQWiG}_{\text{HR}}$. Furthermore, even though the maximal category rate of ESMO does not increase as much as for the other methods, the maximal category rate of ESMO is still the largest. This contradicts the results shown by Dafni et al. (2017), who stated that ESMO has no discriminatory behavior in over- and underpowered trials. This opposite result can be explained by the fact that the simulation study of Dafni et al. (2017) applied different parameter ranges. Furthermore, at the publication time of Dafni et al. (2017), ESMOs current version was v.1.0. However, in this thesis, the updated version (v. 1.1) was used, which uses different thresholds and additionally the absolute benefit rule.
Nevertheless, it is not surprising that all methods have this weakness in case of wrongly assumed treatment effects because the true treatment effect in real studies is not known. Thus, the additional benefit assessment methods cannot penalize over- and underpowered studies appropriately. As a result, the treatment effect assumed for sample size calculation is even more important and furthermore, all methods' application and their results' interpretation should keep this weakness in mind. Nonetheless, in case of overpowered studies, the proposed $\text{Mod-IQWiG}_{\text{HR}}$ method, which uses HR based HR^{+} thresholds instead of RR based thresholds, provides the best behaviour with less increased maximal category rate. However, in most of the other scenarios $\text{Mod-IQWiG}_{\text{HR}}$ might be too conservative, e.g. low TPR values. Furthermore, in case of non-exponential distributed failure times, which still adhere to the proportional hazard assumption, ESMO shows non-desired susceptible results meaning changed category distribution. Especially in case of Gompertz distribution with increasing hazards over time, the rate of the maximal category decreases drastically.
In this thesis, it was also investigated whether delayed treatment effects, which is a kind of non-proportional hazards, do influence the additional benefit assessment methods. The results show that all methods with ordinal outcome (ESMO, IQWiG_{RR}, and $\text{Mod-IQWiG}_{\text{HR}}$) have a shift in category proportion with reduced maximal category and ASCO (continuous outcome) has a reduced score compared to the Standard Scenario (proportional hazards), which is a desired behavior as the proportional hazard assumption is violated. ESMOs reduction is, however, the lowest and thus is again the most liberal one. Important to note is that in this scenario only situations with $\text{med}_{\text{trt}} \gg \text{med}_{\text{ctl}}$ have been considered because ESMO assigns only the lowest category when med_{trt} and med_{ctl} are similar. Thus, similar med_{trt} and med_{ctl} have not been investigated to not penalize ESMO by its design.

Another assumption of the Cox proportional hazard model is non-informative censoring. In case of informative censoring a desired behaviour of the methods would be to have a decreased score/category because the results of the Cox proportional hazard model might be biased. In the event that the control group has a higher censoring rate than the treatment group ($p_C^{ctl} < p_C^{trt}$) and vice versa ($p_C^{ctl} > p_C^{trt}$), Mod-IQWiG$_{\text{HR}}$ shows the most desirable behaviour as the maximal category is decreased ($p_C^{ctl} > p_C^{trt}$) or at least only slightly increased ($p_C^{ctl} < p_C^{trt}$).

4.5 Limitation and directions for further research

In this thesis, extensive simulation studies were performed with various scenarios including different censoring rates, treatment effects, allocation ratios, power, failure time distributions as well as scenarios with non-proportional hazards, informative censoring, and over-/underpowered trials. Moreover, the performed simulation studies are reproducible with the provided R-Code at `https://www.github.com/cbuesch/SumulationStudyABAM`. Additional information on R-Code structure and execution of the programs to determine the results of this thesis can be found in Appendix B.

A limitation of the conducted research is that the focus is on the statistical aspects of each additional benefit assessment method and hence most bonus point adjustments were not applied, e.g. toxicity adjustments leading to the application of reduced version of ASCO and ESMO. Hence, an application of the complete methods on real studies could lead to different results and conclusions. For example, the liberality of ESMO could be reduced by the bonus point adjustments. Nevertheless, this thesis also investigates the general concept of the statistical quantities (HR$^-$, HR-PE, HR$^+$), which is not affected by this limitation.

Furthermore, the maximal categories of ESMO (substantial improvement) and IQWiG$_{\text{RR}}$ (major added benefit) were assumed to be comparable. As already outlined in the Introduction (see Section 1.1), category 4 and 5 of ESMO are defined as "substantial improvement". As category 5 can only be achieved by non-statistical bonus point adjustments, ESMO categories ranged from 1 to 4, leading to comparable maximal categories between ESMO and IQWiG$_{\text{RR}}$. Nevertheless, it can still be argued that this assumption might have led to false conclusions in this research. This thesis, however, still presents other results without this assumption (correlation, ROC including corresponding TPR and FRP, corresponding ASCO

values for IQWiG_{RR}, Mod-IQWiG_{HR}, and ESMO categories), which support the conclusions drawn based on this assumption.

In addition, the applied data generation of Simulation 1 introduces bias into the HR and corresponding CI estimation. Since, however, a combination of administrative censoring (not dependent on the event time) and exponentially distributed censoring achieving specific censoring rate (dependent on the event time) was implemented in Simulation 1, the introduced bias was reduced. Furthermore, this bias does not affect the method comparison to a substantial degree because it affects all compared methods equally. Simulation 2, where the generated censoring times are independent of the event times, proves the robustness of Simulation 1 as it shows similar results.

A validation examination of the additional benefit assessment method comparison was not assessed because no gold standard method exist, i.e. a definition of a truly deserved maximal category classification is missing. Thus, future research might focus on defining a gold standard. One possibility would be to include perceptions and opinions of patients as the additional benefit of new treatments should include at least partly the end user.

Further research should also focus on exploration of the different results in category proportions of ESMO, especially in case of Gompertz failure time distribution. One possible reason for this behaviour is ESMOs combination of absolute and relative benefit rule. The AUC and TPR/FPR results show that HR^{-}, which is used by ESMOs relative benefit rule, is not influenced by different failure time distributions. Hence, it can be speculated that the absolute benefit rule causes ESMOs susceptibility to Gompertz failure time distribution. Nevertheless, this behaviour is still astonishing and thus future research should be conducted.

As mentioned above the outlined simulation studies covered a wide range of different real-life study situations including non-proportional hazards as in the field of immunoncology treatments the proportional hazard assumption is often not fulfilled. In these cases estimation of HR, HR^{-}, and HR^{+} are strongly influenced by the FU time. In this thesis, the FU time is defined to be dependent on the median survival time of the control group ($\text{FU}=2\cdot\text{med}_{\text{ctl}}$) leading to always the same proportion between survival and FU time. Hence, future research should also focus on different scenarios with non-proportional hazards.

Furthermore, the provided ASCO cutoff values to achieve corresponding ESMO, IQWiG_{RR},

and Mod-IQWiG$_{\text{HR}}$ categories should be verified using real-world data where the complete methods are applied. This would enable a more realistic interpretation of the cutoff values.

4.6 Conclusion

In conclusion, IQWiG$_{\text{RR}}$ and ESMO show a high positive association for moderate treatment effects. ASCO and IQWiG$_{\text{RR}}$ as well as ASCO and Mod-IQWiG$_{\text{HR}}$ show a high positive association over the whole range of treatment effects.

Moreover, ESMO excessively awards its maximal category over the whole range of treatment effects and hence cannot distinguish between small and large treatment effects. ASCO and IQWiG$_{\text{RR}}$ have a more conservative behaviour. Different violated assumptions such as non-proportional hazards, over-/underpowered studies, and informative censoring, do not lead to penalization in ESMOs grading; e.g. the maximal category rate is not reduced. Overall, the used thresholds of ESMO are chosen too liberal, which leads to high FPR and easily achieved maximal category. ASCO and IQWiG$_{\text{RR}}$ have a more desirable behaviour, though both methods also portray unwanted behaviour in certain cases. In most cases Mod-IQWiG$_{\text{HR}}$, which uses different thresholds but the same statistical quantity for categorizations as the original IQWiG$_{\text{RR}}$, provides a better solution in case of violated assumptions, i.e. no excessive increase of the maximal category. In most other scenarios where no assumptions are violated it can be too conservative, resulting in a low TPR and a low maximal category rate. Furthermore, ESMO shows an increased rate of maximal category in case of different failure time distributions, which still adhere to the proportional hazard assumptions and hence no changes in the category distribution should be expected. The other methods (ASCO, IQWiG$_{\text{RR}}$, and Mod-IQWiG$_{\text{HR}}$) do not show this undesired behaviour. Hence, ESMO is the most liberal method.

Nonetheless, under the condition that appropriate thresholds are chosen, HR^{-}, which ESMO uses among other things, is the best statistical quantity to assess additional benefit assessment. Even HR-PE provides better properties than HR^{+}. As no gold standard for additional benefit assessment exists, validation examination was performed with different definitions of a truly deserved maximal category classification.

To improve practical comparison between the additional benefit assessment methods without the need to apply all methods, this thesis has proposed ASCO cutoff values which correspond

to the respective categories of the current methods.

Overall, this thesis demonstrates that HR^- instead of HR^+ should be used or the current thresholds should be at least adjusted to optimize the true positive and false positive rate for additional benefit assessment. Thus, in future this research can be used as a guide for improvements of the methods and contributes to the enhancement of additional benefit assessment.

Chapter 5

Summary

In the process of the development of a new treatment, many requirements for market authorization must be met through various stages. After approval, the additional benefit of a new treatment is compared to already established treatments. This assessment can decide on the amount of reimbursement of the new treatment on the market and create transparency for patients regarding the treatments' medical effectiveness and toxicity.

In case of non-curative or advanced diseases like cancer, three different additional benefit assessment methods have been developed. The European Society for Medical Oncology (ESMO) and Institute for Quality and Efficiency in Health Care (IQWiG) constructed methods with an ordinal outcome. For the main classification, IQWiG compares the upper limit of the 95% hazard ratio (HR) confidence interval (CI) (HR^+) against relative risk (RR) based thresholds and ESMO uses mainly the lower limit of the 95% HR-CI (HR^-). The American Society of Clinical Oncology (ASCO) defined a continuous outcome using HR point estimate (PE).

Hence, the main difference of the three methods is the used statistical quantity. There are several points of criticism to each of the assessment methods. For example, the use of the HR-PE for the assessment of the clinical benefit could penalize studies of substantial benefit by ignoring the precision of the estimate. In contrast, the upper or lower limit of the HR-CI considers the variability of the estimate and hence should provide more information.

The aim of this thesis is to obtain a better understanding of the differences between the methods and to answer the question which statistical quantity has the best properties to assess additional benefit. Furthermore, it is investigated which category of ESMO and IQWiG corresponds to which ASCO score in order to achieve an easier comparison between all three methods. To achieve these objectives, this thesis evaluates and compares the above described methods by means of simulation studies comprising different failure time distributions, treatment effects, power, allocation ratios, censoring types, and censoring rates. Furthermore, scenarios with non-proportional hazards, underpowered trials, and overpowered trials are investigated.

The original IQWiG method ($IQWiG_{RR}$) and ESMO show a high positive association for

moderate treatment effects. ASCO and IQWiG_{RR} as well as ASCO and Mod-IQWiG_{HR} (proposed modification of IQWiG_{RR} using HR based thresholds instead of RR based once) show a high positive association over the whole range of treatment effects. Moreover, ESMO excessively awards its maximal category over the whole range of treatment effects (in most scenarios over 80%) and hence cannot distinguish between small and large treatment effects. ASCO and IQWiG_{RR} have a more conservative behaviour.

Different violated assumptions such as non-proportional hazards, over-/underpowered studies, and informative censoring, do not lead to penalization in ESMOs grading; e.g. the maximal category rate is not reduced. Overall, the used thresholds of ESMO for categorization are chosen too liberal, which lead to high false positive rates and easily achievable maximal category grading. ASCO and IQWiG_{RR} have a more desirable behaviour. In most cases Mod-IQWiG_{HR} does provide a better solution in case of violated assumptions, i.e. no excessive increase of the maximal category rate. Nevertheless, in most other scenarios where no assumptions are violated it might be too conservative, i.e. low true positive rate and low maximal category rate.

Furthermore, ESMO shows an even increased rate of maximal category in case of different failure time distributions, which still adhere to the proportional hazard assumptions and hence no changes in the category distribution should be expected. The other methods (ASCO, IQWiG_{RR}, and Mod-IQWiG_{HR}) do not show this undesired behaviour. Hence, ESMO is the most liberal method. Nonetheless, under the condition that appropriate thresholds are chosen, HR^{-}, which ESMO uses among other things, is the best statistical quantity to assess additional benefit. Even HR-PE provides better properties than HR^{+}.

To improve practical comparison between the methods, this thesis proposed ASCO cutoff values which correspond to the respective categories of the current methods: An ASCO score larger than 17, 20, and 23 corresponds to ESMO categories 2, 3, and 4, respectively. ASCO cutoff values of 20 (23) and 36 (45) separate the score of ASCO into the three IQWiG_{RR} (Mod-IQWiG_{HR}) categories "minor", "considerable", and "major added benefit".

Overall, this thesis demonstrates that HR^{-} instead of HR^{+} should be used or the current thresholds should be at least adjusted to optimize the true positive and false positive rate. Thus, in future this research can be used as a guide for improvements of the methods and contributes to the enhancement of additional benefit assessment.

Chapter 6

Zusammenfassung

Bei der Entwicklung neuer Therapien für die Marktzulassung müssen verschiedene Anforderungen erfüllt werden. Nach der Zulassung wird zudem der Zusatznutzen mit bereits etablierten Behandlungen verglichen. Diese Bewertung kann über die Höhe der Erstattung der neuen Behandlung entscheiden und dazu beitragen, die Unsicherheit der Patienten hinsichtlich der Wirksamkeit und Toxizität zu beseitigen.

Im Falle von nicht heilbaren oder fortgeschrittenen Krankheiten wurden drei verschiedene Zusatznutzenbewertungsmethoden entwickelt, welche sich hauptsächlich in der verwendeten statistischen Metrik unterscheiden. Die Europäische Gesellschaft für Medizinische Onkologie (ESMO) und das Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen (IQWiG) haben Methoden mit einem ordinalen Ergebnis konstruiert. Für die primäre kategorielle Bewertung vergleicht IQWiG die obere Grenze des 95% Hazard Ratio (HR) Konfidenzintervall (KI) (HR^+) gegen Relative Risiko (RR) skalierte Schwellenwerte. ESMO verwendet die untere 95% HR-KI Grenze (HR^-). Die Methode der Amerikanischen Gesellschaft für Klinische Onkologie (ASCO) hat ein stetiges Ergebnis, welches den HR Punktschätzer (PS) verwendet. Es gibt mehrere Kritikpunkte für jede der Methoden. Zum Beispiel könnte die Verwendung des HR-PS dazu führen, dass Studien mit wesentlichem Zusatznutzen bestraft werden, weil die Varianz des Schätzers nicht betrachtet wird. Dagegen berücksichtigen HR^+ und HR^- die Variabilität vom HR-PS und sollten daher mehr Informationen enthalten.

Das Ziel dieser Thesis ist es ein besseres Verständnis der Methoden zu erlangen und die Frage zu beantworten, welche statistische Metrik bessere Eigenschaften besitzt, um den Zusatznutzen zu beurteilen. Um einen besseren Vergleich zwischen den verschiedenen Methoden zu ermöglichen wurde zudem untersucht, welche ESMO und IQWiG Kategorie welcher ASCO Punktzahl entspricht. Um diese Ziele zu erreichen, wurden die oben genannten Methoden mit Hilfe von umfangreichen Simulationsstudien evaluiert und miteinander verglichen. Die ursprüngliche IQWiG Methode ($IQWiG_{RR}$) und ESMO zeigen einen großen positiven Zusammenhang bei moderaten Behandlungseffekten. ASCO/$IQWiG_{RR}$ sowie ASCO/Mod-$IQWiG_{HR}$ (vorgeschlagene Änderung von $IQWiG_{RR}$, welche HR- anstatt RR-skalierte Schwel-

lenwerte verwendet) zeigen hingegen einen großen positiven Zusammenhang über die gesamte Breite von Behandlungseffekten. Zudem vergibt ESMO die maximale Kategorie exzessiv über die gesamte Breite von Behandlungseffekten (in den meisten Szenarien über 80% der Fälle) und kann daher zwischen kleinen und großen Behandlungseffekten nicht unterscheiden. ASCO und IQWiG_{RR} haben ein konservativeres Verhalten.

Verschiedene verletzte Annahmen wie über-/unterpowerte Studien, informative Zensierungen und nicht proportionale Hazards werden von ESMO nicht bestraft. Zudem sind die verwendeten Schwellenwerte von ESMO zu liberal, was zu hohen falsch-positiven Raten und einfach zu erreichender maximaler Kategorie führt. ASCO und IQWiG_{RR} weisen dahingegen ein erwünschtes Verhalten auf. In den meisten Fällen mit verletzten Annahmen ist die maximale Kategorie bei Mod-IQWiG_{HR} nicht exzessiv erhöht und daher eine bessere Lösung.

In anderen Szenarien, bei denen keine Annahmen verletzt sind, ist diese Methode aber zu konservativ, was zu einer niedrigen richtig-positiven Rate und geringer Rate an maximaler Kategorie führt. Bei vorliegen von verschiedenen Ereigniszeitverteilungen, welche die proportionale Hazard Annahme einhalten und daher keine Veränderung der Resultate gewünscht ist, zeigt ESMO weiterhin eine erhöhte Rate an maximaler Kategorie. ASCO, IQWiG_{RR}, und Mod-IQWiG_{HR} weisen nicht dieses unerwünschte Verhalten auf. Daher kann man schlussfolgern, dass ESMO die liberalste Methode ist.

Nichtsdestotrotz ist HR^{-}, welche ESMO unter anderem verwendet, die beste statistische Metrik um den Zusatznutzen zu beurteilen, wenn geeignete Schwellenwerte verwendet werden. Selbst HR-PS hat bessere Eigenschaften als HR^{+}.

Um einen besseren Vergleich zwischen den verschiedenen Methoden zu ermöglichen, wurden ASCO Grenzwerte vorgeschlagen, welche den ESMO und IQWiG Kategorien entsprechen: Eine ASCO Punktzahl größer als 17, 20 und 23 entsprechen den ESMO Kategorien 2, 3 und 4. Eine ASCO Punktzahl größer als 20 (23) und 36 (45) entsprechen den IQWiG_{RR} (Mod-IQWiG_{HR}) Kategorien "minor", "considerable" und "major added benefit".

Die Ergebnisse der Dissertation zeigen, dass um die richtig-positive und falsch-positive Rate zu optimieren, HR^{-} anstatt HR^{+} verwendet oder zu mindestens die derzeitigen Schwellenwerte angepasst werden sollten. Daher kann diese Forschung als Hilfestellung bei zukünftigen Verbesserungen der Zusatznutzenbewertungsmethoden verwendet werden und trägt zur Weiterentwicklung der Zusatznutzenbewertung bei.

Chapter 7

References list

Abramowitz, M. and Stegun, I. A. (1965). **Handbook of mathematical functions**. Dover books on intermediate and advanced mathematics. Dover Publ., New York.

Al-Sarraf, M., LeBlanc, M., Giri, P. G., Fu, K. K., Cooper, J., Vuong, T., Forastiere, A. A., Adams, G., Sakr, W. A., Schuller, D. E., and Ensley, J. F. (1998). **Chemoradiotherapy versus radiotherapy in patients with advanced nasopharyngeal cancer: phase III randomized intergroup study 0099.** JCO, 16(4):1310–1317, doi: 10.1200/JCO.1998.16.4.1310.

Askey, R. A. and Daalhuis, A. B. O. (2010). **Generalized hypergeometric functions and Meijer G-function.** In NIST Handbook of Mathematical Functions, edited by F. Olver, D. Lozier, R. Boisvert and C. Clark, Cambridge University Press. pg 403-418.

Becker, D. J., Lin, D., Lee, S., Levy, B. P., Makarov, D. V., Gold, H. T., and Sherman, S. (2017). **Exploration of the ASCO and ESMO value frameworks for antineoplastic drugs**. JCO, 13(7):e653–e665, doi: 10.1200/JOP.2016.020339.

Brent, G. D. and Brent, R. P. (1974). **Algorithms for minimization without derivatives**. Math Comput, 28(127):865–866, doi: 10.2307/2005713.

Büsch, C. A., Kirchner, M., Behnisch, R., and Kieser, M. (2024). **A comparison of additional benefit assessment methods for time-to-event endpoints using hazard ratio point estimates or confidence interval limits by means of a simulation study**. Med Decis Making, 44(4):365–379, doi: 10.1177/0272989X241239928.

Büsch, C. A., Krisam, J., and Kieser, M. (2022). **A comprehensive comparison of additional benefit assessment methods applied by institute for quality and efficiency in health care and european society for medical oncology for time-to-event endpoints after significant phase III trials - a simulation study**. Value Health, 25(11):1853–1862, doi: 10.1016/j.jval.2022.05.015.

Chang, C.-H., Yang, J.-T., and Lee, M.-H. (2015). **A novel "maximizing Kappa" approach for assessing the ability of a diagnostic marker and its optimal cutoff value**. J Biopharm Stat, 25(5):1005–1019, doi: 10.1080/10543406.2014.920347.

Cheng, S., McDonald, E. J., Cheung, M. C., Arciero, V. S., Qureshi, M., Jiang, D., Ezeife, D., Sabharwal, M., Chambers, A., Han, D., Leighl, N., Sabarre, K.-A., and Chan, K. K. W. (2017). **Do the american society of clinical oncology value framework and the european society of medical oncology magnitude of clinical benefit scale measure the same construct of clinical benefit?** J Clin Oncol, 35(24):2764–2771, doi: 10.1200/JCO.2016.71.6894.

Cherny, N., Dafni, U., Bogaerts, J., Latino, N., Pentheroudakis, G., Douillard, J.-Y., Tabernero, J., Zielinski, C., Piccart, M., and de Vries, E. (2017). **ESMO-magnitude of clinical benefit scale version 1.1**. Ann Oncol, 28(10):2340–2366, doi: 10.1093/annonc/mdx310.

Cherny, N., Sullivan, R., Dafni, U., Kerst, J., Sobrero, A., Zielinski, C., de Vries, E., and Piccart, M. (2015). **A standardised, generic, validated approach to stratify the magnitude of clinical benefit that can be anticipated from anti-cancer therapies: the european society for medical oncology magnitude of clinical benefit scale (ESMO-MCBS)**. Ann Oncol, 26(8):1547–1573, doi: 10.1093/annonc/mdv249.

Cherny, N. I., de Vries, E. G. E., Dafni, U., Garrett-Mayer, E., McKernin, S. E., Piccart, M., Latino, N. J., Douillard, J.-Y., Schnipper, L. E., Somerfield, M. R., Bogaerts, J., Karlis, D., Zygoura, P., Vervita, K., Pentheroudakis, G., Tabernero, J., Zielinski, C., Wollins, D. S., and Schilsky, R. L. (2019). **Comparative assessment of clinical benefit using the ESMO-magnitude of clinical benefit scale version 1.1 and the ASCO value framework net health benefit score**. J Clin Oncol, 37(4):336–349, doi: 10.1200/JCO.18.00729.

Cohen, J. (1960). **A coefficient of agreement for nominal scales**. Educ Psychol Meas, 20(1):37–46, doi: 10.1177/001316446002000104.

Cox, D. R. (1972). **Regression models and life-tables**. J R Stat Soc Series B Stat Methodol, 34(2):187–220, doi: 10.1111/j.2517-6161.1972.tb00899.x.

Dafni, U., Karlis, D., Pedeli, X., Bogaerts, J., Pentheroudakis, G., Tabernero, J., Zielinski, C. C., Piccart, M. J., de Vries, E. G. E., Latino, N. J., Douillard, J.-Y., and Cherny, N. I. (2017). **Detailed statistical assessment of the characteristics of the ESMO magnitude of clinical benefit scale (ESMO-MCBS) threshold rules**. ESMO Open, 2(4):e000216–e000216, doi: 10.1136/esmoopen-2017-000216.

Danzon, P. M. and Taylor, E. (2010). **Drug pricing and value in oncology**. Oncologist, 15(S1):24–31, doi: 10.1634/theoncologist.2010-S1-24.

Del Paggio, J., Sullivan, R., Hopman, W., and Booth, C. (2018). **Re-aligning the ASCO and ESMO clinical benefit frameworks for modern cancer therapies**. Ann Oncol, 29(3):773–774, doi: 10.1093/annonc/mdx721.

Filzmoser, P., Fritz, H., and Kalcher, K. (2022). **pcaPP: Robust PCA by projection pursuit**. R package version 2.0-2, `https://CRAN.R-project.org/package=pcaPP`.

Jackson, C. H. (2016). **flexsurv: A platform for parametric survival modeling in R**. J Stat Softw, 70(8):1–33, doi: 10.18637/jss.v070.i08.

Kassambara, A. (2023). **ggpubr: 'ggplot2' based publication ready plots**. R package version 0.6.0, `https://CRAN.R-project.org/package=ggpubr`.

Kolonko, M. (2008). **Stochastische simulation**. Vieweg+Teubner Verlag, Wiesbaden, doi: 10.1007/978-3-8348-9290-4.

Meyer, D., Zeileis, A., and Hornik, K. (2022). **vcd: Visualizing categorical data**. R package version 1.4-10, `https://cran.r-project.org/web/packages/vcd/index.html`.

Morris, T. P., White, I. R., and Crowther, M. J. (2019). **Using simulation studies to evaluate statistical methods**. Stat Med, 38(11):2074–2102, doi: 10.1002/sim.8086.

Muhonen, T., Joensuu, H., and Pfeiffer, P. (2015). **Comment on ESMO magnitude of clinical benefit scale**. Ann Oncol, 26(12):2504–2504, doi: 10.1093/annonc/mdv384.

Mukaka, M. M. (2012). **Statistics corner: A guide to appropriate use of correlation coefficient in medical research**. Malawi Med J, 24(3):69–71.

Piessens, R. (1983). **QUADPACK: a subroutine package for automatic integration**. Springer, Berlin; Heidelberg.

R Core Team (2021). **R: A language and environment for statistical computing**. R version 4.2.1, `https://www.R-project.org/`.

RStudio Team (2021). **Rstudio: Integrated development environment for R**. RStudio version 2022.07.2, `https://posit.co/`.

Schnipper, L. E., Davidson, N. E., Wollins, D. S., Blayney, D. W., Dicker, A. P., Ganz, P. A., Hoverman, J. R., Langdon, R., Lyman, G. H., Meropol, N. J., Mulvey, T., Newcomer, L., Peppercorn, J., Polite, B., Raghavan, D., Rossi, G., Saltz, L., Schrag, D., Smith, T. J., Yu, P. P., Hudis, C. A., Vose, J. M., and Schilsky, R. L. (2016). **Updating the american society of clinical oncology value framework: Revisions and reflections in response to comments received**. J Clin Oncol, 34(24):2925–2934, doi: 10.1200/JCO.2016.68.2518.

