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

Data generation (Appendix)

Simulation design

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 simulation design 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.

- Number of iterations for each scenario: 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.

- 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 (HR $\cdot$ HR_{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:

    $$\text{med}_C \overset{!}{=} \frac{\ln(2)}{\lambda_C} \Rightarrow \lambda_C = \frac{\ln(2)}{\text{med}_C}$$

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

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

    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:

    $$f_C \sim \text{weibull}(\lambda_C, k_C),$$
    $$f_T \sim \text{weibull}(\lambda_T, k_T)$$

    1. λ_C was calculated using the assumed med_C:

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

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

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

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

    where 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 <u>Gompertz</u> distributed failure times, med_C, designHR, and trueHR were fixed to a specific value to calculate the required parameters:

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

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

  1. b_C was calculated using med_C:

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

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

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

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

  where 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 <u>piece-wise exponentially</u> distributed failure times with an additional late treatment effect for the treatment group, 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 probability 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.

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

  1. Generate uniformly distributed administrative censoring times $\mathsf{Cens}_{\mathrm{AC}}$:

  $$\mathsf{Cens}_{\mathrm{AC}} \sim \mathcal{U}(a) + \mathrm{FU},$$

  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 $\mathsf{cens\_rate}$ is achieved:

  $$\mathsf{Cens}_{\mathrm{SC}} \sim \exp(\lambda_{\mathrm{cens}}),$$

  where λ_{cens} is calculated for every failure time t separately so that the specific censoring proportion is met:

  $$P(\mathrm{T}_{\mathrm{cens}_{\mathrm{SC}}} \leq t) = 1 - \exp(-\lambda_{\mathrm{cens}} \cdot t) \overset{\mathrm{def}}{=} \mathsf{cens\_rate}_{\mathrm{needed}}$$
  $$\rightarrow \lambda_{\mathrm{cens}} = -\frac{\ln(1 - \mathsf{cens\_rate}_{\mathrm{needed}})}{t},$$

  where $\mathsf{cens\_rate}_{\mathrm{needed}}$ is the specific censoring proportion still needed to achieve an overall censoring proportion of $\mathsf{cens\_rate}$. Hence,

  $$\mathsf{cens\_rate}_{\mathrm{needed}} = \frac{\mathsf{cens\_rate} \cdot (n_{AC} + n_{SC}) - n_{AC} \cdot \mathsf{cens\_rate}_{\mathrm{AC}}}{n_{SC}},$$

  where n_{AC} and $\mathsf{cens\_rate}_{\mathrm{AC}}$ 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.

- <u>Sample size calculations</u> were performed with the approach of Schoenfeld (1981, 1983):

  1. Calculate the required number of events:

  $$d = \frac{(1+r)^2}{r} \cdot \frac{(\mathrm{z}_{1-\frac{\alpha}{2}} + \mathrm{z}_{1-\beta})^2}{(\ln(\mathsf{designHR}))^2},$$

  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 $\mathrm{z}_{1-\frac{\alpha}{2}}$ as well as $\mathrm{z}_{1-\beta}$ 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:

  $$P_{\mathrm{AC}}(D) = 1 - \frac{1}{6 \cdot (1+r)} \cdot \left[\begin{array}{c} \exp(-\lambda_C) \cdot \mathrm{FU} + r \cdot \exp(-\lambda_T \cdot \mathrm{FU}) \\ +4 \cdot \Big(\exp(-\lambda_C \cdot (\frac{a}{2} + \mathrm{FU})) + r \cdot \exp(-\lambda_T \cdot (\frac{a}{2} + \mathrm{FU})) \Big) \\ + \exp(-\lambda_C \cdot (a + \mathrm{FU})) + r \cdot \exp(-\lambda_C \cdot (a + \mathrm{FU})), \end{array} \right],$$

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

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

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

- Software: The simulation was performed using the software R version 4.0.5, with packages “survival” and “flexsurv” for data generation.

