A comprehensive comparison of additional benefit assessment methods applied by IQWiG and ESMO for time-to-event endpoints after significant phase III trials – A simulation study

Appendix

ADEMP-Strcuture

In the following, the ADEMP structure proposed by Morris et al.[4] of our simulations is outlined, which "provides a structured approach of the planned simulation study involving Aims, Data-generating mechanisms, Estimands, Methods, and Performance measures ("ADEMP"). Furthermore, additional information and the R-Code of the simulation study is available at github.com/cbuesch/IQWiGvsESMO.

Aim:

This simulation study aims to compare the statistical aspects of the additional benefit assessment methods of IQWiG and ESMO in an overall survival setting. Especially the question which fundamental statistical idea (upper vs. lower CI limit) might be better for determination of additional added benefit, is of particular interest.
Further information about these methods itself are provided in the manuscript and the publications of IQWiG's[7] and ESMO's[2, 1] methods.

Data-generating mechanisms:

Since the additional benefit assessment is performed on the basis of conducted and significant phase III trials, these kind of trials have to be simulated to apply the methods of IQWiG and ESMO. Hence, failure time, censoring time, sample size calculation, and the number of iterations for each trial scenario had to be determined / calculated to simulate realistic phase III trials. In the following, the data generation mechanism of the simulations is summarized including the required parameters and choice of distributions.

- Data generation for each trial: To simulate time-to-event data of a two-arm randomized clinical phase III trial, the following algorithm was used:

  1. Set seed.
  2. Generate failure times T with the failure times f_S and f_T for the control/standard and treatment group, n_S and n_T times, respectively.
  3. Generate right-censoring time C for each patient and select the minimum out of C and T for the final observed time-to-event data.

  Therefore, the final data tuple is of the form $(\min(T, C), \mathbb{1}(T \leq C))$, where the first and second entry represent the event time and cause of event (failure or censoring), respectively.

- Seeds of the simulations: To achieve comparability between the different scenarios, the same integer numbers were used as seeds at the beginning of the 10,000 iterations in each

scenario. Therefore, 10,000 integer numbers were once randomly and without replacement drawn out of a sample ranging from 1 to 1 billion.

- Failure time distribution of control and treatment group (f_C and f_T):
  - In case of exponentially distributed failure times, the median overall survival time of the control group (med_C), designHR, and trueHR ($\mathrm{HR} \cdot \mathrm{HR}_{\text{var}}$) were fixed to a specific value to calculate the required λ_C and λ_T of the exponential distribution:

$$f_C \sim \exp(\lambda_C), \qquad f_T \sim \exp(\lambda_T)$$

    1. λ_C was calculated using the assumed med_C:

$$\mathrm{med}_C = \frac{\ln(2)}{\lambda_C} \Rightarrow \lambda_C = \frac{\ln(2)}{\mathrm{med}_C}$$

    2. λ_T was calculated using the trueHR and the proportional hazards assumption:

$$\mathrm{trueHR} = \frac{h_T(t)}{h_C(t)} = \frac{\lambda_T}{\lambda_C} \overset{1.}{\Rightarrow} \lambda_T = \mathrm{trueHR} \cdot \frac{\ln(2)}{\mathrm{med}_C},$$

    where at "1." the calculated conversion of λ_C was used.

    The parameter HR_{var} is needed to illustrate scenarios with incorrect assumed treatment effects for sample size calculation, i.e. designHR $\neq$ trueHR.

  - In case of Weibull distributed failure times, med_C, designHR, and trueHR were fixed to a specific value to calculate the required parameters:

$$f_C \sim \mathrm{weibull}(\lambda_C, k_C),$$

$$f_T \sim \mathrm{weibull}(\lambda_T, k_T)$$

    1. λ_C was calculated using the assumed med_C:

$$\mathrm{med}_C = \frac{(\ln(2))^{1/k_C}}{\lambda_C} \Rightarrow \lambda_C = \frac{(\ln(2))^{1/k_C}}{\mathrm{med}_C}$$

    2. λ_T was calculated using the trueHR and the proportional hazards assumption:

$$\mathrm{trueHR} = \frac{h_T(t)}{h_C(t)} = \frac{\lambda_T^{k_T} \cdot k_T \cdot t^{k_T-1}}{\lambda_C^{k_C} \cdot k_C \cdot t^{k_C-1}} \overset{(*)}{=} \frac{\lambda_T^k}{\lambda_C^k}$$

$$\overset{1.}{\Rightarrow} \lambda_T = \left(\frac{\mathrm{trueHR} \cdot \ln(2)}{\mathrm{med}_C^k} \right)^{\frac{1}{k}},$$

    where at "1." the calculated conversion of λ_C was used and at $(*)$ the shape parameter k was chosen to be identical for both treatment groups to achieve a constant hazard ratio over time ($k_C = k_T$). The parameter HR_{var} was needed to illustrate scenarios with incorrect assumed treatment effects for the sample size calculation, i.e. designHR $\neq$ trueHR.

