Request for CHMP Qualification Opinion

Clinically interpretable treatment effect measures based on recurrent event endpoints that allow for efficient statistical analyses

Mouna Akacha, Bruce Binkowitz, Frank Bretz, Arno Fritsch, Philip Hougaard, Antje Jahn, Franco Mendolia, Henrik Ravn, James Roger, Patrick Schlömer, Heinz Schmidli, Jiawei Wei

January 11, 2018

Contents

1 Executive summary

The objective of this submission is to seek a qualification opinion on recurrent event endpoints for clinical trials where recurrent events are clinically meaningful and where treatments are expected to impact the first as well as subsequent events. We claim that clinically interpretable treatment effect measures (estimands) based on recurrent event endpoints can be defined along with statistical analyses that are more efficient than those targeting treatment effect measures based on the first event only.

Recurrent events refer to the repeated occurrence of the same type of event over time for the same patient, thereby characterizing the disease burden or progression. Recurrent event endpoints are well established in indications where the rate of terminal events (e.g. death) is very low. Examples include relapses in multiple sclerosis (CHMP, 2015), exacerbations in pulmonary diseases (e.g. chronic obstructive pulmonary disease (CHMP, 2012a) and asthma (CHMP, 2010a)), headache attacks in migraine (CHMP, 2007, 2016a), hypoglycemia episodes in diabetes mellitus (CHMP, 2012b), and seizures in epileptic disorders (CHMP, 2010b, 2016b). In these chronic diseases, time-to-first-event endpoints that focus on the treatment effect on the first event are clinically less meaningful and hence rarely used. Experience with recurrent event endpoints is more limited in indications where the rate of terminal events is high. For example, current practice in chronic heart failure suggests that the primary analysis is based on a time-to-first-event endpoint (e.g. first occurrence of heart failure hospitalizations or cardiovascular death), although the clinical meaningfulness of recurrent heart failure hospitalizations is acknowledged in e.g. CHMP (2017).

The primary interest in trials using recurrent event endpoints is usually to understand how treatment affects the occurrence of recurrent events. This raises the question how to measure a treatment effect under the repeated occurrence of an event, which in turn depends critically on the underlying scientific question (Glynn and Buring, 1996): Different endpoints and treatment effect measures (i.e. different estimands; see ICH (2017)) can be considered. Depending on the specific setting, some estimands may be more appropriate than others. For example, accounting for the interplay between recurrent events and terminal events, such as death, is important in indications where the rate of terminal events is high. At the same time, inappropriate statistical approaches are often used to compare event rates

without being transparent about the target of inference (i.e. the estimand) and acknowledging the implicit scientific question of interest. A discussion on the use of recurrent event endpoints in different clinical trial settings is of broad scientific interest.

Depending on the clinical trial setting (e.g. with or without terminal events), different treatment effect measures can be considered. We do not seek to recommend a specific choice, but rather discuss the value and limitations of different treatment effect measures and their associated statistical analyses for recurrent events. We provide a thorough review of the statistical and clinical literature on recurrent events and present the results of extensive simulations studies to support the intended claim.

2 Statement of need

In this section, we outline the need for a qualification opinion about clinically interpretable treatment effect measures based on recurrent events. We first motivate this need by discussing the complex setting of clinical trials in chronic heart failure, where both recurrent hospitalizations and death are relevant when defining treatment effects (Section 2.1). Section 2.2 reviews several examples of diseases where recurrent event endpoints are well established and the rate of death is low in typical clinical trials. Section 2.3 discusses relevant statistical considerations when defining treatment effect measures based on recurrent events. Finally, Section 2.4 outlines the scope of this qualification opinion request.

2.1 Motivation for this request

With the availability of new treatments in the past decades, some diseases, such as heart failure (HF), were converted from short-term fatal diseases to chronic diseases. Traditional endpoints used in HF trials include 'time-to-disease-related-mortality' or a composite of 'time to the first event of either disease-related morbidity or mortality'. These endpoints have limitations as they do not capture the chronic nature of the disease which manifests in recurrent events (e.g. recurrent hospitalizations for HF) which in turn are an important indicator for the disease progression and the health status of patients. Thus, there is a need to tailor these traditional endpoints to best

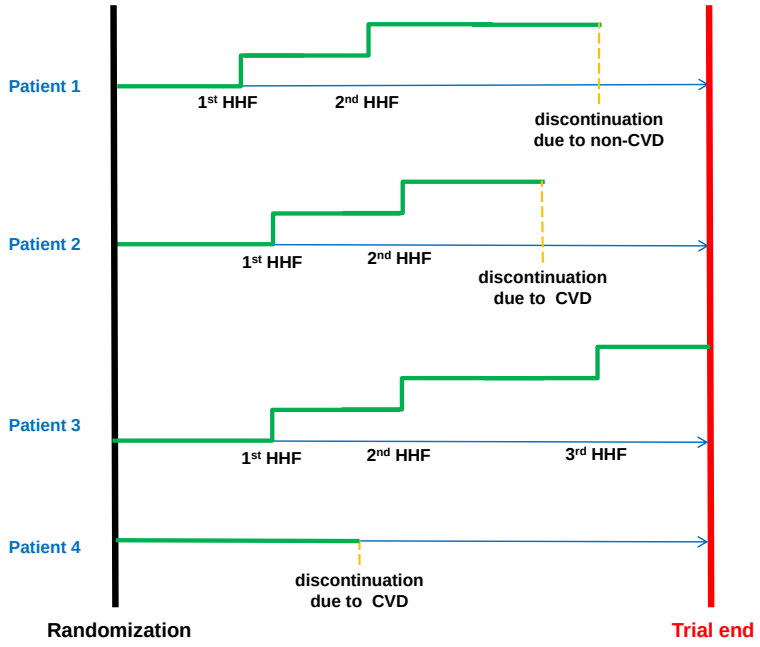
reflect the disease characteristics under chronic conditions.

Time to the first composite event of cardiovascular death (CVD) and hospitalization for heart failure (HHF) is used in many trials that have changed the practice of cardiovascular (CV) medicine. It is more specific than previous endpoints (e.g. time-to-death) and avoids competing risk and multiplicity problems (e.g. through the use of two endpoints, time-to-death and time-to-first-hospitalization). However, it ignores all HHF that occur after the first event despite the fact that these events reflect clinically meaningful information. In addition, improved medical care results in decreasing event rates. Therefore, the sample sizes needed for classical disease-related mortality and morbidity trials have increased to an extent that it becomes more and more challenging to conduct adequately powered trials. In contrast, including all recurrent HHF information is expected to better characterize the disease burden as HHF are an important indicator for disease progression, ultimately leading to clinically more meaningful treatment effect measures and better statistical efficiency (in terms of statistical power).

The recent CHMP (2017) guideline recognizes the clinical meaningfulness of recurrent HHF in patients with chronic heart failure (CHF) to better characterize their disease burden. At the same time, it is acknowledged that despite their importance, recurrent event endpoints are rarely used in CHF clinical trials compared to time-to-first-HHF (Collins et al., 2013; Zannad et al., 2013). Recurrent HHF are mostly used as secondary or exploratory endpoints although case studies highlighting the use or potential value of their repeated occurrence do exist; see e.g. the CHAMPION (Abraham et al., 2011) and PARAGON (Solomon et al., 2017) trials in HF, but also CHMP (2017) and Rogers et al. (2014a). One considerable challenge in analyzing recurrent event data in indications, where the rate of death is high, arises due to competing risks as the determinants of e.g. HHF and death share the same risk factors.

