PROTOCOL: A SIMULATION STUDY TO EVALUATE THE PERFORMANCE CHARACTERISTICS OF STATISTICAL METHODS FOR THE ANALYSIS OF TIME-TO-EVENT DATA UNDER NON-PROPORTIONAL HAZARDS

CONFIRMS Consortium

Revision 2 - 2022-10-14

INTRODUCTION

While well-established methods for time-to-event data are available when the proportional hazards assumption holds, there is no consensus on the best approach under non-proportional hazards (NPH). However, a wide range of parametric and non-parametric methods for testing and estimation in this scenario have been proposed.

We have performed a systematic review of the scientific literature to identify available options for methods for testing and estimation under NPH. In CONFIRMS (2022a) we performed a systematic literature search for methodological approaches for any NPH scenario, any model class and not restricted to a specific disease area. Our review complements previous review articles that have focused mostly on quantitative comparisons for specific NPH scenarios, e.g. Li et al. (2015), for a specific method class, e.g. Rauch et al. (2018), or for NPH situations in specific disease areas, e.g. oncology (Ananthakrishnan et al. (2021)).

We identified a large number of articles that compare newly proposed methods to alternatives using simulation studies. However, we suspect that the simulation assumptions could have been chosen in order to demonstrate superiority of the new method (Boulesteix et al. (2020)). In particular, we identified only a few review articles that compared a broader set of methods using simulation studies without the objective to demonstrate the advantages of a newly proposed method. Although these articles cover a broad range of NPH settings they were limited to testing procedures and guidance on the choice of the test is not consistent between studies. We did not identify any reviews that provide a neutral comparison of methods with respect to effect estimation. Consequently, we will perform a comprehensive simulation study to evaluate the performance characteristics for a selection of analysis methods identified in our review, including methods for testing and estimation.

In addition we performed a review of past marketing authorization procedures, where NPH were discussed in the EMA European Assessment Report (EPAR) (CONFIRMS (2022b)). We identified 16 procedures reporting results from 18 distinct trials. The majority of the corresponding treatments are from the oncology domain with exceptions from influenza (human), multiple sclerosis, conscious sedation, and cardiovascular disease. We managed to extract a wealth of results from the included studies. This includes: sample sizes, median survival times, and hazard ratios. In addition, we digitised survival curves, allowing us to reconstruct individual patient data close to the actual observations from specific examples. Consequently, we obtained a wide range of data from actual trials that will form the basis for the assumptions underlying the distributional scenarios for this simulation study. Especially, reconstructed individual patient level data will be useful to construct case studies to illustrate eventual findings, based on re-analyses of the data and resampling from the extracted survival and censoring distribution.

The various types of NPH issues identified and the related discussions in the identified EPARs will be useful to derive regulatory recommendations on how to address the corresponding implications in future marketing authorization procedures.

The simulation study is planned following the clinical scenario evaluation approach by Benda et al. (2010) and Friede et al. (2010), as well as recommendations provided in Morris et al. (2019). The simulation study will cover a range of different assumptions regarding the underlying statistical model and parameters, investigate different clinical trial design options and compare the performance characteristics of a broad set of statistical methods for the analysis of time to event data under NPH.

Objectives

The objective of this simulation study is to evaluate the operating characteristics of selected statistical methods for testing and estimation in clinical trials with time-to-event endpoints under NPH, considering a wide range of plausible distributional assumptions and for a number of typical design options, and thereby addresses Objective 2 (see Main Study Protocol - Section 1.3) of the overarching research project.

In a next step, we will illustrate and complement the results from the simulation study using case studies, for which individual level patient data are reconstructed and re-analysed using potentially also additional methods not considered in this simulation study. This will form the basis to address the remaining Objective 3 which aims to "assess the regulatory acceptability of these methods for clinical trials that are pivotal for drug development and benefit-risk assessment and derive recommendations" (Main Study Protocol - Section 1.3).

In addition to facilitate the performance of the simulation study we will implement a software package to facilitate simulation of time-to-event data, which will be developed and published as an open-source software package using the statistical programming language R (R Core Team (2022)),

Clinical Scenario Evaluation Framework

The clinical scenario evaluation (CSE) framework includes three key elements:

1. **Assumptions** define the distributional scenarios and corresponding data generating process and represent external conditions that are not under the control of the experimenter. The assumptions are sometimes also referred to as disease-specific features; some of which can be estimated from previous studies and some might need to be assumed.
2. **Options** define the design choices, analysis methods as well as other aspects that are under the control of the experimenter although some options might be constrained by the infrastructure, resources or the healthcare environment
3. **Metrics** define the set of statistics quantifying the operating characteristic of a specific option, which are to be computed using simulation.

The combinations of assumptions and options define scenarios, for which the metrics are to be evaluated in the simulation study. For the present study the assumptions define specific distributions with NPH (e.g. delayed treatment effects). Here the assumptions are selected based on clinical relevance, taking into account findings from the EPAR Review (CONFIRMS 2022b). Options mainly represent different clinical trial designs (e.g. fixed sample design), and different analysis methods (e.g. weighted log-rank tests). The metrics are meant to reflect in particular regulatory considerations on the validity of the statistical procedure including bias, type I error rate and coverage probabilities of confidence intervals.

