ScienceDirect

Contents lists available at sciencedirect.com
Journal homepage: www.elsevier.com/locate/jval

Health Policy Analysis

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

Christopher A. Büsch, MSc, Johannes Krisam, PhD, Meinhard Kieser, PhD

ABSTRACT

Objectives: After a successful Marketing Authorization Application for clinical trials with time-to-event endpoints, the degree of the added benefit from new treatments remains unknown and needs to be assessed. Unfortunately, until now no clear definition for added benefit determination of a treatment exists. Nevertheless, European authorities / societies have developed 2 "additional benefit assessment" methods, which have up to now not been compared: the European Society for Medical Oncology (ESMO) developed a dual rule considering relative and absolute benefit. The German Institute for Quality and Efficiency in Health Care (IQWiG) developed a method using upper 95% hazard ratio confidence interval.

Methods: We evaluate and compare both methods in an extensive simulation study including different censoring rates, failure time distributions, and treatment effects for sample size calculation. The methods' performance is assessed via Receiver Operating Characteristic curves, Spearman correlation, and percentage of achieved maximal scores.

Results: The results show that IQWiG's method has in many situations a lower maximal scoring proportion than ESMO's rule, that is, up to 28.5% versus 94.7%. Various failure time distributions lead to strongly changed maximal scoring percentages for ESMO. High positive correlation between the methods is present for moderate treatment effects.

Conclusions: IQWiG's method is usually more conservative than ESMO's. ESMO's rule tends to be more susceptible for various failure time distributions. Using the lower confidence interval limit seems to be a better solution resulting in a higher true-positive rate without increasing the false-positive rate. Thus, IQWiG's method might need to be adapted accordingly to achieve a better overall classification.

Keywords: additional benefit assessment, clinical phase III trials, oncology trials, survival analysis.

VALUE HEALTH. 2022; 25(11):1853–1862

Introduction

Pharmaceutical companies search for new drugs against specific diseases, for example cancer, which have to be investigated with respect to their quality, safety, and efficacy, preventing a nonbeneficial drug to get on the market. Most efficacy endpoints in cancer trials are time-to-event endpoints such as overall survival or progression-free survival. Therefore, a statistically significant log-rank test result is usually one of many requirements for submission of a Marketing Authorization Application to an appropriate authority, which decides whether the drug is allowed to enter the market. Nevertheless, the degree of the added benefit from new and effective treatments with time-to-event endpoints derived from clinical trials remains unknown at this stage and needs to be assessed. Unfortunately, until now no clear definition and hence no gold-standard for added benefit determination of a treatment exists. However, European authorities/societies have developed 2 different benefit assessment scores for time-to-event endpoints.

In Germany, the Federal Joint Committee defines the additional benefit of new drugs, which forms the basis of negotiations on the reimbursement price. For its decision, it commissions the Institute for Quality and Efficiency in Health Care (IQWiG) to evaluate the additional benefit of new drugs. Thus, the classification takes an important role in how much a new treatment is worth in economical terms. When evaluating time-to-event endpoints, the IQWiG makes use of the upper limit of the 95% hazard ratio (HR) confidence interval (CI), which is compared with specific thresholds categorizing the new treatment into 3 categories: major, considerable, and minor added benefit[1] (Fig. 1). The resulting categories can afterward still be adjusted to reflect other important endpoints such as toxicity and quality of life of the new treatment. Furthermore, the utilized thresholds were calculated assuming binomially distributed data using the relative risk (RR);

Figure 1. Detailed illustration of ESMO's, IQWiG's and modified IQWiG's clinical additional benefit assessment methods for overall survival / time-to-event endpoints.

ESMO's categories					
	1 (low benefit)	2 (low benefit)	3 (low benefit)	4 (substantial improvement	5 (substantial improvement)
$med_C \leq 12$	$HR^- > 0.7$ OR gain < 1.5	$HR^- \leq 0.65$ AND gain ϵ [1.5, 2) OR $HR^- \epsilon$ (0.65, 0.7] AND gain ≥ 1.5	$HR^- \leq 0.65$ AND gain ϵ [2, 3)	$HR^- \leq 0.65$ AND gain ≥ 3 OR Increase in 2-year survival ≥ 10 %	Only achievable with toxicity, QoL or other bonus point adjustments
$med_C \epsilon$ (12, 24]	$HR^- > 0.75$ OR gain < 1.5	$HR^- \leq 0.7$ AND gain ϵ [1.5, 3) OR $HR^- \epsilon$ (0.7, 0.75] AND gain ≥ 1.5	$HR^- \leq 0.7$ AND gain ϵ [3, 5)	$HR^- \leq 0.7$ AND gain ≥ 5 OR Increase in 3-year survival ≥ 10 %	
$med_C > 24$	$HR^- > 0.75$ OR gain < 4	$HR^- \leq 0.7$ AND gain ϵ [4, 6) OR $HR^- \epsilon$ (0.7, 0.75] AND gain ≥ 4	$HR^- \leq 0.7$ AND gain ϵ [6, 9)	$HR^- \leq 0.7$ AND gain ≥ 9 OR Increase in 5-year survival ≥ 10 %	

IQWiG's categories		
minor added benefit	considerable added benefit	major added benefit
$HR^+ \epsilon [0.95, 1)_{RR}$	$HR^+ \epsilon [0.85, 0.95)_{RR}$	$HR^+ < 0.85_{RR}$

Mod-IQWiG's$_{HR}$ (modified IQWiG method)		
minor added benefit	considerable added benefit	major added benefit
$HR^+ \epsilon [0.93, 1)_{HR}$	$HR^+ \epsilon [0.79, 0.93)_{HR}$	$HR^+ < 0.79_{HR}$

ESMO indicates European Society for Medical Oncology; gain, absolute difference in median survival times (in months); HR, hazard ratio; HR^+, estimated upper 95% confidence interval limit of the hazard ratio; HR^-, estimated lower 95% confidence interval limit of the hazard ratio; IQWiG, Institute for Quality and Efficiency in Health Care; IQWiG's, IQWiG's method using relative risk scaled thresholds; Mod-IQWiG'sHR, modified IQWiG method using upper confidence interval limit based on IQWiG's thresholds (transformation into HR-scaled thresholds using the conversion formula proposed by VanderWeele[2]); med_C, estimated median survival time in the control group (in months); QoL, quality of life; RR, relative risk.

ie, a true RR of 0.5 was defined as an effect of "major" extent for the outcome all-cause mortality.[1] These RR-scaled thresholds are also used for time-to-event data, which are frequently present in oncology trials. To investigate the possible influence of differently scaled thresholds on the grading of added benefit for new treatments, we transformed the provided RR-scaled thresholds into HR-scaled thresholds using the conversion formula proposed by VanderWeele.[2]