– In case of Gompertz distributed failure times, $\mathbf{med}_C$, designHR, and trueHR were fixed to a specific value to calculate the required parameters:

$$f_C \sim \mathbf{gompertz}(a_C, b_C),$$

$$f_T \sim \mathbf{gompertz}(a_T, b_T)$$

1. b_C was calculated using $\mathbf{med}_C$:

$$\mathbf{med}_C = \frac{1}{a_C} \cdot \ln\left(1 + \frac{a_C}{b_C} \cdot \ln(2)\right) \Rightarrow b_C = \frac{a_C \cdot \ln(2)}{\exp(\mathbf{med}_C \cdot a_C) - 1}$$

2. b_T was calculated using the trueHR and the proportional hazards assumption:

$$\mathbf{trueHR} = \frac{h_T(t)}{h_C(t)} = \frac{b_T \cdot \exp(a_T \cdot x)}{b_C \cdot \exp(a_C \cdot x)} \overset{(*)}{=} \frac{b_T}{b_C}$$

$$\overset{1.}{\Rightarrow} b_T = \frac{\mathbf{trueHR} \cdot a \cdot \ln(2)}{\exp(\mathbf{med}_C \cdot a) - 1},$$

where at "1." the calculated conversion of λ_C was used and at $(*)$ the shape parameter a was chosen to be same for both treatment groups to achieve a constant hazard ratio over time ($a_C = a_T$). The parameter HR_{var} was needed to illustrate scenarios with incorrect assumed treatment effects for the sample size calculation, i.e. designHR $\neq$ trueHR.

– In the case of piece-wise exponentially distributed failure times with an additional late treatment effect for the treatment group, $\mathbf{med}_C$, designHR, and trueHR were fixed to a specific value. To achieve a late treatment effect for the treatment group, a piece-wise exponential distribution was chosen:

$$\text{F}_C(x) = 1 - \exp(-\lambda_C \cdot x),$$

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

where start_T 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): Assuming an uniform random variable U on the interval [0,1], $X := \text{F}_T^{-1}(U)$ is F_T distributed, meaning $\mathbb{P}(X \leq t) = \text{F}_T, t \in \mathbb{R}$. Therefore, the inversion of the cumulative distribution $\text{F}_T(x)$ is given by

$$\text{F}_T^{-1}(y) = \begin{cases} \frac{\ln(1-y)}{-\lambda_C} & , y \in [0, 1 - \exp(-\lambda_C \cdot \text{start}_T)] \\ \frac{\ln(1-y) + \lambda_C \cdot \text{start}_T}{-\lambda_T} + \text{start}_T & \text{otherwise.} \end{cases}$$

Additionally, λ_C and λ_T were defined in the same way as in the standard exponential case.

- Censoring times ($\mathbf{Cens}$): To simulate a realistic trial, administrative censoring as well as random exponential censoring with a specific censoring rate of $\mathbf{cens\_rate}$ were generated:

  1. Generate uniformly distributed administrative censoring times $\mathbf{Cens}_{\text{AC}}$:
  $$\mathbf{Cens}_{\text{AC}} \sim \mathcal{U}(a) + \text{FU},$$
  where a represents the accrual time and FU the follow-up time.
  2. Generate specific censoring rate times for the remaining simulated events, where the administrative censoring times were not smaller than the simulated event times, so that an overall censoring proportion of $\mathbf{cens\_rate}$ is achieved:
  $$\mathbf{Cens}_{\text{SC}} \sim \exp(\lambda_{\text{cens}}),$$
  where λ_{cens} is calculated for every failure time t separately so that the specific censoring proportion is met:
  $$P(\mathbf{T}_{\text{cens}_{\text{SC}}} \leq t) = 1 - \exp(-\lambda_{\text{cens}} \cdot t) = \mathbf{cens\_rate}_{\text{needed}}$$
  $$\rightarrow \lambda_{\text{cens}} = -\frac{\ln(1 - \mathbf{cens\_rate}_{\text{needed}})}{t},$$
  where $\mathbf{cens\_rate}_{\text{needed}}$ is the specific censoring proportion still needed to achieve an overall censoring proportion of $\mathbf{cens\_rate}$. Hence,
  $$\mathbf{cens\_rate}_{\text{needed}} = \frac{\mathbf{cens\_rate} \cdot (n_{AC} + n_{SC}) - n_{AC} \cdot \mathbf{cens\_rate}_{\text{AC}}}{n_{SC}},$$
  where n_{AC} and $\mathbf{cens\_rate}_{\text{AC}}$ are the sample size and rate of censored patients due to administrative censoring (step 1), respectively. In addition, n_{SC} is the remaining sample size of patients which can be censored by the specific censoring rate in step 2.