Assumptions

The general approach of this simulation study will be to simulate individual participant data (IPD) encompassing recruitment times, event times, censoring times, as well as additional features such as progression-, or study-withdrawal times, and biomarker status, where applicable.

Simulation of data will follow three approaches:

1. For the main part we assume piecewise constant hazards with potentially different hazard rates in different treatment arms, over certain time periods, following specific events or according to biomarker status (see Section *Piecewise Constant Hazard* below).

2. More sophisticated parametric, pharmacometric models (see Section *Joint Models* below) may be used to generate specific scenarios of interest.

3. In addition, we will reconstruct individual participant data from (IPD) Kaplan Meier plots published in selected EPARs to sample from examples of actual trials (see *Bootstrap from IPD* below).

Further details on how the three approaches will be implemented and parameter values will be chosen to achieve survival characteristics (e.g. crossing survival curves, powers within sample size range) of selected broad assumption sets are given in the corresponding sections below.

With this simulation study we intend to explore the following broad assumptions sets:

- Proportional hazards, as a reference scenario,
- Late separation of survival curves,
- Crossing hazard curves
- Differential treatment effect in a subgroup (subgroups will be assumed to be known or unknown)
- Change of treatment effect following an intercurrent event (e.g. treatment switching)

Parameter ranges will be derived so that they match the broad assumption sets above and cover the characteristics of the studies observed in the EPAR review. Table 1 provides a list of parameters and corresponding settings that will be explored for each scenario. Parameter settings are formulated in terms of broad overall features of the resulting survival distributions to match characteristics of relevant clinical scenarios taking into account findings from the EPAR review (CONFIRMS (2022b)).

The parameters of the data generating process under the considered approaches (e.g. piecewise constant hazards) will be calibrated to match these overall features. For certain scenarios, e.g. crossing hazard curves, where the hazard ratio before and after time of onset need to be fixed, additional degrees of freedom in terms of parameter settings may remain and will be chosen based on plausible assumptions, taking the results from the EPAR review into account. Reasonable simplifications will be made that cover the relevant parts of the parameter space. The general approach on how this will be achieved is described in Column *Implementation* of Table 1 with additional details provided in the Section *Piecewise Constant Hazards*.

Not all scenarios will be considered for *Joint Models* and *Bootstrap from IPD*. The purpose of the *Joint Model* will be to generate scenarios where non-proportional hazards are caused by more complex mechanisms (e.g. delayed onset of treatment effect and change of treatment effect after disease progression, e.g., due to treatment switching). The purpose of *Bootstrap from IPD* will be to sample from survival distributions close to actual trial data and evaluate and compare different methods in such a setting.

While parameter ranges are chosen based on results from the EPAR Review (CONFIRMS 2022b), in order to reflect realistic settings of time-to-event data under NPH, certain combinations of different parameter settings may lead to implausible or irrelevant scenarios. Similarly, it is not guaranteed that the resulting parameter space covers all relevant settings. Should we identify corresponding scenarios e.g. during calibration, or software testing, additional scenarios may be added or removed. Any corresponding changes to the overall assumption set will be documented and justified in the study report.

Table 1: Assumption set describing the distributional scenarios considered for the simulation study.

Assumption	Parameters	Ranges	Implementation
Control arm hazards	hazard function	mild, moderate, aggressive	Will be calibrated to achieve plausible ranges with median time-to-event of 36 months (mild), 12 months (moderate) and 6 months (aggressive) as observed in the EPAR review (CONFIRMS (2022b) Fig. 2).
Effect sizes	depends on scenario and estimand	no, small, moderate, large	Parameters governing the difference in survival curves will be set to provide about 50% (small), 80% (moderate), and 90% (large) power under the reference scenario (PH) with log-rank test. For scenarios under NPH parameters will be chosen to yield identical median survival times as in the reference (PH) scenario.
Non-informative study-withdrawal	proportion of study-withdrawal	no study-withdrawal, medium, or substantial study-withdrawal	Study withdrawal will be simulated in addition to administrative censoring at study end using exponential distribution. Corresponding rate parameters will be calibrated to yield ~10% or ~30% of subjects with event times censored due to study withdrawal (medium, substantial study withdrawal).
Delayed onset of treatment effect	Time of onset	1 to 9 months	Corresponding parameters will be calibrated to yield survival functions that separate at time-points compatible with ranges observed in the EPAR review (CONFIRMS (2022b), Table 7)
	HR after onset	According to effect size	See Effect Sizes above.
	Time of onset	1 to 9 months	Corresponding parameters will be calibrated to yield survival functions that cross at time-points compatible with ranges observed in the EPAR review (CONFIRMS (2022b), Table 7)
Crossing hazard functions	HR before crossing	1.5 to 3	This is based on data from 3 studies included in CONFIRMS (2022b) - Supplement 1, where corresponding estimates were reported.
	HR after crossing	According to effect size (see above)	See Effect Sizes above.
Changing hazards following intercurrent event	progression rate	10%/10%, 10%/20%, 20%/20%	Corresponding rate parameters will be calibrated to e.g. yield about 10% progressors in the treatment and 20% in the control group (10%/20%). Similar to the Teriflunomide example, below.