Different outcome measures are available to account for the repeated occurrences of the same type of event over time, such as counting the number of HHF, counting the number of 'bad' events (HHF and CVD), ranking according to a patient's journey or defining a suitable utility function. Figure 1 visualizes the differences between two of the former outcome measures, namely "number of HHF" and "number of 'bad' events". Patient 1 experienced two 'bad' events in form of two HHF and his life is terminated by experiencing a non-CVD. Patient 2 experienced three 'bad' events in form of two HHF

Figure 1: Visualization of four distinct life history processes. CVD: cardiovascular death.



and a fatal event in form of a CVD. In contrast, Patient 3 also experienced three 'bad' events, but in form of three HHF, therefore remaining in the trial until its end, while being alive. Finally, Patient 4 experienced only one 'bad' event, a fatal event early in the trial in form of a CVD. This example also illustrates that the event count may be low for two very different reasons: Either because the risk of experiencing the event is low or because the patient has died early and therefore did not experience many events.

A large number of statistical analysis methods for recurrent event endpoints is available, also in the presence of competing risks. However, concerns and questions remain, especially about the treatment effect being estimated by the various methods. As pointed out by Anker and McMurray (2012), "*the complexity of these tests is beyond the understanding of most clinicians, and the differences between and advantages and disadvantages of all the methods available are unclear to us.*" Likewise, Claggett et al. (2013) argued that "*... the act of appropriately condensing, summarizing and evaluating that infor-*

mation in a clinically meaningful manner becomes increasingly difficult" and Anker et al. (2016) asked "*how to interpret results if recurrent event analysis results differ substantially in magnitude or direction from time-to-first-event analysis?*" The root cause for many of these challenges in interpretation is the competing terminal event of death as its occurrence precludes the occurrence of any other event of interest. For example, in a trial in which the primary outcome is time-to-CVD, the non-CVD is a competing terminal event: A patient who dies of cancer is no longer at risk of experiencing CVD. Regardless of how long the duration of follow-up is extended, a patient will obviously not be observed to die of CV causes once he or she has died of cancer. In clinical trials, where patients are equally randomized to test and control treatment, a selection effect occurs as patients dying from non-CV causes are no more contributing further data. This may create an imbalance if treatments have different effects on the risk of non-CVD. A selection effect also occurs if for each patient only the first event (e.g. first HHF) is considered, and data after an event are discarded; see e.g. Appendix C.

In summary, statistical methods are often applied in complex settings, such as CHF, for which the interpretation of the treatment effect is not clear, leading to the question about the targeted treatment effect of greatest relevance to regulatory and clinical decision making (CHMP, 2017). Triggered by the recent ICH (2017) guideline, it is desirable to condense the relevant information into a clinically meaningful and interpretable measure of the treatment effect, the estimand. This includes a transparent description of how to capture key information like the target population (attribute A), the variable of interest (attribute B), intercurrent events that occur after treatment initiation and either preclude observation of the variable or affect its interpretation (e.g. treatment switching or death; attribute C) and an appropriate summary level (attribute D); see ICH (2017) for a more detailed description of attributes A to D. Accordingly, current practice needs to be reversed: First an agreement on a clinically meaningful estimand of primary interest is needed, which then informs choices about trial design, data collection, and statistical analysis. This should lead to clinical trials resulting in informative and interpretable treatment effects and hence facilitate decisions by regulators, clinicians and patients.

2.2 Recurrent event endpoints in clinical practice

Recurrent events are common in medical research, yet the best ways to measure their occurrence remains subject of discussion. An early argument for the greater importance of event rates, rather than only first events, was provided by Cumming et al. (1990) in their trials of falls. The cumulative risk of fractures increases with each fall; hence the number of falls is a more specific indicator of risk rather than whether one has fallen. A high rate of recurrent falls may especially increase the risk of injury. A focus on only those who fall at least once can blur important distinctions between groups when one group has an increased risk of recurrence relative to the other.

Recurrent event endpoints are well established in indications where repeated occurrences of the same type of event are clinically meaningful, treatments are expected to impact the first as well as subsequent events and where the rate of terminal events, such as death, is low in typical clinical trial settings. In such indications, selecting an appropriate treatment effect measure is important for benefit-risk assessments and in determining whether the treatment is actually modifying the disease course. A treatment effect measure with poor reliability or interpretability may lead to inaccurate results or improper use of treatments. In the following, we briefly review diseases where recurrent event endpoints are routinely used in clinical trials.

Relapsing-remitting multiple sclerosis (RRMS): The most common form of multiple sclerosis is characterized by recurrent acute episodes of neurological abnormalities (relapses), which are followed by complete or partial recovery. The treatment objective in RRMS is typically to prevent or reduce the frequency of new relapses, and to delay worsening of disability. Relapse-related outcomes are important because prevention of relapses benefits patients immediately; see also CHMP (2015). The most common primary variable used in the recent past has been the annualized relapse rate (ARR) which is the average number of relapses in one year (Lavery et al., 2014; van Munster and Uitdehaag, 2017); see also Table 1. The ARR is a recurrent event endpoint which takes into account that patients may relapse repeatedly, and, by reporting the relapse rate per year, depends less on the follow-up time of patients during a clinical trial. Secondary relapse-related variables in RRMS trials often include time-to-first-relapse, the number (%) of relapse-free patients, severity of relapses, relapses with complete or partial recovery, and relapses leading to hospitalizations (D'Souza et al., 2008), although each of these

variables has its own limitations. The analysis of the time-to-first-relapse is inefficient because the information following the first relapse is ignored. The number and proportion of relapse-free patients may be misleading because it depends on the time that patients were observed. Also, this variable does not distinguish between patients who have one relapse and those who have several, which may lead to incorrect conclusions if the treatment fails to influence the first relapse but reduces the risk of subsequent relapses (Glynn and Buring, 1996). Severity of relapses, completeness of recovery after a relapse and relapses leading to hospitalizations target at rather specific aspects of the treatment effect and thus fail to characterize more broadly the disease burden or progression.

Asthma: Another indication where recurrent event endpoints are well established. It is a chronic inflammatory disorder of the airways caused by the interaction of genetic and environmental factors. The disease is characterized by variable and recurring symptoms, airflow obstruction, bronchial hyper responsiveness and underlying inflammation. The GINA (2017) report on asthma management and prevention recognizes that patients can experience episodic flare-ups (exacerbations) of asthma that may be life-threatening and each exacerbation carries a significant burden to patients and the community. It continues stating that the long-term goals of asthma management are to achieve good symptom control, and to minimize the future risk of exacerbations, fixed airflow limitations and occurrence of adverse events. Similarly, CHMP (2010a) recommends the exacerbation rate as a clinically relevant endpoint to assess treatment in asthma patients. The statistical methods used to analyze this endpoint (as percentage of patients, annualized rate, time-to-first-event) should be justified. The trial length should be of sufficient duration to capture these events and dependent on the study treatment as well as the disease severity in the patient population. In standard Phase III trials, the duration is often one year to balance out seasonal effects that have a major impact in asthma.

Chronic obstructive pulmonary disease (COPD): A respiratory disorder characterized by airflow limitation, which is not fully reversible. The airflow limitation is usually progressive and is associated with an abnormal inflammatory response in the lungs to noxious particles or gases, primarily caused by cigarette smoking. COPD patients often suffer from acute exacerbations (i.e. a sudden worsening of symptoms). As exacerbations are a major cause of

Table 1: Primary variables for selected late-stage clinical trials in RRMS. CDMS: Clinically Definite MS, EDSS: Expanded Disability Status Scale