Schnipper, L. E., Davidson, N. E., Wollins, D. S., Tyne, C., Blayney, D. W., Blum, D., Dicker, A. P., Ganz, P. A., Hoverman, J. R., Langdon, R., Lyman, G. H., Meropol, N. J., Mulvey, T., Newcomer, L., Peppercorn, J., Polite, B., Raghavan, D., Rossi, G., Saltz, L., Schrag, D., Smith, T. J., Yu, P. P., Hudis, C. A., Schilsky, R. L., and of Clinical Oncology, A. S. (2015). **American society of clinical oncology statement: A conceptual framework to assess the value of cancer treatment options**. J Clin Oncol, 33(23):2563–2577, doi: 10.1200/JCO.2015.61.6706.

Schoenfeld, D. A. (1981). **The asymptotic properties of nonparametric tests for comparing survival distributions**. Biometrika, 68(1):316–319, doi: 10.1093/biomet/68.1.316.

Schoenfeld, D. A. (1983). **Sample-size formula for the proportional-hazards regression model**. Biometrics, 39(2):499–503, doi: 10.2307/2531021.

Skipka, G., Wieseler, B., Kaiser, T., Thomas, S., Bender, R., Windeler, J., and Lange, S. (2016). **Methodological approach to determine minor, considerable, and major treatment effects in the early benefit assessment of new drugs**. Biom J, 58(1):43–58, doi: 10.1002/bimj.201300274.

Svensson, E. (2000a). **Comparison of the quality of assessments using continuous and discrete ordinal rating scales**. Biom J, 42(4):417–434, doi: 10.1002/1521-4036(200008)42:4<417::AID-BIMJ417>3.0.CO;2-Z.

Svensson, E. (2000b). **Concordance between ratings using different scales for the same variable**. Stat Med, 19(24):3483–3496, doi: 10.1002/1097-0258(20001230)19:24<3483::AID-SIM786>3.0.CO;2-A.

Therneau, T. M. (2023). **A package for survival analysis in R**. R package version 3.5-7, `https://CRAN.R-project.org/package=survival`.

Therneau, T. M. and Grambsch, P. M. (2000). **Modeling survival data**. Springer, New York, Berlin, Heidelberg.

Thiele, C. and Hirschfeld, G. (2021). **cutpointr : Improved estimation and validation of optimal cutpoints in R**. J Stat Softw, 98(11):1–27, doi: 10.18637/jss.v098.i11.

VanderWeele, T. J. (2020). **Optimal approximate conversions of odds ratios and hazard ratios to risk ratios**. Biometrics, 76(3):746–752, doi: 10.1111/biom.13197.

Weeks, J. C., Catalano, P. J., Cronin, A., Finkelman, M. D., Mack, J. W., Keating, N. L., and Schrag, D. (2012). **Patients' expectations about effects of chemotherapy for advanced cancer**. N Engl J Med, 367(17):1616–1625, doi: 10.1056/NEJMoa1204410.

West, H., McCleod, M., Hussein, M., Morabito, A., Rittmeyer, A., Conter, H. J., Kopp, H.-G., Daniel, D., McCune, S., Mekhail, T., Zer, A., Reinmuth, N., Sadiq, A., Sandler, A., Lin, W., Ochi Lohmann, T., Archer, V., Wang, L., Kowanetz, M., and Cappuzzo, F. (2019). **Atezolizumab in combination with carboplatin plus nab-paclitaxel chemotherapy compared with chemotherapy alone as first-line treatment for metastatic non-squamous non-small-cell lung cancer (IMpower130): a multi-**

centre, randomised, open-label, phase 3 trial. Lancet Oncol, 20(7):924–937, doi: 10.1016/S1470-2045(19)30167-6.

Wickham, H. (2016). **ggplot2: elegant graphics for data analysis**. Springer International Publishing, Cham, doi: 10.18637/jss.v035.b01.

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T. L., Miller, E., Bache, S. M., Müller, K., Ooms, J., Robinson, D., Seidel, D. P., Spinu, V., Takahashi, K., Vaughan, D., Wilke, C., Woo, K., and Yutani, H. (2019). **Welcome to the tidyverse**. J Open Source Softw, 4(43):1686, doi: 10.21105/joss.01686.

Wild, C., Grössmann, N., Bonanno, P., Bucsics, A., Furst, J., Garuoliene, K., Godman, B., Gulbinovič, J., Jones, J., Pomorski, M., and Emprechtinger, R. (2016). **Utilisation of the ESMO-MCBS in practice of HTA**. Ann Oncol, 27(11):2134–2136, doi: 10.1093/annonc/mdw297.

Chapter 8

Personal contribution and publications

Partial of this dissertation have already been published in the following articles:

1. **Büsch, C. A.**, Krisam, J., and Kieser, M. (2022). **A comprehensive comparison of additional benefit assessment methods applied by institute for quality and efficiency in health care and european society for medical oncology for time-to-event endpoints after significant phase III trials - a simulation study.** Value Health, 25(11):1853–1862, doi:10.1016/j.jval.2022.05.015

2. **Büsch, C. A.**, Kirchner, M., Behnisch, R., and Kieser, M. (2024). **A comparison of additional benefit assessment methods for time-to-event endpoints using hazard ratio point estimates or confidence interval limits by means of a simulation study.** Med Decis Making, 44(4):365–379, doi:10.1177/0272989X241239928.

Publication 1 gives a detailed insight in the comparison of the additional benefit assessment methods of IQWiG and ESMO. In addition, which statistical quantity has the best properties to assess additional benefit is investigated. Hence, the description of the additional benefit assessment methods (Section 2.2) and of Simulation 1 (Section 2.3.1) of this thesis have already been published in this article. Furthermore, the results of Simulation 1 (Section 3.1) and conclusions drawn (Chapter 4) are also depicted in this publication. The manuscript has been written by the lead author but may contain comments and corrections from the co-authors and the reviewers.
Publication 2 gives a detailed overview of all three additional benefit assessment methods (ASCO, IQWiG and ESMO) and provides the first comparison between IQWiG and ASCO. In addition, in order to achieve an easier comparison between all three methods an investiga-

tion is performed in which corresponding ASCO scores are determined, which correspond to the categories of ESMO and IQWiG. Hence, the description of the additional benefit assessment methods (Section 2.2) and of Simulation 2 (Section 2.3.2) of this thesis have already been published in this article. Furthermore, the results of Simulation 2 (Section 3.2) and conclusions drawn (Chapter 4) are also depicted in this publication. Rouven Behnisch as co-author of this publication had the idea for the censoring data generation for independent administrative censoring and independent right-censoring with an overall targeted censoring proportion of p_C without introducing bias to the HR estimation. The manuscript has been written by the lead author but may contain comments and corrections from the co-authors and the reviewers.

Conference contributions

The presenting author is underlined.

1. **Büsch C.A.**, Krisam J., Kieser M. (2021). **Assessment of additional benefit for time-to-event endpoints after significant phase III trials – investigation of ESMO and IQWiG approaches.** 67th Biometrisches Kolloquium. (virtual) Presentation. Münster, Germany.

2. **Büsch C.A.**, Krisam J., Kieser M. (2021). **Additional benefit method assessment for time-to-event endpoints – A comparison of ESMOs and IQWiGs approach.** 42nd Annual Conference of the International Society (ISCB 2021). (virtual) Presentation. Lyon, France.

Appendix A

Additional results

In this Appendix, additional results of both simulation studies showing no influence compared to the Standard Scenario are illustrated. The list of figures shown at the beginning of this thesis provides a convenient overview of all figures provided in this Appendix.

The Appendix is structured as follows: In Section A.1 the results of Simulation 1 are outlined. Firstly, the relationship between the additional benefit assessment methods is shown by displaying pairwise Spearman and/or Kendall-τ_b correlations between the methods using heatmaps for each simulation scenario. Secondly, the description of the additional benefit assessment methods are shown for each simulation scenario, where the ASCO score distribution is illustrated using boxplots separated into the categories of ESMO, IQWiG_{RR}, $\text{Mod-IQWiG}_{\text{HR}}$, and overall, respectively (y-axis). Thirdly, the results of each specific simulation scenario are shown including the relationship between methods (line figures of Spearman correlation) and AUC values of ROC curves (nested loop plots).

In Section A.2 the results of Simulation 2 are outlined in the same way described above.

A.1 Simulation 1

Figure 30: *Pairwise Kendall-τ_b correlation of the additional benefit assessment methods (x-axis) for the different scenarios (y-axis) where all sub-scenarios were combined for correlation calculation.*

A.1.1 Standard Scenario

A.1.1.1 Relationship between methods

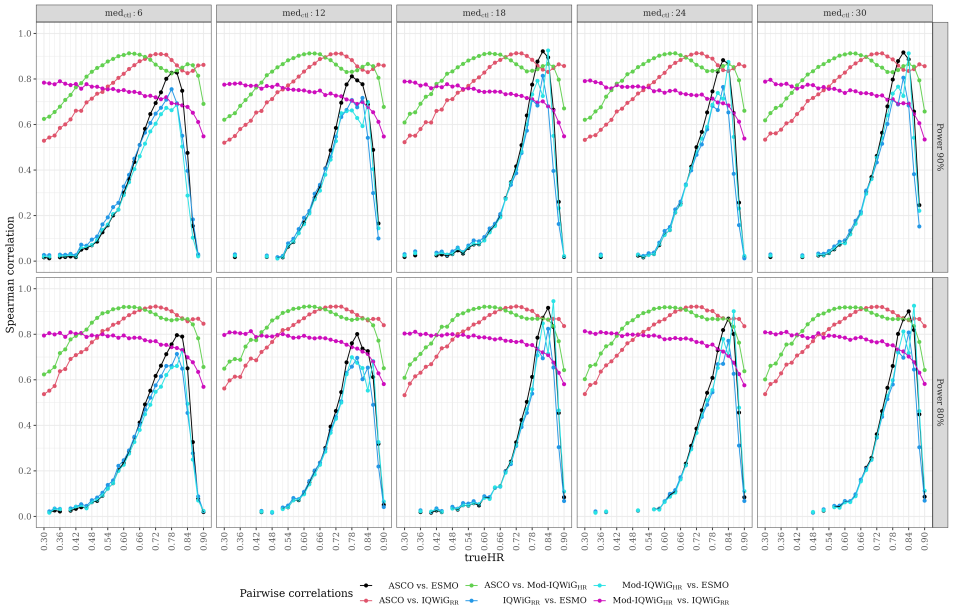

Figure 31: *Pairwise Spearman correlation (y-axis) separated by trueHR (x-axis) of Standard Scenario with p_C=20%.*

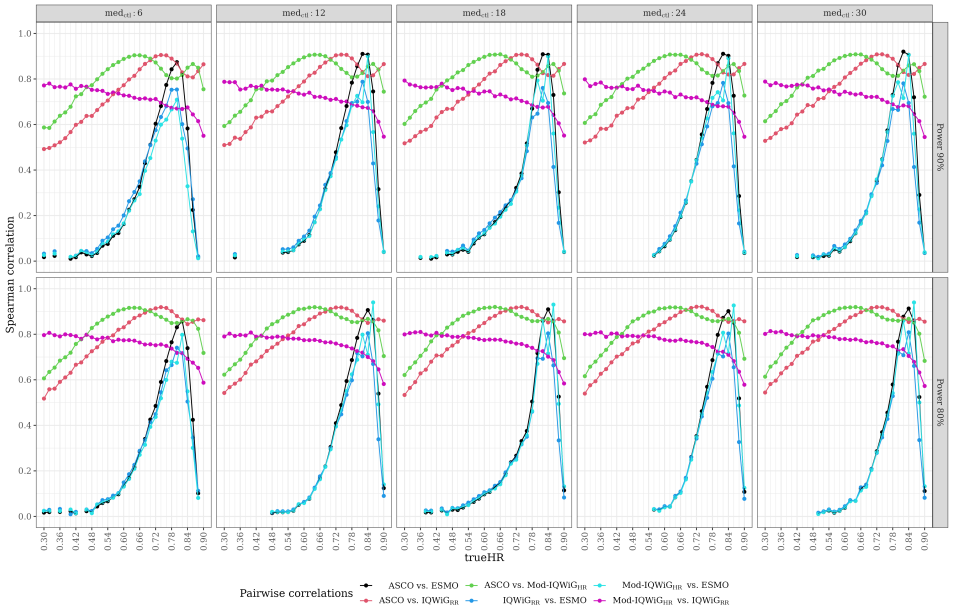

Figure 32: *Pairwise Spearman correlation (y-axis) separated by trueHR (x-axis) of Standard Scenario with p_C=40%.*

A.1.1.2 Using constant sample size:

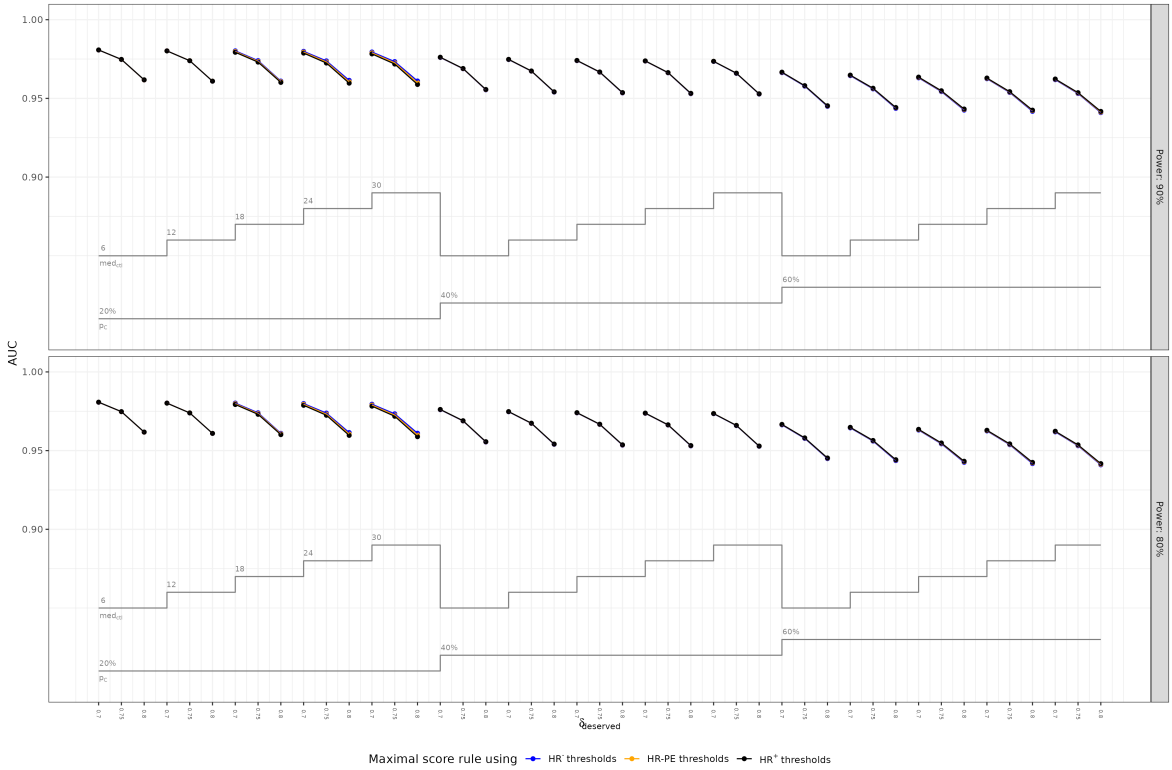

Figure 33: *AUC of ROC curves (y-axis) separated by $\delta_{deserved}$ (x-axis) of Standard Scenario for each sub-scenario with constant sample size.*

A.1.2 Scenario 2 (incorrect assumed designHR for sample size calculation)

A.1.2.1 Relationship between methods

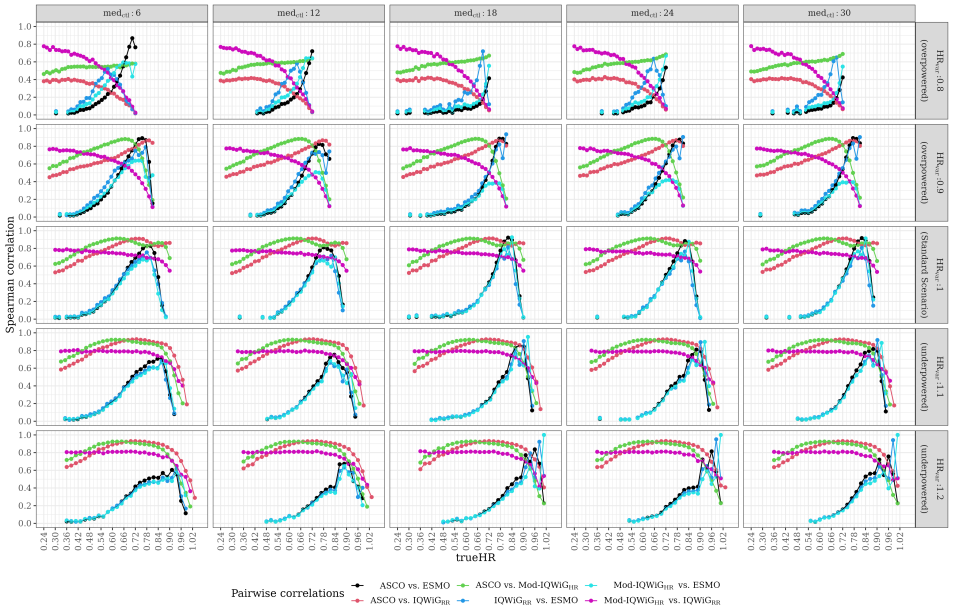

Figure 34: *Pairwise Spearman correlation (y-axis) separated by trueHR (x-axis) of Scenario 2 with p_C=20% and 90% power.*

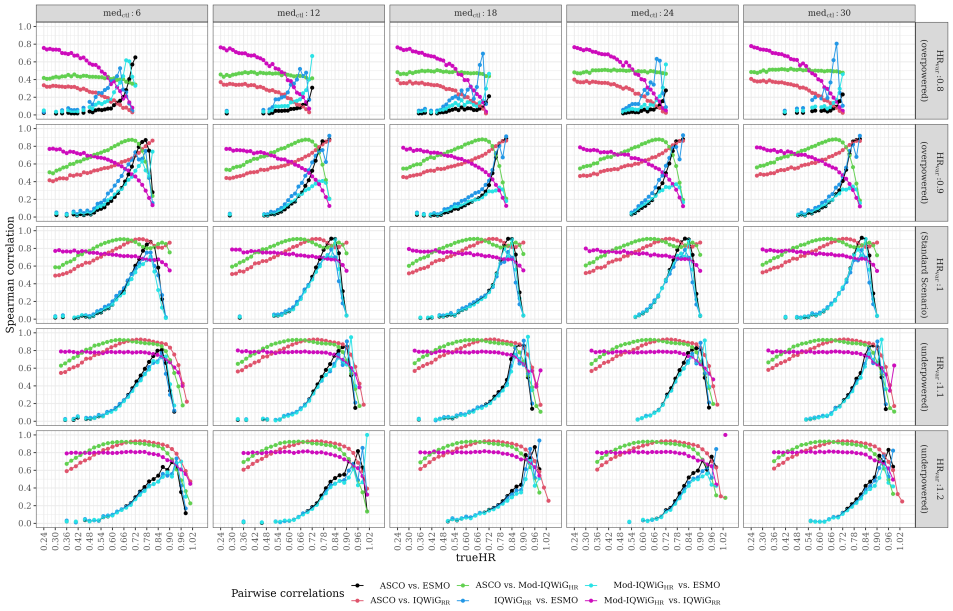

Figure 35: *Pairwise Spearman correlation (y-axis) separated by trueHR (x-axis) of Scenario 2 with p_C=40% and 90% power.*

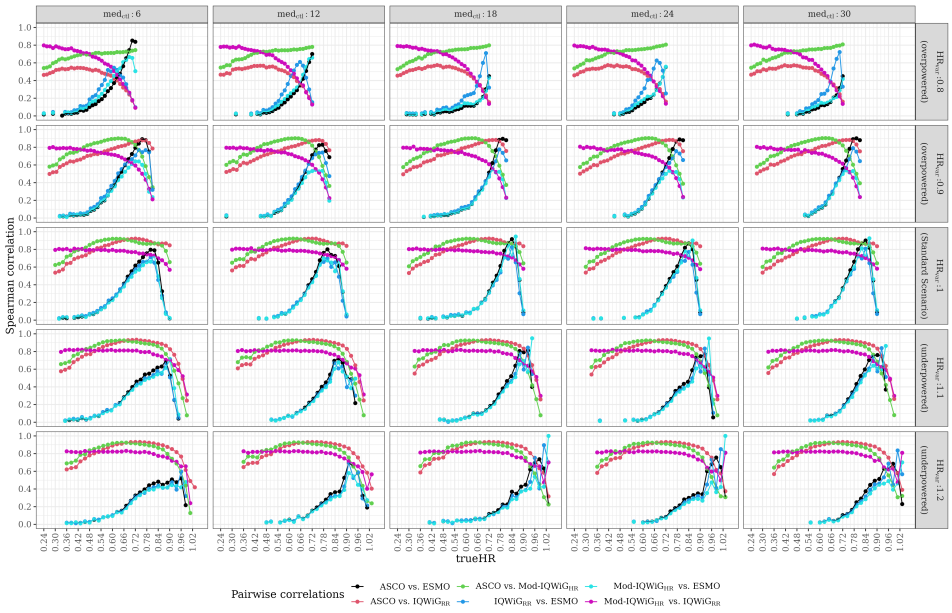

Figure 36: *Pairwise Spearman correlation (y-axis) separated by trueHR (x-axis) of Scenario 2 with p_C=20% and 80% power.*

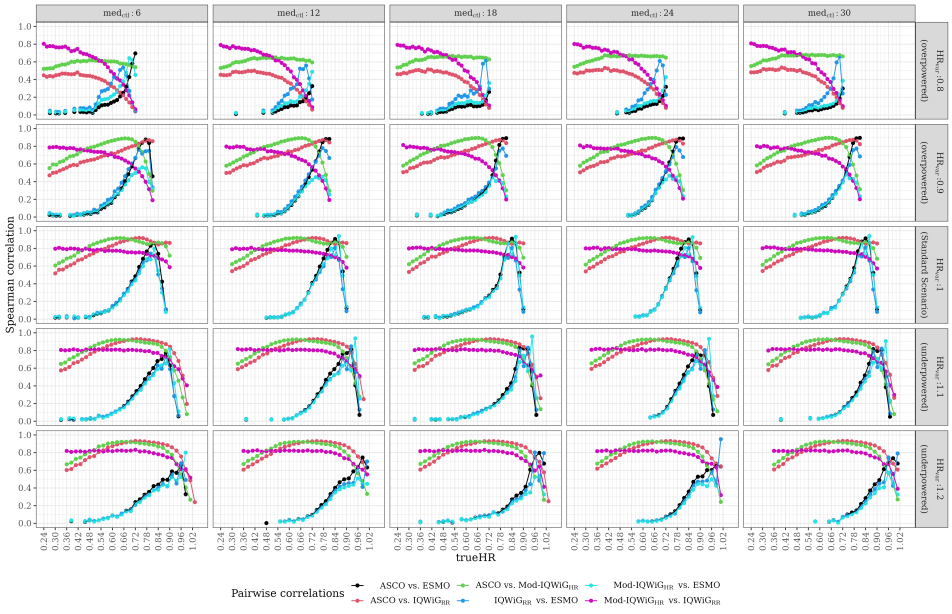

Figure 37: *Pairwise Spearman correlation (y-axis) separated by trueHR (x-axis) of Scenario 2 with p_C=40% and 80% power.*

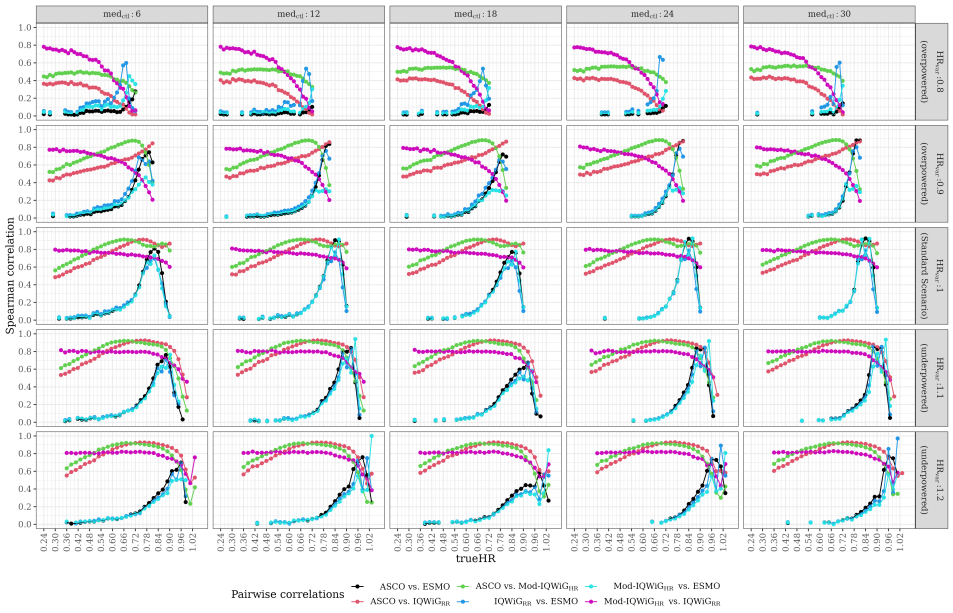

Figure 38: *Pairwise Spearman correlation (y-axis) separated by trueHR (x-axis) of Scenario 2 with p_C=60% and 80% power.*

A.1.2.2 AUC

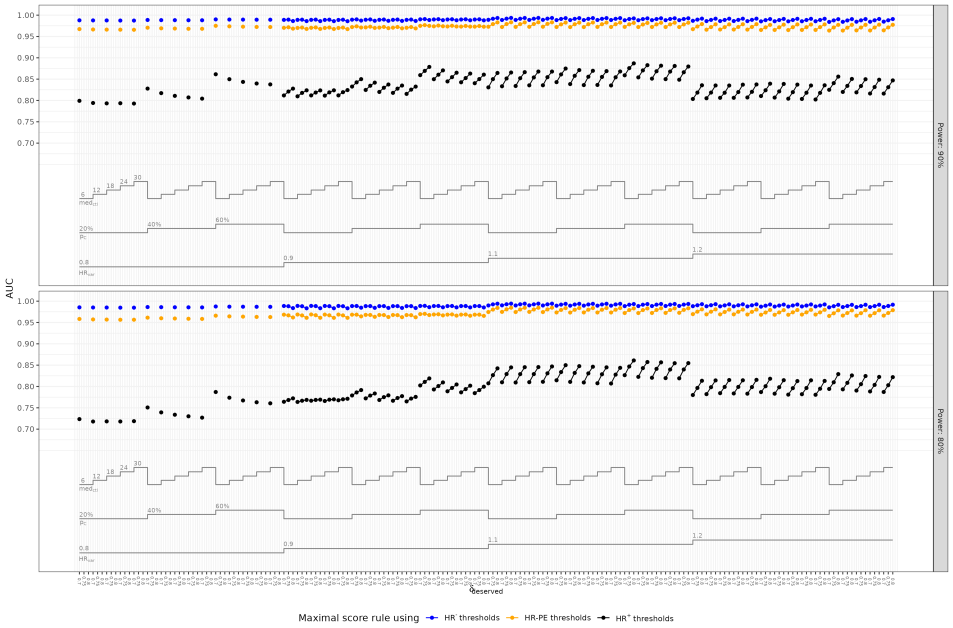

Figure 39: *AUC of ROC curves (y-axis) separated by $\delta_{deserved}$ (x-axis) of Scenario 2 for each sub-scenario.*