On a European level, the European Society for Medical Oncology (ESMO) has developed the Magnitude of Clinical Benefit Scale version 1.1 using a dual rule to compute a preliminary scale. The dual rule consists of the relative and absolute benefit achieved by the new treatment compared to specific thresholds categorizing new drugs into 4 categories. The relative and absolute benefit are assessed by the lower limit of the 95% HR-CI (HR^-), and the observed absolute difference in median survival times (gain), respectively. These thresholds vary based on the observed median survival times in the control group (med_C); for example, if gain $<$ 1.5 months (and $HR^- > 0.7$) for $med_C \leq 24$ or gain $<$ 4 (and $HR^- > 0.7$) for $med_C > 24$, only category 1 can be achieved (Fig. 1, column "1 (low benefit)"). Additionally, a maximum preliminary category can be achieved if the survival rate increases by 10% or more at key milestones. After the rating of the dual rule, the resulting preliminary scale can be adjusted to reflect the toxicity and quality of life of the new treatment. Therefore, ESMO's method in the non-curative setting with overall survival as primary endpoint classifies new treatments into 5 clinical benefit categories (Fig. 1), where grades 5 and 4 represent substantial and grades 3 to 1 low benefit.[3,4] ESMO's classification has a more indirect influence on the treatments price, meaning that "drugs, which obtain the highest scores on the scale, will be emphasized in the ESMO guidelines, with the hope that they will be rapidly endorsed by health authorities across the European Union."[3]

All measures used by both methods are calculated from the trials of the new medication. Additionally, in both scientific publications of these 2 methods,[1,3,4] it is stated that the upper/lower limit of the CI takes the variability of the estimate into account and hence should provide more information than the point estimate (PE). Moreover, Skipka et al[1] indicate that the PE might be biased for trials that were discontinued at a preplanned interim analysis. Nevertheless, as mentioned in 2 letters to the editor,[5,6] the use of the lower CI limit could lead to a higher probability of a better

grade because it may reward studies with a smaller sample size and, hence, a wider CI. This issue in ESMO's method has been addressed by inclusion of the required absolute benefit thresholds.[3,4]

Until now, only the usage of the lower CI limit compared with the PE was performed,[7] whereas a comparison of the 2 above mentioned methods, to understand the differences between the benefit assessment rules, has not been performed. Especially the question on which fundamental statistical idea (upper vs lower CI limit) might be better for determination of additional added benefit needs to be verified so that new drugs are fairly classified. This article tries to eliminate this knowledge gap and creates a systematic and detailed overview of the differences between the statistical parts of the assessment methods.

Methods

An extensive simulation study was performed by generating different scenarios of phase III trials to provide the aspired detailed overview between the 2 methods. The data generation was performed by standard time-to-event data mechanism: Failure and censoring times were generated n_{obs} times, and the minimum of both values was used as event time for the analysis. The sample size n_{obs} for each simulated trial was derived using Schoenfeld's approach[8,9] to ensure a specific power for a two-sided log-rank test at a significance level of 5% for the treatment effect designHR. Three different distributions with proportional hazards were used for failure time generation: exponential, Weibull, and Gompertz. During data generation, a fixed value for the median survival time in the control group (med_C) and the treatment effect (trueHR) were used to derive the required parameters of the failure time distributions. In order to assess both overpowered and underpowered studies with incorrect assumed treatment effects, the factor HR_{var} was defined for deviance between designHR and trueHR (trueHR = designHR·HR_{var}). To achieve proportional hazards for Weibull and Gompertz distributions, the shape parameter of each of the distributions was fixed to 2 different values causing the hazard function to increase/decrease over time.

Overall, the following scenarios were used, where each subscenario was generated 10 000 times:

1. *Standard scenario (scenario 1):* exponentially distributed failure times using designHR ∈ {0.3, 0.32, …, 0.9}, designHR = trueHR, med_C ∈ {6, 12, 18, 24, 30 months}, power of 80% and 90%; leading to (31 • 5 • 2 • 3=) 930 subscenarios. The fixed parameters med_C, designHR, and trueHR (HR • HR_{var}) were used to calculate the required parameters λ_C and λ_T of the exponential distribution (see Appendix in Supplemental Materials found at https://doi.org/10.1016/j.jval.2022.05.015 for further information).
2. *Incorrect assumed treatment effect (scenario 2):* Overpowered/underpowered studies using the same parameters as scenario 1; except designHR ≠ trueHR using HR_{var} ∈ {0.8, 0.9, 1.1, 1.2}; leading to (31 • 5 • 2 • 3 • 4=) 3720 subscenarios.
3. *Two different parameter distributions (scenario 3):* Same parameters as scenario 1, except using different failure time distributions in compliance with proportional hazards.
   - *Scenario 3a:* Weibull distributions using shape ∈ {0.5, 1.5}; leading to (31 • 5 • 2 • 3 • 2=) 1860 subscenarios
   - *Scenario 3b:* Gompertz distributions using shape ∈ {-0.2, 0.2}; leading to (31 • 5 • 2 • 3 • 2=) 1860 subscenarios
4. *Non-proportional hazards/non-constant HR (scenario 4):* Delayed treatment effect for the treatment group using piece-wise exponential failure time distributions; leading to (31 • 5 • 2 • 3=) 930 subscenarios. To achieve a late treatment effect for the treatment group, a piece-wise exponential distribution was chosen:

$$F_C(x) = 1-exp(-\lambda_C \cdot x),$$

$$F_T(x) = \begin{cases} 1-exp(-\lambda_C \cdot x) & , x \in [0,\ start_T] \\ 1-exp(-\lambda_C \cdot start_T \cdot exp(-\lambda_T \cdot (x-start_T)) & , otherwise \end{cases}$$

where F_C and F_T are the cumulative distribution functions of the treatment and control group, $\lambda_C > 0$ and $\lambda_T > 0$ are the parameters of the corresponding exponential distributions, and $start_T$ ($= \frac{1}{3} \cdot med_C$) is the time point of treatment effect start for the treatment group. The failure times of the treatment groups were generated using the inversion method by Kolonko (chapter 8).[10] Hence, proportional hazards were assumed before and after $start_T$. Additionally, λ_C and λ_T were defined the same way as in the standard scenario 1 (see Appendix in Supplemental Materials found at https://doi.org/10.1016/j.jval.2022.05.015 for further information).

In every scenario, a combination of administrative censoring (accrual time of 2 years and follow-up time of 2·med_C) and exponential censoring was used, aiming for an overall censoring rate of at least 20%, 40%, or 60%. The Appendix in Supplemental Materials found at https://doi.org/10.1016/j.jval.2022.05.015 gives further information about the data generation mechanism.