Assumption	Parameters	Ranges	Implementation
	hazard after progression	mild, aggressive	Corresponding rate parameters will be calibrated to reduce median survival times by 80% (mild) and 50% (aggressive) depending on the control group hazard. HR will be assumed to be 1 after progression. Settings (esp. mild) also covers treatment switching (see also Estimands below).
Biomarker subgroups	Subgroup prevalence	10% to 50%	Prevalence of biomarker positive subjects will be set to match corresponding proportions observed in the EPAR Review (CONFIRMS (2022b) - Figure 3)
	HR in subgroup relative to overall population	90%-70%	Corresponding, parameters will be calibrated to yield hazard ratios for the biomarker positive group in the range of .9 to .7 relative to the overall hazard ratio to match examples observed in the EPAR Review (CONFIRMS (2022b) - Figure 3)

Piecewise Constant Hazards

Piecewise constant hazards are a simple yet flexible way to model time to event data with non-proportional hazards, where cumulative hazards and survival functions can be expressed explicitly. For each subject the hazard to experience the event of interest is assumed to be constant over predefined periods of time. The hazard may differ between subjects in different treatment groups, or biomarker subgroups. Hazards may change to a different fixed value at a given time-point (e.g. at the time of treatment effect onset). Assuming that hazards change after a random time (which is itself defined via a piecewise constant hazard), the impact of disease progression on the hazard can be modelled. With mixtures of distributions from those models, differential treatment effects in subgroups and treatment switching after disease progression can be modelled. Derivation of the formulas and algorithms are given in Ristl et. al 2020 who also provide an R implementation.

IPD with parameters calibrated to match assumptions provided in Table 1 above will be modelled as follows:

Delayed onset of treatment effect will be modelled by a constant baseline hazard in the control group and a piecewise constant hazard in the treatment group. The hazard in the treatment group is the same as in the control group in the first time interval and smaller in the second time interval.

Crossing hazard curves will be modelled by a constant baseline hazard in the control group and a piecewise constant hazard in the treatment group. The hazard in the treatment group is larger than the hazard in the control group in the first time interval and smaller than the hazard in the control group in the second time interval.

Changing hazards following an intercurrent event (e.g. treatment discontinuation or switching) will be modelled with a (potentially different) hazard for the event of interest in both arms and a (potentially different) hazard for time to intercurrent event in both arms. The hazard for the event of interest following the intercurrent event is the same in both groups and larger than the hazard before the event. This model covers intercurrent events like treatment discontinuation or switching (e.g. caused by a disease progression). See also Section *Estimands* below for further Details..

Biomarker subgroups will be simulated with (potentially different) constant hazards in biomarker negative subjects in the control and treatment group and different hazards for biomarker positive subjects, again (potentially) different between the treatment and control group. Biomarker status will be sampled for each subject from a binomial distribution using frequencies matching the range of prevalence given in Table 1.

Additionally, non-informative study-withdrawal will be modelled using the same constant hazard in both treatment groups. In addition, due to different options concerning recruitment speed and sample size, different proportions of administrative censoring at study end will be obtained (see also subsection *Design* below). Note that since either event-time is independent of treatment and covariates (e.g. biomarker subgroup), such that an analysis that assumes non-informative censoring will be appropriate under the resulting independent censoring. Only intercurrent events may result in informative censoring which will be addressed as part of the estimand strategy (see Section *Estimand*, below).

To match event-time distributions of scenarios under NPH to the reference scenario under proportional hazards, first parameters defining the latter (i.e. PH) will be chosen to provide a given power to the log-rank test with the particular design options (e.g. sample size, recruitment speed). In a second step, the parameters of the hazard functions of a corresponding scenario under NPH will be chosen to yield identical median survival times in each treatment group.

In summary, the simulated individual participant data will consist of study arm, time of event or censoring status. Additionally time of intercurrent event and biomarker subgroup will be simulated for the respective scenarios. Please see also Section *Simulation Parameters*, below, for an exhaustive list of the resulting simulation scenarios.

Joint Models

"Semi-mechanistic" models from the pharmacometric literature (population PKPD models, also sometimes referred to as joint models) will be used to combine some of the scenarios described above to simulate studies with individual patient data and different sample sizes. In the context of PKPD models, parametric survival models with time varying exposure metrics like drug exposure (PK) as factors in the hazard can often result in non-proportional hazards between treatment groups. The parametric models describe base hazard functions that may change over time (e.g., Weibull or Gompertz hazard functions) and may use other predictors such as drug PK models, other disease status models and/or patient covariates that may change over time to drive changes in the hazard functions (see Holford (2013)). One potential case study be used as a basis for simulation of trial data to be analysed with the analysis methods developed and investigated in this work could be the use of Avelumab in metastatic merkel cell and advanced urothelial carcinoma, which has been shown to have time changing clearance (Wilkins et al. 2019). Population PKPD simulations of exposure-safety and exposure-efficacy (both time-to-event PD endpoints) of Avelumab may result in non-proportional hazards in various design scenarios (Novakovic et al. 2020).