A.1.3 Scenario 3 (different failure time distributions)

A.1.3.1 Relationship between methods

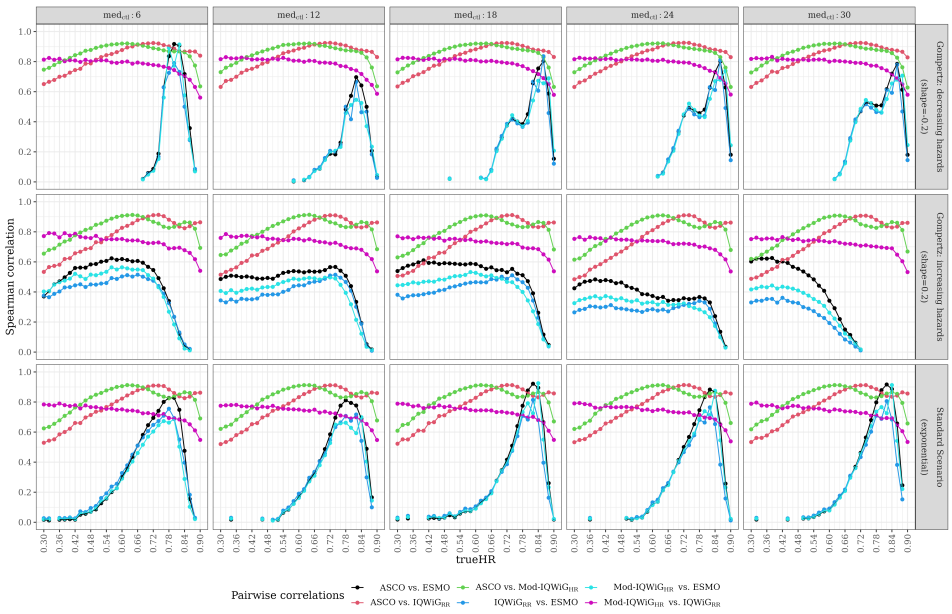

Figure 40: *Pairwise Spearman correlation (y-axis) separated by trueHR (x-axis) of Scenario 3 with Gompertz failure time distribution, p_C=20%, and 90% power.*

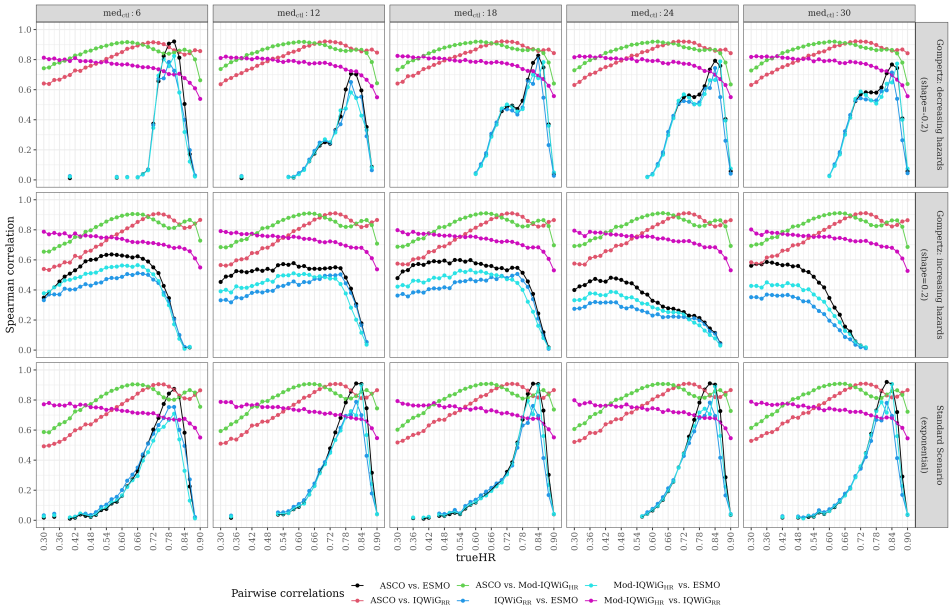

Figure 41: *Pairwise Spearman correlation (y-axis) separated by trueHR (x-axis) of Scenario 3 with Gompertz failure time distribution, p_C=40%, and 90% power.*

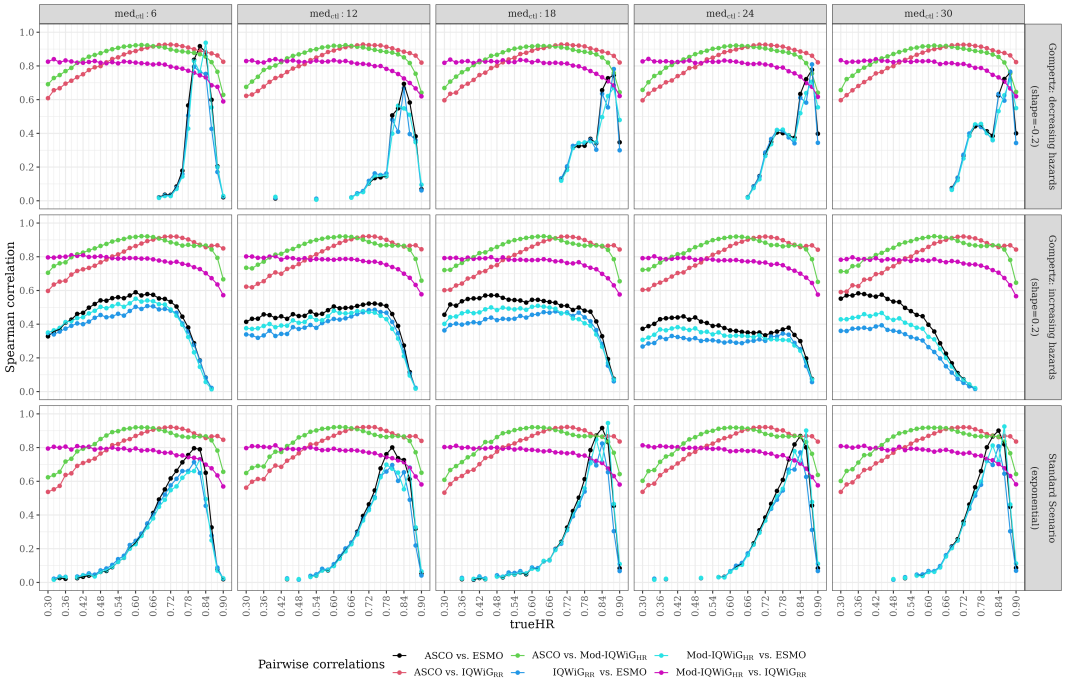

Figure 42: *Pairwise Spearman correlation (y-axis) separated by trueHR (x-axis) of Scenario 3 with Gompertz failure time distribution, p_C=20%, and 80% power.*

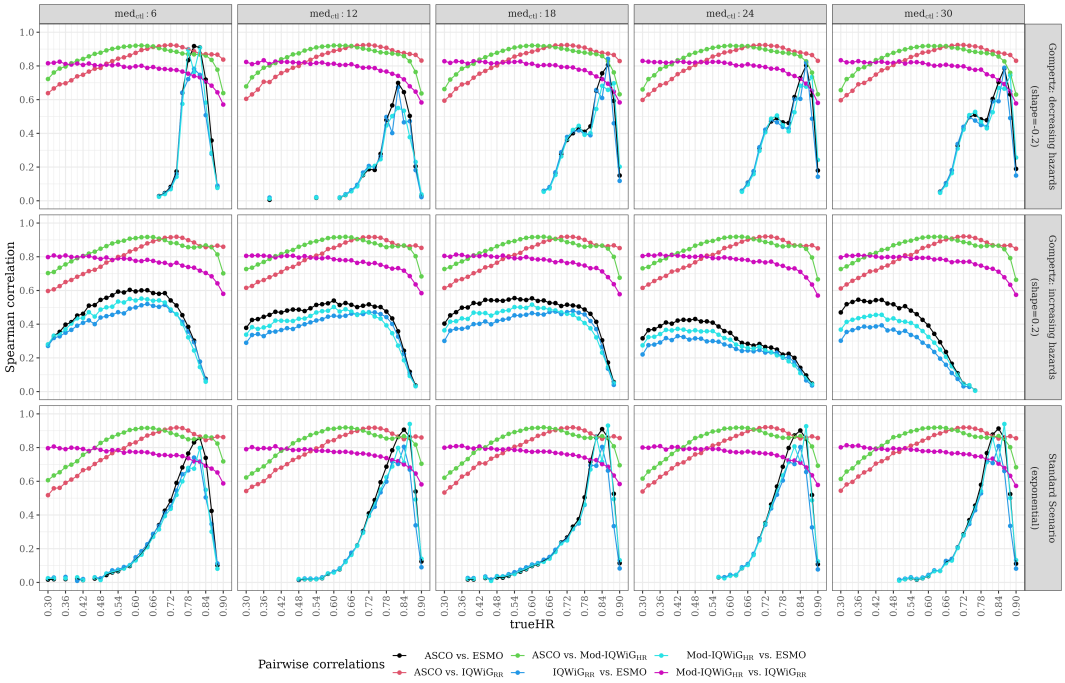

Figure 43: *Pairwise Spearman correlation (y-axis) separated by trueHR (x-axis) of Scenario 3 with Gompertz failure time distribution, p_C=40%, and 80% power.*

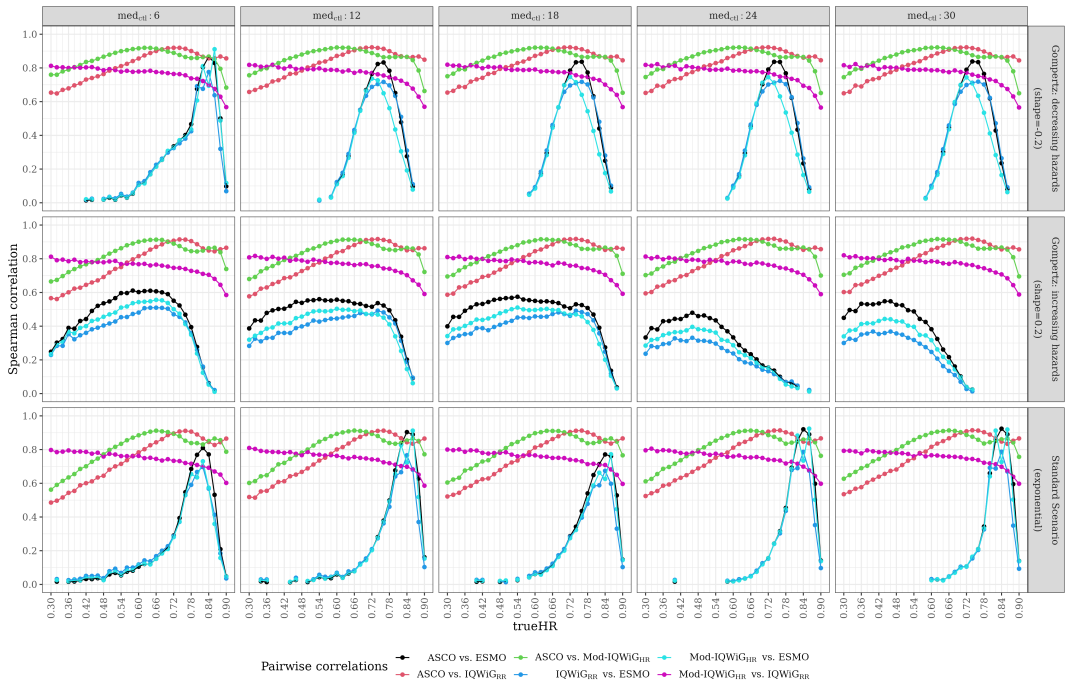

Figure 44: *Pairwise Spearman correlation (y-axis) separated by trueHR (x-axis) of Scenario 3 with Gompertz failure time distribution, p_C=60%, and 80% power.*

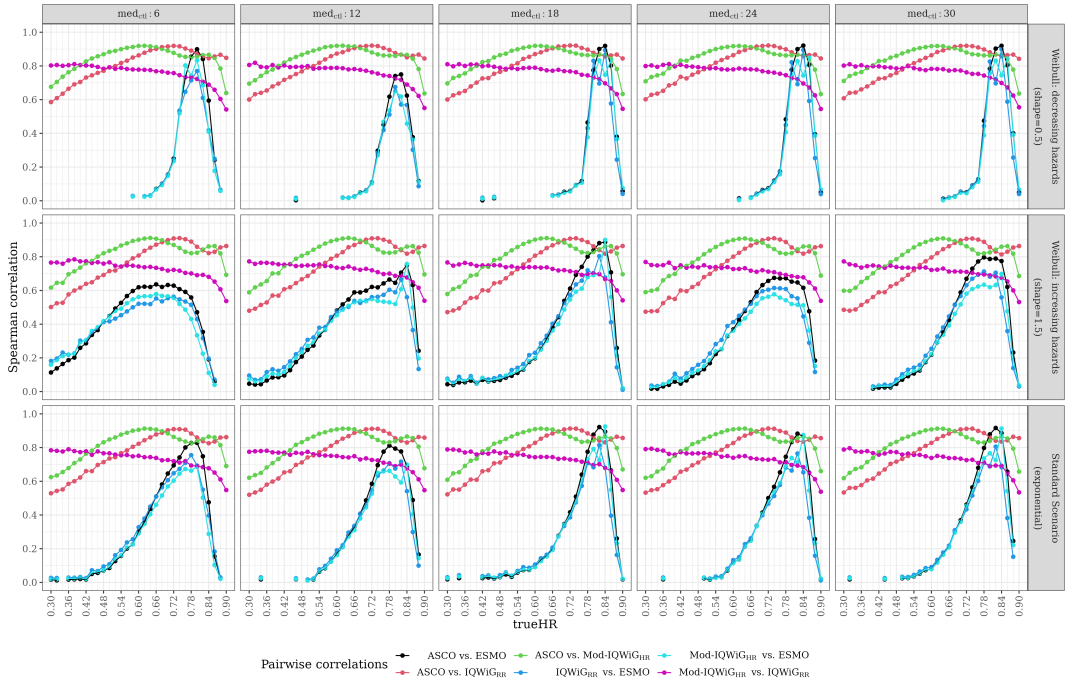

Figure 45: *Pairwise Spearman correlation (y-axis) separated by trueHR (x-axis) of Scenario 3 with Weibull failure time distribution, p_C=20%, and 90% power.*

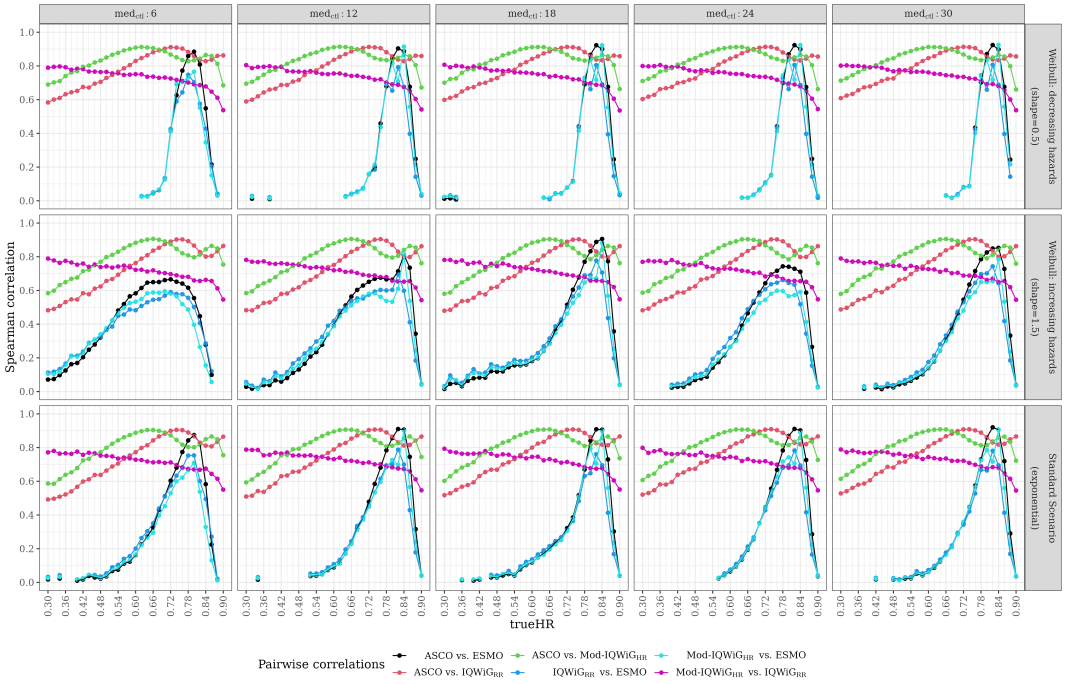

Figure 46: *Pairwise Spearman correlation (y-axis) separated by trueHR (x-axis) of Scenario 3 with Weibull failure time distribution, p_C=40%, and 90% power.*

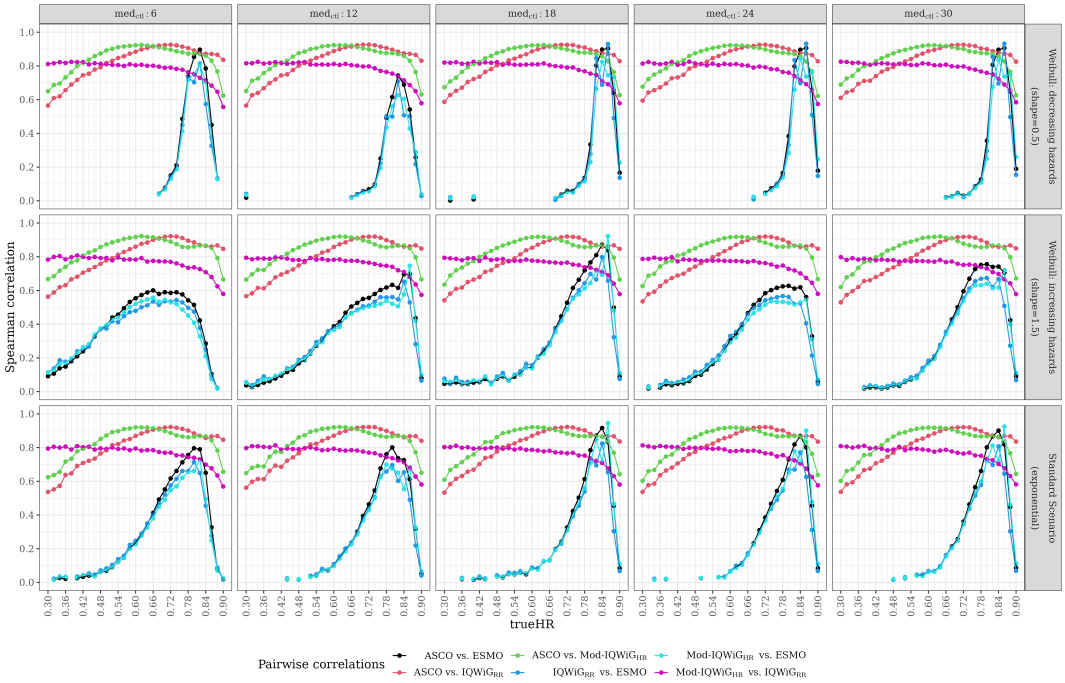

Figure 47: *Pairwise Spearman correlation (y-axis) separated by trueHR (x-axis) of Scenario 3 with Weibull failure time distribution, p_C=20%, and 80% power.*

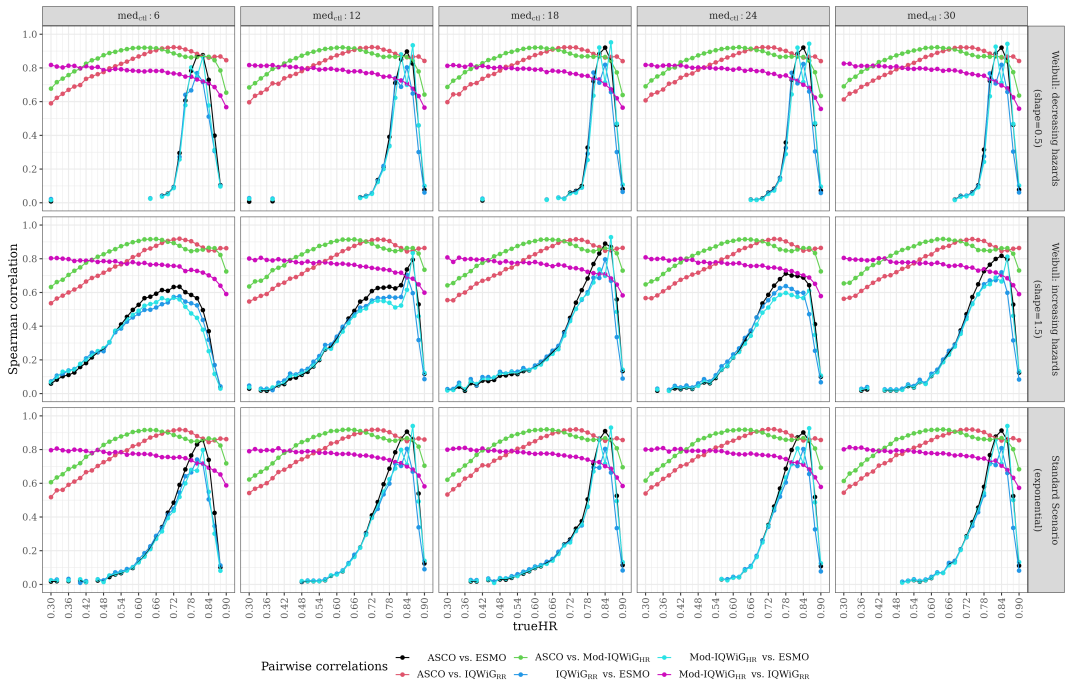

Figure 48: *Pairwise Spearman correlation (y-axis) separated by trueHR (x-axis) of Scenario 3 with Weibull failure time distribution, p_C=40%, and 80% power.*

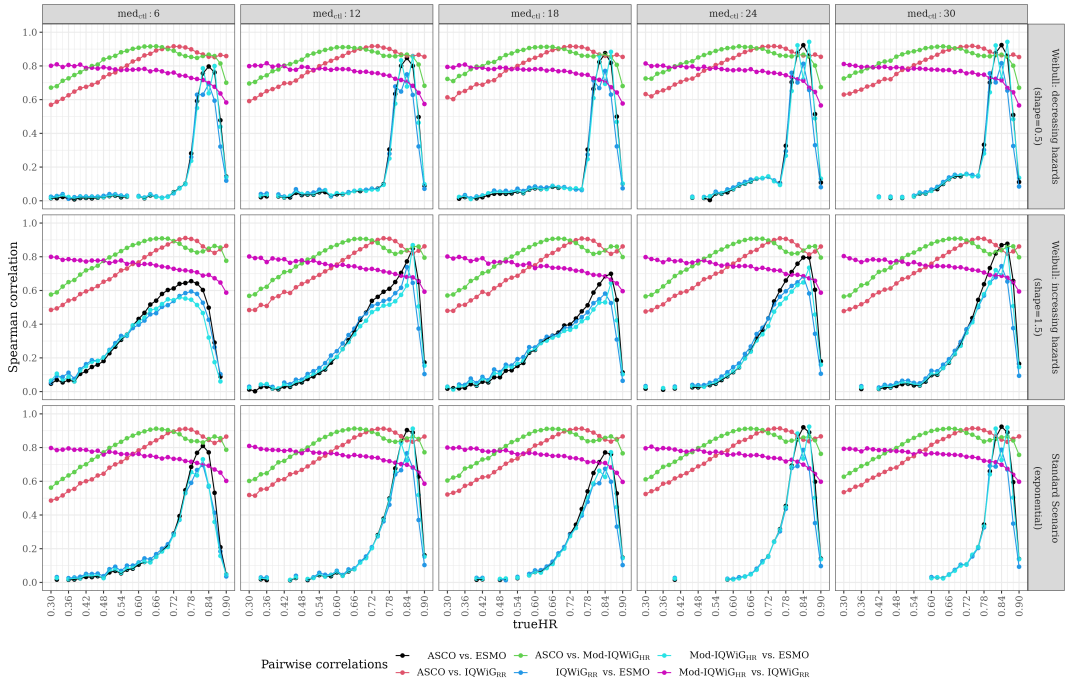

Figure 49: *Pairwise Spearman correlation (y-axis) separated by trueHR (x-axis) of Scenario 3 with Weibull failure time distribution, p_C=60%, and 80% power.*

A.1.3.2 AUC

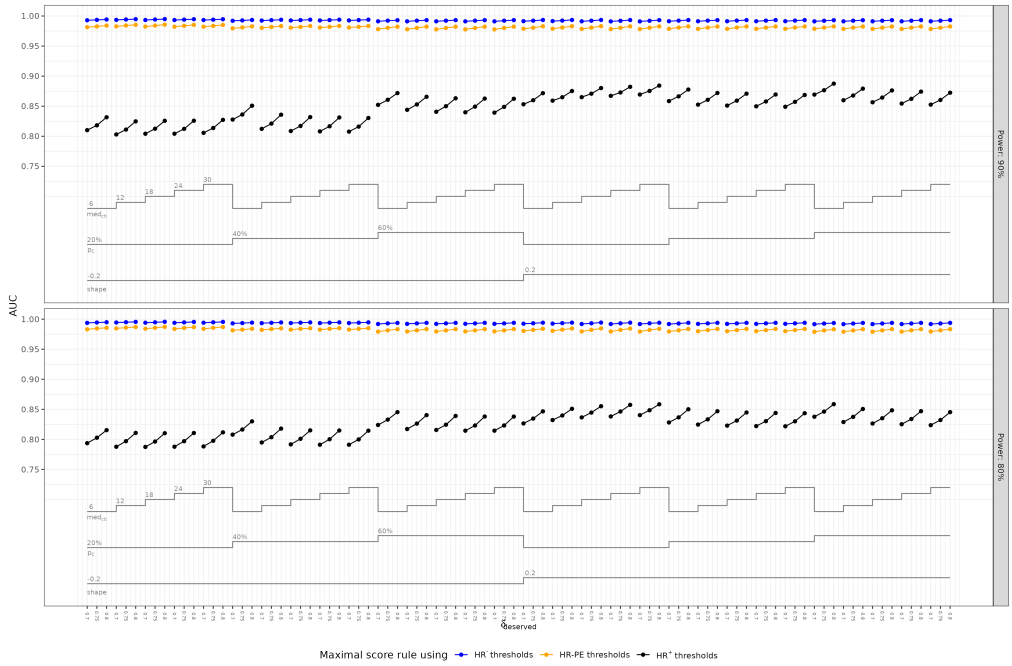

Figure 50: *AUC of ROC curves (y-axis) separated by $\delta_{deserved}$ (x-axis) of Scenario 3 with Gompertz failure time distribution for each sub-scenario.*