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

  1. Calculate the required number of events:
  $$d = \frac{(1+r)^2}{r} \cdot \frac{(\mathbf{z}_{1-\frac{\alpha}{2}} + \mathbf{z}_{1-\beta})^2}{(\ln(\mathbf{designHR}))^2},$$
  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 ($r = n_T/n_C$), designHR is the expected hazard ratio / treatment effect, and $\mathbf{z}_{1-\frac{\alpha}{2}}$ as well as $\mathbf{z}_{1-\beta}$ are the standard normal percentiles.
  2. Calculate the probability of an event $P(D)$ and divide the number of required events d by this probability to get the required sample size N. For the combination of administrative censoring (AC) and a specific censoring proportion (SC), $P(D)$ is calculated the following way:
  $$P_{\text{AC}}(D) = 1 - \frac{1}{6 \cdot (1+r)} \cdot \left[\begin{array}{c} \exp(-\lambda_C) \cdot \text{FU} + r \cdot \exp(-\lambda_T \cdot \text{FU}) \\ +4 \cdot \left(\exp(-\lambda_C \cdot (\frac{a}{2} + \text{FU})) + r \cdot \exp(-\lambda_T \cdot (\frac{a}{2} + \text{FU})) \right) \\ + \exp(-\lambda_C \cdot (a + \text{FU})) + r \cdot \exp(-\lambda_C \cdot (a + \text{FU})), \end{array} \right],$$

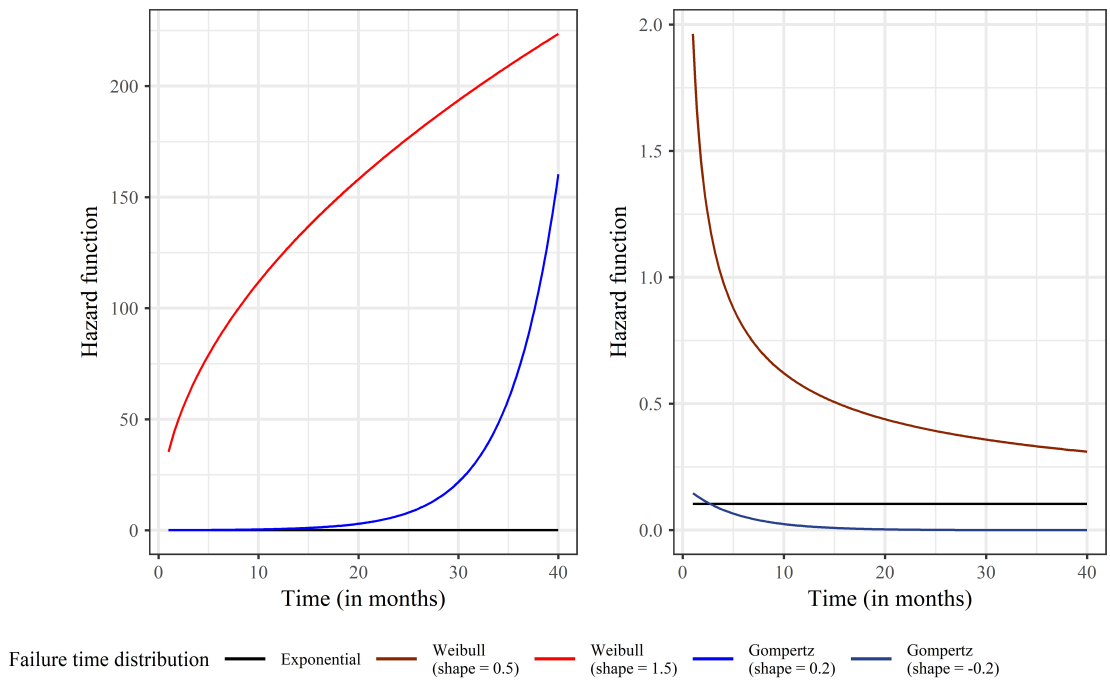
$P_{\text{SC}} = 1 - \text{cens\_rate}$

$$\Rightarrow \text{If } P_{\text{AC}}(D) < P_{\text{SC}}(D) \text{ then } P(D) = P_{\text{AC}}(D)$$