Study name	**Reference**	**Primary variable**
EVIDENCE	Panitch et al. (2005)	Proportion of patients who remained relapse-free
BENEFIT	Kappos et al. (2006)	(*i*) Time to CDMS (*ii*) Time to MS according to the McDonald criteria
CHAMPIONS	CHAMPIONS Study Group (2006)	Rate of development of CDMS
AFFIRM	Polman et al. (2006)	(*i*) Rate of clinical relapse at one year (*ii*) Rate of sustained progression of disability, as measured by the EDSS, at two years
SENTINEL	Rudick et al. (2006)	(*i*) Rate of clinical relapse at one year (*ii*) Cumulative probability of sustained disability progression, as measured by the EDSS, at two years
REGARD	Mikol et al. (2008)	Time to first relapse over 96 weeks
BEYOND	O'Connor et al. (2009)	Relapse risk
PreCISe	Comi et al. (2009)	Time to CDMS
TRANSFORMS	Cohen et al. (2010)	Annualized relapse rate
FREEDOMS	Kappos et al. (2010)	Annualized relapse rate
TEMSO	O'Connor et al. (2011)	Annualized relapse rate
DEFINE	Gold et al. (2012)	Proportion of patients who had a relapse by 2 years
CONFIRM	Fox et al. (2012)	Annualized relapse rate
FREEDOMS II	Calabresi et al. (2014b)	Annualized relapse rate
ADVANCE	Calabresi et al. (2014a)	Annualized relapse rate
TOWER	Confavreux et al. (2014)	Annualized relapse rate
DECIDE	Kappos et al. (2015)	Annualized relapse rate
OPERA I and II	Hauser et al. (2017)	Annualized relapse rate

morbidity, mortality, and the need for hospitalization or urgent care, Mahler and Criner (2007) conclude that "*an exacerbation in a patient with COPD has been considered analogous to an acute coronary event in a patient with coronary heart disease.*" Accordingly, CHMP (2012a) recognizes that the rate of moderate or severe exacerbations is a clinically relevant endpoint related to the associated morbidity and mortality and the usually significantly increased health-care requirement. The frequency and/or severity of exacerbations are important outcome measures that should be considered in COPD trials (Keene et al., 2008a,b). Such measures can include reduction in the number of exacerbations, annual rate and severity of exacerbations. Time-to-first-exacerbation might also be considered. If one of these measures is chosen as the primary efficacy endpoint, the others should be assessed also to ensure that improvement in one endpoint does not result in worsening in another. The frequency of exacerbations should normally be assessed over a period of at least one year due to seasonal variation in exacerbation rates.

Migraine: A primary headache disorder characterized by recurrent headaches that are moderate to severe. Typically, the headaches affect one half of the head, are pulsating in nature, and last from two to 72 hours. Recognizing the recurrent nature of the disease manifestations, CHMP (2007) recommends the frequency of attacks within a pre-specified period as the primary endpoint in migraine prophylaxis trials. Likewise, the related CHMP (2016a) concept paper suggests the choice of primary (migraine days vs headache days vs number of attacks) and secondary endpoints (symptom severity) as a critical item for discussion.

Epilepsy: A group of neurological disorders characterized by epileptic seizures, i.e. episodes that can vary from brief nearly undetectable to long periods of vigorous shaking. These episodes can result in physical injuries including occasionally broken bones. In epilepsy, seizures tend to recur and as a rule, have no immediate underlying cause. Conversely, isolated non-recurring seizures that are provoked by a specific cause (e.g. poisoning) are not deemed to represent epilepsy. CHMP (2010b) recommends that the assessment of efficacy should be based primarily upon seizure frequency and/or occurrence. The related CHMP (2016b) concept paper suggests a revision of the trial design in the add-on setting as a critical item for discussion in an update of its original guideline, e.g. validity and acceptability of a time-to-first-event approach as alternative endpoint and consequences for trial duration.

2.3 Statistical considerations on recurrent event analyses

The recent ICH (2017) guideline highlights the importance of defining suitable estimands. The choice of the primary estimand will usually be the main determinant for aspects of trial design and conduct, and guide the decision on an appropriate statistical analysis method targeting this estimand. In this section we briefly discuss statistical considerations relevant in this context.

2.3.1 Estimand

In any clinical trial with recurrent event endpoints the scientific question of interest has to be clearly stated, leading to a suitable choice of the primary estimand, under particular consideration of the therapeutic and experimental context. For example, Kuramoto et al. (2008) listed the following questions of potential interest:

- Does treatment decrease the event number over the trial period compared to control?
- How many events does treatment prevent, on average, compared to control?
- What is the treatment effect on the number of subsequent events among those who experienced the preceding event?
- What is the treatment effect on the number of higher-order events, e.g. third event, compared to control?

This list shows the importance of pre-specifying and choosing a clinically interpretable estimand, and also emphasizes the difference between the first event and what can happen when further events are considered.

When choosing an estimand, events that occur after treatment initiation need particular attention as they may lead to confounding. While randomized trials are expected to be free from baseline confounding, such 'intercurrent events' (ICH, 2017) will likely complicate the description and interpretation of treatment effects. Examples of intercurrent events include the use of an alternative treatment (e.g. rescue medication, prohibited medication, or subsequent line of therapy), as well as the patient's discontinuation from

treatment or even treatment switching, and, of course, terminal events such as death. In Section 3 we provide a detailed description of estimands based on recurrent event and time-to-first-event endpoints.

2.3.2 Trial design

An estimand should be understandable to a broader audience, including practicing clinicians and patients. Hence, when defining an estimand, one will typically refer to a certain time window for comparing treatment and control. For example, in psoriasis trials the response rate at week 12 and in diabetes trials the change from baseline in HbA1c at week 24 are typically of primary interest. Similarly for recurrent event data, the treatment effect for a fixed time period (e.g. two years) will be easiest to communicate. A clinical trial design where each patient is followed for the same time (fixed follow-up time of e.g. two years) would be adequate in such settings.

An alternative trial design is to follow patients until terminating the trial at some point in calendar time (flexible follow-up time). The advantage of this second design is that the patients enrolled first may be followed for a relatively long time, thus yielding long-term information without delaying the trial end. Statistical analyses typically allow to take into account data from patients with different follow-up times.

In both design options, patients are censored as they are no longer followed after some time. In the design with fixed follow-up time, the censoring happens at e.g. two years, while in the design with flexible follow-up time, censoring occurs at trial end. Censoring always implies a loss of information because we do not know what happens to patients after censoring.

A patient may also be censored due to other reasons, e.g. when withdrawing the consent to participate in the trial. This type of censoring is more difficult to address in the statistical analysis as it may be related to treatment; see e.g. NRC (2010).

Censoring should not be confused with death and other terminal events: After death no events will occur and therefore following the patient is logically the same as not following the patient. In other words, we do not lose any information on the patient by not following him after death, because there is no information that can be lost. Note that other events than death can lead to similar implications and conceptual challenges. For example, intake

of rescue medication may make it irrelevant to follow the disease process further for certain estimand strategies. In this sense, one may consider the use of rescue medication as a terminal event. Such terminal events must first be addressed at the estimand level. Different strategies to account for the termination of the recurrent event process due to terminal events (here: death) will be discussed in Section 3; see also Hernan and Robins (2018).

2.3.3 Statistical analysis

To be relevant, a statistical analysis method has to target the selected estimand. Additionally, the assumptions made by the analysis method should be plausible. For example, counting the total number of events can result only in non-negative integer values. Such data are non-normally distributed, and the variance varies with the mean. Thus, it is inappropriate to analyze such count data using ordinary linear regression because the linear model assumes homogeneity of variance and could produce meaningless negative predicted values. Some trialists may rescale the counts to a dichotomy (e.g. 'relapsed' versus 'did not relapse') or a set of ordered categories (e.g. 0, 1, 2, and ≥ 3), when defining the estimand. The data may then be analyzed using e.g. a logistic regression. However, reduction of counts into categories wastes information and may lead to a considerable loss in statistical power.

A simple model for analyzing count data is to assume that they are distributed according to a Poisson distribution. However, certain diseases exhibit greater heterogeneity, i.e. variability in event rates between patients, than expected with the Poisson distribution, known as overdispersion. For example, RRMS relapse data often exhibit overdispersion, which can arise in several ways (Wang et al., 2009). It can be a result of heterogeneity among patients, i.e. each patient has a constant relapse rate, but some patients may be more prone than others to relapse, partly due to genetic and environmental differences or unmeasured covariates. Overdispersion can also be a consequence of contagion, i.e. occurrence of a relapse increases a patient's risk for subsequent ones. Both heterogeneity and contagion mechanisms cause statistical correlation between relapses. Specifically, a patient with a history of relapses is likely to continue to experience more relapses, while subjects with no history of relapses tend to experience fewer future relapses. Thus, overdispersed count data are more spread than expected under the assumption of a Poisson distribution. The use of a Poisson model would not be appropriate

in such cases, as this would typically lead to too narrow confidence intervals for the treatment effect. Therefore, statistical models for overdispersed count data are needed.