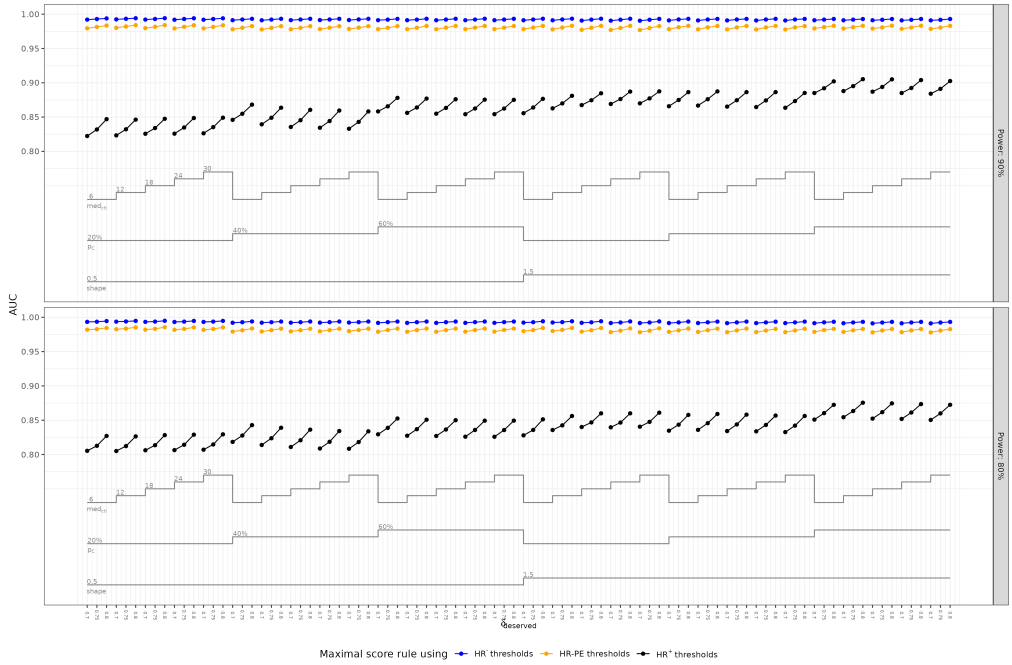

Figure 51: *AUC of ROC curves (y-axis) separated by $\delta_{deserved}$ (x-axis) of Scenario 3 with Weibull failure time distribution for each sub-scenario.*

A.1.4 Scenario 4 (non-proportional hazards)

A.1.4.1 Relationship between methods

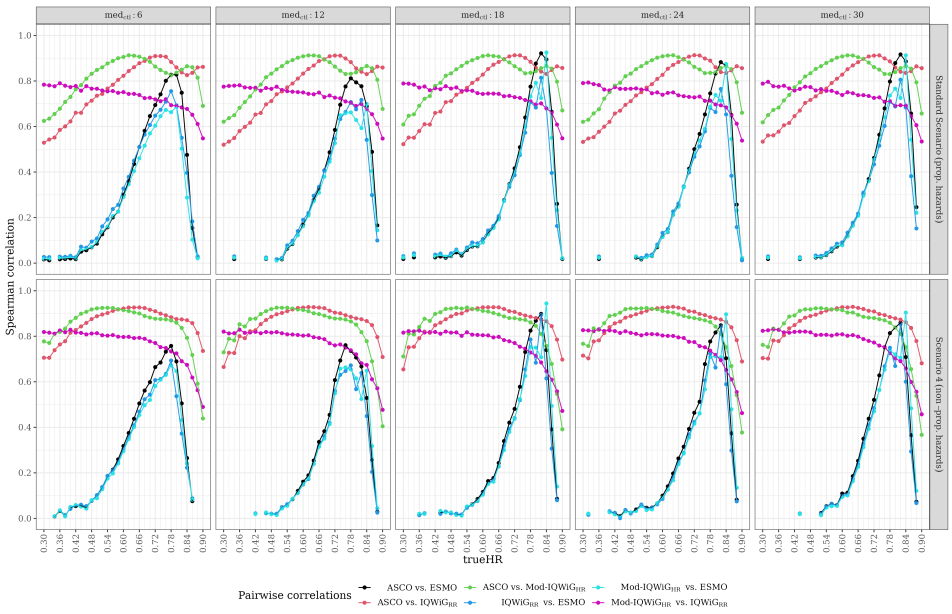

Figure 52: *Pairwise Spearman correlation (y-axis) separated by trueHR (x-axis) of Scenario 4 with p_C=20% and 90% power.*

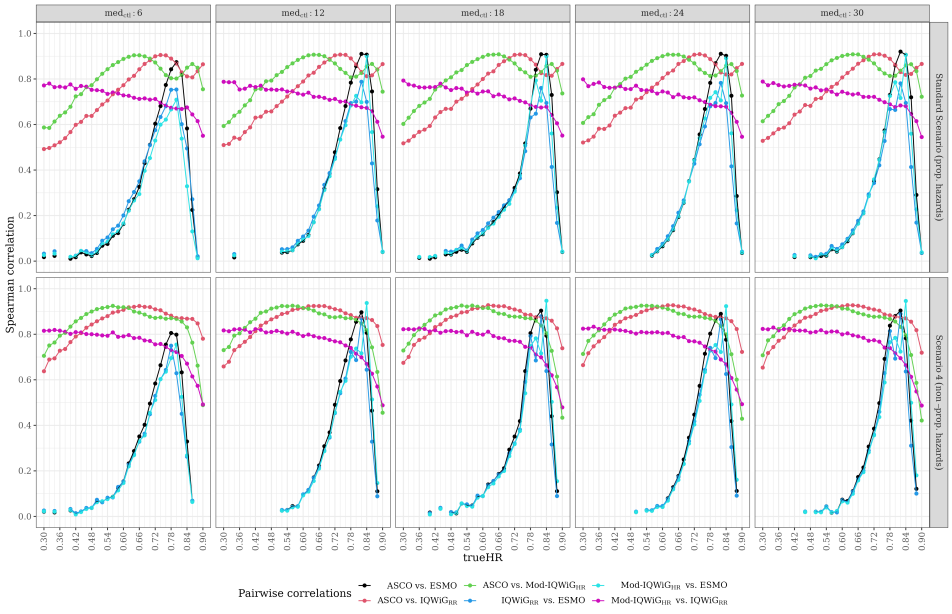

Figure 53: *Pairwise Spearman correlation (y-axis) separated by trueHR (x-axis) of Scenario 4 with p_C=40% and 90% power.*

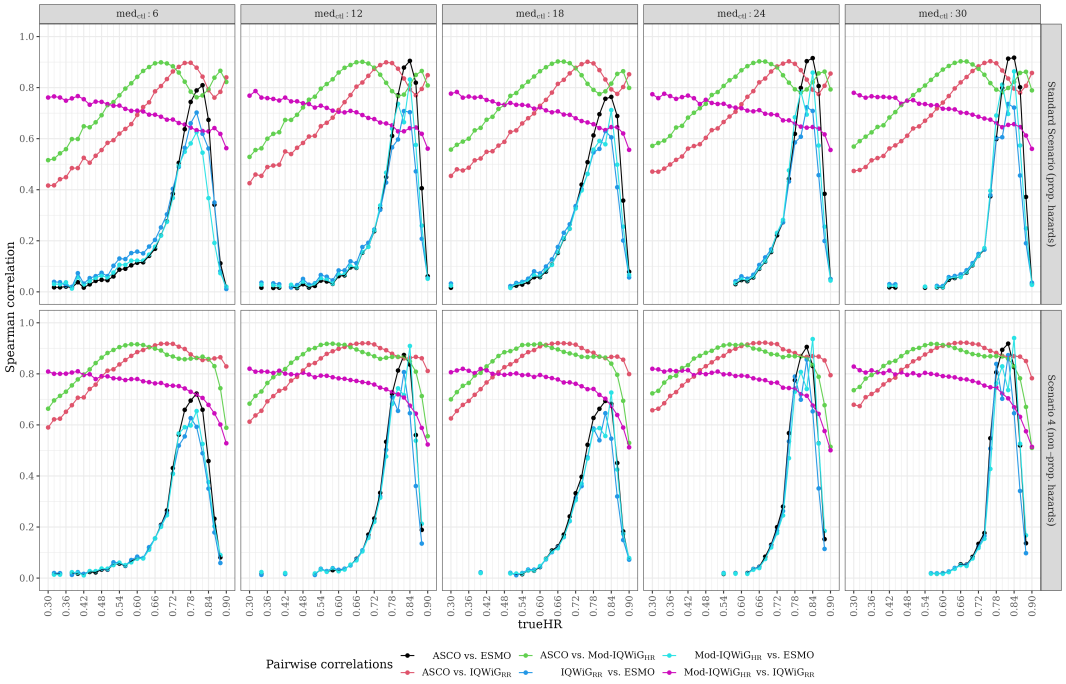

Figure 54: *Pairwise Spearman correlation (y-axis) separated by trueHR (x-axis) of Scenario 4 with p_C=60% and 90% power.*

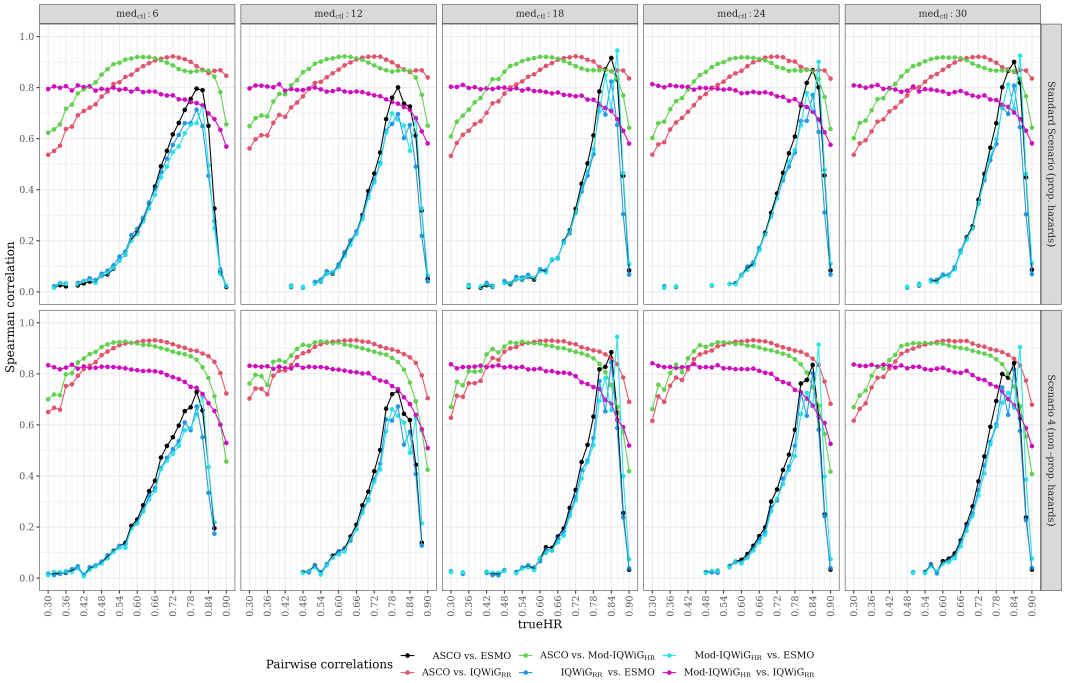

Figure 55: *Pairwise Spearman correlation (y-axis) separated by trueHR (x-axis) of Scenario 4 with p_C=20% and 80% power.*

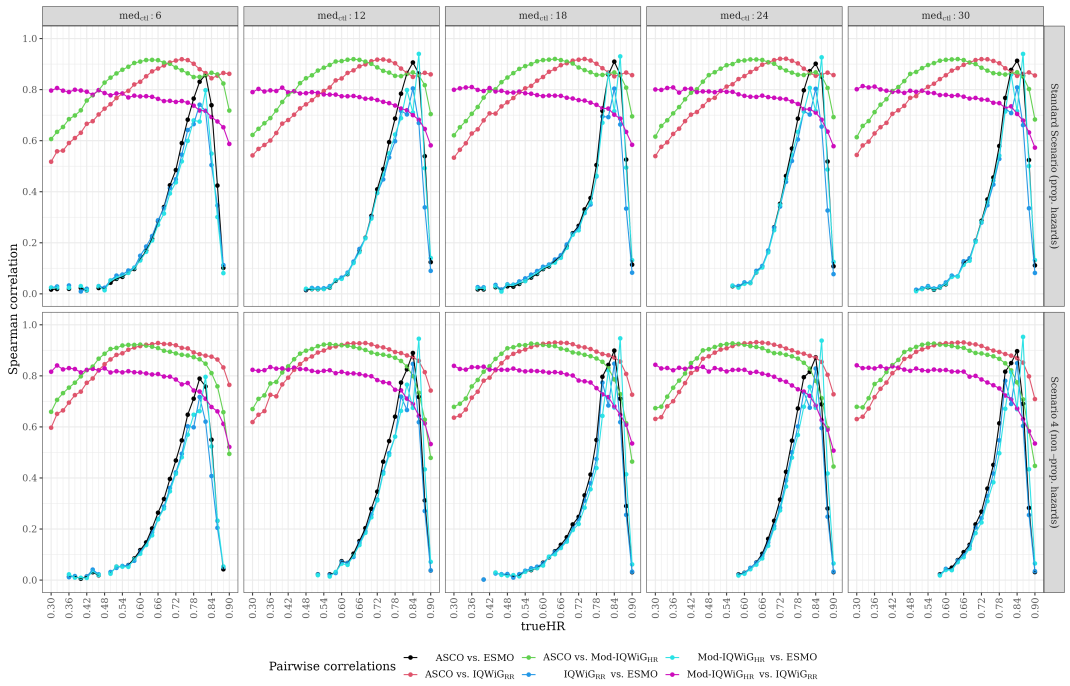

Figure 56: *Pairwise Spearman correlation (y-axis) separated by trueHR (x-axis) of Scenario 4 with p_C=40% and 80% power.*

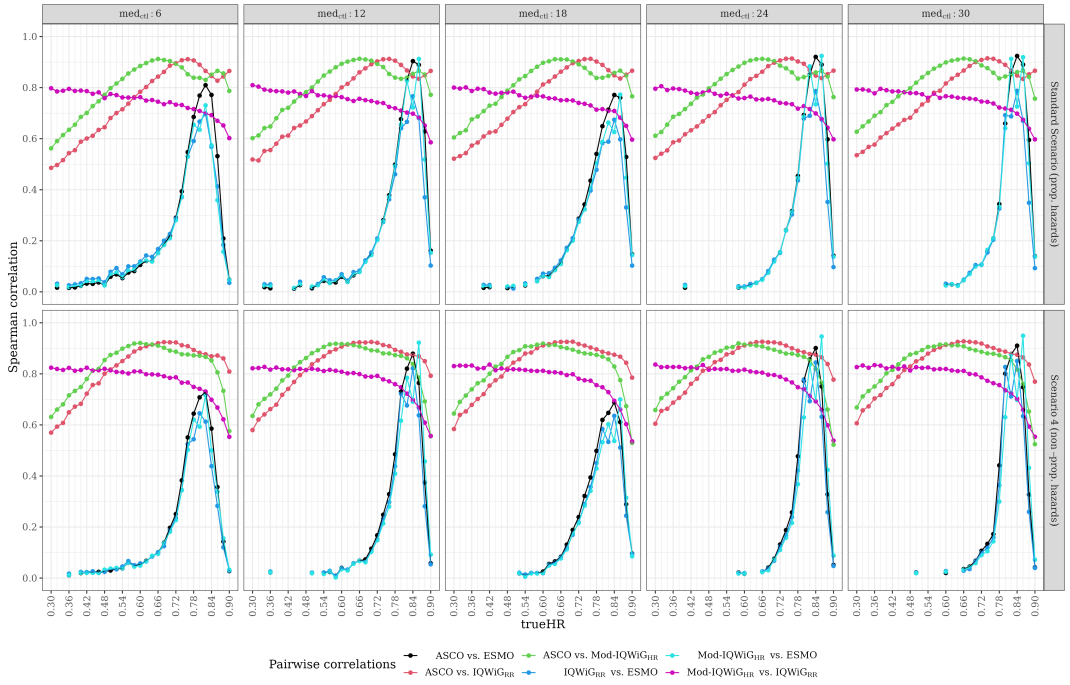

Figure 57: *Pairwise Spearman correlation (y-axis) separated by trueHR (x-axis) of Scenario 4 with p_C=60% and 80% power.*

A.1.4.2 AUC

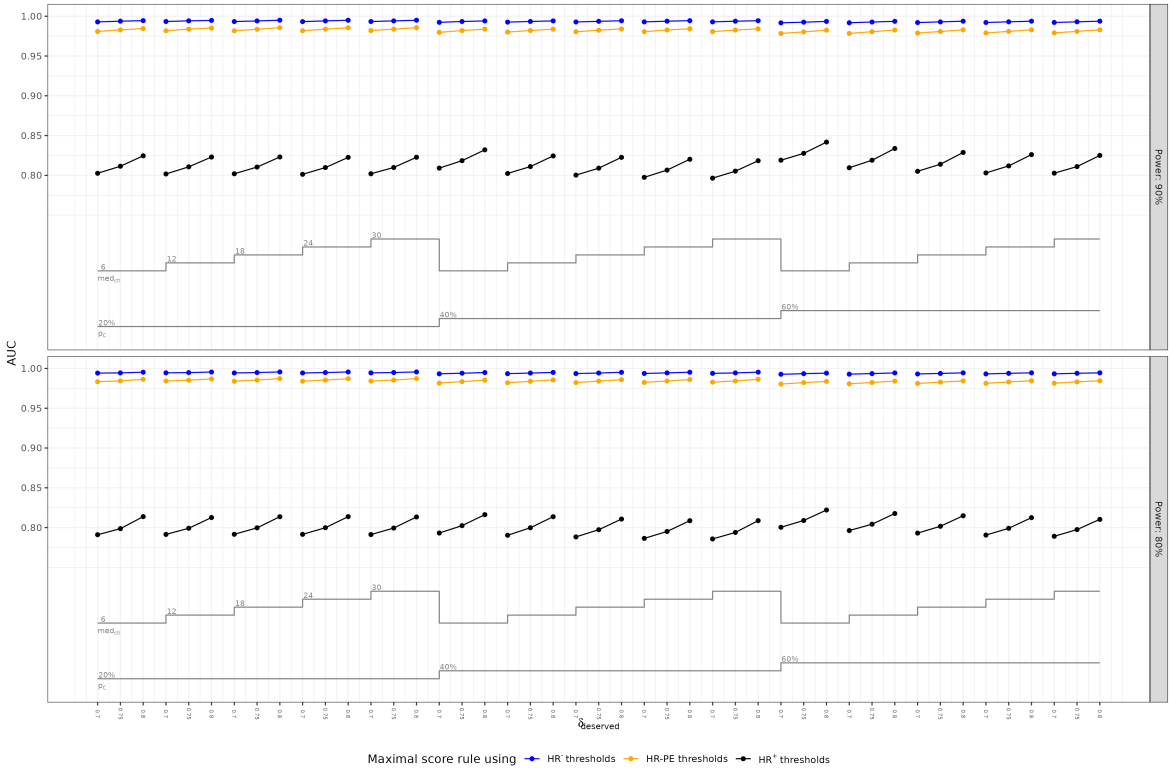

Figure 58: *AUC of ROC curves (y-axis) separated by $\delta_{deserved}$ (x-axis) of Scenario 4 for each sub-scenario.*

A.1.5 Scenario 5 (unequal sample sizes)

A.1.5.1 Relationship between methods

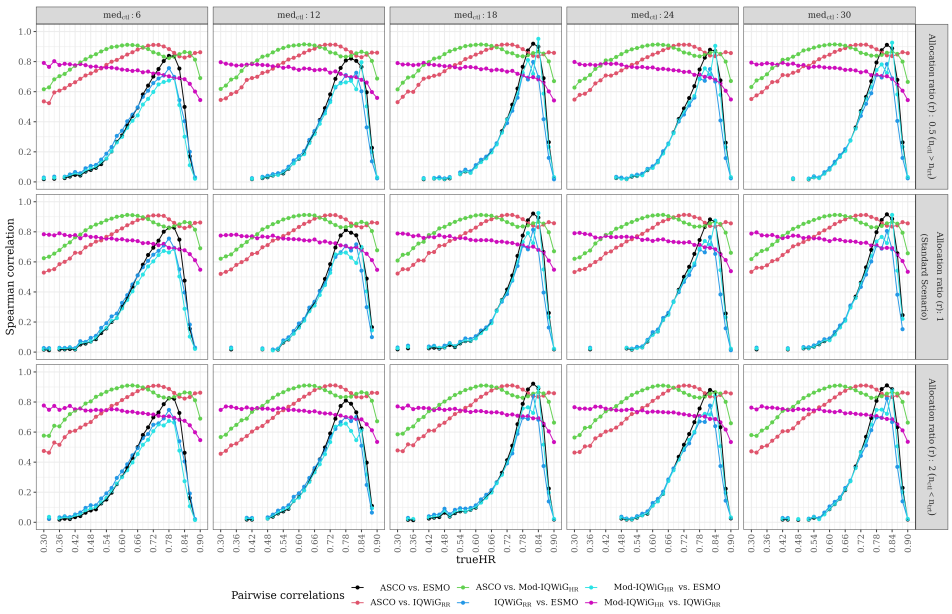

Figure 59: *Pairwise Spearman correlation (y-axis) separated by trueHR (x-axis) of Scenario 5 with p_C=20% and 90% power.*

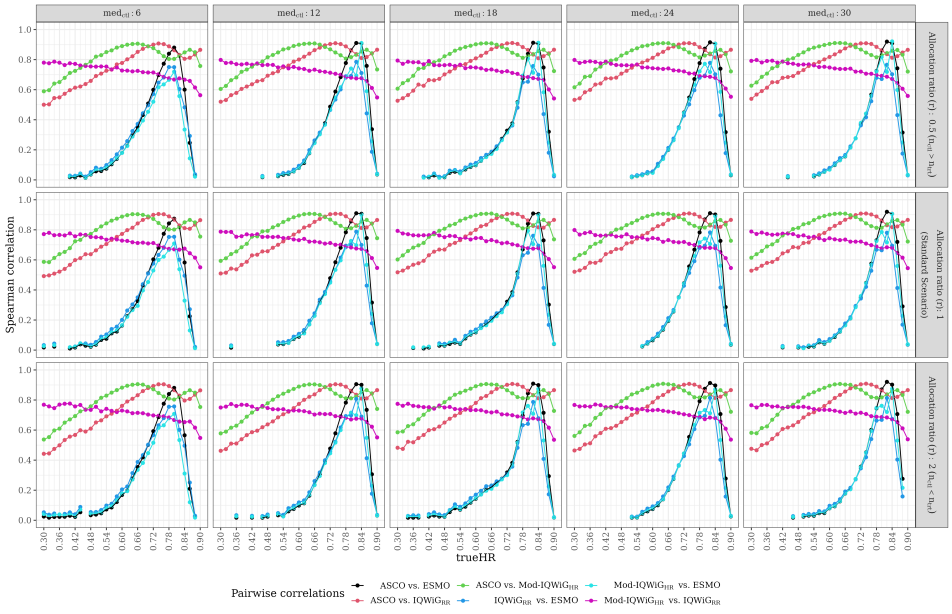

Figure 60: *Pairwise Spearman correlation (y-axis) separated by trueHR (x-axis) of Scenario 5 with p_C=40% and 90% power.*

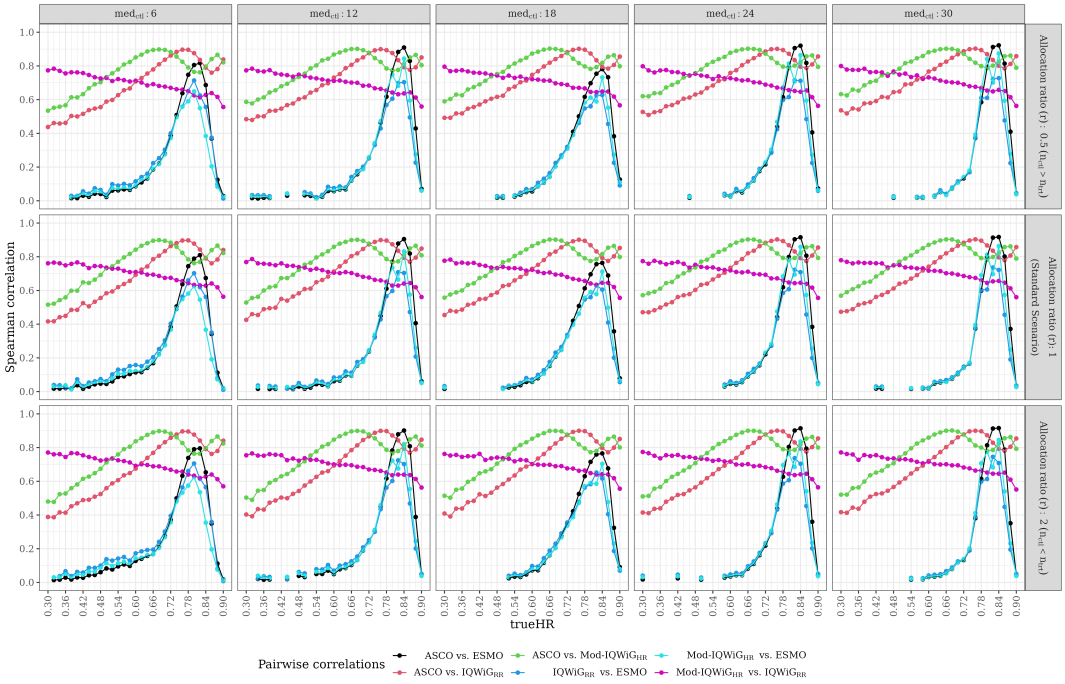

Figure 61: *Pairwise Spearman correlation (y-axis) separated by trueHR (x-axis) of Scenario 5 with p_C=60% and 90% power.*