A second potential case study for simulation uses a multistate model, similar to the piecewise constant hazards approach described above, to describe overall survival in HER2-negative breast cancer patients treated with docetaxel (Krishnan et al. (2021)). However, instead of piecewise constant hazards, the hazards are generally continuous and time changing, with some base hazards described by a time-dependent Weibull function, and some hazards having time-changing functions based on other disease models such as tumour size, which is driven by a PK model for the drug. The example includes disease progression, treatment switching after disease progression, and differential treatment effects. After appropriate choice of the model parameters crossing hazards and/or late hazard separation could be included.

In any simulation of data, such as from the case studies described above, the original models will be used as starting points, where model parameters and structures will be adjusted to simulate data within the limits described in Table 1.

BOOTSTRAP FROM IPD

Additionally data will be generated by resampling from the digitised survival curves from the case studies. Data from Kaplan Meier plots will be digitised interactively, by entering the numbers at risk for different timepoints (where available) and identifying axes and survival curves in plots by mouse clicks. Individual

patient data (event times and censoring status) will be reconstructed with a method described in Guyot et al. (2012). In this way both the empirical survival and censoring distributions from the actual data can be estimated.

From this reconstructed individual participant data, bootstrap samples can be drawn to estimate standard errors of the metrics (e.g. RMST, rejection probability for different test procedures, …). This is equivalent to sampling from the estimated empirical distribution function (Efron (1981)) and computationally more efficient than direct sampling e.g. via inverse transformation. Nevertheless, bootstrapping is computationally less involved. Bias will be estimated in terms of deviation of estimates from the empirical distribution function. Type I error rates will be explored by sampling from the control group only. Because, the control group survival function - reconstructed from published data - represents only one particular example of a specific underlying hazard function and sampling both treatment arms from the same treatment group would necessarily imply proportional (even equal) hazards, the latter may be of interest for analysis methods that rely on parametric assumptions about the hazard function.

As the resulting scenarios refer to actual trial examples. Relevant estimands will be defined for each scenario (see also Section *Estimand*). Consequently, only methods capable of estimating the corresponding estimand will be considered for simulation.

It should be stressed that resulting estimates (e.g. of bias and variance) are with respect to the estimated empirical distribution functions and not the unknown distribution functions from which the corresponding trial data can be considered realisations. Nevertheless, one may still consider the estimated distribution functions relevant instances of realistic data generating processes with NPH.

We plan to include the following case studies corresponding to different sources of non-proportional hazards:

- PEMBROLIZUMAB, EMEA/H/C/003820/II/0065

  Active Substance: Pembrolizumab

  Invented Name: Keytruda

  Procedure Type: Type II Variation - Extension of indication to include, as monotherapy or in combination with Chemotherapy, first-line treatment of recurrent or metastatic head and neck squamous cell carcinoma in adults

  Trial: Keynote-048: Sample Size 882 (301 pembro mono, 281 pembro combo)

  Primary Analysis: OS, Stratified log-rank test, CoxPH

  HR Estimate ITT combo: 0.72, (p=0.00025)

  NPH issues: delayed treatment effect, differential treatment effect (PD L1 CPS)

  Model Diagnostics: Schoenfeld Residuals, log(-log(Survival)), treatment*time interaction

  Subgroup effect (combo) HR: CPS > 1: HR=0.65; CPS > 20: HR=0.60

- NIVOLUMAB EMEA/H/C/003985/II/0080

  Active Substance: Nivolumab

  Invented Name: Opdivo

  Procedure Type: Type II Variation - Extension of Indication for Nivolumab for the treatment of Oesophageal Squamous Cell Carcinoma (OSCC)

Trial ONO-4538-24: Sample Size 419 (210 in treatment group)

Primary Analysis: OS, Stratified log-rank test

HR Estimate ITT: 0.77 (p=0.189)

NPH issues: anticipated delayed treatment effect (immunotherapy vs. chemotherapy)

Model Diagnostics: CoxPH with treatment by time interaction, Stratified piecewise CoxPH, Kernel-based estimation of instantaneous Hazard

Piecewise HR: Mts 0-2: 2.48; Mts:2-3: 1.03; Mts: 3-4: 0.44; Mts 4-5: 0.55; Mts 5-6: 0.77

- TERIFLUNOMIDE, EMEA/H/C/002514/X/0031/G

  Active Substance: Teriflunomide

  Invented Name: Aubagio

  Procedure Type: Type II Variation - Extension of a marketing authorisation for Aubagio to add a new strength, 7 mg film-coated tablet, for use in paediatric patients from 10 years of age and older with relapsing remitting multiple sclerosis (MS).