Each simulated trial was analyzed using a log-rank test, and if significant, IQWiG's and ESMO's methods were applied. Therefore, HR-PEs with corresponding 95% Wald-CIs using Cox regressions, gain, and the 2-, 3-, and 5-year survival increase were calculated (Fig. 1). In rare cases of subscenarios with large treatment effects, the survival curve did not fall below 50% and thus med_C or med_T could not be calculated. To overcome this issue, a conservative approach was implemented, using instead the last present time point (event or censoring) in the survival curve. Moreover, to investigate the potential influence of wrongly RR-scaled thresholds of IQWiG's method (IQWiG), the provided RR-scaled thresholds were transformed into HR-scaled thresholds (Mod-$IQWiG_{HR}$) using the conversion formula by VanderWeele[2] (Fig. 1):

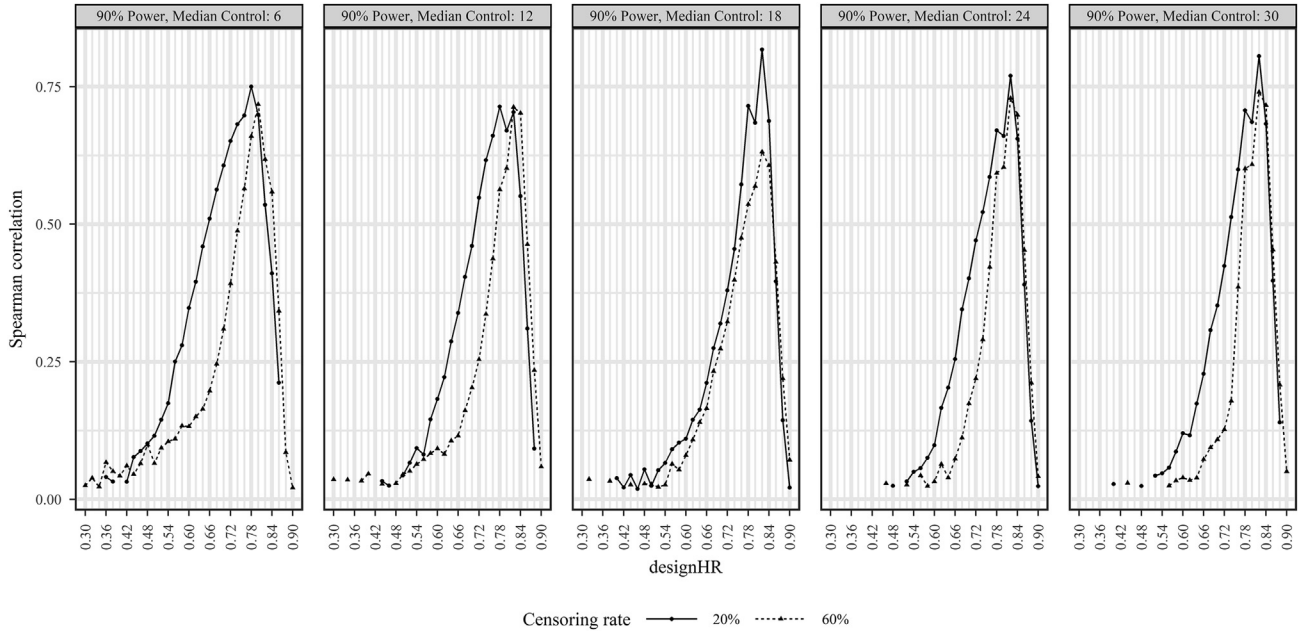
$$RR = \frac{1-0.5^{\sqrt{HR}}}{1-0.5^{\sqrt{1/HR}}}$$

Given that this formula has no analytical solution for HR, we used a numerical approach (optimization) to calculate the HR-scaled thresholds.

In an attempt to achieve a fair comparison, some assumptions needed to be made. Given that this simulation study aims to compare the statistical aspects of the methods in an overall survival setting, ESMO's score 5 was not used because it can only be achieved with additional bonus points adjustments, for example, toxicity improvements. Therefore, ESMO's preliminary scale ranging from 1 to 4 was used, and thus, the maximal scores of both methods are comparable (major added benefit ≈ substantial improvement).

The proportion of simulated trials that achieved maximal score was used as the main metric of comparison. Furthermore, ROC curves were generated comparing different thresholds (ranging

Figure 2. Pairwise Spearman correlation between ESMO's dual rule und IQWiG's additional benefit assessment methods. Illustrated using line charts of the standard scenario (designHR = trueHR) with different underlying median survival times for the control group (6, 12, 18, 24 and 30 months), designHRs (0.3 to 0.9), censoring rates (20% and 60%), and power of 90%. In scenarios with very large treatment effects, only the same score was assigned; therefore, some correlations could not be calculated and, hence, are missing (eg, bottom right panel).



designHR indicates design hazard ratio, used for sample size calculation; trueHR, true underlying hazard ratio for data generation.

from 0.2 to 1) as definition of maximal additional benefit classification for the HR-PE, as well as for HR^- and HR^+ after a statistically significant trial (HR-PE). For each of these thresholds, all simulated subscenarios with designHR ranging from 0.3 to 0.9 were used for calculating the True-Positive Rate (TPR) and False-Positive Rate (FPR). In this context, a true-positive event means that a treatment is deservedly classified as maximal score by the additional benefit assessment method, whereas a false-positive event denotes that a treatment is not deservedly classified as maximal score by the additional benefit assessment method. To calculate TPR and FPR, a ground truth was needed, and since there is no gold-standard method available, a maximal score was assumed to be justified if trueHR $< \delta_{deserved}$ for different cutoffs values of $\delta_{deserved}$ (0.5, 0.6, 0.7, and 0.8). For the assessment of the correspondence between the 2 methods, pairwise Spearman correlation was calculated using the interpretation provided by Mukaka as a third metric for comparison,[11,12] examining the complete range of categories for the 2 methods.

The simulation study was performed using the software R[13] version 4.0.0, with packages "ggplot,"[14] "survival,"[15,16] and "flexsurv"[17] (R-Code and additional information including ADEMP structure proposed by Morris et al[18] available at github.com/cbuesch/IQWiGvsESMO).