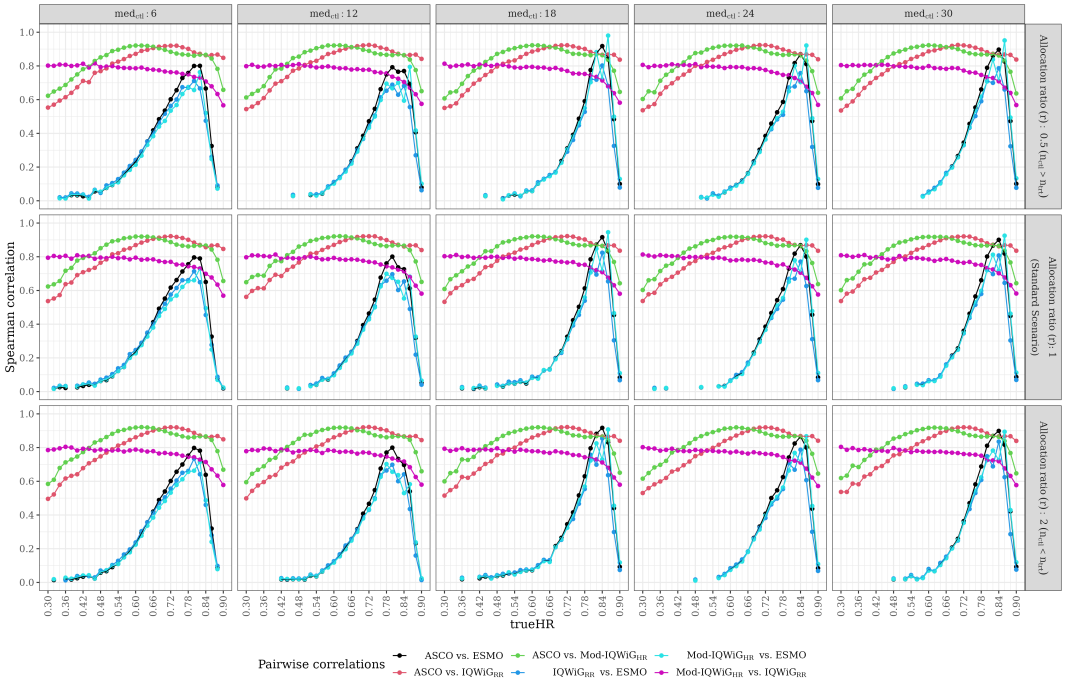

Figure 62: *Pairwise Spearman correlation (y-axis) separated by trueHR (x-axis) of Scenario 5 with p_C=20% and 80% power.*

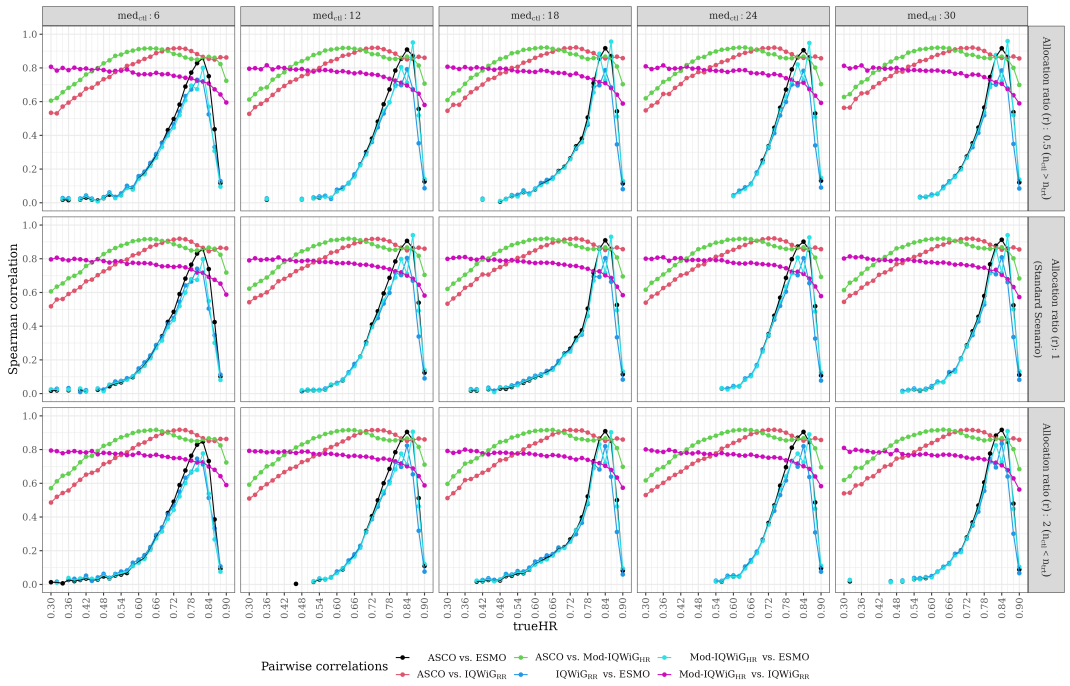

Figure 63: *Pairwise Spearman correlation (y-axis) separated by trueHR (x-axis) of Scenario 5 with p_C=40% and 80% power.*

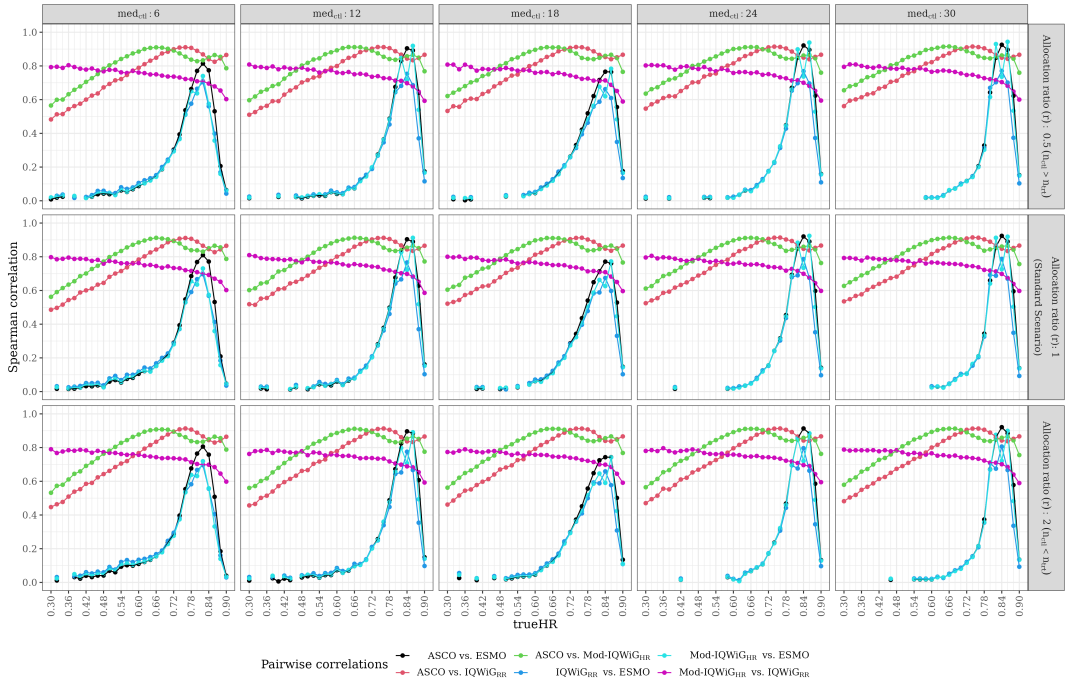

Figure 64: *Pairwise Spearman correlation (y-axis) separated by trueHR (x-axis) of Scenario 5 with p_C=60% and 80% power.*

A.1.5.2 AUC

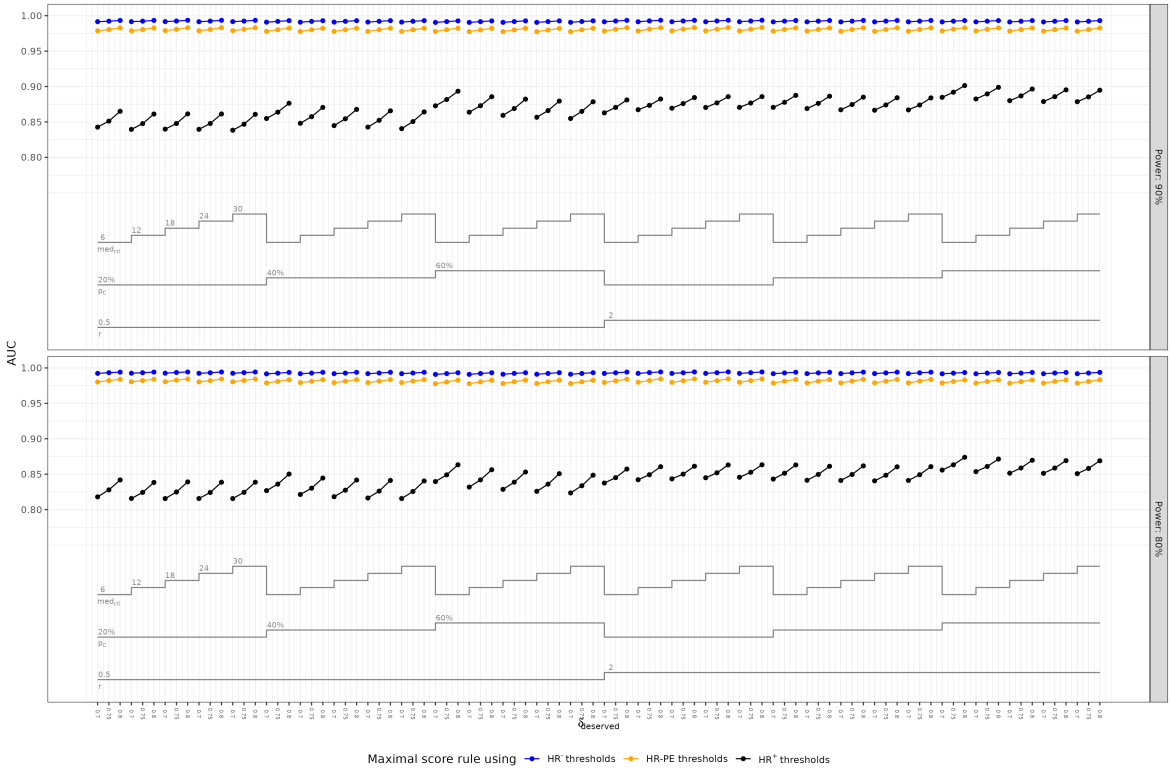

Figure 65: *AUC of ROC curves (y-axis) separated by $\delta_{deserved}$ (x-axis) of Scenario 5 for each sub-scenario.*

A.1.6 Scenario 6 (only exponential distributed censoring)

A.1.6.1 Relationship between methods

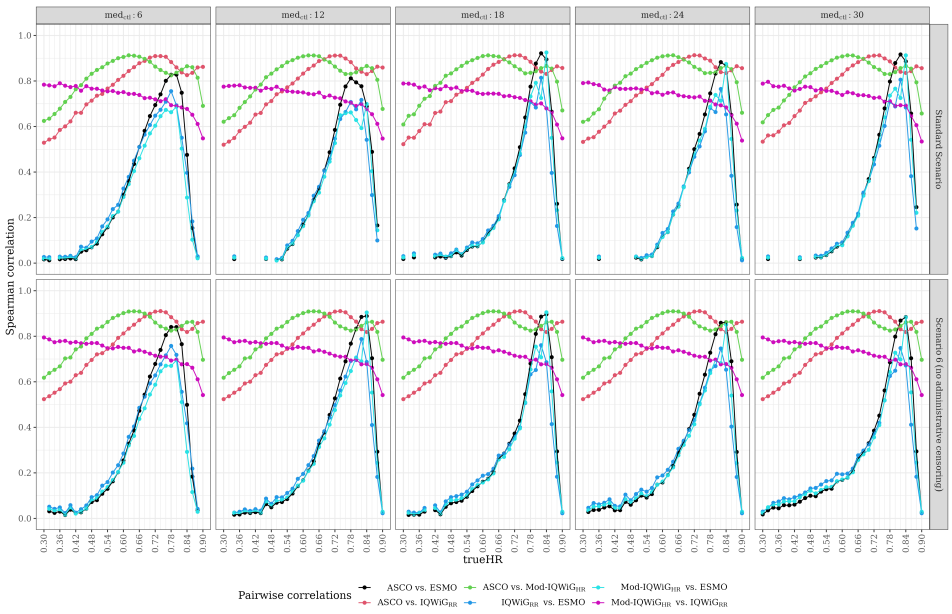

Figure 66: *Pairwise Spearman correlation (y-axis) separated by trueHR (x-axis) of Scenario 6 with p_C=20% and 90% power.*

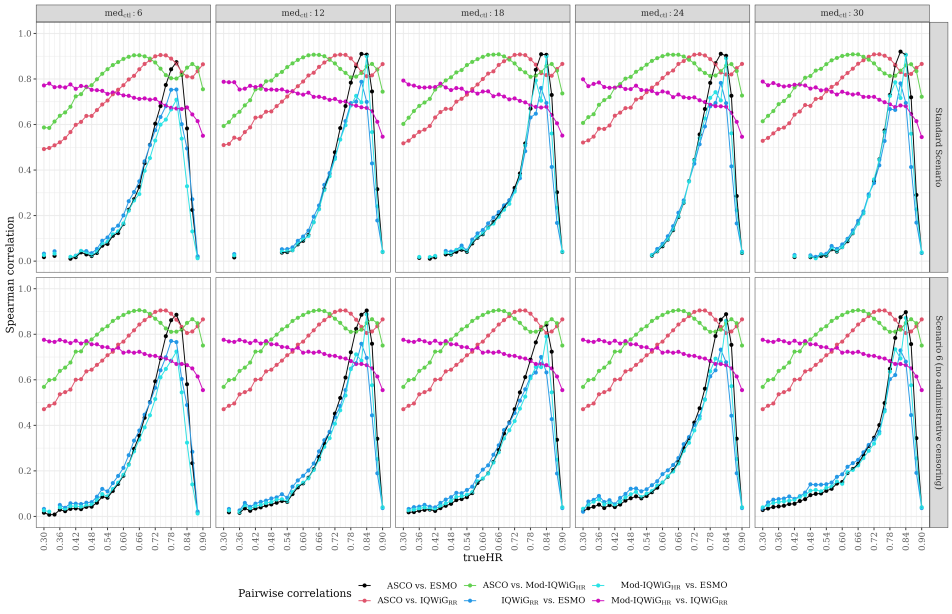

Figure 67: *Pairwise Spearman correlation (y-axis) separated by trueHR (x-axis) of Scenario 6 with p_C=40% and 90% power.*

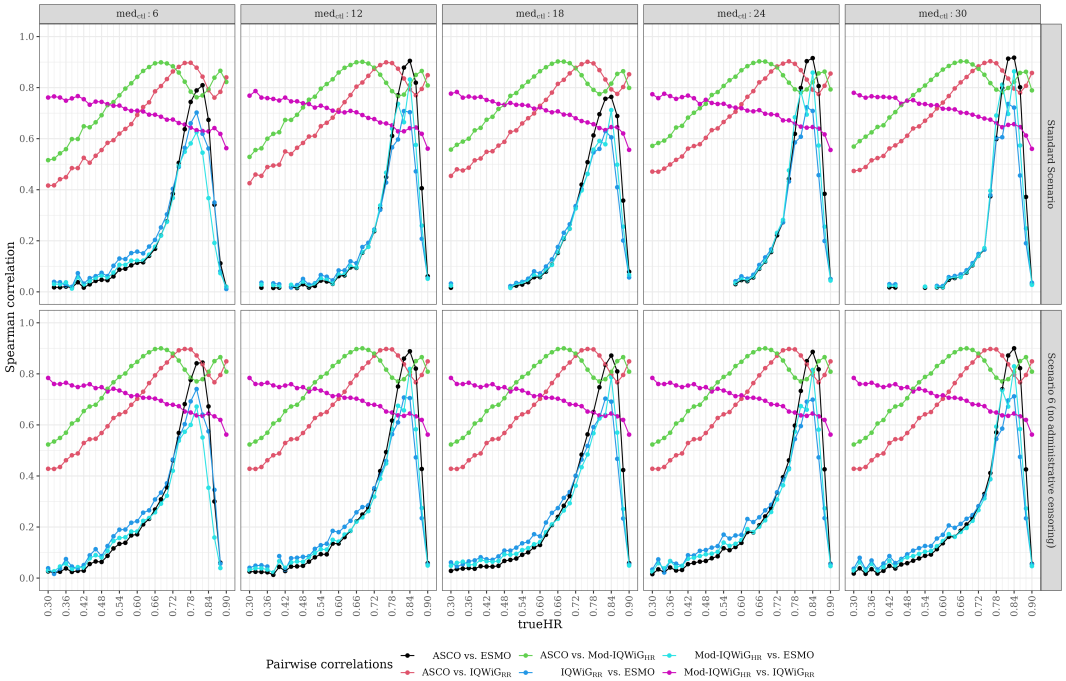

Figure 68: *Pairwise Spearman correlation (y-axis) separated by trueHR (x-axis) of Scenario 6 with p_C=60% and 90% power.*

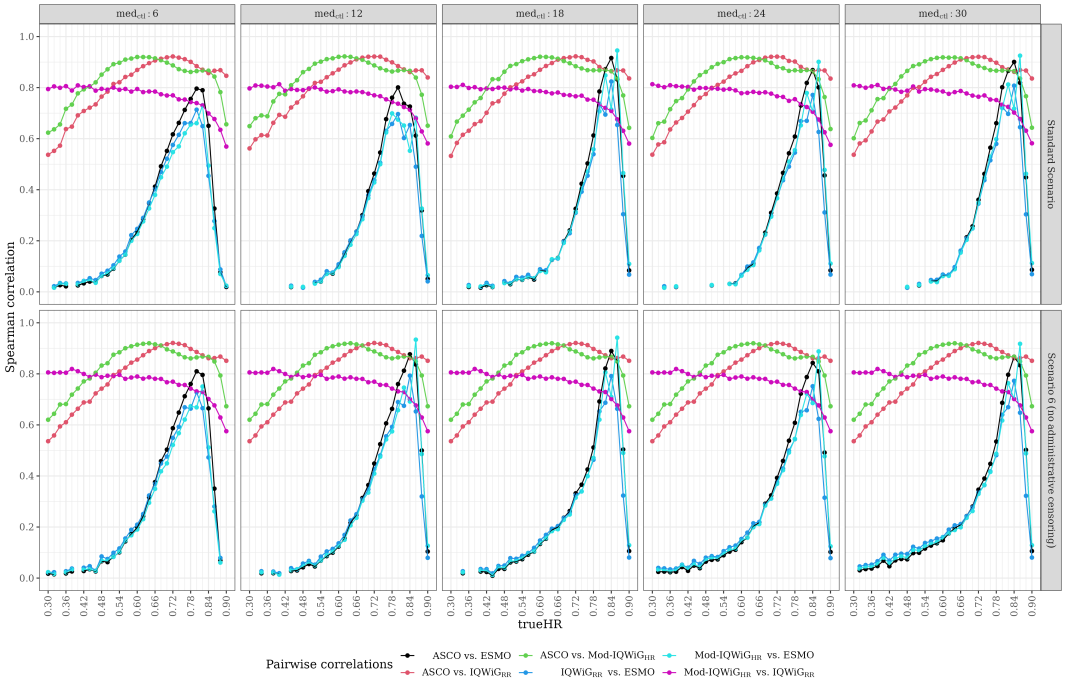

Figure 69: *Pairwise Spearman correlation (y-axis) separated by trueHR (x-axis) of Scenario 6 with p_C=20% and 80% power.*

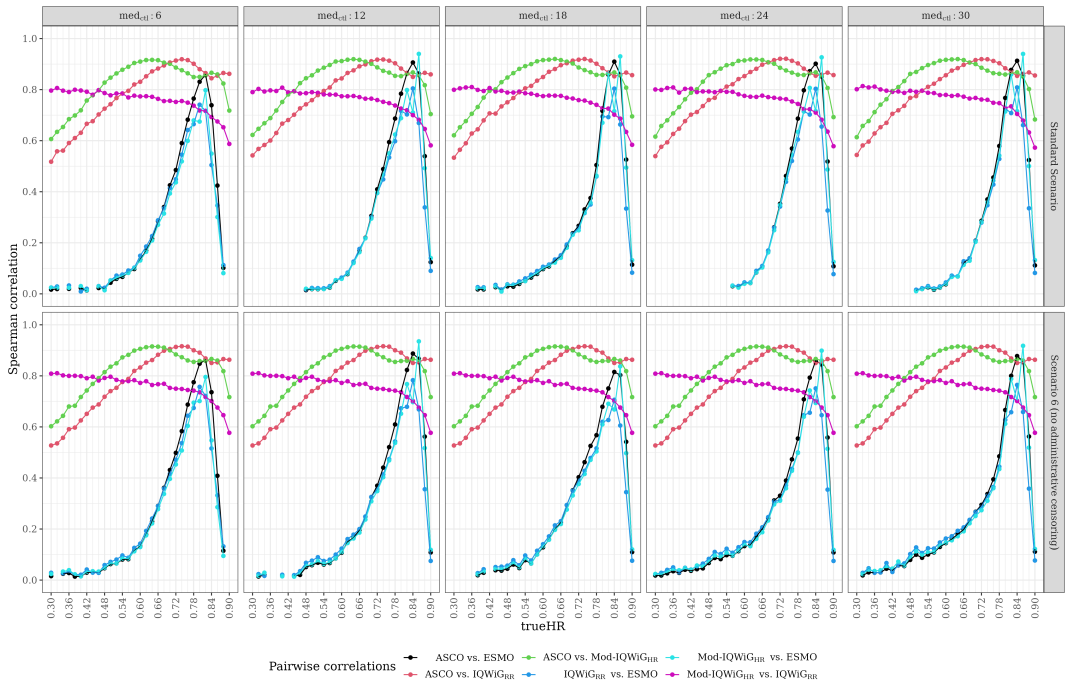

Figure 70: *Pairwise Spearman correlation (y-axis) separated by trueHR (x-axis) of Scenario 6 with p_C=40% and 80% power.*

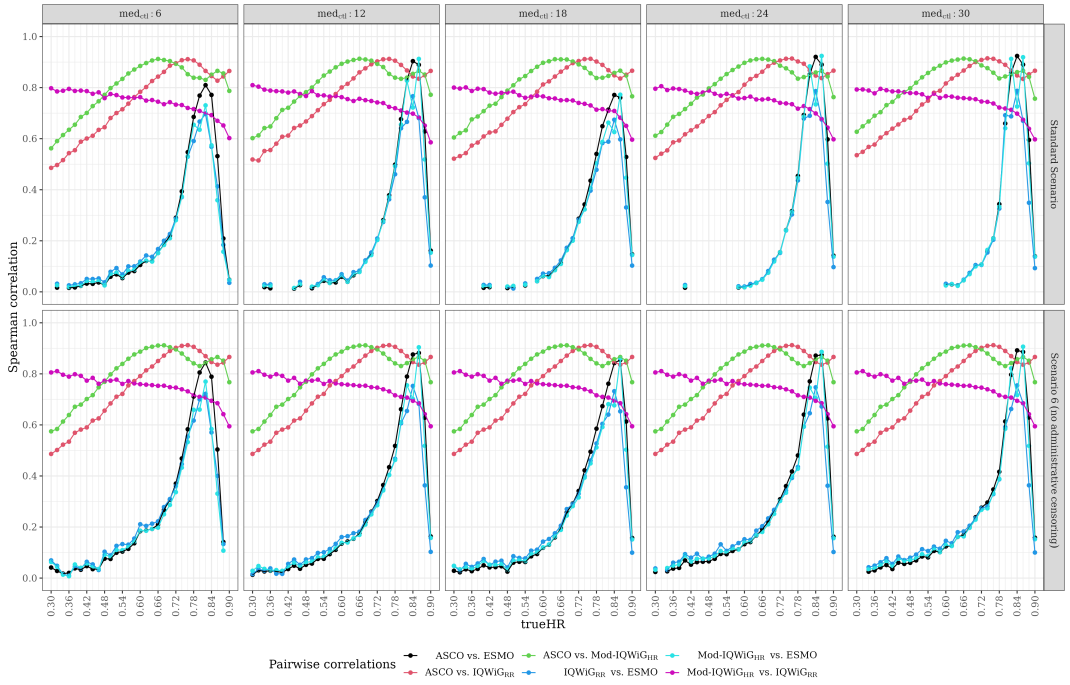

Figure 71: *Pairwise Spearman correlation (y-axis) separated by trueHR (x-axis) of Scenario 6 with p_C=60% and 80% power.*

A.1.6.2 AUC

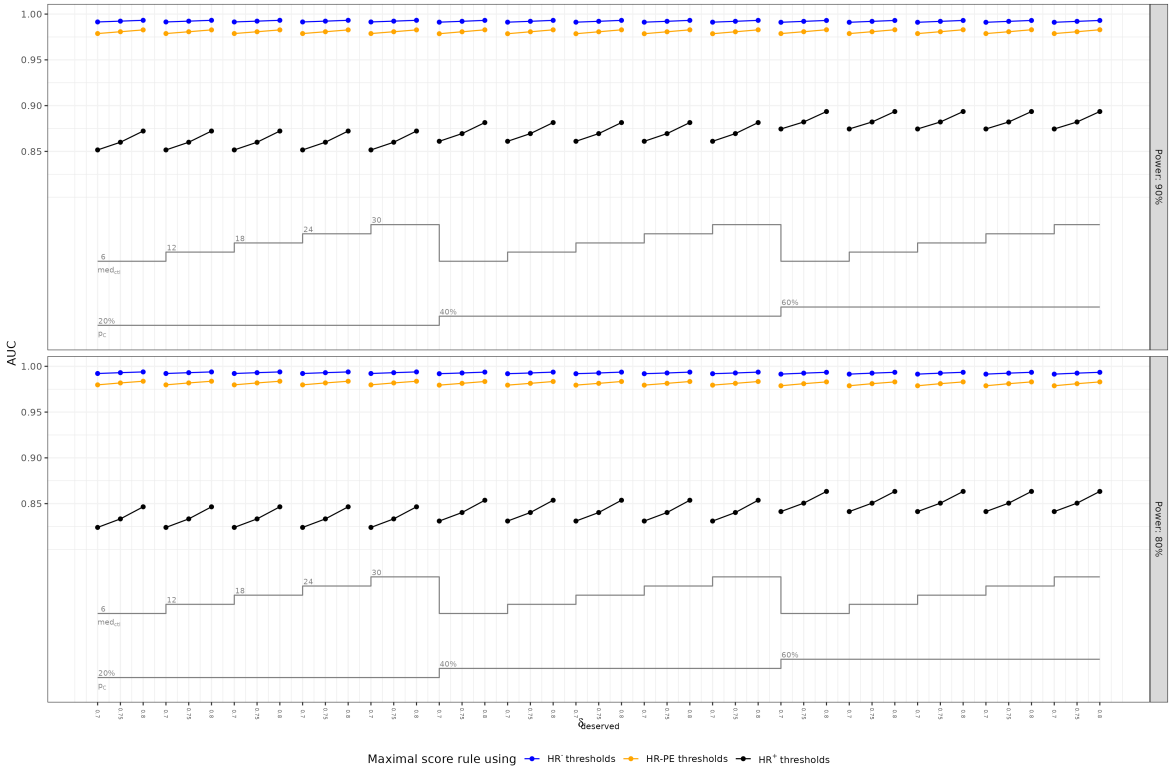

Figure 72: *AUC of ROC curves (y-axis) separated by $\delta_{deserved}$ (x-axis) of Scenario 6 for each sub-scenario.*