Accounting for possible overdispersion is an important statistical consideration for recurrent event data, but not the only one. Dependent on the specific assumptions, different regression models for the recurrent event data can be formulated. For example, under the proportional rate assumption fully parametric models such as the Poisson model and the negative binomial (NB) model (accounting for overdispersion) could be considered. We refer to Appendix A.2 for a technical introduction of several classes of statistical models for analyzing recurrent event data. As all these models are based on some assumptions, it is important to assess the robustness across a range of plausible assumptions via a thorough sensitivity analysis, see Section 3. General references on statistical considerations for recurrent event data include Cook and Lawless (2007), Hougaard (2000), and Therneau and Grambsch (2000).

2.4 In-scope and out-of-scope of this request

We claim that treatment effect measures can be defined based on recurrent event endpoints that are clinically interpretable and allow for efficient statistical analyses. To support this claim, we investigate different estimands and associated analysis methods for recurrent events. However, it is not our objective to suggest, create, or validate new endpoints.

Depending on the clinical trial setting, different treatment effect measures (estimands) can be considered. We do not seek to recommend a specific choice, but rather discuss the value and limitations of different treatment effect measures and their associated statistical analyses. As discussed, different ways of including more information than just the first recurrent event are possible. For example, in CHF one may count the number of HHF, count the number of HHF and CVD, perform some ranking according to a patient's journey or define a suitable utility function. In this request, we focus on the first two measures only. The other two are out-of-scope as are aspects like quality of life or functional status of the patients.

The concepts surrounding recurrent events can also be used to evaluate safety, to assess risk versus benefit and quantify health economic value. Recurrent events often occur in all those settings, but the focus of this request is on

efficacy as seen in clinical trials, particularly on clinically interpretable treatment effects and on the efficiency of statistical analyses in an efficacy setting.

In many chronic disease trials, patients are at risk of different types of recurrent events. For example, transient ischemic attacks may be classified according to location in CV trials and migraines may be differentiated by severity in neurological trials. Also, the duration of event conditions could be an important aspect, especially if there is considerable variation in the duration or if some episodic conditions last for a long time. For the purpose of this request, methodological discussions as to whether and how to account for the duration or severity of events are out-of-scope. Instead, we refer the methodologically interested reader to Cook and Lawless (2007, Chapter 6).

Also, we do not discuss how to define an event in a clinically meaningful way. For example, in RRMS relapses are generally defined as neurologic symptoms lasting more than 24 hours which occur at least 30 days after the onset of a preceding event (Kappos et al., 2006), though definitions can vary by trial which will not be discussed in this request.

The work presented in this request, and the examples cited, focus on large Phase III confirmatory trials. Nevertheless, the described concepts are important in early drug development as well. Determination of an interpretable clinical endpoint is equally relevant in early phases and should be an important building block of a clinical development program leading to Phase III. Recognition of the importance of a recurrent event process early helps in preparation for larger, later stage trials. Later sections in this request include simulations that examine the properties of different methods across various sample sizes. The properties shown with the smaller samples may be useful for guiding design and analysis for early phase and/or smaller trials. Nevertheless, the primary scope of this request is on confirmatory trials.

Although we briefly mention sensitivity analyses and the handling of missing data, a thorough discussion of these aspects is out-of-scope of this request.

The objective through the following sections is to demonstrate that recurrent event data collected from clinical trials can be used in a 'better way' compared to the current practice. That is, we address the question whether more valuable information can be included when drawing inference on treatment effects. This qualification opinion request will answer 'yes' to that question, but first an agreement on the estimand of primary interest is needed (ICH, 2017). Statistical approaches need to be aligned to the estimand of choice and

robustness of conclusions ought to be assessed through a sensitivity analysis. Section 3 outlines estimands for recurrent event endpoints and also considers estimands which focus on the first event only. The setting of interest is that of a chronic disease, where a new treatment is investigated in terms of reducing disease burden. Section 4 provides case studies for RRMS and CHF to illustrate the various estimands introduced in Section 3 in situations without and with competing terminal events (death), respectively, although the key considerations are more broadly applicable. Section 5 then addresses the efficiency comparison of time-to-first-event with recurrent event analyses through two comprehensive simulation studies, each motivated by the case studies described in Section 4, and covering a wide range of practical scenarios. Section 6 concludes this request with a summary of the key findings. Detailed technical results and the complete results of the simulation studies are left for the appendix.

3 Estimands based on recurrent event and time-to-first-event endpoints

The recent ICH (2017) guideline emphasizes that trial objectives and statistical approaches should be aligned by clearly defining the estimand of interest. An estimand defines what is to be estimated to address a specific scientific question of interest. In clinical trials we are usually interested in estimating treatment effects with respect to the variable of interest (e.g. HHF and CVD in CHF). However, intercurrent events occurring after randomization, such as treatment discontinuation, non-CVD or rescue medication intake, may complicate both the definition and estimation of relevant treatment effects. Such events need to be taken into account when defining the estimand of interest.

An estimand can generally be described through the following four attributes:

(A) *Population:* As reflected through the inclusion/exclusion criteria of a given trial, e.g. RRMS patients with at least one documented relapse in the year preceding enrollment;

(B) *Variable:* As required to address the scientific question. For ease of interpretation and communication, the variable will typically refer to a specific time window, e.g. number of relapses up to two years;

(C) *Intercurrent events:* Specification of how to account for intercurrent events to reflect the scientific question of interest, e.g. whether these are ignored;

(D) *Summary measure:* For the variable which provides a basis for a comparison between different treatment conditions, e.g. difference in variable means.

An estimator defines the specific analysis method according to which the estimand is to be estimated from the trial data. When defining an estimator, assumptions will typically have to be made and it is essential to conduct a sensitivity analysis in the form of a structured and targeted sequence of analyses. These analyses should use estimators that focus on the identical estimand as the primary estimator, allowing investigation of robustness to model assumptions and data limitations. In contrast, supplementary analyses are concerned with different estimands that help putting the results into a broader perspective, e.g. to investigate the treatment effect on other relevant aspects of a disease. The estimand framework helps distinguishing between the target of estimation (trial objectives, estimand), method of estimation (estimator, estimate, measures of uncertainty), and sensitivity analysis.

Estimands with a causal interpretation are of main interest and would typically be preferred, as also emphasized in NRC (2010): "*Estimation of the primary (causal) estimand, with an appropriate estimate of uncertainty, is the main goal of a clinical trial.*" Causal estimands are often defined using the potential outcome framework, considering how the outcome of treatment compares to what would have happened to the same patients under different treatment conditions; see e.g. Little and Rubin (2000), Imbens and Rubin (2015), and Hernan and Robins (2018).

In the following, we present estimands for recurrent event endpoints and also consider estimands when the focus lies on the first event only. The setting of interest is that of a chronic disease, e.g. CHF or RRMS, where a new treatment is investigated in terms of reducing the disease burden. Reductions in disease-related events, such as HHF or CVD for CHF or relapses for RRMS, would be clinically relevant. The main scientific question concerns the comparison of test versus control treatment, and is best addressed by a randomized controlled clinical trial.

3.1 Settings without terminal events

In many therapeutic areas with disease-related recurrent events, the rate of death during a clinical trial is low, such as in RRMS or asthma. In this section, we focus on the setting where death or other terminal events are not considered to be a relevant intercurrent event. Instead, we focus on the intercurrent event of treatment discontinuation.