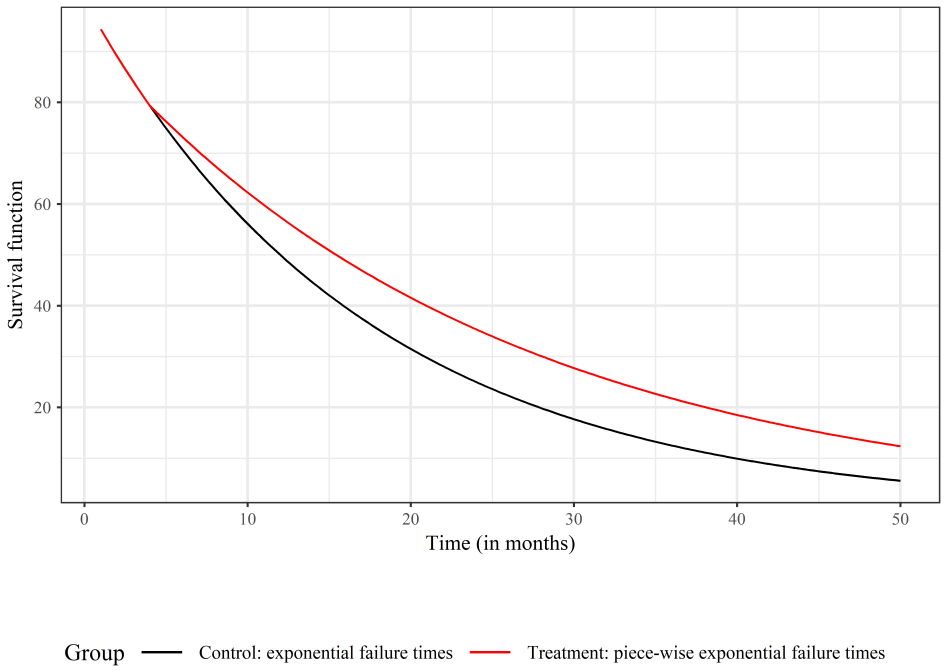
$$\Rightarrow \text{If } P_{\text{AC}}(D) > P_{\text{SC}}(D) \text{ then } P(D) = P_{\text{SC}}(D).$$

- Scenarios / specification of parameters
  An extensive simulation study was performed to provide the aspired detailed overview between the two methods by generating different scenarios of phase III trials:

  1. *Standard Scenario (Scenario 1):* Exponentially distributed failure times using
     - $\text{med}_C \in \{6, 12, 18, 24, 30\}$
     - $\text{designHR} \in \{0.3, 0.32, 0.34, ..., 0.86, 0.88, 0.9\}$
     - $\text{HR}_{\text{var}} = 1$
     - $\beta \in \{0.1, 0.2\}$
     - $\alpha = 0.05$ (two-sided)
     - $r = 1$
     - Combination out of administrative censoring (accrual time of 2 years and a follow-up time of $2 \cdot \text{med}_C$) and exponential censoring so that $\text{cens\_rate} \in \{0.2, 0.4, 0.6\}$ was achieved.
  2. *Incorrect assumed treatment effect (Scenario 2):* To achieve over and under powered studies, the same parameters were used as in Scenario 1, except designHR$\neq$ trueHR was chosen: $\text{HR}_{\text{var}} \in \{0.8, 0.9, 1.1, 1.2\}$.
  3. *Two different parameter distributions (Scenario 3):* Weibull and Gompertz distributions were used instead of exponential distribution as failure time distributions. To achieve proportional hazards for Weibull and Gompertz distributions, the shape parameter of each distribution was fixed to two different values, causing the hazard function to increase/decrease over time. An example using a designHR of 0.9 (designHR=trueHR) and med_C of 6 months can be found in Appendix Figure 1.
  4. *Non-proportional hazards (Scenario 4):* Delayed treatment effect using piece-wise exponential failure time distributions.
     For this objective, the same parameters as in Scenario 1 were used. The underlying distribution of the treatment group was chosen to be exponential using the distribution parameter λ_C of the control group until start_T and λ_T after start_T (delayed treatment effect).
     ESMO's dual rule uses the gain of the new treatment (absolute difference of median treatment outcomes) to establish the different categories. Hence, if $\text{med}_T \approx \text{med}_C$, the method would only assign the lowest score to a new treatment. To not penalize ESMO's method for its construction, start_T was set to $\frac{1}{3}$ of the assumed median survival time of the control group (med_C) for the simulations ($\text{start}_T \ll \text{med}_C$). An example using a designHR of 0.7 (designHR=trueHR), med_C of 12 months and start_T of 4 months can be found in Appendix Figure 2.