A.1.7 Scenario 7 (informative censoring)

A.1.7.1 Relationship between methods

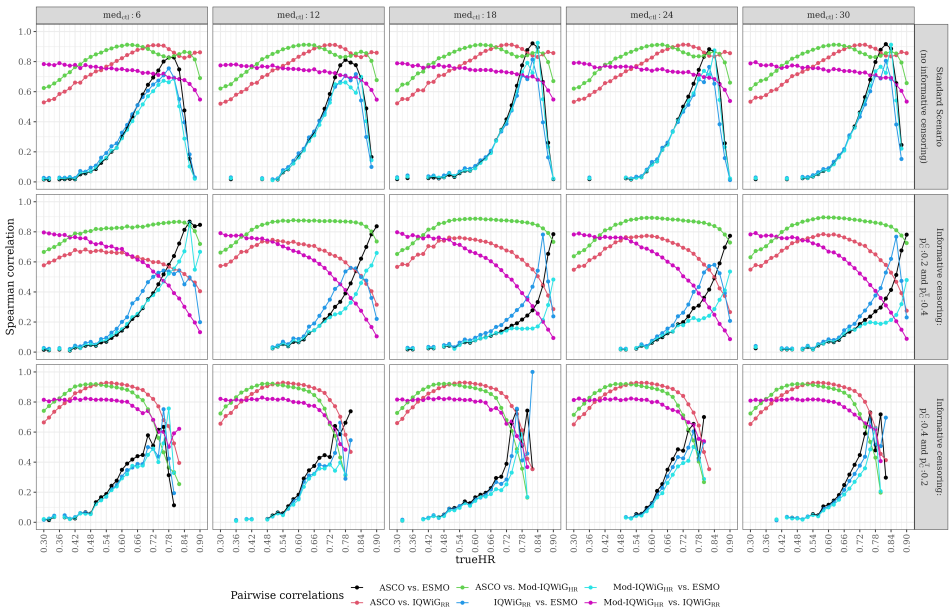

Figure 73: *Pairwise Spearman correlation (y-axis) separated by trueHR (x-axis) of Scenario 7 with p_C=20% and 90% power.*

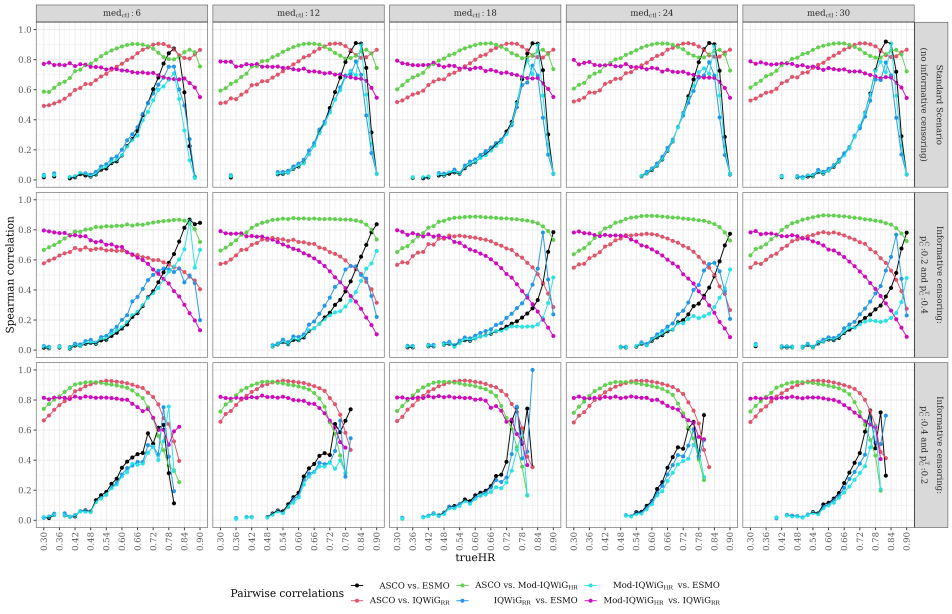

Figure 74: *Pairwise Spearman correlation (y-axis) separated by trueHR (x-axis) of Scenario 7 with p_C=40% and 90% power.*

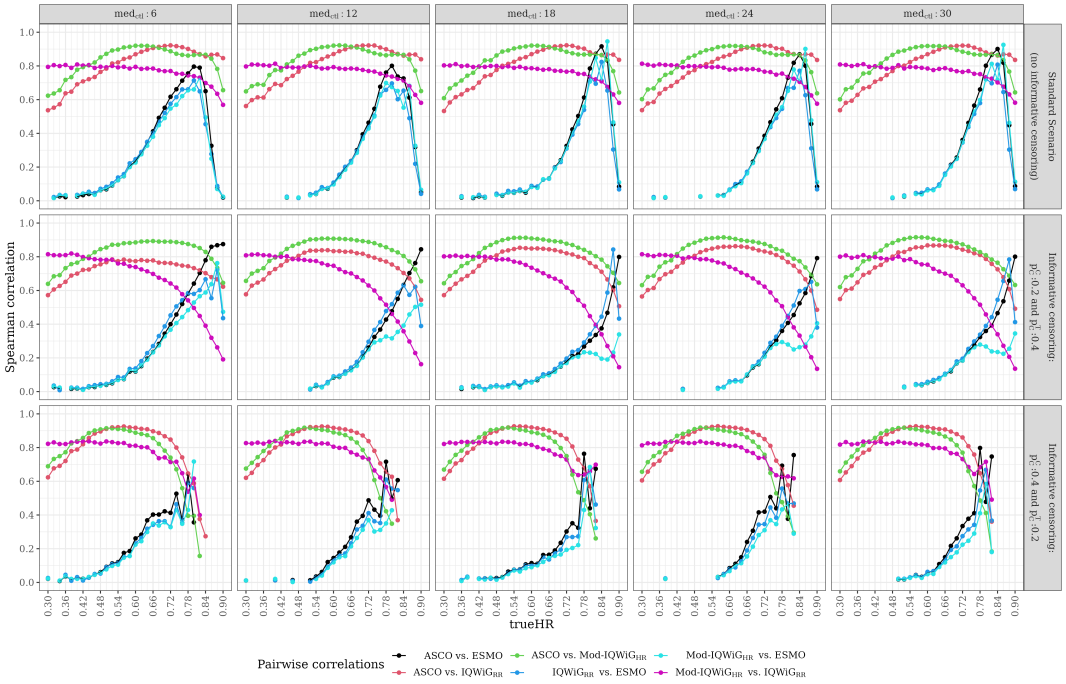

Figure 75: *Pairwise Spearman correlation (y-axis) separated by trueHR (x-axis) of Scenario 7 with p_C=20% and 80% power.*

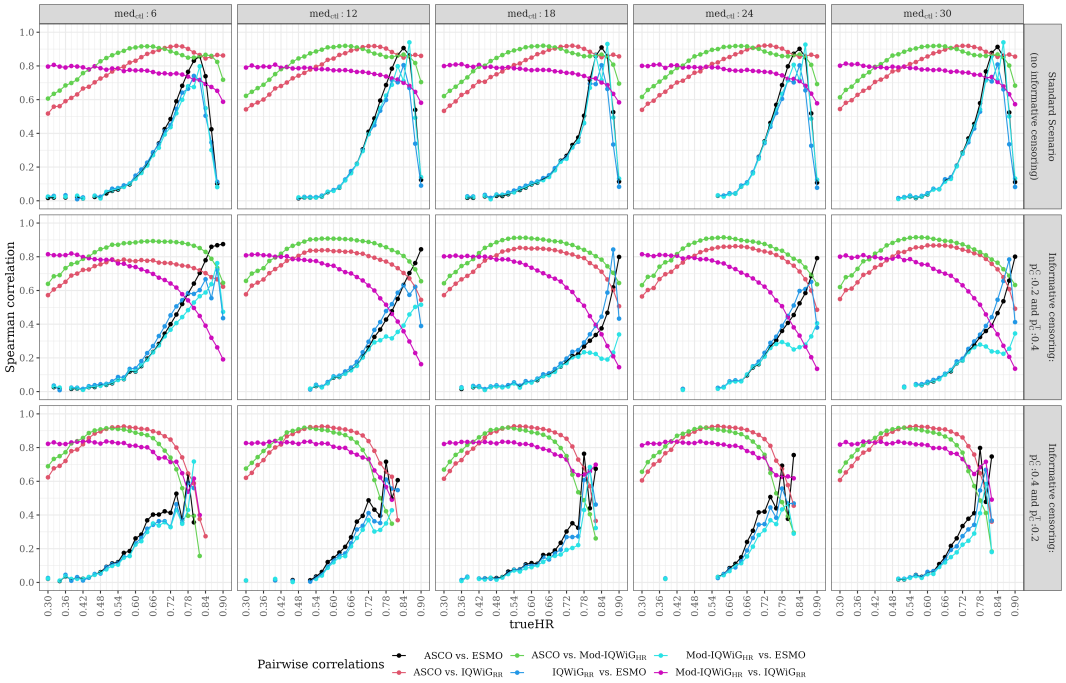

Figure 76: *Pairwise Spearman correlation (y-axis) separated by trueHR (x-axis) of Scenario 7 with p_C=40% and 80% power.*

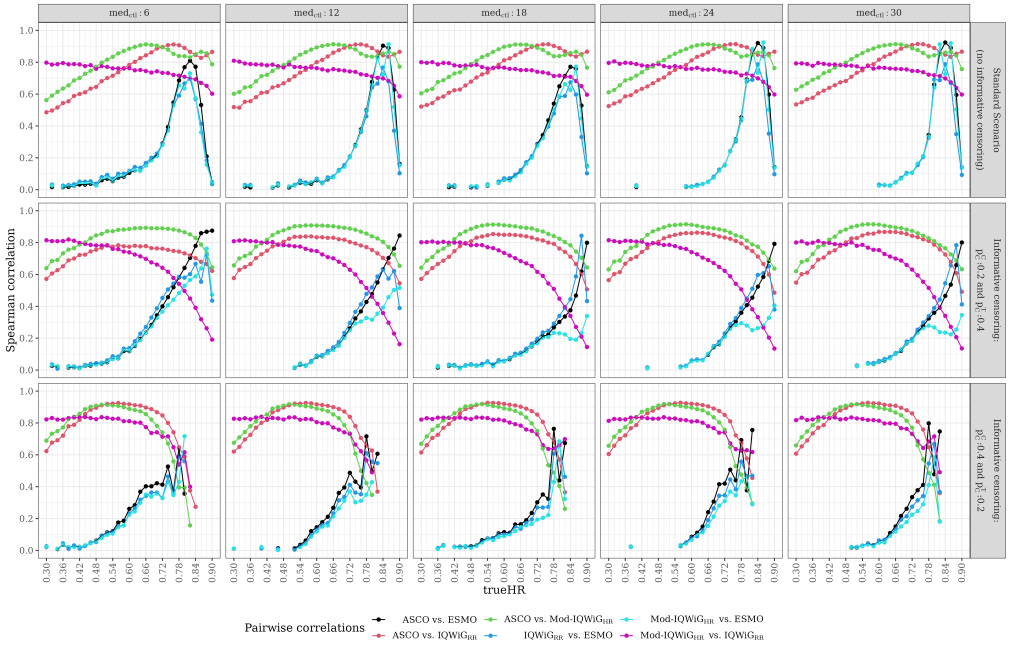

Figure 77: *Pairwise Spearman correlation (y-axis) separated by trueHR (x-axis) of Scenario 7 with p_C=60% and 80% power.*

A.1.7.2 AUC

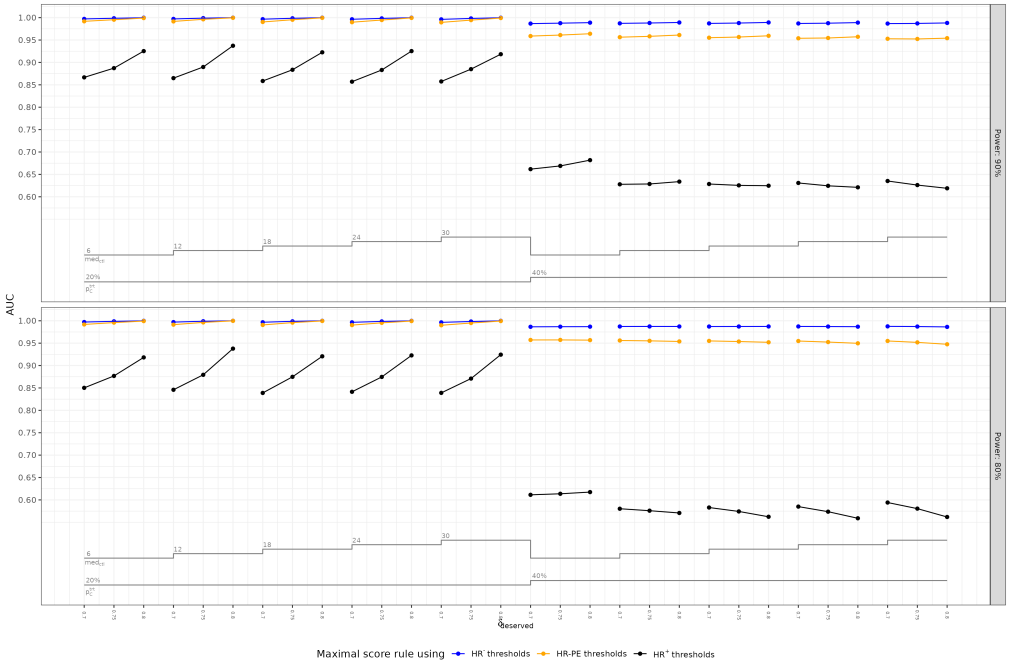

Figure 78: *AUC of ROC curves (y-axis) separated by $\delta_{deserved}$ (x-axis) of Scenario 7 for each sub-scenario.*

A.2 Simulation 2

Figure 79: *Pairwise Spearman correlation of the additional benefit assessment methods (x-axis) for the different scenarios (y-axis) where all sub-scenarios were combined for correlation calculation (Simulation 2).*

Figure 80: *Pairwise Kendall-τ_b correlation of the additional benefit assessment methods (x-axis) for the different scenarios (y-axis) where all sub-scenarios were combined for correlation calculation (Simulation 2).*

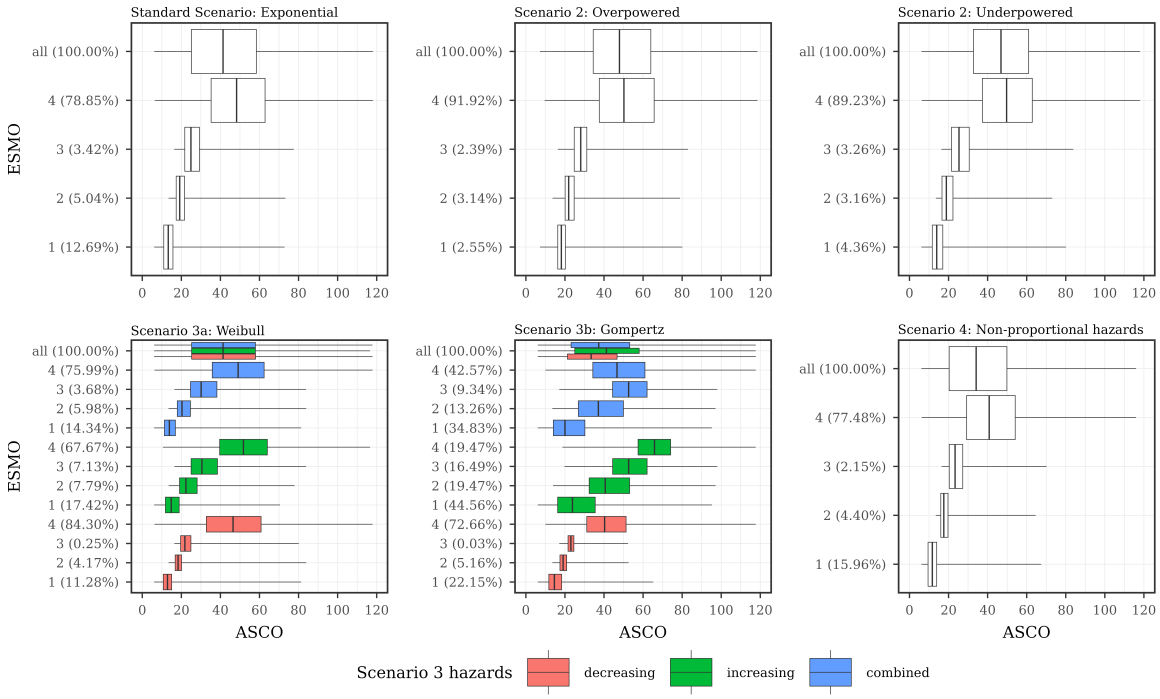

Figure 81: *Description of ASCO score distribution (x-axis) separated into the categories of ESMO and overall (y-axis) using boxplots (Simulation 2).*

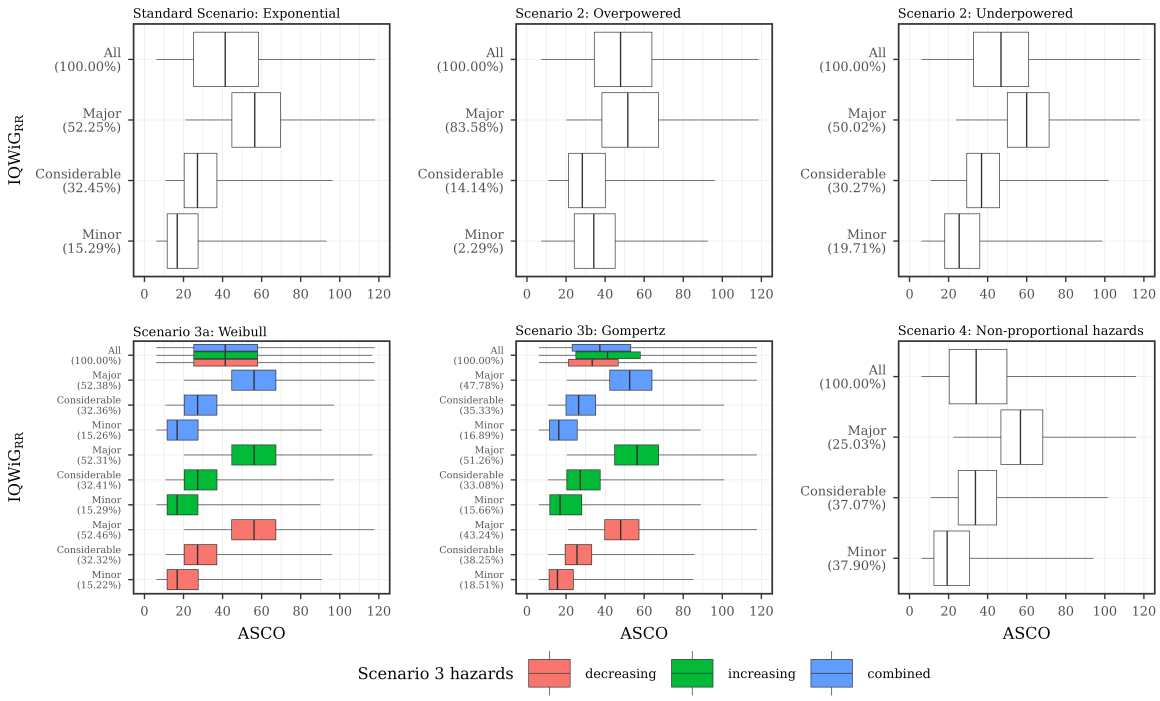

Figure 82: *Description of ASCO score distribution (x-axis) separated into the categories of $IQWiG_{RR}$ and overall (y-axis) using boxplots (Simulation 2).*

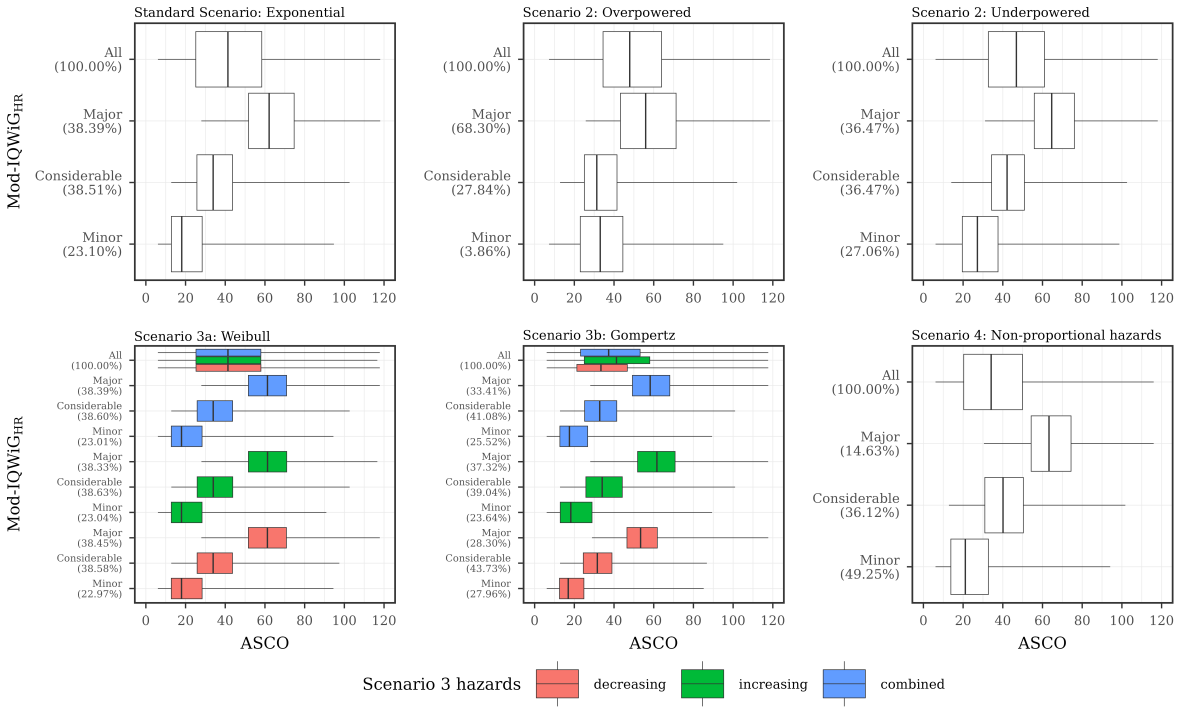

Figure 83: *Description of ASCO score distribution (x-axis) separated into the categories of Mod-IQWiG$_{HR}$ and overall (y-axis) using boxplots (Simulation 2).*

A.2.1 Standard Scenario

A.2.1.1 Relationship between methods

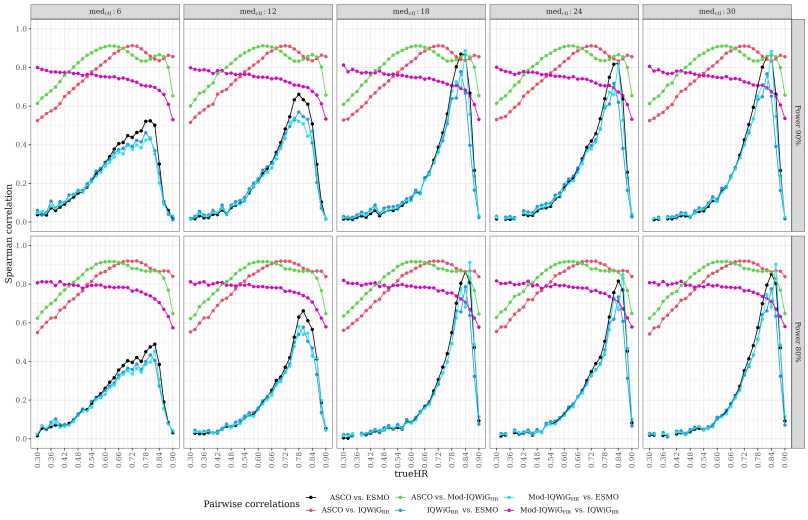

Figure 84: *Pairwise Spearman correlation (y-axis) separated by trueHR (x-axis) of Scenario 2 with p_C=60% (Simulation 2).*

A.2.1.2 HR-PE, HR$^-$, and HR$^+$ estimations with sample size calculation and with constant sample size

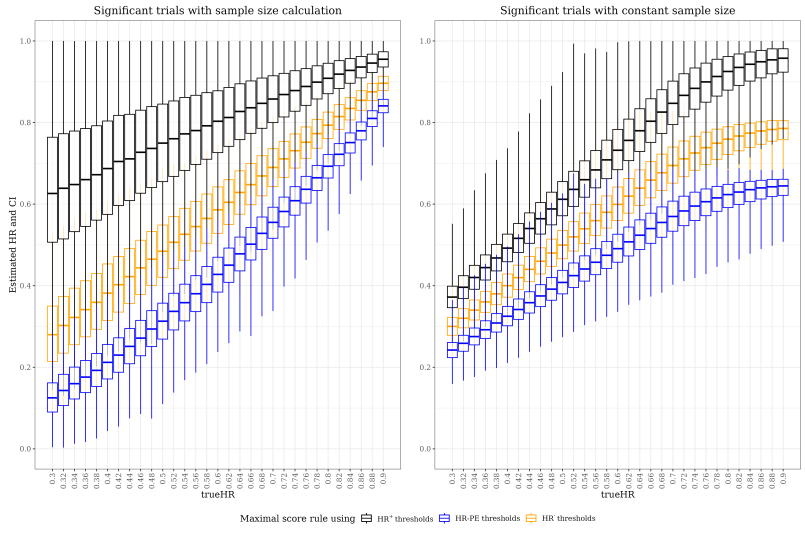

Figure 85: *Description of HR-PE, HR$^-$, and HR$^+$ estimation distribution (y-axis) separated by trueHR (x-axis) using boxplots with sample size calculation (left panel) and with constant sample size (right panel) of the Standard Scenario with 90% power, med_{ctl}=6, and p_C=60% (Simulation 2).*

A.2.2 Scenario 2 (incorrect assumed designHR for sample size calculation)