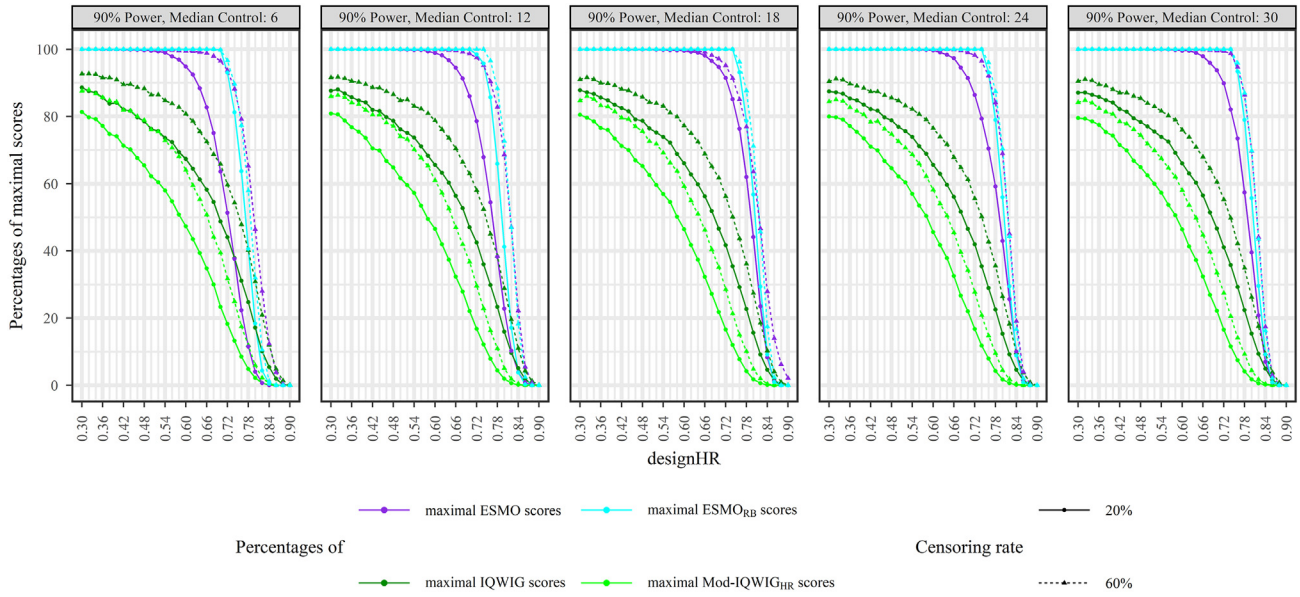
Results

In the standard scenario, ESMO's and IQWiG's methods show low to high positive correlation (Fig. 2). Overall, smaller censoring and decreasing underlying treatment effect (designHR) lead to stronger correlation between ESMO's and IQWiG's method; for example, in case of med_C = 12 months (Fig. 2, second panel), censoring rates of 20% and 60% lead to an on average correlation of 0.30 and 0.20, respectively. Nevertheless, the correlation peaks at "moderate" treatment effects to approximately 0.75, for example, designHR = 0.78, top left panel with 20% censoring rate. Figure 3, which illustrates the proportion of maximal scores of each method including Mod-IQWiG's$_{HR}$, IQWiG's and ESMO's relative benefit rule, and ESMO's full method, explains that, at these values for designHR, the high positive correlation is due to the similar maximal scoring proportions. Furthermore, Figure 3 shows that ESMO's method has a higher proportion of maximal scores (almost all being maximal) for large treatment effects (low values of designHR), which is the reason for the lack of correlation. For smaller treatment effects (large designHRs), IQWiG's method appears to have a higher proportion (Fig. 3, dark-green above violet line; designHR $\approx$ 0.74 or larger in left panel with 20% censoring rate) with an overall maximal scoring proportion being small, and hence, the value of the designHR, where IQWiG's intersects with ESMO (Fig. 3), is the same designHR where the correlation reaches its maximum (Fig. 2). This designHR changes with the underlying med_C of the trial. In other words, the higher the med_C of the study, the smaller the treatment effect has to be for the ESMO method to be more conservative than the IQWiG's method.

The assumed med_C only slightly influences the correlation pattern. The main difference can be seen in med_C equal to 6 months versus the other scenarios (Fig. 2, left panel vs other panels). In this subscenario, the correlation curve tends to substantially increase earlier than in scenarios with $med_C > 6$ months. In cases of large treatment effects, the correlation was not always computable because one of the methods always assigned the same rating (see trueHR < 0.54 top right panel).

Figure 3 further illustrates that, with the implementation of the absolute benefit rule, ESMO's dual rule achieves a reduction, for example, 83.1% to 55.2% in the scenario with designHR of 0.78 (Fig. 3; censoring rate 40%, med_C = 12). Nevertheless, this

Figure 3. Proportion of maximal scores of all significant trials for Mod-IQWiG's$_{HR}$ (light-green) IQWiG's (dark-green), ESMO's relative benefit rule (blue), and ESMO's dual rule (violet). Illustrated using line charts for different subscenarios of the standard scenario 1 (designHR = trueHR) with different underlying median survival times for the control group (6, 12, 18, 24, and 30 months), designHRs (0.3-0.9), censoring rates (20% and 60%), and power of 90%.



designHR indicates design hazard ratio, used for sample size calculation; ESMO, European Society for Medical Oncology; ESMO$_{RB}$, ESMO's method using only the relative benefit rule; IQWiG, Institute for Quality and Efficiency in Health Care; IQWiG's, IQWiG's method using relative risk scaled thresholds; Mod-IQWiG's$_{HR}$, modified IQWiG method using upper confidence interval limit based on IQWiG's thresholds (transformation into HR-scaled thresholds using the conversion formula proposed by VanderWeele[2]); trueHR, true underlying hazard ratio for data generation.

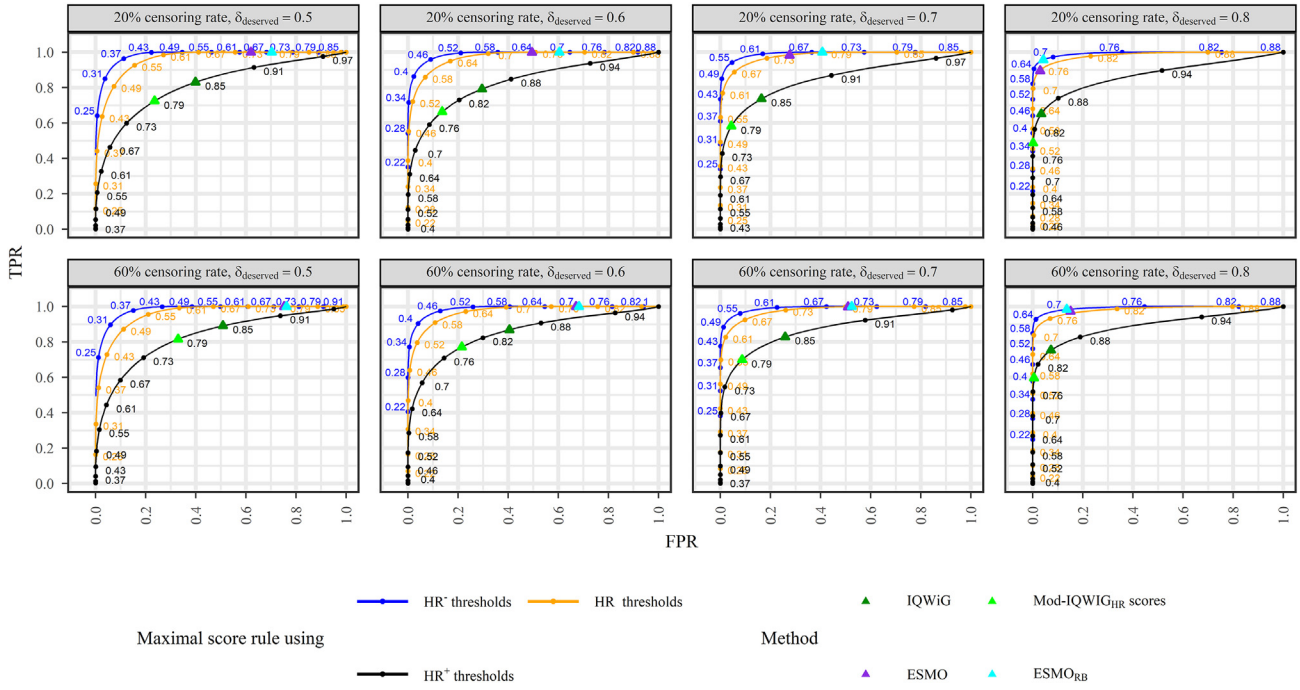
reduction occurring for the absolute benefit rule is less pronounced in scenarios with higher censoring rates (violet above or equal to blue line). IQWiG's rule and Mod-IQWiG's$_{HR}$ rule tend to be more conservative in comparison to ESMO's method in many scenarios. Overall, in scenarios with large effects, ESMO's dual rule shows liberal results as it assigns higher scores than IQWIG's approach. For example, a maximal score rating proportion of 42.5% is given by IQWiG's rule compared with 78.6% of ESMO's dual rule (Fig. 3; designHR = 0.72, censoring rate 20%, med$_C$ = 12). Conversely, Mod-IQWiG's$_{HR}$ is more conservative than all other methods, showing the lowest proportion of maximal scores over all scenarios.