  Trial TERIKIDS: Sample Size 166 (109 in treatment group)

  Primary Analysis: Time to first relapse, Stratified log-rank test

  HR Estimate ITT: 0.66 (p=0.2949)

  NPH issues: unclear, appears to be a delayed treatment effect, censoring prior to rescue

  Model Diagnostics: visual inspection of survival curves

  Sensitivity Analysis: Alternative estimand considering high MRI activity (trigger for rescue) as event

For the last example, the EPAR reports KM curves from which IPD event- and censoring times can be reconstructed. To simulate the effect of progression with subsequent treatment switching the distribution of subjects with progression will be modelled using marginal progression rates reported in the EPAR, as well as, KM plots from an analysis of time to first clinical relapse or high MRI activity (event to trigger rescue) published in the Supplemental material of Chitnis et al. (2021).

OPTIONS

Options refer to parameters considered to be under control of the experimenters. We consider sample size and follow-up to be options under the control of the experimenter. However, the settings will be chosen in relation to specific assumptions, such as baseline hazard, censoring distribution, and effect size, in order to obtain study designs with relevant properties in terms of Type I error and Power. In Section *Estimand* we specify the different estimand options that will be considered in the simulation study. Section *Study Design* then provides parameter ranges specifying the size, duration and type of trial. Section *Analysis Methods* specifies the analysis methods considered in this study and for which the operating characteristics under various assumptions and design options will be evaluated. Finally, Section *Metrics* defines the metrics that will be used to evaluate the performance characteristics of the different analysis methods under various assumptions and for different design options.

ESTIMAND

According to the ICH E9 (R1) addendum on estimands and sensitivity analysis in clinical trials to the guideline on statistical principles for clinical trials (EMA/CHMP/ICH/436221/2017) trial planning should proceed in sequence with the definition of a suitable estimand immediately following the specification of trial objectives. As we consider scenarios with and without different types of intercurrent events different estimand strategies may be considered relevant. It should be noted that the relevance of a specific estimand is not primarily determined by the data generating process. Specifically, as argued above, scenarios simulating the occurrence of an intercurrent event can be considered representative, e.g., where the intercurrent event is treatment switching or treatment discontinuation.

Regarding the estimand attribute treatment, we consider treatments that induce different shapes of the hazard curve. Regarding the attribute population, we consider settings, where there are predictive or prognostic sub-populations. To address intercurrent events, we consider different estimand strategies in selected specific scenarios. Especially, we consider trials where a composite estimand strategy is used and the intercurrent event (as treatment switching due to progression) is considered a treatment failure and the composite endpoint is defined as the time to event (of interest) or the time to the intercurrent event, whichever occurs earlier. Furthermore, a treatment policy strategy will be considered, where the time to event under the treatment policy (of which the intercurrent event, as, e.g. a treatment switch, is part of) is the estimands of interest. By addressing both estimands, conclusions can be drawn for either setting using the same simulated data.

In addition, depending on the analysis method, different population summary measures will be considered (e.g. hazard ratio for CoxPH model). See Table 3 for a complete list of analyses methods and corresponding summary measures. This shall enable an evaluation of estimand strategies targeting different population summaries.

STUDY DESIGN

Study design options mainly comprise the number of subjects to be included, the duration during which respective subjects are recruited, and the targeted number of overall events at which time the study is stopped. In addition to fixed sample designs, we consider group-sequential designs, where an interim analysis with the option to stop the study for success is performed once a certain number of events have been observed. Consequently, the number(s) of events at which the interim analyses will be conducted, as well as, the alpha spending function have to be chosen.

With respect to recruitment patterns, we assume that subjects are enrolled at uniformly distributed times over a given timespan. We assume the studies to be event driven, such that studies end when a pre-specified number of events is observed. All participants who did not have an event up to this time are censored. In other words: the study ends at the smallest time T such that the number of participants whose event time plus time of enrollment is smaller than T is larger than the pre-specified number of events. All patients for whom the time of enrollment plus the time of event is larger than T are censored.

See Table 2 for a complete list of the parameters that will be considered in the simulation study for each design option.

Table 2: Option set defining the different clinical trial designs considered for the simulation study.

Option	Parameters	Ranges	Comments
Recruitment speed	timespan of recruitment	18 months, 30 months	Corresponds to the range of recruitment period for studies reported in CONFIRMS (2022b) - Supplement 1
Number of Patients	number of subjects recruited	300, 500, 1000, 1500	Range for small studies to large studies in broad indications, as observed in the EPAR Review - excluding the largest study in cardiovascular domain with ~13,000 subjects (CONFIRMS (2022b), Table 6)
Number of events	number of observed events after which the study is stopped	depending on effect size and power scenario	
Interim Analyses	time(s) of interim analyses	no IA, 50% of events	
	α spending function	O'Brien-Fleming type alpha spending function	

Analysis methods

With respect to analysis methods, we distinguish between reference methods (not necessarily suitable under NPH) and alternative methods (taking into account NPH) to be evaluated in this simulation study. In addition we distinguish test and estimation methods. In Table 3 we list the different procedures that will be used to analyse the simulated data-sets and for which operating characteristics will be evaluated. The table provides names, short descriptions and intended use of the corresponding procedures. For methods that permit estimation of a clinically meaningful summary measure (e.g. HR, RMST) the respective measure is provided. For methods primarily focused on statistical testing based on statistics with no clear clinical interpretation (e.g. log-rank test, max-combo test) this is indicated, as well.