A.2.2.1 Relationship between methods

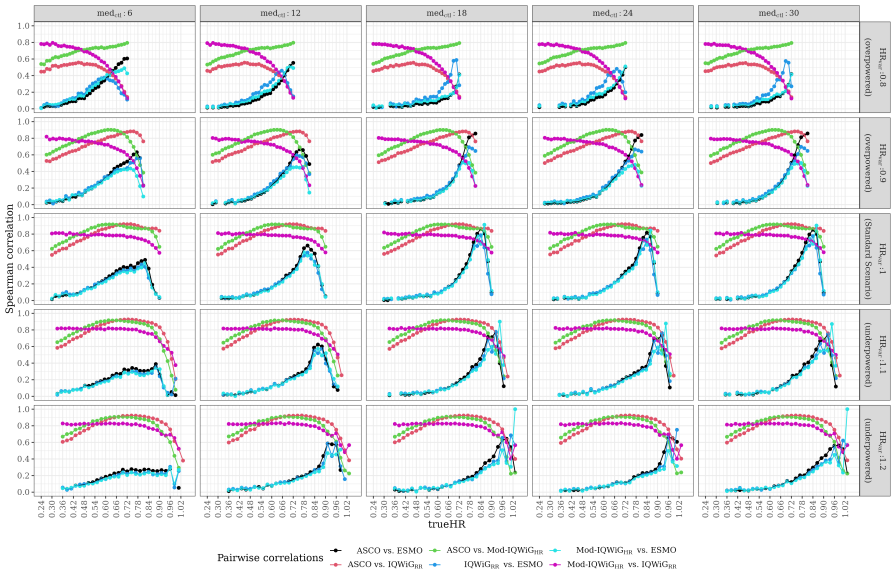

Figure 86: *Pairwise Spearman correlation (y-axis) separated by trueHR (x-axis) of Scenario 2 with 80% power (Simulation 2).*

A.2.2.2 AUC

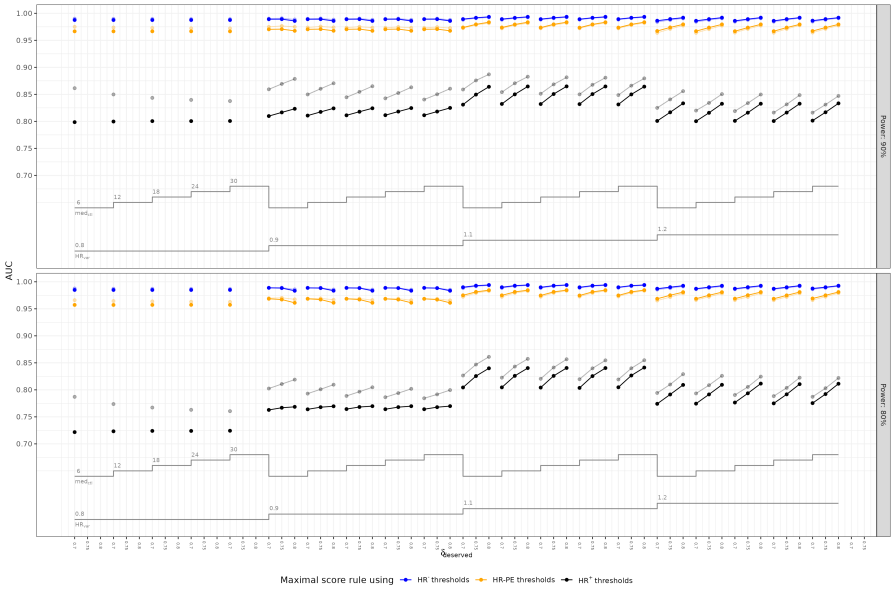

Figure 87: *AUC of ROC curves (y-axis) separated by $\delta_{deserved}$ (x-axis) of Scenario 2 for each sub-scenario (Simulation 2).*

A.2.3 Scenario 3 (different failure time distributions)

A.2.3.1 Relationship between methods

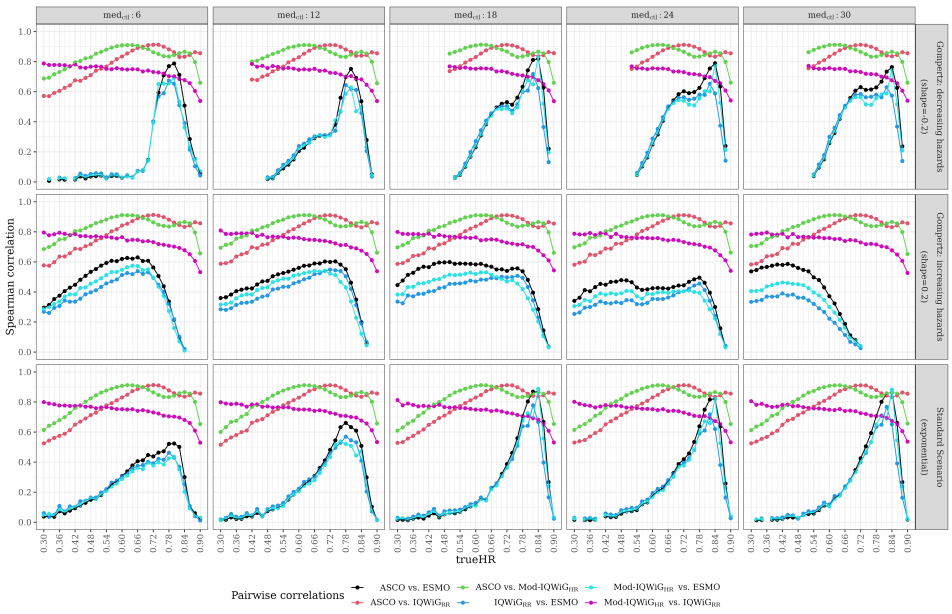

Figure 88: *Pairwise Spearman correlation (y-axis) separated by trueHR (x-axis) of Scenario 3 with Gompertz failure time distribution and 90% power (Simulation 2).*

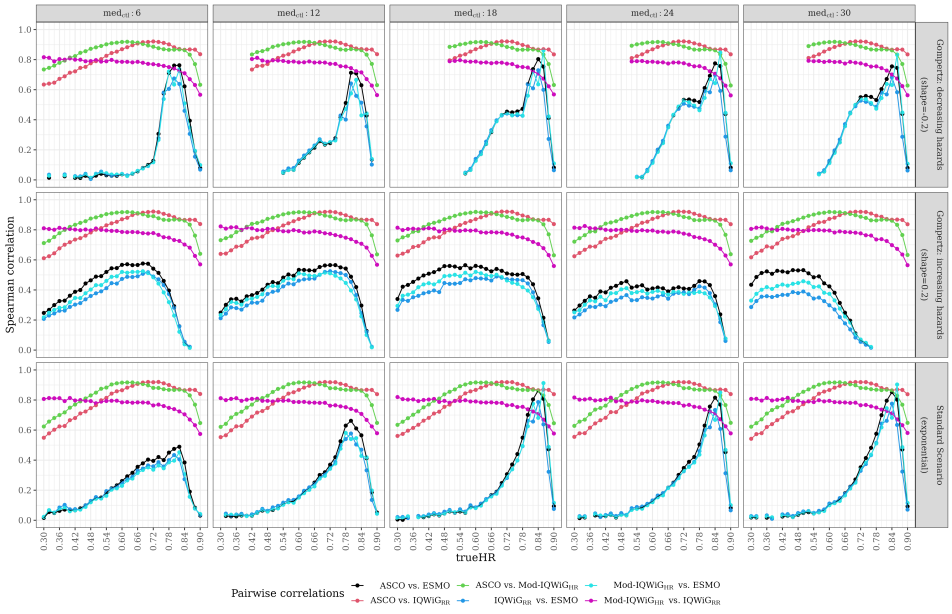

Figure 89: *Pairwise Spearman correlation (y-axis) separated by trueHR (x-axis) of Scenario 3 with Gompertz failure time distribution and 80% power (Simulation 2).*

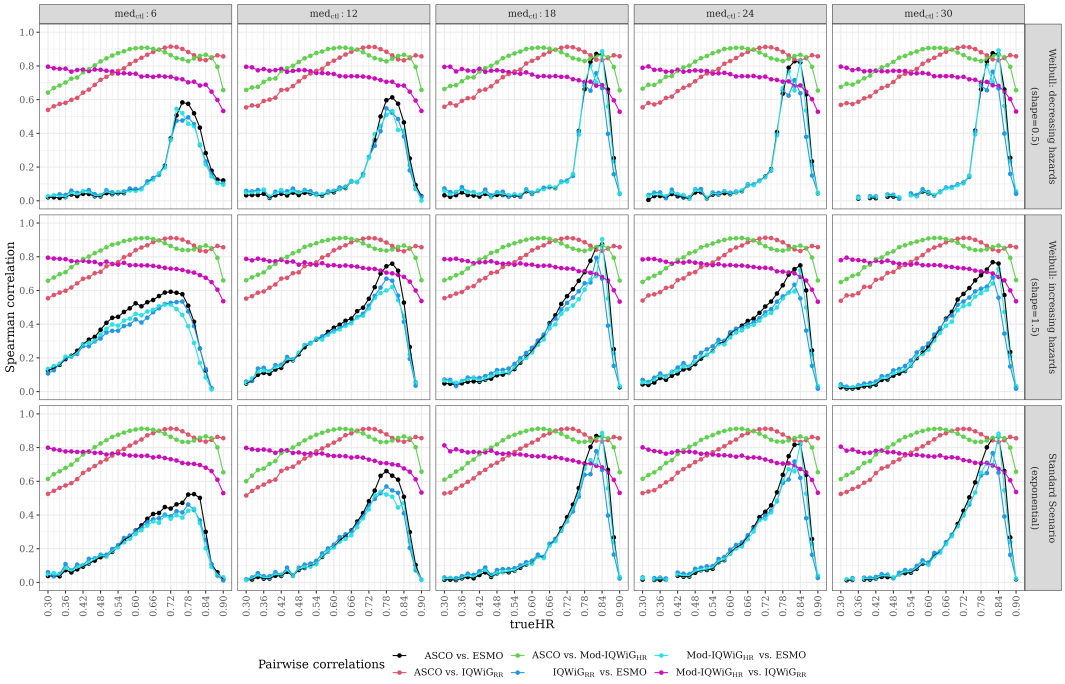

Figure 90: *Pairwise Spearman correlation (y-axis) separated by trueHR (x-axis) of Scenario 3 with Weibull failure time distribution and 90% power (Simulation 2).*

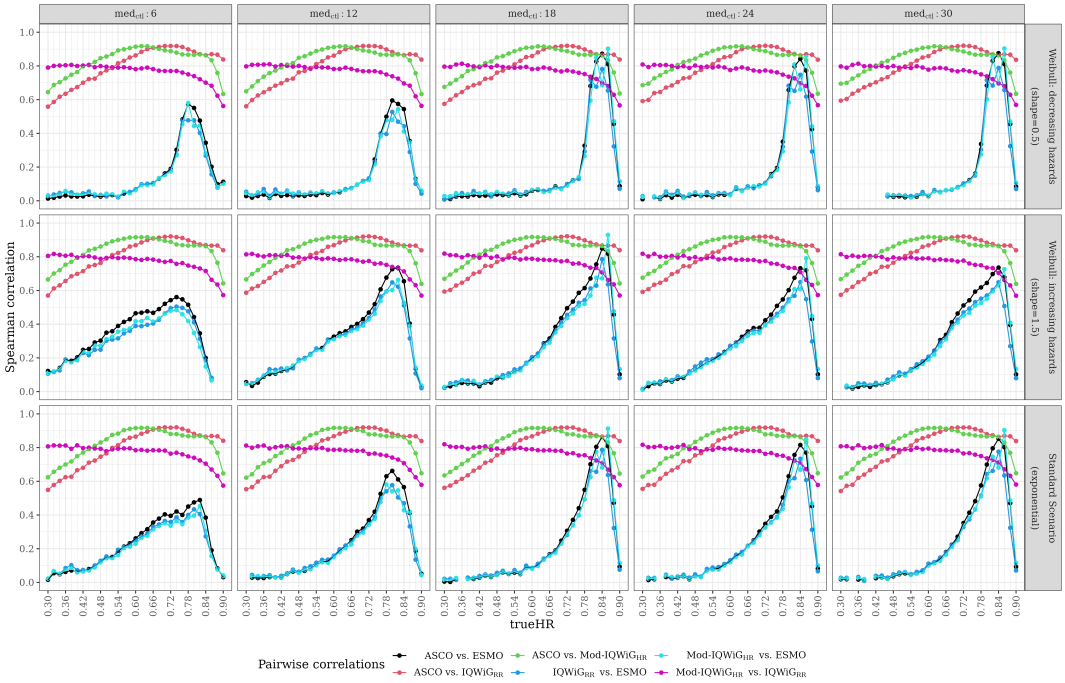

Figure 91: *Pairwise Spearman correlation (y-axis) separated by trueHR (x-axis) of Scenario 3 with Weibull failure time distribution and 80% power (Simulation 2).*

A.2.3.2 AUC

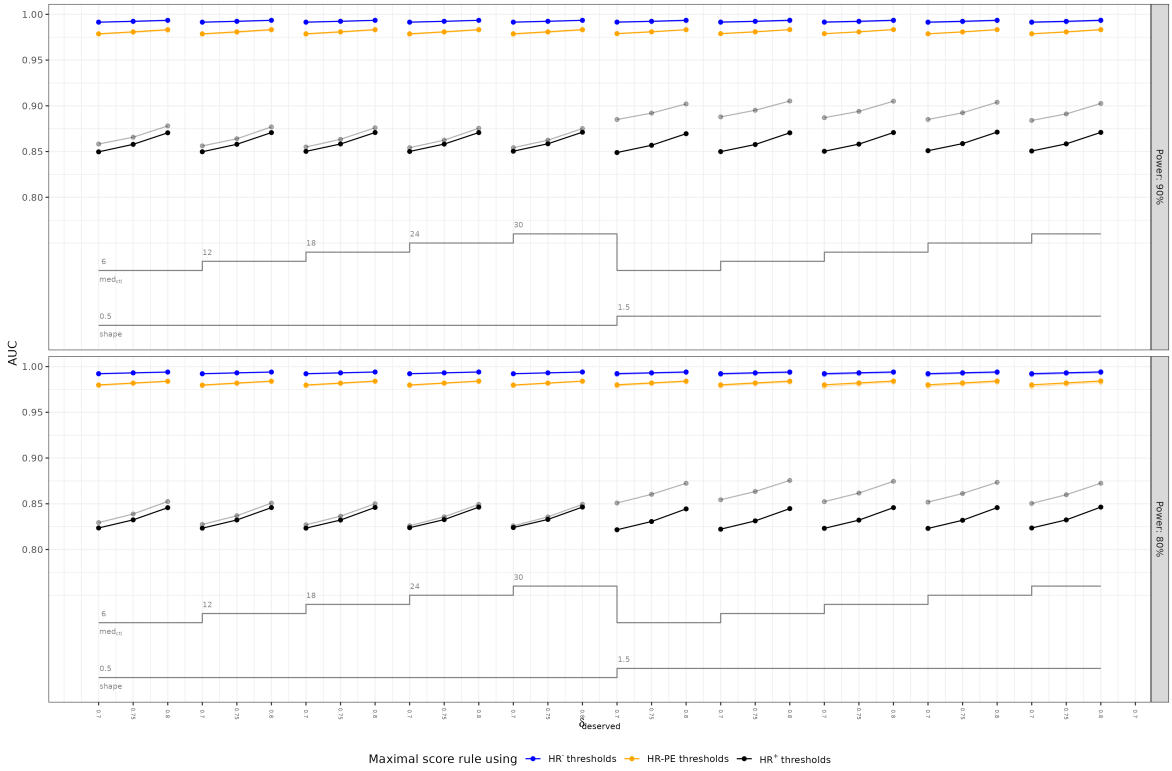

Figure 92: *AUC of ROC curves (y-axis) separated by $\delta_{deserved}$ (x-axis) of Scenario 3 with Weibull failure time distribution for each sub-scenario (Simulation 2).*

A.2.4 Scenario 4 (non-proportional hazards)

A.2.4.1 Relationship between methods

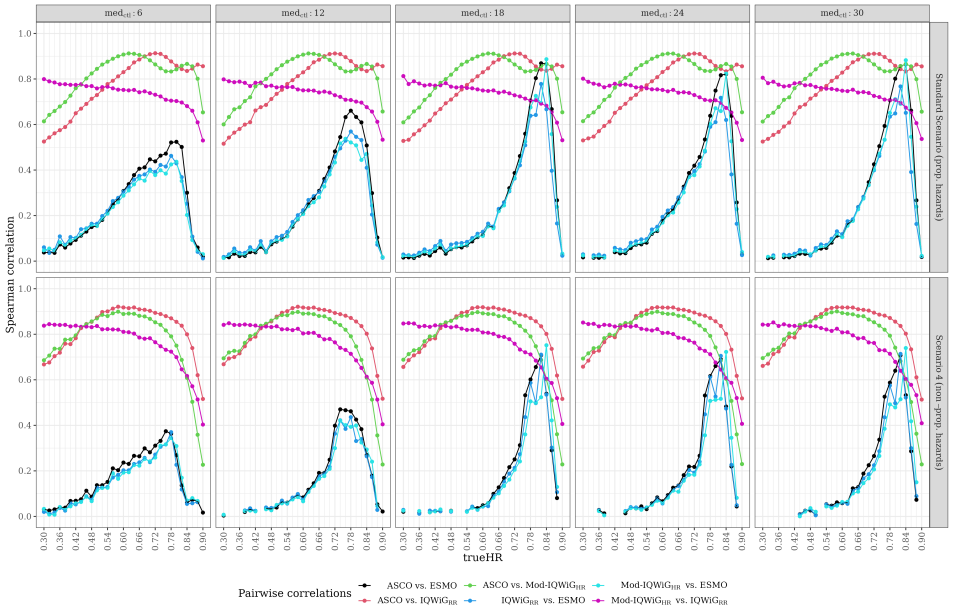

Figure 93: *Pairwise Spearman correlation (y-axis) separated by trueHR (x-axis) of Scenario 4 with 90% power (Simulation 2).*

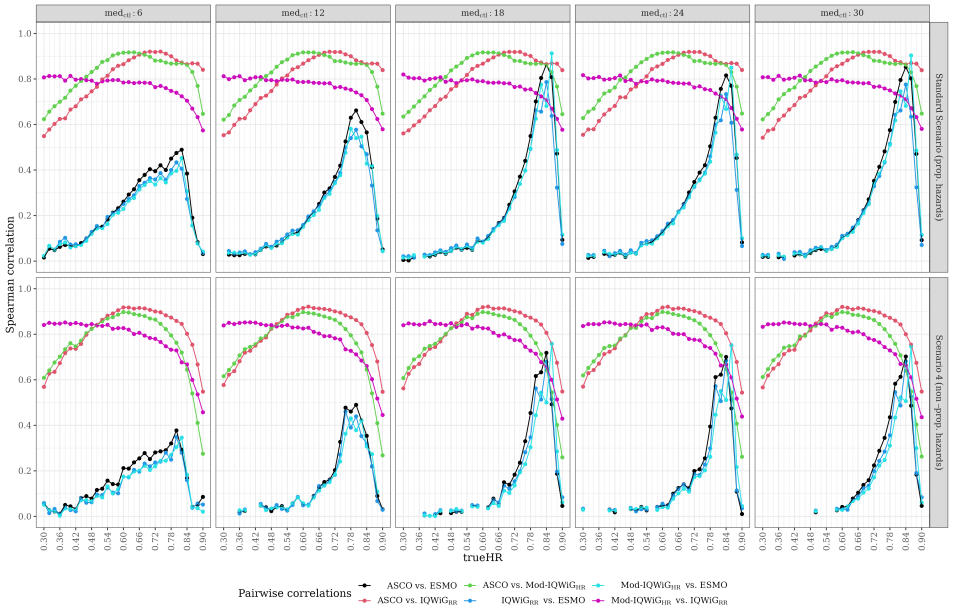

Figure 94: *Pairwise Spearman correlation (y-axis) separated by trueHR (x-axis) of Scenario 4 with 80% power (Simulation 2).*

A.2.4.2 AUC

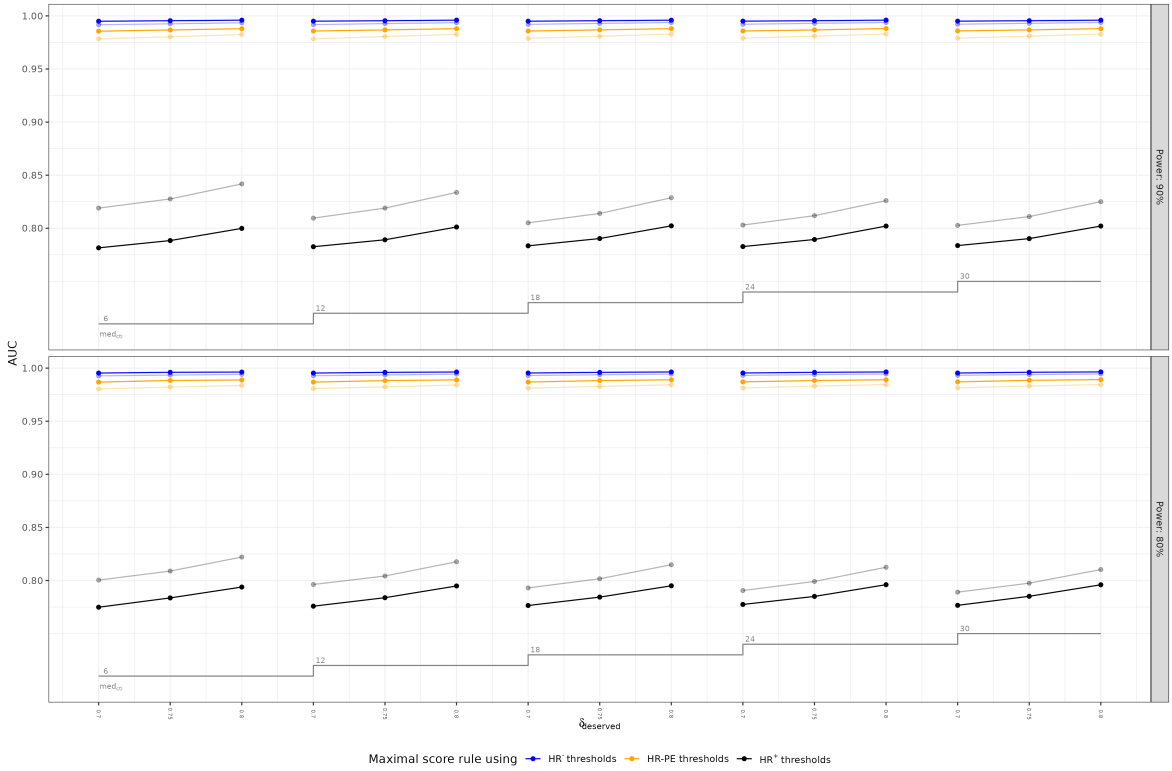

Figure 95: *AUC of ROC curves (y-axis) separated by $\delta_{deserved}$ (x-axis) of Scenario 4 for each sub-scenario (Simulation 2).*

Appendix B

Information to reproduce simulation studies

In this Appendix, additional information on R-Code structure and execution of the programs to determine the results of this thesis for both simulation studies are shown. The R-Code is provided at `https://www.github.com/cbuesch/SumulationStudyABAM`.
The following R-Packages need to be installed before R-Code usage: *tidyverse, stringr, cutpointr, vcd, data.table, survival, flexsurv, irr, foreach*, and *doParallel*. This can be done by using the code provided hereafter:

```
install.packages(c("tidyverse", "stringr", "cutpointr", "vcd", "data.table",
                   "survival", "flexsurv", "irr", "foreach", "doParallel"))
```

The provided programs are using the doParallel package for parallel computing in windows systems. In case a unix-like system is used, the package *doMC* instead of *doParallel* needs to be installed. Furthermore, all occurrences of

```
cl <- makeCluster(num.cl)
registerDoParallel(cl)
```

need to be replaced by

```
registerDoMC(num.cl)
```

Moreover, all lines containing

```
stopCluster(cl)
```

need to be removed. In addition, the running time of the programs (especially the data generation and data analysis) can take even with multiple cores for parallel computing several days. Therefore, to reduce the running time, 40 cores or more are recommended. Another solution is to reduce the number of iterations (n.sim), which decreases, however, the precision of the simulation study.

In the following, several R-Code-Scripts are explained, which can be found in the folder "RPrograms" at `https://www.github.com/cbuesch/SumulationStudyABAM`:

- Costume functions for data generation and analysis of both simulation studies (*Sim1_0_Functions_DataGeneration.R*, *Sim2_0_Functions_DataGeneration.R*, *Sim1_0_Functions_Analysis.R*, and *Sim2_0_Functions_Analysis.R*):
  These four scripts contain all needed functions for the data generation and analysis. Hence, they need to be loaded before conducting the simulations.

- Data generation of both simulation studies (*Sim1_1_DataGeneration.R* and *Sim2_1_DataGeneration.R*):
  These scripts conduct the data generation for all scenarios of both simulation studies. At the beginning the working directory needs to be set, where the file *Sim1_Simulation.seed* and the script *Sim1_0_Functions_DataGeneration.R* for Simulation 1 or the file *Sim2_Simulation.seed* and the script *Sim2_0_Functions_DataGeneration.R* for Simulation 2 are saved. Furthermore, as mentioned the number of cores (num.cl) and number of iterations (n.sim) need to be kept in mind so that the running time is not increased too much. At last the folder path for saving of the generated data needs to be specified.

- Data analysis of both simulation studies (*Sim1_2_DataAnalysis.R* and *Sim2_2_DataAnalysis.R*):
  These scripts conduct the data analysis of the generated data for all scenarios of both simulation studies. At the beginning the working directory needs to be set, where the script *Sim1_0_Functions_Analysis.R* for Simulation 1 or where the script *Sim2_0_Functions_Analysis.R* for Simulation 2 is saved. Furthermore, as mentioned the number of cores (num.cl) and number of iterations (n.sim) need to be kept in mind so that the running time is not increased too much. At last the folder path, where the generated data was saved (path_data), and the folder path, where the analysed data should be saved (path_ana) needs to be specified.

- Optimal cutoff cutoff determination of both simulation studies (*Sim1_3_DataAnalysis_ASCO_CutoffValues.R* and *Sim2_3_DataAnalysis_ASCO_CutoffValues.R*):
  These scripts conduct the investigating which ESMO/IQWiG$_{\text{RR}}$/Mod-IQWiG$_{\text{HR}}$ cat-

egory correspond to which ASCO score for both simulation studies. At the beginning the folder path, where the generated data and analysed data was saved, needs to be specified. As mentioned the number of cores (num.cl) and number of iterations (n.sim) need to be kept in mind so that the running time is not increased too much.

Appendix C

Derivation of censoring distribution parameter λ_C of Simulation 2

For Simulation 2 the censoring time distribution for the data generation was defined as a combination of independent administrative censoring and independent right-censoring with an overall targeted censoring proportion of p_C without introducing bias to the HR estimation. Hence, the distributions of administrative censoring, independent right-censoring, and failure times needed to be independent from each other, yielding equation 2.7 for p_C:

$$\begin{aligned} p_C &= \int \left(\int_0^{dur-a} f_T(t) F_C(t) dt + S_T(dur - a) \right) f_{Acc}(a) da \\ &= \int \int_0^{dur-a} f_T(t) F_C(t) f_{Acc}(a) dt da + \int S_T(dur - a) f_{Acc}(a) da. \end{aligned}$$

More information regarding derivation of p_C can be found in Section 2.3.2.2.