Moreover, only ESMO's method changes with different assumed med$_C$ whereas IQWiG's method does not depend on med$_C$ (Fig. 1). Therefore, IQWiG's and Mod-IQWiG's$_{HR}$ rule stay similar in each row of Figures 2 and 3. Furthermore, in subscenarios with larger simulated censoring rates all methods achieve a higher proportion of maximal scores (Fig. 3, squares above triangles above circles). Nonetheless, ESMO's dual rule does not increase further because it already achieves 100% with small censoring rates.

To answer the question on which fundamental statistical idea (upper vs lower CI limit) might be better for determination of added benefit, ROC curves of the standard scenario were generated. Figure 4 illustrates subscenarios with med$_C$ of 12 months and censoring rate of 20% (first row) and 60% (second row). Focusing only on the CI limits after a statistically significant trial, the lower CI limit (HR^-, blue curve) always lies closer to the perfect classifier (point in the top left corner with coordinates (0,1)) and hence provides a better solution, if appropriate thresholds are chosen. This means that using HR^- with a threshold close to 1 leads to a large TPR as well as FPR and hence is not close to the perfect classifier. Nevertheless, a threshold of 0.43 provides TPR and FPR closer to the perfect classifier over a range of different $\delta_{deserved}$ values (Fig. 4; FPR = 0.2229, TPR = 0.9990 for $\delta_{deserved}$ = 0.5 and 20% censoring; FPR = 0.0002, TPR = 0.7345 for $\delta_{deserved}$ = 0.7 and 20% censoring). Furthermore, even using the HR-PE (yellow curve) provides a better ROC curve than using HR^+ (black curve). If the particular methods by ESMO and IQWiG are compared (triangles), IQWiG's methods always shows a lower FPR and can thus again, as for Figure 3, be interpreted as the more conservative method. ESMO's dual rule also shows very high FPRs, which can be seen as too liberal. Nevertheless, TPR is lower for IQWiG's method than ESMO's dual rule. All other methods (ESMO$_{RB}$, ESMO, IQWiG's) have a higher FPR and hence are more liberal than Mod-IQWiG's$_{HR}$ method. It also can be seen that the results for 20% and 60% censoring rates are very similar. Hence, the censoring rate does not influence the above-described behaviors, if this aspect is included in the sample size calculation.

The results for simulations with incorrect assumed treatment effects for sample size calculation (scenario 2) are illustrated in bar charts of maximal score proportions in Figure 5. The first row shows results for subscenarios with underestimated treatment effects (overpowered trials), the second row shows results for subscenarios with correct assumed treatment effects (standard scenario 1), and the last 2 rows show results for overestimated treatment effects (underpowered trials). It can again be seen that ESMO's dual rule often has a reduced proportion of maximal score compared with the ESMO`s$_{RB}$ rule (violet vs blue). As analogously shown by Figure 3, this is due to the sign of an achieved reduction by the implementation of the absolute benefit rule. In underpowered trials with large treatment effects (designHR < 0.8), Mod-IQWiG's$_{HR}$ and IQWiG's show a larger reduction of the maximal score proportions than ESMO's dual rule compared

Figure 4. ROC curves comparing different thresholds (ranging from 0.2 to 1) as definition of maximal additional benefit classification for the HR point estimate (yellow line) as well as for HR^- (blue line) and HR^+ (black line). For each of these thresholds, all simulated subscenarios with designHR ranging from 0.3 to 0.9 were used for TPR and FPR calculation. A true-positive event means that a treatment is deservedly classified as maximal score by the additional benefit assessment method, whereas a false-positive event denotes that a treatment is not deservedly classified as maximal score by the additional benefit assessment method. A maximal score is assumed to be justified if trueHR $<$ $\delta_{deserved}$ holds for different cutoffs values of $\delta_{deserved}$ (0.5, 0.6, 0.7, and 0.8). In addition, ESMO's dual rule (violet), ESMO's$_{RB}$ rule (blue), Mod- IQWiG's$_{HR}$ (light-green), and IQWIG's (dark-green) method are illustrated with triangles.



designHR indicates design hazard ratio, used for sample size calculation; ESMO, European Society for Medical Oncology; ESMO$_{RB}$, ESMO's method using only the relative benefit rule; HR, hazard ratio; HR^+, estimated upper 95% confidence interval limit of the hazard ratio; HR^-, estimated lower 95% confidence interval limit of the hazard ratio; IQWiG, Institute for Quality and Efficiency in Health Care; IQWiG's, IQWiG's method using hazard ratio scaled thresholds; Mod-IQWiG's$_{HR}$, modified IQWiG method using upper confidence interval limit based on IQWiG's thresholds (transformation into HR-scaled thresholds using the conversion formula proposed by VanderWeele[2]); ROC, receiver operating characteristic; trueHR, true underlying hazard ratio for data generation.

with the standard scenario. For example, in scenarios with designHR = 0.7, HR_{var} = 1.1 and censoring rate of 60%, ESMO's dual rule and IQWiG's exhibit a reduction of 4.1% (98.7%-94.6%) and 31.9% (63.0%-31.1%), respectively. For small treatment effects, all methods achieve similar results. On the contrary, in overpowered trials all methods show an increased maximal score proportion. Especially for IQWiG's rule in subscenarios with small treatment effects (HR_{var} = 0.9 and designHR = 0.9), the highest increased proportion compared to the standard scenario can be observed and is thus the most liberal one, for example, 69.8% (69.9%-0.1%) inflation in the 60% censoring scenarios. Mod-IQWiG's$_{HR}$ rule, in contrast, is the most conservative method in this subscenario with proportions still very close to 0%. In overpowered scenarios with larger treatment effects, the methods achieve rather similar results as ESMO's dual rule already achieves a maximal scoring proportion of 100% in the standard scenario (middle row) and, hence, cannot further increase. The differences between ESMO's and IQWiG's method are due to 2 reasons. First, ESMO's method additionally incorporates the absolute benefit rule, whose reduction impact in case of underpowered studies is not as large anymore as for correctly powered studies. Second, as illustrated by Figure 3, IQWiG's threshold choice for the HR^+ cutoff point for maximal scores is chosen more conservatively than ESMO's threshold choice for the HR^- (violet vs blue). Note that, for a HR_{var} of 0.8, results were barely affected by variation in the other parameters, being usually 100% for the ESMO methods and 80% to 100% for IQWiG's and Mod-IQWiG's$_{HR}$ methods.