3.1.1 Recurrent event endpoints

We focus here on the commonly used 'treatment policy' and 'hypothetical' estimands, but also briefly discuss alternative estimands.

3.1.1.1 Treatment policy estimand The treatment policy estimand refers to the effect of the initially assigned treatment and not the effect of the treatment eventually received. The treatment policy estimand is often considered to be of interest, and is closely related to the intent-to-treat principle (ICH, 1998). The following four attributes characterize this estimand:

(A) *Population:* Usually defined through appropriate inclusion/exclusion criteria to reflect the targeted patient population for approval. It may sometimes also be defined based on data collected in a run-in period (before randomization) if the aim is e.g. to focus on the patient population which can tolerate control and/or test treatment.

(B) *Variable:* Number of recurrent events up to a certain follow-up time (e.g. two years). The choice of the time window is to a certain degree arbitrary and balances feasibility with the desire to assess the treatment effect sufficiently well.

(C) *Intercurrent events:* Regardless of whether or not an intercurrent event had occurred. Note that disregarding intercurrent events such as treatment discontinuation may lead to difficulties in the clinical interpretation of the treatment effect, especially if many patients discontinue treatment, or if discontinuations are strongly imbalanced between groups. For example, if patients discontinue study treatment for lack of efficacy, and then take another approved treatment, the treatment policy estimand compares 'test treatment followed by approved treatment'

against 'control treatment followed by approved treatment'. If more patients on test treatment discontinue, an ineffective test treatment may appear to be effective. Similarly, if more patients on control discontinue, an effective test treatment may appear to be ineffective. It is thus advisable to collect information on any treatments being used after study treatment discontinuation.

(D) *Summary measure:* Often the expected number of events in the follow-up time (e.g. two years), which may also be expressed as an annualized rate. This summary measure can be interpreted without assumptions on how the events are generated. Comparisons of test versus control treatment could be based on the ratio (or difference) of the expected number of events. Note that the treatment effect could be explained to patients as 'Prescription of the test treatment is expected to decrease the number of events within the next two years by 30% compared to prescription of the control treatment.' Other summary measures for the number of events (or the annualized rate) could be chosen, such as the median number of events, if considered clinically meaningful.

A randomized clinical trial, where each patient is followed-up for exactly the same time, is an appropriate design to address a treatment policy estimand.

The expected number of events in test and control treatment (for the selected time window of e.g. two years) can be easily estimated if no data are missing, i.e. if all patients are followed as required for two years. The estimate is simply the average number of events in each group. Hence, the treatment effect may be estimated as the ratio

$$\frac{\text{average number of events in test treatment}}{\text{average number of events in control treatment}}$$

and may be expressed as a percentage reduction of events on test treatment compared to control.

For inference, statistical models are typically applied which require further (testable) assumptions. For example, a standard NB regression is often used to model the number of relapses in RRMS, i.e. a time-homogenous NB model with a constant marginal event rate (Appendix A.2.2.4). The maximum-likelihood estimate for the rate ratio population parameter in the NB model is numerically identical to the ratio of average event numbers (if no data are missing, and hence all patients have the same follow-up time). Hence, the

NB model provides a valid point estimate for the treatment policy estimand, even if model assumptions are not correct (if there are no missing data).

The Anderson-Gill model with robust variance estimator proposed by Lin et al. (2000) (LWYY, Appendix A.2.3.1) can also be used for inference. LWYY gives the same point estimate as NB and hence is also an appropriate estimator for the treatment policy estimand (if there are no missing data). Other methods, such as those described in Wei et al. (1989) (WLW, Appendix A.2.3.2) or Prentice et al. (1981) (PWP, Appendix A.2.2.3), do not provide estimates of the treatment policy estimand, and hence should not be used for analysis in the context considered here. Although WLW and PWP could be considered as supplementary analyses, the treatment effects implied by these methods remain difficult to interpret.

A statistical hypothesis test is valid if the type I error rate is controlled at a pre-specified significance level. Under the (strict) null hypothesis of identical recurrent event data processes, some of the assumptions made for the statistical analysis are always correct. This holds e.g. for the constant rate ratio assumption for treatment versus control made by NB and LWYY. However, other assumptions such as the distributional assumption of a NB counting process made by the NB model may be incorrect. Note that a hypothesis test may still control the type I error rate (and hence be valid) even when the statistical model used to derive it is not fully appropriate.

To limit the assumptions for estimation, recurrent event information for the entire follow-up time (e.g. two years) is required. However, for various reasons patients may drop out early from the trial, leading to a missing data problem. Such missing data will have to be imputed implicitly or explicitly. Importantly, the imputed data should be in line with the treatment policy estimand. Generally, untestable assumptions will be required for such implicit or explicit imputations of missing data. The robustness of conclusions across a range of assumptions can be assessed with a sensitivity analysis. We refer to NRC (2010) for a more detailed discussion on missing data issues.

In general, when defining estimands, one main analysis method should be defined which can be accompanied by a sensitivity analysis. For example, the NB model paired with an appropriate imputation method to handle missing data could be chosen as the main analysis. In a sensitivity analysis, the assumptions made for the main analysis and the missing data handling approach can be varied across a range of plausible assumptions.

3.1.1.2 Hypothetical estimand An alternative treatment effect refers to the hypothetical setting where all patients stay on the initially assigned treatment for the intended duration. Sometimes such an estimand is of scientific interest, especially if treatment discontinuation could be avoided in practice. If, however, many patients are expected to discontinue due to adverse events or other tolerability issues then the question, what the effect would be had these patients continued their treatment, appears to be of limited clinical and scientific value.

The hypothetical estimand has the same attributes (A) population, (B) variable, and (D) summary measure as the treatment policy estimand. However, intercurrent events are handled differently:

(C) *Intercurrent events:* The hypothetical setting is of interest where the intercurrent event of treatment discontinuation would not occur, i.e. patients would continue their treatment for the intended duration.

A randomized clinical trial with fixed or flexible follow-up time would be an appropriate design. Recurrent event information after discontinuation of study treatment does not have to be collected for the main analysis, although it may in many cases be important from a safety perspective to follow-up on patients after discontinuation.

In contrast to the treatment policy estimand, stronger assumptions are needed to obtain consistent point estimates and to perform inference for the hypothetical estimand. Statistical models will typically be used to implicitly or explicitly impute data after treatment discontinuation. For example, a standard NB model (Appendix A.2.2.4) censors the patient at time of treatment discontinuation and uses an offset of log(discontinuation time). The standard time-homogeneous NB model assumes that a patient's recurrent event rate does not change with time, and also that treatment discontinuation is not informative. By 'not informative' we mean that information after treatment discontinuation for a given patient can be appropriately predicted based on the observed data of that given patient and other similar patients. Similarity is established in terms of the same baseline characteristics featured in the model and the observed recurrent event history up to the point of treatment or study discontinuation. If these assumptions are appropriate, the estimated rate ratio parameter of the NB model will be consistent with the hypothetical estimand. Note that semi-parametric NB models do not rely on the assumption of a constant event rate (Cook and Lawless, 2007).

LWYY (Appendix A.2.3.1) allows that recurrent event rates change with time, but requires stronger assumptions than NB regarding treatment discontinuation. LWYY implicitly predicts missing data based on the baseline characteristics featured in the model but does not include information on the observed recurrent event data process after randomization (while NB does). Again, if assumptions are appropriate, LWYY provides consistent estimates for the estimand of interest.

In summary, either NB or LWYY may be selected for the main analysis. The robustness of conclusions to alternative assumptions can be investigated with a sensitivity analysis. In line with our discussion for the treatment policy estimand, WLW and PWP do not provide consistent estimates of the hypothetical estimand but could be considered as supplementary analyses.