While some methods will be included in the simulation study where individual patient data are simulated according to distributional assumptions listed above, others are only considered for reanalysis of data in case studies that will be developed to illustrate the main conclusions of the simulation study. The corresponding methods are indicated for re-analysis of data sets in the table below. As those methods may not target a single summary measure (e.g. KM Plots) the corresponding column is left empty.

Table 3: List of analysis methods considered for the simulation study.

Method	Description	Use	Summary Measure	Implementation/calculation
Log-Rank test	Current standard: hypothesis test	Reference method - testing	NA - testing only	nph::logrank.test
Difference in median survival time	Current standard to contextualize effect size	Reference method - estimation	Difference in median survival times	nph::nphparams
Cox PH regression model	Current standard: 95% CI for HR	Reference method - estimation	hazard ratio	survival::coxph
Weighted Logrank tests	Gehan-Wilcoxon test, Fleming-Harrington family with weights (ρ, γ): (0,0), (0,1),(1,1),(1,0), modestly weighted log-rank test (Magirr (2019))	Alternative method - testing	NA - testing only	nph::logrank.test
MaxCombo-Tests	Maximum test combining the four Fleming-Harrington weighted tests, above	Alternative method - testing	NA - testing only	nph::logrank.maxtest
Milestone Survival Probabilities based on KM	Milestones: 6 months, 12 months	Alternative method - estimation	Milestone survival probabilities	KM curves and corresponding confidence intervals at given timepoints survival::survfit
Weighted Cox Regression (RMST)	Time range: 6 months, 12 months	Alternative method - estimation	restricted mean survival time	survival::survfit
Weighted Cox Regression (aHR)	Time range: 6 months, 12 months	Alternative method - estimation	Average hazard ratio	coxphw::coxphw
Piecewise exponential model	Choice of time intervals: 3 months, 12 months	Alternative method - testing	NA - for testing only (yields estimates for each interval)	pch::pchreg
Parametric survival model(s)	Standard predefined pharmacometric models (see Holford (2013)), utilizing ensemble modeling (model averaging).	Alternative method - estimation	All - using parametric models the summary measure (and estimand) can be chosen as part of the model specification	Using dedicated modelling software, e.g. NONMEM
Kaplan-Meier (KM)	Standard approach/well known	For re-analysis of data sets	NA - For illustration purposes in case studies	
Median survival	Standard approach to describe groupwise survival	For re-analysis of data sets	NA - For illustration purposes in case studies	

Method	Description	Use	Summary Measure	Implementation/calculation
Time varying coefficients	treatment coefficient via interaction with basis functions; change point of treatment coefficient. E.g., Time-dependent Cox model,	For re-analysis of data sets	NA - For illustration purposes in case studies	
Short and long-term HR	Yang-Prentice (2015): restricted shape of HR(t), easier to estimate and interpret than time varying treatment coefficient.	For re-analysis of data sets	NA - For illustration purposes in case studies	
Aalen additive hazards model	Martinussen & Pipper, (2014) estimates the Causal Odds of Concordance which is equivalent to the aHR in RCTs without covariate adjustment	For re-analysis of data sets	Causal Odds of Concordance	

METRICS

The metrics to evaluate the performance of different methods will be:

- Probability to reject the null hypothesis (type I error rate, power)
- Bias and MSE (mean squared error) for the estimation of parameters of interest.
- When confidence intervals are available, coverage probabilities and half-width of confidence intervals for parameter of interest

Specifically, for statistical hypothesis tests the probability of rejection in each scenario and parameter set will be the primary metric of evaluation. We distinguish between Type I error rates if the scenario and parameters belong to the null hypothesis and power if the scenario and parameters belong to the alternative. In case methods indicated for estimation (see Table 3, above) provide statistical hypothesis tests (e.g. based on confidence intervals) corresponding rejection probabilities will be provided as well. In addition, for scenarios evaluating method performance in group sequential designs, the proportion of early rejections and the average sample size will be reported.

For estimators the bias and mean squared error will represent the primary metrics of evaluation. In addition, for methods that provide confidence intervals for the parameters, the width and coverage probabilities of corresponding intervals will be reported.

In general, for dichotomous outcomes (e.g. test decisions) the absolute and relative frequencies of replications will be reported along with confidence intervals as a measure of (simulation) uncertainty. For continuous metrics, Monte Carlo standard errors will be reported.

EVALUATION METHODS

Methods will be evaluated by generating data according to the parametric scenarios or resampling from reconstructed data from the case studies and then applying each method. The number of replications will be determined by the Monte Carlo standard error of the rejection probability of a test, assuming the worst case of 50% coverage probability. For example requiring a standard error of 1 percentage point requires 2500 replications, according to:

$$n_{sim} = \frac{0.5 \cdot (1-0.5)}{0.01^2} = 2500.$$

For the evaluation of Type I error rates close to 0.025 one-sided, 2500 replicates would provide standard errors around 0.003. Consequently, estimated Type I error rates above 0.03 (i.e. $\sim 0.025 + 1.96 \cdot 0.003$) would give strong indication for an inflation of the Type I error rate. Consequently, we plan to simulate at least 2500 replications per scenario.