The impact of various distributions (scenario 3) and non-proportional hazards / non-constant HRs (scenario 4) are illustrated in Figure 6, which shows the proportion of maximal scores. IQWiG's approach using HR^+ (green lines) shows very similar results for Weibull and Gompertz distribution compared to the standard scenario (left panel). Contrarily, ESMO's reduction due to the absolute benefit rule either drastically decreases the score in cases of increasing hazards (Gompertz, shape of 0.2; Weibull, shape of 1.5) or even slightly increases it in cases of decreasing hazards (Gompertz, shape of −0.2; Weibull, shape of 0.5). Hence, this method is very susceptible to the underlying distribution. In cases with delayed treatment effects and thus non-proportional hazards / non-constant HR, ESMO's method shows only slightly reduced scores compared to proportional hazards. Contrarily, IQWiG's method reduces the maximal scoring proportion by almost 20%, which may be due to combination of penalization for non-proportional hazards and a weaker overall treatment effect in these scenarios.

Furthermore, scenarios with less powered studies, that is, 80%, and censoring rate of 40% show similar results as described earlier.

Discussion

The performed simulation study provides valuable insight into the differences between the benefit assessment scales for time-to-event outcomes of ESMO's and IQWiG's approach. The results for

Figure 5. Proportion of maximal scores of all significant trials for Mod-IQWiG's$_{HR}$ (light-green) IQWiG's (dark-green), ESMO's relative benefit rule (blue), and ESMO's dual rule (violet). Illustrated using bar charts for different subscenarios of scenario 2 (designHR ≠ trueHR) with a median survival time for the control group of 12 months, designHRs (0.3, 0.5, 0.7, and 0.9), censoring rates (20% and 60%) and power of 90%.

designHR indicates design hazard ratio, used for sample size calculation; ESMO, European Society for Medical Oncology; ESMO$_{RB}$, ESMO's method using only the relative benefit rule; HR^+, estimated upper 95% confidence interval limit of the hazard ratio; HR^-, estimated lower 95% confidence interval limit of the hazard ratio; HR_{var}, factor for deviance between designHR and trueHR; IQWiG, Institute for Quality and Efficiency in Health Care; IQWiG's, IQWiG's method using hazard ratio scaled thresholds; Mod-IQWiG's$_{HR}$, modified IQWiG method using upper confidence interval limit based on IQWiG's thresholds (transformation into HR-scaled thresholds using the conversion formula proposed by VanderWeele[2]); ROC, Receiver Operating Characteristic; trueHR, true underlying hazard ratio for data generation (trueHR = designHR·HR_{var}).

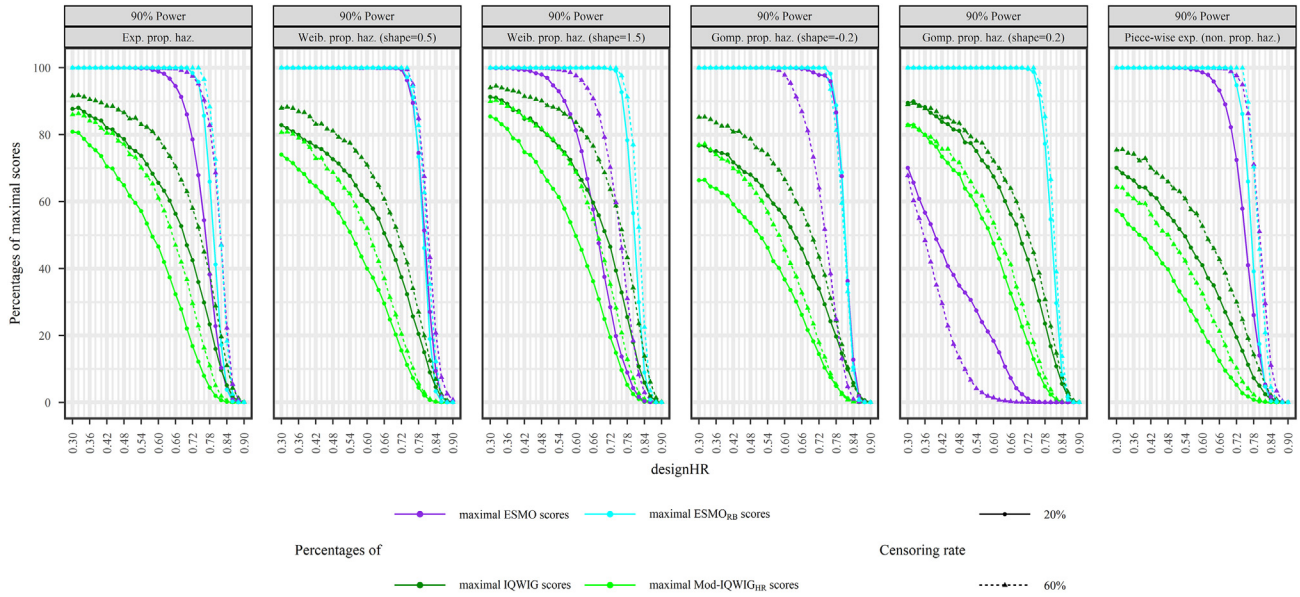
the Spearman correlation clearly depict a low positive relationship between ESMO's dual rule and IQWiG's in most scenarios. Only moderate treatment effects lead to similar results for both methods, whereas results differ otherwise.

Across all simulated scenarios, IQWiG's and Mod-IQWiG's$_{HR}$ rules usually provide a lower proportion of maximal scores, making the methods more conservative than ESMO's dual rule. Furthermore, Mod-IQWiG's$_{HR}$ rule is even more conservative than IQWiG's. Nevertheless, ESMO achieves a downgrading of the relative benefit assessment with the help of its absolute benefit rule, which makes the method less liberal. Particularly in trials with small treatment effects, ESMO's dual rule can even be slightly more conservative than IQWiG's approach. Moreover, the censoring rate strongly influences both methods, assigning larger censoring rates with larger proportions of maximal scores (Fig. 3). This is a consequence of more information being available with a larger censoring rate due to a larger sample size leading to a narrower estimated CI and thus to a larger proportion of maximal scores, as compared to the same scenario with less censoring. Note that, in our simulation, we included the censoring rate in our sample size calculation, leading to differing sample sizes between these 2 scenarios. Accordingly, this resulted in the (on average) same number of events in both scenarios. Thus, a larger censoring rate is accompanied with more patients, which are censored over the course of the simulated study. This means that additional patients belong to the "patients at risk" population. Therefore, comparing these 2 scenarios, the one with a larger censoring rate does have more information because it was included in the sample size calculation.