3.1.1.3 Other estimands and additional considerations In settings without terminal event, the treatment policy and hypothetical estimands discussed above seem to be the most commonly used estimands in current practice. For example, a treatment policy strategy could be used for some intercurrent events (e.g. patients discontinuing treatment due to adverse events), while a hypothetical strategy could be used for others (e.g. patients discontinuing treatment due to perceived lack of efficacy). However, alternative estimands may also be considered. The ICH (2017) guideline discusses various strategies for selecting estimands, and these considerations straightforwardly apply to the case of recurrent events without terminal events.

3.1.2 Time-to-first-event endpoints

Disease-related events such as relapses in RRMS occur repeatedly for the same patient. However, sometimes events after the first event are ignored, and comparison of test and control treatment is based on a time-to-first-event variable, e.g. time-to-first-relapse for RRMS. In this section we focus on such time-to-first-event endpoints. Note that the general framework in ICH (2017) also applies to time-to-first-event endpoints, although they were not discussed explicitly. Here, we highlight some points which need special attention.

Similar to the case of recurrent event endpoints, we discuss only the treatment policy and hypothetical estimands. These estimands would consider the same population (attributes A) and handling of intercurrent events (attribute C)

as in the recurrent event setting discussed in Section 3.1.1. The variable of interest (attribute B) now becomes the time-to-first-event up to a specified follow-up time (e.g. two years).

However, the selection of an appropriate summary measure (attribute D) is more challenging. The hazard ratio (HR) of a Cox proportional hazards model is typically used in time-to-first-event settings to summarize a treatment effect. However, the HR does not always allow for a causal interpretation (Aalen et al., 2015), which seems undesirable. Additionally, the HR is difficult to interpret if the proportional hazards assumption does not hold. Alternative summary measures such as the event-free probability at the follow-up time or the restricted mean survival time would admit a causal interpretation; see e.g. Royston and Parmar (2011), Uno et al. (2014, 2015), Pak et al. (2017), and Rufibach (2017). These can also be interpreted without reference to a particular statistical model.

For the analysis, a standard Cox proportional hazards regression is commonly used regardless of the estimand of interest. When interest lies in the hypothetical estimand then the patient should be censored at the time of treatment discontinuation. A standard Cox model applied to such data targets the hypothetical estimand (with the HR summarizing the treatment effect), if the model is appropriate and the treatment discontinuations are not informative. In the presence of between-patient heterogeneity not accounted for by the covariates (common in many diseases such as RRMS), the proportional hazards assumption typically does not hold, and hence the resulting HR estimate would be difficult to interpret. For the estimation of the treatment policy estimand, the data after treatment discontinuation would not be censored. In this setting, the assumptions of the Cox proportional hazards model are likely to be violated leading again to difficulties in interpretation.

Despite the issues with the use of Cox proportional hazards regression for estimation (e.g. lack of causal interpretation, difficult to communicate to non-statisticians, proportional hazards assumption not plausible in relevant settings), associated statistical hypothesis tests are closely linked to the log-rank test and hence valid, at least if treatment discontinuations are independent of the time to the first event.

3.2 Settings with terminal events

Serious chronic diseases such as CHF are characterized through recurrent disease-related morbidity events, e.g. HHF. Patients with such diseases also have an appreciable risk for disease-related deaths, e.g. CVD. Disease-related deaths are terminal events, i.e. events that terminate the recurrent event process such that no more events can occur afterwards. Additional terminal events such as disease-unrelated deaths further complicate this setting.

Terminal events pose conceptual challenges when drawing conclusions from associated clinical trials. We would like to acknowledge that patients who die can no longer experience any morbidity events. While disease-related deaths preclude all future morbidity events, patients in less serious conditions may remain on trial and experience many morbidity events. Thus, simply counting the number of events may not be sufficient. The event count could be low for two very different reasons, either because the risk of experiencing the event is low, or because the patient has died and therefore not experienced many events. A key question is thus how to account for the intercurrent event of death. When answering this question it is important to keep in mind that "*truncation by competing events raises logical questions about the meaning of the causal estimand that cannot be bypassed by statistical techniques*" (Hernan and Robins, 2018). Note that the question above also applies to other settings than recurrent events in the presence of mortality, e.g. longitudinal biomarkers when mortality is appreciable or semi-competing risk problems.

In the following, we present several estimands that make use of the recurrent morbidity event information up to e.g. two years of follow-up. For ease of exposition, we focus on only one intercurrent event, namely disease-related death. For more than one type of intercurrent event, including disease-unrelated deaths, we refer to Section 3.2.1.6.1. Considerations on suitable estimators and sensitivity analyses are kept short, especially if the estimators were already discussed in Section 3.1.

3.2.1 Recurrent event endpoints

Motivated by ICH (2017), we discuss five estimand strategies: treatment policy, composite, hypothetical, principal strata, and while-alive. The treatment policy and hypothetical strategies were already discussed for recurrent event endpoints without terminal events in Section 3.1.1, where the only

intercurrent event under consideration was treatment discontinuation. The implications for a terminal intercurrent event such as death are entirely different, and hence we discuss these two strategies again in this context. While we describe these five strategies, we are not suggesting that they are all clinically meaningful. Moreover, we emphasize that an investigation of recurrent event endpoints in the presence of disease-related death is incomplete without considering in addition an estimand that focuses on disease-related death itself. As mentioned before, we consider a world in which only one intercurrent event can occur (disease-related death) for the purpose of clarity. Other intercurrent events (e.g. treatment discontinuation or disease-unrelated death) are expected not to occur.

3.2.1.1 Treatment policy estimand A treatment policy estimand would ignore intercurrent events. However, in our case the intercurrent event is disease-related death, which cannot be ignored as it makes further recurrent events impossible. The same issue also occurs for other types of endpoints, not just for recurrent event endpoints. For example, if the variable is clinical response at one year, then this value does not exist for a patient who dies earlier. Hence, a treatment policy estimand is not suitable for terminal intercurrent events, as also indicated in ICH (2017).

3.2.1.2 Composite estimand The composite strategy includes the intercurrent event of disease-related death in the variable definition. There are various ways how this could be done, and we focus here on one specific composite estimand for illustration. Alternative estimands falling into this category are discussed in Section 3.2.1.6.2.

The specific composite estimand considered here may be described as follows:

(A) *Population:* Defined through appropriate inclusion/exclusion criteria to reflect the targeted patient population for approval;

(B) *Variable:* Number of unfavorable events including disease-related morbidity events (e.g. HHF) and disease-related death (e.g. CVD) up to two years;

(C) *Intercurrent events:* The intercurrent event of disease-related death is captured through the variable definition;

(D) *Summary measure:* Expected number of unfavorable events at two years; comparisons between test and control treatment could be based on the ratio or difference (Section 3.1.1).

By focusing on the unfavorable event count at two years and noting that no unfavorable events can occur after death, this estimand focuses on the naive unfavorable event count up to two years. As previously discussed, this unfavorable event count may be low for two very different reasons, either because the risk of experiencing the unfavorable events is low or because the patient has died early and therefore could not experience additional unfavorable events. The clinical meaningfulness of this estimand is therefore debatable unless it is complemented with an estimand that focuses on the time to disease-related death; see Section 3.1.2.

For the variable defined in (B), a death event is implicitly considered to be equivalent to a disease-related morbidity event. Alternatively, a higher weight could be given to death. For example, one could define a variable as the number of disease-related morbidity events (e.g. HHF) up to two years for patients who do not die within two years, and 24 (equivalent to monthly hospitalizations) for patients who die within two years. We will not discuss such alternative options in the following.

A design that targets this specific composite estimand is a randomised parallel group design where patients are followed up for two years or until death.

In general, the same statistical considerations as laid out in Section 3.1.1 also apply when analyzing the composite estimand above. However, in addition it is important to acknowledge that patients who die for a disease-related cause can no longer experience morbidity events thereafter. This can be done by censoring the patients at the end of two years rather than at time of death. Methods that target this specific composite estimand include non-parametric methods (Ghosh and Lin, 2000) and semi-parametric models (Ghosh and Lin, 2002) which make the constant mean ratio assumption. More recently, Mao and Lin (2016) described how the semi-parametric LWYY proportional mean model (Appendix A.2.3.1) can be used with censoring at trial end. Of these models one can be chosen as the main analysis, while others could play a role in a sensitivity analysis.