Scenarios / specification of parameters

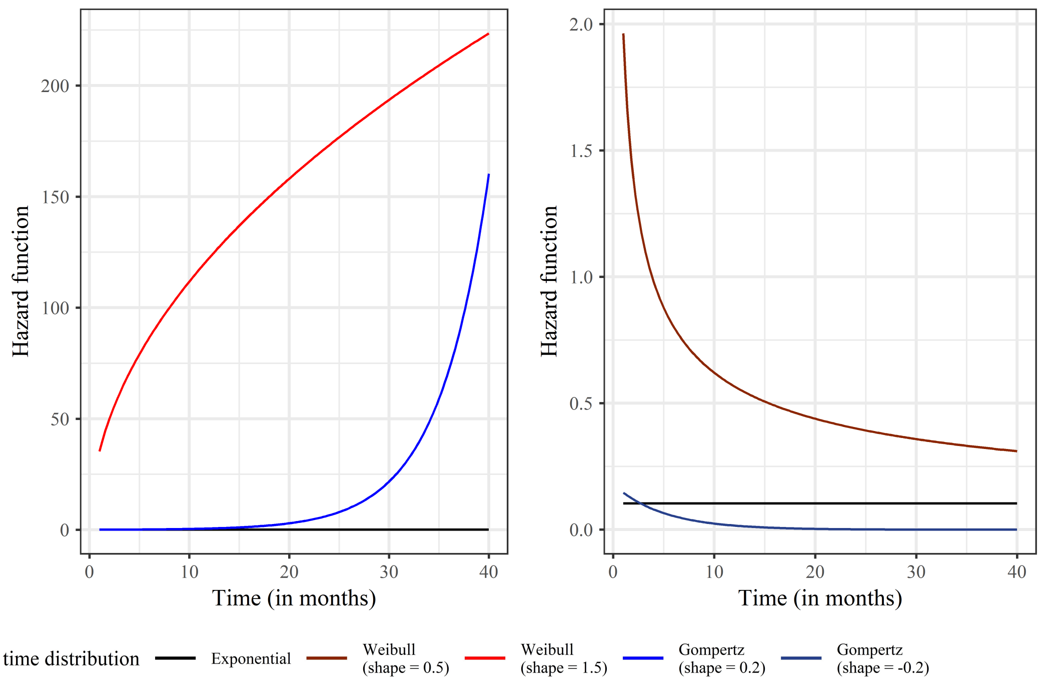
An extensive simulation study was performed to provide the aspired detailed overview between the two methods by generating different scenarios of phase III trials:

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.

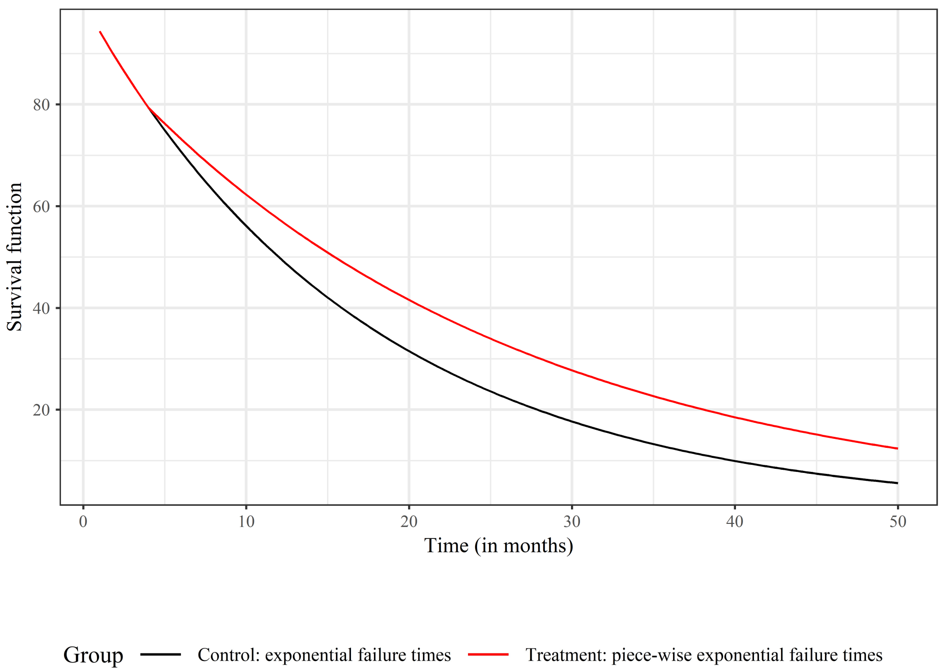
2. *Incorrect assumed treatment effect (Scenario 2):* To achieve over and under powered studies, the same parameters were used as in Scenario 1, except designHR$\neq$ trueHR was chosen: $\text{HR}_{\text{var}} \in \{0.8, 0.9, 1.1, 1.2\}$.

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 med$_T \approx$ 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 (start$_T \ll$ 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 ($\text{start}_T = \frac{1}{3} \cdot \text{med}_C$).

References

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

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

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