In addition, the ROC curves (Fig. 4) showed that categorizing the additional benefit based on HR^- provides a better TPR as well as FPR than using HR^+. This is a result of the estimated HR^- increasing faster over the range of simulated HRs (0.3-0.9) compared with the HR^+ estimates. Subsequently, it is easier to find a cutoff value that categorizes the simulated trials into deserving or not deserving a maximal grading. If no sample size calculation is performed and hence a fixed sample size is present, the estimated HR^- and HR^+ would increase similarly over the range of simulated HRs and hence the ROC curves of HR^- and HR^+ would be the same. In real study application, however, a sample size calculation is mandatory due to various reasons, such as ethics, time, and costs. Then again, the choice of the threshold for categorization is very important; otherwise, unacceptably large FPR or low TPR would occur. For example, using the current ESMO threshold for its dual rule, ESMO's dual rule shows very poor FPR, leading to easily achievable maximal scores. Hence, if one uses HR^- instead of HR^+, one can select a threshold to achieve an FPR rate similar to one of the IQWiG methods while simultaneously achieving a higher TPR rate, similar to ESMO's dual rule.

In practice, trials are often planned with an erroneously overestimated or underestimated treatment effect, leading to over- or underpowered trials and resulting in higher or lower percentages of maximal scores for the methods, which could possibly lead to deliberate overpowering to achieve a higher additional benefit grading. In this case, ESMO's dual rule behaves a bit more conservatively than IQWiG's method. Nevertheless, the institutions that perform the assessment for additional benefit should closely monitor this aspect to avoid the exploitation of this

Figure 6. Proportion of maximal scores of all significant trials for Mod-IQWiG's$_{HR}$ (light-green) IQWiG's (dark-green), ESMO's relative benefit rule (blue) and ESMO's dual rule (violet). Illustrated using line charts for different subscenarios of scenario 3 (various failure time distributions) and scenario 4 (delayed treatment effect: non-proportional hazards / non-constant hazard ratio) with a median survival time for the control group of 12 months, designHRs (0.3-0.9), censoring rates (20% and 60%), and power of 90%.



designHR indicates design hazard ratio, used for sample size calculation; ESMO, European Society for Medical Oncology; ESMO$_{RB}$, ESMO's method using only the relative benefit rule; Exp. prop. haz., exponential distributed failure times with proportional hazards (standard scenario 1); Gomp. prop. haz., Gompertz distributed failure times with proportional hazards (scenario 3); IQWiG, Institute for Quality and Efficiency in Health Care; IQWiG's, IQWiG's method using relative risk scaled thresholds; Mod-IQWiG's$_{HR}$, modified IQWiG method using upper confidence interval limit based on IQWiG's thresholds (transformation into HR-scaled thresholds using the conversion formula proposed by VanderWeele[2]); Piece-wise exp. (non. prop. haz.), piece-wise exponential distributed failure times with non-proportional hazards / non-constant hazard ratio leading to a delayed treatment effect (scenario 4); trueHR, true underlying hazard ratio for data generation; Weib. prop. haz., Weibull distributed failure times with proportional hazards (scenario 3).

weakness. A possible different solution might be Mod-IQWiG's$_{HR}$, which reduces the probability of maximal scores to a moderate amount; however, in most other cases, it is very conservative.

Furthermore, the simulation study showed that ESMO's dual rule is susceptible to the underlying failure time distribution. In case of non-exponential distributions such as Gompertz and Weibull, ESMO's absolute benefit rule drastically decreases or increases the proportion of maximal score, which is unfavorable considering proportional hazards were still present in these scenarios, thus making similar results as in the standard exponential case desirable. IQWiG's method is not affected by non-exponential distributions and therefore provides a better solution. In case of non-proportional / non-constant HR, such as trials with delayed treatment effects, both methods have a reduced proportion of maximal scores, which is a desirable conservative behavior. Nevertheless, ESMO's maximal score proportion is only slightly reduced and thus IQWiG's method performs more favorably in this case. Note that, in cases of delayed treatment effect with similar median survival times in both treatment groups, ESMO's method will directly assign only the lowest score (Fig. 1). In our simulation study, only situations with $med_T >> med_C$ have been considered so that ESMO's method is not being punished by its design.

Conclusions

IQWiG's and ESMO's additional benefit rules show a high positive association for moderate treatment effects, assigning similar scoring distributions, yet our research shows that IQWiG's additional benefit assessment method is more conservative than ESMO's dual rule in most scenarios. Especially with various underlying failure time distributions such as Weibull and Gompertz still adhering to the proportional hazard assumption, ESMO's rule tends to be more susceptible leading to an extremely high or low percentage of maximal scores. In contrast, IQWiG's method provides similar results as with exponentially distributed failure times.

The concern that the use of the lower CI limit could lead to a larger probability of higher grades,[5,6] possibly crediting studies with smaller sample size and hence with a wider CI, is partly justified. In scenarios with large treatment effects, ESMO's relative benefit rule, which uses the HR^-, has a higher probability of maximal grades than IQWIG's method, which uses the HR^+. Nevertheless, the higher probabilities also depend on the choice of the thresholds rather than on HR^- and HR^+ alone. Furthermore, our research confirms Dafni et al's[7] statement that the HR^- should be used in preference to the PE. Nevertheless, they also state that no approach can be perfect, which can be supported by our findings; ie, different thresholds need to be chosen for different definitions of justified maximal grading ($\delta_{deserved}$, Fig. 4). Furthermore, our research shows that the PE might be superior compared with the HR^+ and might hence be a valid alternative. Therefore, the assumption that the CI provides more information than using the PE of the HR,[1,3,4,7] as it considers the variability of the HR estimate, cannot be confirmed. Besides the comparison of HR^- and the PE our research also includes HR^+ in the comparison. Hence, our investigation takes a deeper look into the comparison of ESMO's dual rule and IQWiG's method.

Our research has several strengths including an extensive simulation study covering a wide range of censoring rates, failure time distributions, and treatment effects. We investigate which

statistical measure suits the additional benefit assessment better, and thus, our research is a first detailed insight in the comparison of IQWiG's and ESMO's method. Furthermore, we constructed our research reproducibly. Nevertheless, 3 limitations are present as well. First, we only investigated the statistical aspect of the methods and hence did not include bonus point adjustments. Second, we assumed that the maximal scores of the methods can be considered as equal. It can be argued that this assumption is not reasonable and hence our conclusion might not be fair. Nevertheless, our research still features conclusions without this assumption (ROC, correlation), which do support the findings of the results with this assumption. Finally, our censoring mechanism is partly depending on the event times (see Appendix in Supplemental Materials found at https://doi.org/10.1016/j.jval.2022.05.015) and consequently introduces a small bias to the HR estimation. Given that this introduced bias is affecting all compared methods equally, the method comparison described in this article is not substantially affected. In addition, there are still a number of gaps in knowledge around the additional benefit assessment, which should be addressed in future research:

- In case of Gompertz (shape = 0.2) distributed failure times, ESMO's method shows a very different behavior compared with other distributions. This difference might be influenced by ESMO's absolute benefit rule, because the relative benefit rule (Fig. 6, blue line), which uses HR^{-}, is not affected by different underlying failure time distributions. The same can be seen for IQWiG's method, which also only depends on the 95% HR-CI and is not substantially affected by different failure time distributions. Nevertheless, the considerable difference of ESMO's method in case of Gompertz (shape = 0.2) compared with other distributions is still quite astonishing and the reasons behind it are not completely clear.
- Using the HR-PE compared with the HR^{+} for an additional benefit assessment might be a valid alternative if the PE is not biased as, for example, for trials that were stopped early at a preplanned interim analysis. Hence, the additional benefit assessment method developed by the American Society of Clinical Oncology, which uses the HR-PE creating a continuous assessment of a new cancer treatment,[19,20] should be included in future research.
- The threshold of 0.43 for the perfect classifier using the lower CI limit after a significant trial is solely based on the ROC curves (Fig. 4) and hence still needs to be further investigated in different scenarios with various design aspects, for example, sample sizes and effect sizes.
- If the proportional hazards assumption is violated, then the HR-PE, along with its CI, will be strongly influenced by the follow-up time. This aspect is explored partly by scenario 4 in our research but still requires further investigation, especially since the follow-up time depends on the assumed med_C ($FU = 2 \cdot med_C$) in our simulations. Furthermore, in the area of immuno-oncology treatments, this field might be especially of interest because a proportional hazards violation is suspected.

In summary, focusing on the statistical aspect of the methods, ESMO's dual rule tends to be more liberal and more susceptible to various failure time distributions than IQWiG's method. Furthermore, ESMO's poor FPR and the characteristic of not being able to distinguish between treatments with decent effects, leading to generously assigning of maximal scores, need to be taken in consideration in further versions of this method. Nonetheless, when solely regarding the CI limits as decision criterion, HR^{-} seems to be a better solution leading to a higher TPR than HR^{+} without increasing the FPR, if appropriate thresholds are chosen. Thus, IQWiG's method might need to be adapted accordingly to achieve a better overall classification.

Supplemental Materials

Supplementary data associated with this article can be found in the online version at https://doi.org/10.1016/j.jval.2022.05.015.

Article and Author Information

Accepted for Publication: May 11, 2022

Published Online: June 28, 2022

doi: https://doi.org/10.1016/j.jval.2022.05.015

Author Affiliations: Institute of Medical Biometry (IMBI), Heidelberg University, Heidelberg, Germany (Büsch, Krisam, Kieser).

Correspondence: Christopher A. Büsch, MSc, Department of Medical Biometry, Institute of Medical Biometry (IMBI), Heidelberg University, Im Neuenheimer Feld 130.3, D-69120 Heidelberg, Germany. Email: buesch@imbi.uni-heidelberg.de

Author Contributions: *Concept and design:* Büsch, Kieser
Acquisition of data: Büsch
Analysis and interpretation of data: Büsch, Krisam, Kieser
Drafting of the manuscript: Büsch, Krisam, Kieser
Critical revision of the paper for important intellectual content: Büsch, Kieser
Statistical analysis: Büsch, Kieser
Supervision: Krisam, Kieser

Conflict of Interest Disclosures: The authors reported no conflicts of interest.

Funding/Support: The authors received no financial support for this research.

REFERENCES

1. Skipka G, Wieseler B, Kaiser T, et al. Methodological approach to determine minor, considerable, and major treatment effects in the early benefit assessment of new drugs. *Biom J*. 2016;58(1):43–58.
2. VanderWeele TJ. Optimal approximate conversions of odds ratios and hazard ratios to risk ratios. *Biometrics*. 2020;76(3):746–752.
3. Cherny NI, Sullivan R, Dafni U, et al. 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) [published correction appears in *Ann Oncol*. 2017;28(11):2901-2905]. *Ann Oncol*. 2015;26(8):1547–1573.
4. Cherny NI, Dafni U, Bogaerts J, et al. ESMO-Magnitude of Clinical Benefit Scale version 1.1. *Ann Oncol*. 2017;28(10):2340–2366.
5. Muhonen T, Joensuu H, Pfeiffer P. Comment on ESMO magnitude of clinical benefit scale. *Ann Oncol*. 2015;26(12):2504.
6. Wild C, Grössmann N, Bonanno PV, et al. Utilisation of the ESMO-MCBS in practice of HTA. *Ann Oncol*. 2016;27(11):2134–2136.
7. Dafni U, Karlis D, Pedeli X, et al. Detailed statistical assessment of the characteristics of the ESMO Magnitude of Clinical Benefit Scale (ESMO-MCBS) threshold rules. *ESMO Open*. 2017;2(4):e000216.
8. Schoenfeld DA. The asymptotic properties of nonparametric tests for comparing survival distributions. *Biometrika*. 1981;68(1):316–319.
9. Schoenfeld DA. Sample-size formula for the proportional-hazards regression model. *Biometrics*. 1983;39(2):499–503.
10. Kolonko M. Inversionsmethode. In: *Stochastische Simulation: Grundlagen, Algoritmen und Anwendungen*. 1st ed. Wiesbaden, Germany: Springer; 2008:85–97.
11. Mukaka MM. Statistics corner: a guide to appropriate use of correlation coefficient in medical research. *Malawi Med J*. 2012;24(3):69–71.
12. Hinkle DE, Wiersma W, Jurs SG. *Applied Statistics for the Behavioral Science*. 5th ed. Boston, MA: Houghton Mifflin; 2003.
13. R: a language and environment for statistical computing. *The R Project for Statistical Computing*. R Core Team. 2020. https://www.R-project.org/. Accessed June 20, 2022.
14. Wickham H, Navarro D, Pedersen TL. *ggplot2: Elegant Graphics for Data Analysis*. 1st ed. New York, NY: Springer-Verlag; 2016.

15. Therneau T. A package for survival analysis in R. R package version 3.2-11. 2021. https://CRAN.R-project.org/package=survival. Accessed June 20, 2022.
16. Therneau TM, Grambsch PM. *Modeling Survival Data: Extending the Cox Model*. 1st ed. New York, NY: Springer-Verlag; 2000.
17. Jackson CH. flexsurv: a platform for parametric survival modeling in R. *J Stat Softw*. 2016;70:i08.
18. Morris TP, White IR, Crowther MJ. Using simulation studies to evaluate statistical methods. *Stat Med*. 2019;38(11):2074–2102.
19. Schnipper LE, Davidson NE, Wollins DS, et al. American Society of Clinical Oncology statement: a conceptual framework to assess the value of cancer treatment options. *J Clin Oncol*. 2015;33(23):2563–2577.
20. Schnipper LE, Davidson NE, Wollins DS, et al. Updating the American Society of Clinical Oncology value framework: revisions and reflections in response to comments received. *J Clin Oncol*. 2016;34(24):2925–2934.