Appendix Figure 1: Hazard functions of exponential, Weibull, and Gompertz distribution with assumed parameters for a designHR = 0.9, designHR=trueHR and $\text{med}_C = 6$ months



Appendix Figure 2: Survival functions of piece-wise exponential distribution with late treatment effect, assuming a designHR of 0.7, designHR=trueHR, $\text{med}_C = 12$ months and $\text{start}_T = 4$ months ($\text{start}_T = \frac{1}{3} \cdot \text{med}_C$).

- Software: The simulation was performed using the software R version 4.0.5, with packages "tidyverse", "survival" and "flexsurv" for data generation and analysis. Additionally, the package "ggpubr" was used for illustrations.

Estimands:
Since in this simulation study the aim is not a simple parameter of the data generating model, we are more interested in different targets:

1. Estimation of the maximal scoring rate θ_{max}
2. Selection of best "model" / fundamental statistical idea (upper vs. lower CI limit)
3. Estimation of the correspondence between the method θ_{corr}

Methods:
Since the additional benefit assessment is performed on the basis of conducted and significant phase III trials, each simulated trial was analyzed using a log-rank test and, if significant, the additional benefit assessment methods of IQWiG and ESMO (including dual rule and relative benefit rule ESMO$_{RB}$) were applied. In an attempt to achieve a fair comparison, some assumptions needed to be made. Since this simulation study aims to compare the statistical aspects of the methods, ESMO's score 5 was not used because it can only be achieved with additional bonus points adjustments, e.g. toxicity improvements. Therefore, ESMO's preliminary scale ranging from 1 until 4 was used.
Moreover, to investigate the potential influence of wrongly RR-scaled thresholds of IQWiG's method, the provided RR-scaled thresholds were transformed into HR-scaled thresholds (Mod-IQWiG$_{HR}$) using the conversion formula by VanderWeele[8]:

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

Since this formula has no analytical solution for HR, we used a numerical approach (optimization) to calculate the HR-scaled thresholds.
To apply the above mentioned methods, HR-PEs with corresponding 95% Wald-CIs using Cox regressions, gain and the 2-, 3- and 5-year survival increase were calculated. In rare cases of sub-scenarios 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. Further information of how the methods are calculated is provided in the manuscript, especially Table 1.

Performance measures and n_{sim}:

Performance measures:

1. To estimate the maximal scoring rate of each sub-scenario and each method, the proportion is used:
$$\widehat{\theta_{\text{max}}} = \frac{\text{Number of maximal scores}}{\text{Number of significant trials}},$$
where a maximal score is defined differently for each method:
   - IQWiG and Mod-IQWiG$_{HR}$: "major added benefit"
   - ESMO and ESMO$_{RB}$: 4

2. Selection of best "model" / fundamental statistical idea (upper vs. lower CI limit):
ROC curves were generated comparing different thresholds (ranging from 0.2 to 1) as definition of maximal additional benefit classification for the HR-PE, as well as for HR^- and HR^+. For each of these thresholds, all simulated sub-scenarios with designHR ranging from 0.3 until 0.9 were used for calculating the True Positive Rate (TPR) and False Positive Rate (FPR). In this context, a true positive and false positive event means that a treatment is deservedly classified as maximal score, or, respectively, is not deservedly classified as maximal score. 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 cut-offs values of $\delta_{deserved}$ (0.5, 0.6, 0.7, and 0.8).
Example for TPR and FPR calculation:
Lets assume a sub-scenario of Standard Scenario 1 with designHR = 0.80, 90% power, censoring rate of 20%, and $\text{med}_C = 12$ months as well as a method using solely the upper CI limits with threshold of 0.85 as decision criterion for a maximal score (hence, if a significant trial has a upper CI of smaller or equal to 0.85 a maximal score is given - as in IQWiG's method). In this specific case, IQWiG's method scored 1440 out of 9012 significant trails the maximal grading. If we further assume a $\delta_{deserved} = 0.7$ we would have 1440 false positives (FP), 7572 true negatives (TN) as well as no true positives (TP) and false negatives (FN) because the designHR is larger than $\delta_{deserved}$. The cross table below illustrates these numbers:

		Deserved maximal score ($\delta_{deserved} = 0.7$)	
		yes	no
Achieved maximal score with threshold of 0.85	yes	0	1440
	no	0	7572

Since no TP and FN are present, TPR cannot be calculated. Thus, we used all designHRs ranging from 0.3 until 0.9, leading to non-zero TP and FN cases and thus in this sub-scenario to the following cross table:

		Deserved maximal score ($\delta_{deserved} = 0.7$)	
		yes	no
Achieved maximal score with threshold of 0.85	yes	139911	75571
	no	49769	14824

Hence, FPR and TPR can now be estimated:

$$\widehat{\text{FPR}} = \frac{\text{FP}}{\text{TN} + \text{FP}} = \frac{14824}{75571 + 14824} = 0.1640$$

$$\widehat{\text{TPR}} = \frac{\text{TP}}{\text{TP} + \text{FN}} = \frac{139911}{139911 + 49769} = 0.7376$$

Another reason why this approach makes sense is that the additional benefit assessment method should be used for all new drugs including ones with large and small treatment effects.

3. To estimate the correspondence between the methods pairwise Spearman correlation was calculated examining the complete range of categories for the two methods:

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

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

Number of iterations for each scenario (n_{sim}):
In each simulated scenario, it was sought to achieve a standard deviation of 0.25% for the coverage probability of a maximal ESMO grade or maximal IQWiG grade assuming constant error variance. For each iteration of a scenario, the confidence interval either covers the true value or it does not. Thus, an indicator variable was defined:

$$Y_i = \begin{cases} 1, & \text{if the maximal grade is given,} \\ 0, & \text{if it is not given.} \end{cases}$$

Therefore, Y_i is a Bernoulli variable and the coverage probability $\mathbb{E}(Y_i) = p$ can be estimated by the sample proportion. Since the variance of a Bernoulli variable is given by $p \cdot (1-p)$ and the simulations generated independent and identically distributed Bernoulli variables, the variance of the simulation-based estimate of p is $\frac{p \cdot (1-p)}{n_{sim}}$, where n_{sim} is the number of iterations. In fact, it can be shown that

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

Therefore, to achieve a variance less than some pre-specified threshold δ, n_{sim} can be calculated using

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

Setting the threshold δ to $2.5 \cdot 10^{-5}$, which corresponds to a standard deviation of 0.5%, results in a number of $n_{sim} = 10,000$ iterations for each scenario.

References

[1] N. I. Cherny, U. Dafni, J. Bogaerts, et al. Esmo-magnitude of clinical benefit scale version 1.1. *Annals of Oncology*, (28):2340–2366, 2017.

[2] N. I. Cherny, R. Sullivan, U. Dafni, 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). *Annals of Oncology*, (26):1547–1573, 2015.

[3] M. Kolonko. *Stochastische Simulation: Grundlagen, Algoritmen und Anwendungen*. Springer: Vieweg+Teubner, GWV Fachverlage GmbH, Wiesbaden 2008, 1 edition, 2008.

[4] T.P. Morris, I.R. White, and M.J. Crowther. Using simulation studies to evaluate statistical methods. *Statistics in Medicine*, (38):2074–2102, 2019.

[5] D.A. Schoenfeld. The asymptotic properties of nonparametric tests for comparing survival distributions. *Biometrika*, (68):316–319, 1981.

[6] D.A. Schoenfeld. Sample-size formula for the proportional-hazards regression model. *Biometrics*, (39):499–503, 1983.

[7] G. Skipka, B. Wieseler, T. Kaiser, et al. Methodological approach to determine minor, considerable, and major treatment effects in the early benefit assessment of new drugs. *Biometrical Journal*, (58):43–58, 2016.

[8] T. J. VanderWeele. Optimal approximate conversions of odds ratios and hazard ratios to risk ratios. *Biometrics Biometric Methodology*, (76):746–752, 2020.