The performance of the methods with respect to their target estimand will be reported in tables and graphically. For all scenarios and options each metric will be reported. Monte Carlo standard errors will be reported to account for uncertainty due to a finite number of simulations (compare Table 6. from Morris et al. (2019)).

The results will be presented as tables and figures.

The tables will be structured so that the scenarios correspond to lines and methods or options will correspond to columns. Different performance metrics can be presented in different tables or in rows in each table cell. For example, in the case of estimation one table for bias and mean squared error and a second table for length and coverage probability of confidence intervals.

Bivariate Plots will include one line per method and one parameter of the scenario on the x-axis and the value of one performance measure on the y-axis. Other parameter values will be kept fixed or varied over multiple plots (facets).

Figures for tests will contain the rejection probability on the y-axis. Parameter values that are in the null-hypothesis will be highlighted, nominal alpha level will be added as a horizontal line.

Additionally figures for each summary statistic will be created that show the true value of each summary statistic and the value of each method targeting this summary statistic. Presentation of the different scenarios will be the same as in the figures of performance measures.

The complete results will be provided as listings and figures in pdf format, as well as, in a machine readable digital format (csv, rda) that can be filtered, processed and visualized e.g. using functionality of our simulation framework package or other third-party software (e.g. generic shiny apps to explore simulation results). In addition, we will create a summary report presenting the most relevant outcomes of the simulation study relying mostly on figures, which are capable of presenting and contrasting the results of a large number of settings within the broader assumption sets in a single plot. For example, for the assumption set relating to delayed onset of treatment effect, the Type I error rate of corresponding methods could be shown on the y-axis, ranging the time of onset on the x-axis in a plot matrix of facets with effect size and amount of non-informative censoring in rows and columns. Results for different baseline hazards could then be shown only for a specific setting to illustrate its impact on the operating characteristics, or if little impact is observed this can be explained in the text. However, in case it is found that the baseline hazard has a large impact, it may be used as a dimension in the figure with another less impactful parameter only presented for illustration. The final decision on which results to show as part of a compact simulation report, will depend on the results and cannot be foreseen in full detail at this time.

Software

The Simulations will be performed in R (R Core Team (2022)) making use of the **SimDesign** package (Chalmers and Adkins (2020)) to initialise random number generators, save random seeds and dispatch the computations and collect the results.

Functions to generate the data, apply the reviewed methods and aggregate the results will be written by the consortium. For data generation of piecewise constant hazards and hazards changing after a random time the **nph** package (Ristl et al. (2021)) will be used. Other packages might be used for additional data generating models if necessary. The different scenarios will be implemented in functions calling the data generating functions from other packages with the respective parameters or own code.

Additionally functions to draw bootstrap samples from available or reconstructed individual patient data will be implemented. For the Joint Models the data will be generated with a different software package. The `SimDesign` package provides functionality to write import functions to read in those datasets instead of generating datasets in R. In case externally generated data need to be enriched with additional features (e.g. recruitment time, study-withdrawal time) corresponding functionality will be implemented with the import functions.

Implementations of the data analysis methods applied to the generated data will be used from different packages where they are already implemented. Wrapper functions compatible with the `SimDesign` framework will be provided.

All R functions and documentation will be published in an R package, the code to reproduce the simulation study and tables and graphs from the report will be published as one or more vignettes to the R package. Code used to generate data from other scenarios will also be made available.

Simulation Parameters

The simulations will explore all options for the study design for all parameters of the different scenarios. The parameters for the options are given in Table 4, the parameters for the scenarios are given in Table 5.

This will give 16 options for 4 assumption sets with 144, 288, 54, 324 combinations of parameters each yielding 12960 simulations in total. Each simulation will be run 2500 times, 10 methods described in Table 3 will be applied to each simulated dataset.

Table 4: List of parameters for options.

Option	Parameters	Degrees of freedom
recruitment speed	18, 30 months	2
number of patients	300, 500, 1000, 1500	4
interim analysis	none, O'Brien Fleming type boundaries after 50% of events	2

Table 5: List of parameters for assumptions.

Assumption set	Parameter	Degrees of freedom
Delayed onset of treatment	control arm hazard	3
	non informative censoring	3

Assumption set	Parameter	Degrees of freedom
	time of onset	4
	effect size	4
Crossing hazard curves	control arm hazard	3
	non informative censoring	3
	time of onset	4
	HR before crossing	2
	effect size	4
Changing hazards after intercurrent event	control arm hazard	3
	non informative censoring	3
	progression rate	3
	hazard after progression	2
Biomarker subgroup	control arm hazard	3
	non informative censoring	3
	prevalence	3

Assumption set	Parameter	Degrees of freedom
	effect size in overall population	4
	HR in subgroup relative to overall population	3

REFERENCES

Ananthakrishnan, R., Green, S., Previtali, A., Liu, R., Li, D., and LaValley, M. (2021), "Critical review of oncology clinical trial design under non-proportional hazards," *Critical Reviews in Oncology/Hematology*, Elsevier, 162, 103350.