In the performed simulation study (Simulation 2) the censoring time distribution C is assumed to be exponential distributed ($C \sim \text{Exp}(\lambda_C)$). Furthermore, three different failure time distributions are assumed, leading to different p_C equations. Thus, in the following each failure time distribution, its effect on censoring probability p_C and how the censoring distribution parameter λ_C can be obtained, is outlined:

For improved readability, in the following part of this Appendix the abbreviations "1" and "2" were used instead of "ctl" and "trt", respectively. Let us assume failure times T_1 and T_2 with the same underlying distribution for the control and treatment group of a phase III clinical trial, respectively. Since the additional benefit assessment methods are only applied after a significant trial, the simulated trial must have a positive treatment effect. Hence, the parameters of the underlying distributions of T_1 and T_2 are different, leading to a hyper-distributed failure time distribution $T = p_1 \cdot T_1 + p_2 \cdot T_2$, where p_1 and p_2 are the probability of being allocated to the control and treatment group, respectively. In the following, the

implications on parameter definition of the censoring distribution assuming different failure time distributions is described:

- *Assuming hyper-exponential distributed failure time T:*
  In case of a hyper-exponential failure time distribution for $T_1 \sim \text{Exp}(\lambda_1)$ and $T_2 \sim \text{Exp}(\lambda_2)$, the density, distribution, and survival functions are defined as follows:

$$\begin{aligned} F_T(t) &= p_1 \cdot (1 - \exp(-\lambda_1 \cdot t)) + p_2 \cdot (1 - \exp(-\lambda_2 \cdot t)), \\ f_T(t) &= p_1 \cdot (\lambda_1 \cdot \exp(-\lambda_1 \cdot t)) + p_2 \cdot (\lambda_2 \cdot \exp(-\lambda_2 \cdot t)), \\ S_T(t) &= p_1 \cdot \exp(-\lambda_1 \cdot t) + p_2 \exp(-\lambda_2 \cdot t). \end{aligned}$$

  Moreover, based on assumed independent exponential censoring time $C \sim \text{Exp}(\lambda_C)$ and uniform accrual time $Acc \sim \text{Unif}[0, a_{max}]$, the censoring probability p_C can be calculated using formula (2.7):

$$\begin{aligned} p_C &= \int_0^{a_{max}} \int_0^{dur-a} f_T(t) F_C(t) f_{Acc}(a) dt da + \int_0^{a_{max}} S_T(dur - a) f_{Acc}(a) da \\ &= \int_0^{a_{max}} \int_0^{dur-a} (p_1 \cdot (\lambda_1 \cdot \exp(-\lambda_1 \cdot t)) + p_2 \cdot (\lambda_2 \cdot \exp(-\lambda_2 \cdot t))) \\ &\qquad \cdot (1 - \exp(-\lambda_C \cdot t)) \cdot \frac{1}{a_{max}} dt da + \\ &\quad \int_0^{a_{max}} (p_1 \cdot \exp(-\lambda_1 \cdot (dur - a)) + p_2 \exp(-\lambda_2 \cdot (dur - a))) \cdot \frac{1}{a_{max}} da \\ &= \frac{1}{a_{max}} \left(\frac{p_1 \lambda_1}{(\lambda_1 + \lambda_C)^2} \Big(\exp((a_{max} - dur)(\lambda_1 + \lambda_C)) - \exp(-dur(\lambda_1 + \lambda_C)) \Big) + \right. \\ &\qquad \left. \frac{p_2 \lambda_2}{(\lambda_2 + \lambda_C)^2} \Big(\exp((a_{max} - dur)(\lambda_2 + \lambda_C)) - \exp(-dur(\lambda_2 + \lambda_C)) \Big) \right) + \\ &\quad p_1 + p_2 - \frac{p_1 \lambda_1}{\lambda_1 + \lambda_C} - \frac{p_2 \lambda_2}{\lambda_2 + \lambda_C}. \end{aligned}$$

  As last step, this equation needed to be solved for λ_C, to receive the parameter of the exponential distributed independent censoring distribution. As this equation has no mathematical solution, a numerical approximation was used.

- *Assuming hyper-exponential distributed failure time T with delayed treatment effect in the treatment group using piece-wise exponential failure time distributions:*

  In this case, a hyper-exponential failure time distribution is present for $T_1 \sim \text{Exp}(\lambda_1)$ and T_2 being piece-wise exponential distributed, where start_2 is the time point of treatment effect start for the treatment group. Hence, the density, distribution, and survival functions are defined as follows:

$$F_T(x) = \begin{cases} p_1 \cdot (1 - \exp(-\lambda_1 t)) + p_2 \cdot (1 - \exp(-\lambda_2 t)), & \\ & t \in [0, \text{start}_2] \\ p_1 \cdot (1 - \exp(-\lambda_1 t)) + p_2 \cdot \Big(1 - \exp(-\lambda_1 \cdot \text{start}_2) & \cdot \exp(-\lambda_2 \cdot (t - \text{start}_2))\Big), \\ & \text{otherwise}, \end{cases}$$

$$f_T(x) = \begin{cases} p_1 \cdot (\lambda_1 \exp(-\lambda_1 t)) + p_2 \cdot (\lambda_2 \exp(-\lambda_2 t)), & \\ & t \in [0, \text{start}_2] \\ p_1 \cdot (\lambda_1 \exp(-\lambda_1 t)) + p_2 \cdot \Big(\lambda_2 \exp(-\lambda_1 \cdot \text{start}_2) & \cdot \exp(-\lambda_2 \cdot (t - \text{start}_2))\Big), \\ & \text{otherwise}, \end{cases}$$

$$S_T(x) = \begin{cases} p_1 \cdot \exp(-\lambda_1 t) + p_2 \cdot \exp(-\lambda_2 t), & \\ & t \in [0, \text{start}_2] \\ p_1 \cdot \exp(-\lambda_1 t) + p_2 \cdot \Big(\exp(-\lambda_1 \cdot \text{start}_2) & \cdot \exp(-\lambda_2 \cdot (t - \text{start}_2))\Big), \\ & \text{otherwise}. \end{cases}$$

  Moreover, based on assumed independent exponential censoring time $C \sim \text{Exp}(\lambda_C)$ and uniform accrual time $Acc \sim \text{Unif}[0, a_{max}]$, the censoring probability p_C can be calculated using formula (2.7):

$$
\begin{aligned}
p_C &= \int_0^{a_{max}} \int_0^{dur-a} f_T(t) F_C(t) f_{Acc}(a) dt da + \int_0^{a_{max}} S_T(dur-a) f_{Acc}(a) da \\
&= \int_0^{a_{max}} \left[\int_0^{\text{start}_2} f_T^{t\in[0,\text{start}_2]}(t) F_C(t) f_{Acc}(a) dt + \int_{\text{start}_2}^{dur-a} f_T^{otherwise}(t) F_C(t) f_{Acc}(a) dt \right] da + \int_0^{a_{max}} \underbrace{S_T(dur-a)}_{\substack{dur-a>\text{start}_2 \\ \rightarrow S_T^{otherwise}(dur-a)}} f_{Acc}(a) da \\
&= \underbrace{\int_0^{a_{max}} \int_0^{\text{start}_2} \left[p_1 \cdot (\lambda_1 \exp(-\lambda_1 t)) + p_2 \cdot (\lambda_2 \exp(-\lambda_2 t)) \right] \cdot (1 - \exp(-\lambda_C \cdot t)) \cdot \frac{1}{a_{max}} dt da +}_{\text{(A)}} \\
&\quad \underbrace{\int_0^{a_{max}} \int_{\text{start}_2}^{dur-a} \left[p_1 \cdot (\lambda_1 \exp(-\lambda_1 t)) + p_2 \cdot \left(\lambda_2 \exp(-\lambda_1 \cdot \text{start}_2) \cdot \exp(-\lambda_2 \cdot (t - \text{start}_2)) \right) \right] \cdot (1 - \exp(-\lambda_C \cdot t)) \cdot \frac{1}{a_{max}} dt da +}_{\text{(A)}} \\
&\quad \underbrace{\int_0^{a_{max}} \left[p_1 \cdot \exp(-\lambda_1 (dur-a)) + p_2 \cdot \left(\exp(-\lambda_1 \cdot \text{start}_2) \cdot \exp(-\lambda_2 \cdot ((dur-a) - \text{start}_2)) \right) \right] \cdot \frac{1}{a_{max}} da}_{\text{(B)}} .
\end{aligned}
$$

(A) and (B) can be solved as follows:

$$
\begin{aligned}
\text{(A)} &= \frac{1}{a_{max}} \cdot \int_0^{a_{max}} \Bigg(- p_1 \cdot \Big(\exp(-\lambda_1 (dur - a)) - 1 \Big) - p_2 \cdot \Big(\exp(-\lambda_1 \text{start}_2) \cdot (\exp(-\lambda_2 (dur - a - \text{start}_2)) - 1) + \exp(-\lambda_2 \text{start}_2) - 1 \Big) + \\
&\qquad \frac{p_1 \lambda_1}{\lambda_1 + \lambda_C} \cdot \Big(\exp(-(dur - a)(\lambda_1 + \lambda_C)) - 1 \Big) + \\
&\qquad \frac{p_2 \lambda_2}{\lambda_2 + \lambda_C} \cdot \Big(\exp(-\lambda_1 \text{start}_2)(\exp(-\lambda_2 (dur - a - \text{start}_2) - \lambda_C (dur - a)) - \exp(\lambda_C - a)) + \exp(-\text{start}_2 (\lambda_2 + \lambda_C)) - 1 \Big) \Bigg) da \\
&= \frac{1}{a_{max}} \cdot \Bigg(- \frac{p_1}{\lambda_1} \cdot \Big(\exp(-\lambda_1 (dur - a_{max})) - \exp(-\lambda_1 dur) \Big) - \frac{p_2}{\lambda_2} \cdot \exp(-\lambda_1 \text{start}_2) \cdot \Big(\exp(-\lambda_2 (dur - a_{max} - \text{start}_2)) - \exp(-\lambda_2 (dur - \text{start}_2)) \Big) \Bigg) + \\
&\quad \frac{1}{a_{max}} \cdot \Bigg(\frac{p_1 \lambda_1}{(\lambda_1 + \lambda_C)^2} \cdot \Big(\exp(-(dur - a_{max})(\lambda_1 + \lambda_C)) - \exp(-dur(\lambda_1 + \lambda_C)) \Big) + \\
&\qquad \frac{p_2 \lambda_2 \exp(-\lambda_1 \text{start}_2)}{(\lambda_2 + \lambda_C)^2} \cdot \Big(\exp\big(- \lambda_2 (dur - a_{max} - \text{start}_2) - \lambda_C (dur - a_{max}) \big) - \exp\big(- \lambda_2 (dur - \text{start}_2) - \lambda_C dur \big) \Big) \Bigg) + \\
&\quad p_1 - p_2 \cdot \Big(\exp(-\lambda_2 \text{start}_2) - \exp(-\lambda_1 \text{start}_2) - 1 \Big) - \frac{p_1 \lambda_1}{\lambda_1 + \lambda_C} - \frac{p_2 \lambda_2}{\lambda_2 + \lambda_C} \cdot \Big(\exp(-\text{start}_2 (\lambda_1 + \lambda_C)) - \exp(-\text{start}_2 (\lambda_2 + \lambda_C)) + 1 \Big),
\end{aligned}
$$

$$
\begin{aligned}
\text{(B)} &= \frac{1}{a_{max}} \cdot \left[\frac{p_1}{\lambda_1} \cdot \exp(-\lambda_1 (dur - a)) + \frac{p2 \cdot \exp(-\lambda_1 \text{start}_2)}{\lambda_2} \cdot \exp(-\lambda_2 (dur - a - \text{start}_2)) \right]_0^{a_{max}} \\
&= \frac{1}{a_{max}} \cdot \Bigg(\frac{p_1}{\lambda_1} \cdot \Big(\exp(-\lambda_1 (dur - a_{max})) - \exp(-\lambda_1 dur) \Big) + \frac{p_2}{\lambda_2} \cdot \exp(-\lambda_1 \text{start}_2) \cdot \Big(\exp(-\lambda_2 (dur - a_{max} - \text{start}_2)) - \exp(-\lambda_2 (dur - \text{start}_2)) \Big) \Bigg).
\end{aligned}
$$

Hence, combining the results from (A) and (B) yields:

$$
\begin{aligned}
p_C = \frac{1}{a_{max}} \cdot \Bigg(& \frac{p_1\lambda_1}{(\lambda_1+\lambda_C)^2} \cdot \Big(\exp(-(dur - a_{max})(\lambda_1+\lambda_C)) - \exp(-dur(\lambda_1+\lambda_C)) \Big) + \\
& \frac{p_2\lambda_2 \exp(-\lambda_1 \text{start}_2)}{(\lambda_2+\lambda_C)^2} \cdot \Big(\exp\big(-\lambda_2(dur - a_{max} - \text{start}_2) - \lambda_C(dur - a_{max}) \big) - \\
& \exp\big(-\lambda_2(dur - \text{start}_2) - \lambda_C dur \big) \Big) \Bigg) + \\
& p_1 - p_2 \cdot \Big(\exp(-\lambda_2 \text{start}_2) - \exp(-\lambda_1 \text{start}_2) - 1 \Big) - \frac{p_1\lambda_1}{\lambda_1+\lambda_C} - \\
& \frac{p_2\lambda_2}{\lambda_2+\lambda_C} \cdot \Big(\exp(-\text{start}_2(\lambda_1+\lambda_C)) - \exp(-\text{start}_2(\lambda_2+\lambda_C)) + 1 \Big).
\end{aligned}
$$

As last step, this equation needed to be solved for λ_C, to receive the parameter of the exponential distributed independent censoring distribution. As this equation has no mathematical solution, a numerical approximation was used.

- *Assuming hyperweibull distributed failure time T:*

  In case of a hyperweibull failure time distribution for $T_1 \sim \text{Weib}(\lambda_1, k_1)$ and $T_2 \sim \text{Weib}(\lambda_2, k_2)$, the density, distribution, and survival functions are defined as follows:

$$F_T(t) = p_1 \cdot \Big(1 - \exp(-(t\lambda_1)^{k_1})\Big) + p_2 \cdot \Big(1 - \exp(-(t\lambda_2)^{k_2})\Big),$$
$$f_T(t) = p_1 \cdot \Big((k_1\lambda_1 \cdot (t\lambda_1)^{k_1-1} \exp(-(t\lambda_1)^{k_1}))\Big) + p_2 \cdot \Big((k_2\lambda_2 \cdot (t\lambda_2)^{k_2-1} \exp(-(t\lambda_2)^{k_2}))\Big),$$
$$S_T(t) = p_1 \cdot \exp(-(t\lambda_1)^{k_1}) + p_2 \cdot \exp(-(t\lambda_2)^{k_2}).$$

  Moreover, based on assumed independent exponential censoring time $C \sim \text{Exp}(\lambda_C)$ and uniform accrual time $Acc \sim \text{Unif}[0, a_{max}]$, the censoring probability p_C can be calculated using formula (2.7):

$$\begin{aligned}
p_C &= \int_0^{a_{max}} \int_0^{dur-a} f_T(t) F_C(t) f_{Acc}(a) dt da + \int_0^{a_{max}} S_T(dur-a) f_{Acc}(a) da \\
&= \int_0^{a_{max}} \int_0^{dur-a} \left(p_1 \cdot \Big(1 - \exp(-(t\lambda_1)^{k_1})\Big) + p_2 \cdot \Big(1 - \exp(-(t\lambda_2)^{k_2})\Big) \right) \cdot \\
&\qquad (1 - \exp(-\lambda_C t)) \cdot \frac{1}{a_{max}} dt da + \\
&\quad \int_0^{a_{max}} \left(p_1 \cdot \exp(-((dur-a)\lambda_1)^{k_1}) + p_2 \cdot \exp(-((dur-a)\lambda_2)^{k_2}) \right) \cdot \frac{1}{a_{max}} da \\
&= p_1 + p_2 - \frac{1}{a_{max}} \Bigg(\int_0^{a_{max}} \int_0^{dur-a} p_1 k_1 \lambda_1 (t\lambda_1)^{k_1-1} \exp(-(t\lambda_1)^{k_1}) \exp(-\lambda_C t) + \\
&\qquad p_2 k_2 \lambda_2 (t\lambda_2)^{k_2-1} \exp(-(t\lambda_2)^{k_2}) \exp(-\lambda_C t) dt da \Bigg).
\end{aligned}$$

  As last step, this equation needed to be solved for λ_C, to receive the parameter of the exponential distributed independent censoring distribution. As this equation has no mathematical solution, a numerical approximation was used.

- *Assuming hyper-gompertz distributed failure time T:*

  In case of a hypergompertz failure time distribution for $T_1 \sim \text{Gomp}(a_1, b_1)$ and $T_2 \sim \text{Gomp}(a_1, b_2)$, the density, distribution, and survival functions are defined as follows:

$$
\begin{aligned}
F_T(t) &= p_1 \cdot \Big(1 - \exp\Big(-\frac{b_1}{a_1}(\exp(a_1 t) - 1)\Big)\Big) + p_2 \cdot \Big(1 - \exp\Big(-\frac{b_2}{a_2}(\exp(a_2 t) - 1)\Big)\Big), \\
f_T(t) &= p_1 \cdot \Big(b_1 \exp(a_1 t) \exp\Big(-\frac{b_1}{a_1}(\exp(a_1 t) - 1)\Big)\Big) \\
&\quad + p_2 \cdot \Big(b_2 \exp(a_2 t) \exp\Big(-\frac{b_2}{a_2}(\exp(a_2 t) - 1)\Big)\Big), \\
S_T(t) &= p_1 \cdot \exp\Big(-\frac{b_1}{a_1}(\exp(a_1 t) - 1)\Big) + p_2 \cdot \exp\Big(-\frac{b_2}{a_2}(\exp(a_2 t) - 1)\Big).
\end{aligned}
$$

  Moreover, based on assumed independent exponential censoring time $C \sim \text{Exp}(\lambda_C)$ and uniform accrual time $Acc \sim \text{Unif}[0, a_{max}]$, the censoring probability p_C can be calculated using formula (2.7):

$$
\begin{aligned}
p_C &= \int_0^{a_{max}} \int_0^{dur-a} f_T(t) F_C(t) f_{Acc}(a) dt da + \int_0^{a_{max}} S_T(dur-a) f_{Acc}(a) da \\
&= \int_0^{a_{max}} \int_0^{dur-a} \Big[p_1 \cdot b_1 \exp(a_1 t) \exp\Big(-\frac{b_1}{a_1}(\exp(a_1 t) - 1)\Big) + p_2 \cdot b_2 \exp(a_2 t) \exp\Big(-\frac{b_2}{a_2}(\exp(a_2 t) - 1)\Big)\Big] \cdot \\
&\qquad (1 - \exp(-\lambda_C t)) \cdot \frac{1}{a_{max}} dt da + \frac{1}{a_{max}} \int_0^{a_{max}} S_T(dur-a) da \\
&= \frac{1}{a_{max}} \int_0^{a_{max}} \int_0^{dur-a} \Big[p_1 \cdot b_1 \exp(a_1 t) \exp\Big(-\frac{b_1}{a_1}(\exp(a_1 t) - 1)\Big) + p_2 \cdot b_2 \exp(a_2 t) \exp\Big(-\frac{b_2}{a_2}(\exp(a_2 t) - 1)\Big)\Big] dt da - \\
&\quad \frac{1}{a_{max}} \int_0^{a_{max}} \int_0^{dur-a} \Big[p_1 \cdot b_1 \exp(a_1 t) \exp\Big(-\frac{b_1}{a_1}(\exp(a_1 t) - 1)\Big) + p_2 \cdot b_2 \exp(a_2 t) \exp\Big(-\frac{b_2}{a_2}(\exp(a_2 t) - 1)\Big)\Big] \cdot \exp(-\lambda_C t) dt da + \\
&\quad \frac{1}{a_{max}} \int_0^{a_{max}} S_T(dur-a) da \\
&= \frac{1}{a_{max}} \int_0^{a_{max}} \underbrace{\Big[-p_1 \cdot \exp\Big(-\frac{b_1}{a_1}(\exp(a_1 t) - 1)\Big) - p_2 \cdot \exp\Big(-\frac{b_2}{a_2}(\exp(a_2 t) - 1)\Big)\Big]_0^{dur-a}}_{\text{upper limit is equal to } -S_T(dur-a)} da - \\
&\quad \frac{1}{a_{max}} \int_0^{a_{max}} \int_0^{dur-a} \Big[p_1 \cdot b_1 \exp(a_1 t) \exp\Big(-\frac{b_1}{a_1}(\exp(a_1 t) - 1)\Big) + p_2 \cdot b_2 \exp(a_2 t) \exp\Big(-\frac{b_2}{a_2}(\exp(a_2 t) - 1)\Big)\Big] \cdot \exp(-\lambda_C t) dt da + \\
&\quad \frac{1}{a_{max}} \int_0^{a_{max}} S_T(dur-a) da \\
&= -\frac{1}{a_{max}} \int_0^{a_{max}} S_T(dur-a) da + \frac{1}{a_{max}} \int_0^{a_{max}} p_1 + p_2 da - \\
&\quad \frac{1}{a_{max}} \int_0^{a_{max}} \int_0^{dur-a} \Big[p_1 \cdot b_1 \exp(a_1 t) \exp\Big(-\frac{b_1}{a_1}(\exp(a_1 t) - 1)\Big) + p_2 \cdot b_2 \exp(a_2 t) \exp\Big(-\frac{b_2}{a_2}(\exp(a_2 t) - 1)\Big)\Big] \cdot \exp(-\lambda_C t) dt da + \\
&\quad \frac{1}{a_{max}} \int_0^{a_{max}} S_T(dur-a) da
\end{aligned}
$$

$$
\begin{aligned}
&= \frac{1}{a_{max}} \int_0^{a_{max}} p_1 + p_2 da - \\
&\frac{1}{a_{max}} \int_0^{a_{max}} \int_0^{dur-a} \Big[p_1 \cdot b_1 \exp(a_1 t) \exp\Big(- \frac{b_1}{a_1} (\exp(a_1 t) - 1) \Big) + \\
&\qquad\qquad p_2 \cdot b_2 \exp(a_2 t) \exp\Big(- \frac{b_2}{a_2} (\exp(a_2 t) - 1) \Big) \Big] \cdot \exp(-\lambda_C t) dt da \\
&= p_1 + p_2 - \frac{1}{a_{max}} \int_0^{a_{max}} \int_0^{dur-a} \Big[p_1 \cdot b_1 \exp(a_1 t) \exp\Big(- \frac{b_1}{a_1} (\exp(a_1 t) - 1) \Big) + \\
&\qquad\qquad p_2 \cdot b_2 \exp(a_2 t) \exp\Big(- \frac{b_2}{a_2} (\exp(a_2 t) - 1) \Big) \Big] \cdot \\
&\qquad\qquad \exp(-\lambda_C t) dt da.
\end{aligned}
$$

As last step, this equation needed to be solved for λ_C to receive the parameter of the exponential distributed independent censoring distribution. As this equation has no mathematical solution, a numerical approximation was used.

In the performed simulation study, numerical approximation for solving double integrals was performed by an algorithm proposed by Piessens (1983), which is implemented in the integrate function of the stats package in R. Furthermore, minimization without derivatives introduced by Brent and Brent (1974) was applied for solving for λ_C as no mathematical solution exists, which is implemented in the uniroot function of the stats package in R.

Curriculum Vitae

Personal information

Christopher Alexander Büsch, born on the 20th of July 1993 in Neuwied am Rhein, Germany.

Education

Heidelberg University - Doctoral student (Dr. sc. hum.) since 10/2018

University of Applied Sciences Koblenz - 04/2016 - 08/2018
Master of Science: Applied Mathematics

University of Canterbury, Christchurch - Study Abroad 07/2017 - 01/2018

University of Applied Sciences Koblenz - 03/2013 - 04/2016
Bachelor of Science: Biomathematics

Werner-Heisenberg-Gymnasium, Neuwied - A-Level (Abitur) 08/2004 - 03/2013

Professional experience

Heidelberg University - University Hospital since 10/2018
Research Fellow at Institute of Medical Biometry
University of Applied Sciences Koblenz 08/2016 - 01/2018
Scientific assistant

Acknowledgments

First and foremost I would like to thank my supervisor Prof. Dr. Meinhard Kieser for the opportunity to conduct this research and to write this thesis. I greatly value his provided scientific support, constructive suggestions and always positive approach throughout this research.

Furthermore, I would like to thank my current and former colleagues at the Institute for Medical Biometry for the worthwhile atmosphere and moral support. A special thanks goes out to my colleagues Dr. Marietta Kirchner, Rouven Behnisch, Dr. Jan Meis, and Dr. Svenja Seide for the joint in-depth discussions and constructive criticism, which contributed greatly to this work.

Last but definitely not least, I would like to thank my family and family-in-law, especially my wife Laura Büsch-Kievit. Without their emotional support and always believing in me, this work would not have been possible to achieve.

Eidesstattliche Versicherung

1. Bei der eingereichten Dissertation zu dem Thema *Comparison of different additional benefit assessment methods for oncology treatments* handelt es sich um meine eigenständig erbrachte Leistung.

2. Ich habe nur die angegebenen Quellen und Hilfsmittel benutzt und mich keiner unzulässigen Hilfe Dritter bedient. Insbesondere habe ich wörtlich oder sinngemäß aus anderen Werken übernommene Inhalte als solche kenntlich gemacht.

3. Die Arbeit oder Teile davon habe ich bislang nicht an einer Hochschule des In- oder Auslands als Bestandteil einer Prüfungs- oder Qualifikationsleistung vorgelegt.

4. Die Richtigkeit der vorstehenden Erklärungen bestätige ich.

5. Die Bedeutung der eidesstattlichen Versicherung und die strafrechtlichen Folgen einer unrichtigen oder unvollständigen eidesstattlichen Versicherung sind mir bekannt. Ich versichere an Eides statt, dass ich nach bestem Wissen die reine Wahrheit erklärt und nichts verschwiegen habe.

Ort und Datum Unterschrift

Angaben zu verwendeten KI-basierter elektronischer Hilfsmittel

Zur Dokumentation der verwendeten Hilfsmittel ist der schriftlichen Ausarbeitung kein Anhang bezüglich Liste und Beschreibung aller verwendeter KI-basierter Hilfsmittel hinzugefügt, weil keine KI-basierten elektronischen Hilfsmittel verwendet wurden. Mir ist bewusst, dass insbesondere der Versuch einer nicht dokumentierten Nutzung KI-basierter Hilfsmittel als Täuschungsversuch zu werten ist:

Gem. §16 Abs. 2 der Promotionsordnung "Dr. sc. hum.":
"Ergibt sich vor Aushändigung der Promotionsurkunde, dass der Doktorand / Doktorandin bei einer Promotionsleistung getäuscht hat, so kann der Promotionsausschuss diese Promotionsleistung oder alle bisher erbrachten Promotionsleistungen für ungültig erklären. In besonders schweren Fällen kann der Promotionsausschuss die Annahme als Doktorand / Doktorandin endgültig widerrufen."

Ort und Datum Unterschrift