3.2.1.3 Hypothetical estimand We define this estimand by asking the hypothetical question what would have happened had the patient not died

due to a disease-related cause. The hypothetical estimand shares the same attribute (A) as the composite estimand. The other attributes are different:

(B) *Variable:* Number of disease-related morbidity events (e.g. HHF) up to two years;

(C) *Intercurrent events:* The hypothetical setting is of interest where the intercurrent event of disease-related death would not occur;

(D) *Summary measure:* Expected number of disease-related morbidity events at two years; comparisons between test and control treatment could be based on the ratio or difference (Section 3.1.1).

The same design considerations as mentioned for the composite estimand also apply here.

Estimators for the hypothetical effect can be obtained in a variety of ways, dependent on the specifics of the assumed latent process after disease-related death. For certain latent processes, NB (Appendix A.2.2.4) with censoring at the time of disease-related death can be used. Conceptually, this method predicts the latent process for a given patient based on a) baseline covariates included in the model, b) information on the patient's recurrent event process prior to disease-related death and c) similar patients that share similar baseline characteristics and recurrent event data information after randomization. Alternative choices that consider other latent processes include LWYY (Appendix A.2.3.1) with censoring at the time of disease-related death. Here the latent process for a given patient is characterized based on a) baseline covariates included in the model and b) similar patients that share similar baseline characteristics. In contrast to NB, LWYY does not include information on the recurrent events that occur after randomization and prior to death in order to inform the latent process. Yet an alternative approach to predict the latent process is through the use of joint frailty models (Appendix A.2.4.2). They model the recurrent morbidity and the mortality events simultaneously while accounting for the correlation between these two event processes. By linking these two processes, certain assumptions can be captured, e.g. that a larger cumulative number of morbidity events leads to a higher risk of dying; see also Cowling et al. (2006) and Liu et al. (2004).

Any of the aforementioned statistical analyses for this estimand will rest on assumptions about disease-related morbidity events that would have been

observed under the hypothetical setting where patients had not died due to a disease-related cause. Generally, the assumptions needed for such predictions cannot be verified from the observed data.

3.2.1.4 Principal stratum estimand The principal stratum estimand (Frangakis and Rubin, 2002) shares the same attributes (B) and (D) as the hypothetical estimand. The population (principal stratum) is defined as follows:

(A) *Population:* Defined through patients who would not die due to a disease-related cause over a period of two years, regardless of treatment assignment, within the targeted population defined by inclusion/exclusion criteria.

As disease-related deaths do not occur for this principal stratum population, attribute (C) becomes

(C) *Intercurrent events:* The intercurrent event of disease-related death is captured through the population definition.

The principal stratum estimand has a causal interpretation as it refers to the treatment effect in a subgroup properly defined by intercurrent events. However, as disease-related morbidity and mortality events are related, focusing on a population where no patients would die during the trial may often not be of primary clinical interest.

A statistical analysis for this estimand requires causal inference methods; see e.g. Imbens and Rubin (2015) and Hernan and Robins (2018). The robustness of conclusions with respect to the underlying assumptions is assessed by an appropriate sensitivity analysis.

3.2.1.5 While-alive estimand The while-alive or while-on-treatment estimand focuses on the treatment effect while patients are alive or, in other words, while the intercurrent event did not occur. This estimand has the same attribute (A) as the composite estimand. The remaining attributes are defined as follows:

(B) *Variable:* Number of disease-related morbidity events (e.g. HHF) while the patient did not die due to a disease-related cause;

(C) *Intercurrent events:* The intercurrent event of disease-related death is captured through the variable definition;

(D) *Summary measure:* Expected number of disease-related morbidity events divided by the restricted mean survival time (Royston and Parmar, 2011; Uno et al., 2014).

The same design considerations as for the composite estimand can be applied here.

For the statistical analysis, one could use LWYY (Appendix A.2.3.1) with censoring at the time of disease-related death, as this estimator would target the while-alive estimand; see also Section 5.2.2 and Appendix E.3.

3.2.1.6 Additional Considerations

3.2.1.6.1 More than one intercurrent event So far we only considered one intercurrent event, namely disease-related death. However, in practice usually more than one intercurrent event needs to be accounted for. In the case of e.g. CHF trials at least two additional intercurrent events are worth discussing: non-CVD and treatment discontinuation for various reasons. Often, non-CVD are considered to be non-informative, e.g. if treatments are likely to have no effect on them. Thus, different strategies can be plausible and clinically meaningful. In particular the hypothetical, principal stratum and while-alive strategies appear reasonable in this context. As for treatment discontinuation, clinical trials in CHF have traditionally focused on the treatment policy strategy, i.e. the effect of treatment assignment regardless of study treatment discontinuation; see also Section 3.1.1.

For illustration, we describe two estimands that take into account the three intercurrent events CVD, non-CVD, and treatment discontinuation. These two estimands will also be discussed in Sections 4 and 5. The two estimands differ with respect to the variable of interest (attribute B):

- Estimand 1 (HHF): Number of HHF while the patient is alive;
- Estimand 2 (HHF+CVD): Number of unfavorable events, i.e. number of HHF and CVD, up to and including the time of death.

For both estimands, the population (A) is defined by appropriate inclusion/exclusion criteria. In terms of attribute (C), we are interested in the treatment effect regardless of treatment discontinuation and while patients are alive, i.e. they did not die due to any cause. Hence, for the intercurrent event of treatment discontinuation, a treatment policy strategy is used, while for CVD and non-CVD, a while-alive strategy is applied. The summary measure (D) is the event rate while patients are alive which can be expressed as the expected number of disease-related morbidity events divided by the restricted mean survival time. For comparisons between test and control treatment, the summary measure becomes the rate ratio and can be interpreted as

$$\frac{\text{expected number of events per unit time alive in test treatment}}{\text{expected number of events per unit time alive in control treatment}}.$$

3.2.1.6.2 Other estimands So far we focused on estimands that count the number of all disease-related morbidity and potentially all disease-related mortality events. Other estimands could also be considered and can generally be classified into two categories: a) approaches focusing on a hierarchy of variables and b) approaches using a weighted composite variable.

The first class of approaches categorizes patients according to their worst outcome and an agreed hierarchy of variables, e.g. mortality is worse than hospitalization, which in turn is worse than a certain drop in a quality of life index. While experience is limited, there are examples where such approaches were applied CHF trials:

- The score by Packer (2001) categorizes patients according to their clinical course as 'improved', 'unchanged' or 'worse'. This was used e.g. in the REVIVE and RELAX-AHF-Asia trials; see Packer et al. (2013) and Sato et al. (2017). The odds ratio was used as a summary measure by comparing the odds of being in a more favorable category when taking the test treatment relative to the odds of being in a more favorable category when taking the control treatment.

- Felker et al. (2008) and Subherwal et al. (2012) advocate the use of global ranking approaches where all patients are ranked from worst rank to best rank with respect to their clinical experience.

- Pocock et al. (2011) propose the use of win ratios, either in a matched-pair or an unmatched approach. The unmatched approach compares each patient on test treatment with each patient on control treatment based on hierarchically ordered endpoints, such as CVD and HHF, with the option to include recurrent HHF. The summary measure of interest is the win ratio, which in the matched-pair case is the proportions of winners divided by the proportion of losers in the test treatment group. In the unmatched case it is the number of pairwise comparisons where the test treatment wins divided by the number of pairwise comparisons where the test treatment loses; see also Dong et al. (2016).

- Claggett et al. (2014) propose an ordered categorical outcome derived from multiple time-to-event outcomes by creating a sequence of nested composite outcomes. The probability of a patient falling into each possible category at a fixed follow-up time can then be used as basis for the intervention effect: the net probability that a treated patient experiences a better rather than worse categorical outcome compared to a control patient.