Benda, N., Branson, M., Maurer, W., and Friede, T. (2010), "Aspects of modernizing drug development using clinical scenario planning and evaluation," *Drug Inf. J.*, Springer Science; Business Media LLC, 44, 299–315.

Boulesteix, A.-L., Hoffmann, S., Charlton, A., and Seibold, H. (2020), "A replication crisis in methodological research?" *Significance*, Wiley Online Library, 17, 18–21.

Chalmers, R. P., and Adkins, M. C. (2020), "Writing effective and reliable Monte Carlo simulations with the SimDesign package," *The Quantitative Methods for Psychology*, 16, 248–280. https://doi.org/10.20982/tqmp.16.4.p248.

Chitnis, T., Banwell, B., Kappos, L., Arnold, D. L., Gücüyener, K., Deiva, K., Skripchenko, N., Cui, L.-Y., Saubadu, S., Hu, W., and others (2021), "Safety and efficacy of teriflunomide in paediatric multiple sclerosis (TERIKIDS): A multicentre, double-blind, phase 3, randomised, placebo-controlled trial," *The Lancet Neurology*, Elsevier, 20, 1001–1011.

CONFIRMS (2022a), "Report on the systematic literature review: Statistical analysis of trials where nonproportional hazards are expected. Deliverable 2."

CONFIRMS (2022b), "Review of EMA EPARS where nonproportional hazards were identified. Deliverable 2."

Efron, B. (1981). Censored data and the bootstrap. Journal of the American Statistical Association, 76(374), 312-319.

Friede, T., Nicholas, R., Stallard, N., Todd, S., Parsons, N., Valdés-Márquez, E., and Chataway, J. (2010), "Refinement of the clinical scenario evaluation framework for assessment of competing development strategies with an application to multiple sclerosis," *Drug information journal: DIJ/Drug Information Association*, Springer, 44, 713–718.

Guyot, P., Ades, A., Ouwens, M. J., and Welton, N. J. (2012), "Enhanced secondary analysis of survival data: Reconstructing the data from published kaplan-meier survival curves," *BMC medical research methodology*, Springer, 12, 1–13.

Holford, N. (2013), "A time to event tutorial for pharmacometricians," *CPT: pharmacometrics & systems pharmacology*, Wiley Online Library, 2, 1–8.

Krishnan, S. M., Friberg, L. E., Bruno, R., Beyer, U., Jin, J. Y., and Karlsson, M. O. (2021), "Multistate model for pharmacometric analyses of overall survival in HER2-negative breast cancer patients treated with docetaxel," *CPT: pharmacometrics & systems pharmacology*, Wiley Online Library, 10, 1255–1266.

Li, H., Han, D., Hou, Y., Chen, H., and Chen, Z. (2015), "Statistical inference methods for two crossing survival curves: A comparison of methods," *PLoS One*, Public Library of Science San Francisco, CA USA, 10, e0116774.

Magirr, D., & Burman, C. F. (2019). Modestly weighted logrank tests. *Statistics in medicine*, *38*(20), 3782-3790.

Martinussen, T. and Pipper, C.B. (2014), Estimation of Causal Odds of Concordance using the Aalen Additive Model. Scand J Statist, 41: 141-151. https://doi.org/10.1002/sjos.12004

Morris, T. P., White, I. R., and Crowther, M. J. (2019), "Using simulation studies to evaluate statistical methods," *Stat. Med.*, Wiley, 38, 2074–2102.

Novakovic, A. M., J. J. Wilkins, and H. Dai. 2020. "Changing Body Weight–based Dosing to a Flat Dose for Avelumab in Metastatic Merkel Cell and Advanced Urothelial Carcinoma." Clinical. https://ascpt.onlinelibrary.wiley.com/doi/abs/10.1002/cpt.1645.

R Core Team (2022), *R: A language and environment for statistical computing*, Vienna, Austria: R Foundation for Statistical Computing.

Rauch, G., Brannath, W., Brückner, M., and Kieser, M. (2018), "The average hazard ratio–a good effect measure for time-to-event endpoints when the proportional hazard assumption is violated?" *Methods of information in medicine*, Schattauer GmbH, 57, 089–100.

Ristl, R., Ballarini, N., Götte, H., Schüler, A., Posch, M., and König, F. (2021), "Delayed treatment effects, treatment switching and heterogeneous patient populations: How to design and analyze RCTs in oncology," *Pharmaceutical statistics*, 20, 129–145.

Wilkins, J. J., Brockhaus, B., Dai, H., Vugmeyster, Y., White, J. T., Brar, S., Bello, C. L., Neuteboom, B., Wade, J. R., Girard, P., and others (2019), "Time-varying clearance and impact of disease state on the pharmacokinetics of avelumab in merkel cell carcinoma and urothelial carcinoma," *CPT: pharmacometrics & systems pharmacology*, Wiley Online Library, 8, 415–427.