These approaches allow the inclusion of information on changes in symptoms and functional status in addition to clinical outcomes. In particular, events with greater clinical importance can be given greater relative weight. Also, the directional consistency in the components is not crucial as long as everyone agrees on the chosen hierarchy. The latter is very important and could be perceived as a limitation of these approaches as different stakeholder (e.g. patients, clinicians, payers, regulators) may well be interested in different hierarchies. Additional challenges in using such estimands may arise when patients have differential follow-up time as this may not be directly captured in the definition of the variables. For rank-based approaches, communication and interpretation of the treatment effect may also be challenging. Finally, experience with such estimands for confirmatory trials is limited and trial designs may become challenging. For example, there may be lack of historical data to inform the sample size assessment (e.g. baseline rates, correlation between the components, minimum clinically relevant difference).

The second class of approaches uses a weighted composite approach and allows for the inclusion of multiple events while accounting for the relative importance of the individual components through the choice of adequate, perhaps subjective, weights. Estimands that fit into this category have been

used in the past (Taylor et al., 2004; Sampson et al., 2010). However, they sometimes result in estimates that are difficult to interpret; see Taylor et al. (2004). Most importantly, the choice of weights requires agreement of all relevant stakeholders. In the context of CHF, Anker et al. (2016) note that the use of such approaches is "*limited by the lack of consensus on the relative weighting of events and inconsistency across trials.*" A popular approach that fits into this category is based on the days alive and out of hospital and the weighted version of symptom-adjusted days alive and out of hospital (Cleland, 2002). Note that any endpoint incorporating a function of days in the hospital is subject to influences beyond just a patient's condition, as hospital reimbursement policies and local medical practices differ around the world and introduce heterogeneity that may inhibit the ability to detect a treatment effect. An example of this was observed by Pfeffer et al. (2015) in the TOPCAT trial. This issue of regional differences may be addressed, partly if not entirely, by stratification, as long as the treatment differences across strata are not too different to be interpretable.

Alternative approaches to define estimands that do not fall into either of the two classes above may also be valuable. For example, an approach inspired by multi-state survival data methods is to consider the integrated hazards instead of the event counts. In the absence of terminal events, the integrated hazards will agree with the mean number of events, as explained in Appendix A.2.4. In the presence of terminal events, the two quantities do not agree. The mean number of events will be smaller reflecting that patients who died can no longer experience any event. This may make a treatment with high mortality appear better than it deserves. Using the integrated hazard as the target of estimation will avoid such effects. It would be calculated for each treatment so that it can be compared as a measure of the treatment effect.

Utility-based methods are yet another alternative, which seem particularly useful if utilities can be assigned to different life history paths. With a utility-based approach one may not need to distinguish between different causes of death (e.g. disease- or morbidity-related or not) as the utility would be e.g. 0 for days after death from any cause, 1 for living at perfect health and some lower value for each day in which they are suffering from a morbidity-related symptom. This could be viewed as a generalization of the idea of quality-adjusted life years used in economic evaluations.

3.2.2 Time-to-first-event endpoints

In Section 3.2.1 we discussed recurrent event endpoints subject to a competing terminal event. While many clinical trials collect such information, they often only report the results focussing on the first event, e.g. the first morbidity event. In such cases, disease-related and disease-unrelated deaths are competing terminal events. In principle, the same estimands and considerations as discussed in Section 3.2.1 also apply for time-to-first-event endpoints subject to competing terminal events.

While the focus of this request is not on time-to-first-event endpoints, we discuss some related aspects in Appendix B as it may benefit the discussion around competing risk approaches for time-to-first-event endpoints. We focus on the case where interest lies in the time to the disease-related death, e.g. CVD, and where disease-unrelated death is a competing terminal event. All considerations can be applied to the case where morbidity events or a composite event are of main interest and subject to terminal events, e.g. disease-related or unrelated deaths.

Note that for only two recurrent event processes the estimand for a recurrent event analysis is the same as for a time-to-first-event analysis: a) a Poisson model with a proportional rate function, if appropriate, and b) a renewal model with proportional hazards for the times between events, if appropriate. In both cases it is assumed that there is no between-patient heterogeneity. These are extreme settings so that in general a time-to-first-event analysis will target a different estimand than a recurrent event analysis.

4 Case studies

In this section we discuss estimands and analysis methods for two case studies. The first one involves the acyclovir trial in RRMS while the second one is based on the ValHeft trial in CHF.

4.1 Relapsing-remitting multiple sclerosis

Lycke et al. (1996) investigated the effect of an antiviral drug, acyclovir, in patients with RRMS. This was a randomized, placebo-controlled, double-blind clinical trial, where 60 RRMS patients were randomly assigned to test treat-

ment (acyclovir at 800mg, three times daily) or placebo, and then followed for two years. A non-parametric test was used to compare the two groups with respect to relapses. The estimand of main interest was not clearly specified, in particular regarding the summary measure. For our discussion here we focus on the intercurrent event of treatment discontinuation.

4.1.1 Estimands

We consider two hypothetical estimands, which differ with respect to the variable of interest (attribute B) and the summary measure (attribute D).

- Time-to-first-relapse estimand: Suppose we are only interested in the first relapse for each patient, and disregard any following relapses. Time-to-first-relapse up to two years is then the variable of interest (attribute B). The hazard ratio (HR) of a Cox proportional hazards model is used as a summary measure (attribute D), but see Section 3.1.2 for a discussion on the limitations of the HR.
- Number-of-relapses estimand: The variable of interest (attribute B) reflecting disease activity is the number of relapses in the first two years. An interpretable summary measure (attribute D) is the ratio

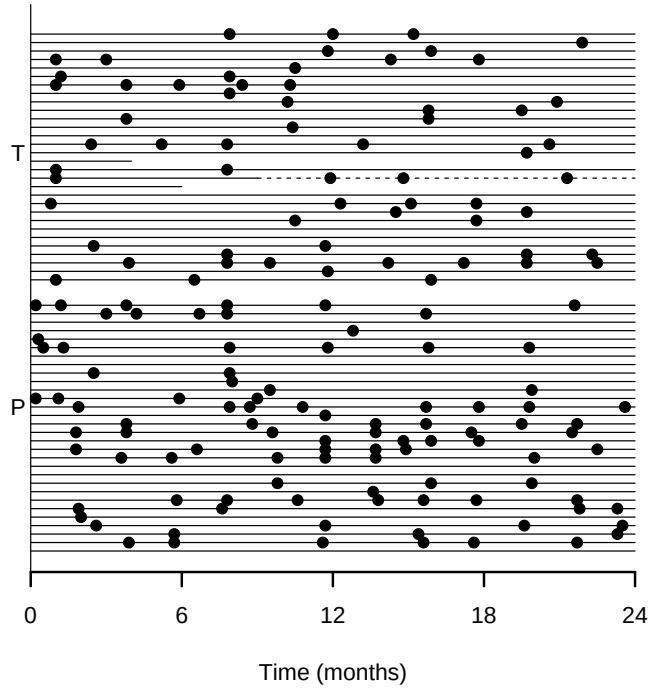
$$\frac{\text{mean number of relapses in test treatment}}{\text{mean number of relapses in placebo}}.$$

For both estimands, the population (attribute A) is defined by the inclusion/exclusion criteria given in Lycke et al. (1996). We are interested in the hypothetical estimands (Section 3.1.1.2) where the intercurrent event of treatment discontinuation would not occur and patients continue their treatment until the end of two years (attribute C).

4.1.2 Analysis

For the hypothetical estimands considered here, relapse data after treatment discontinuation are irrelevant and therefore not included in any analysis, i.e. patients are censored at treatment discontinuation. Missing information for patients that discontinue early from the trial or treatment are implicitly imputed based on the analysis methods discussed below; see Appendix A.1.4 for further discussion.

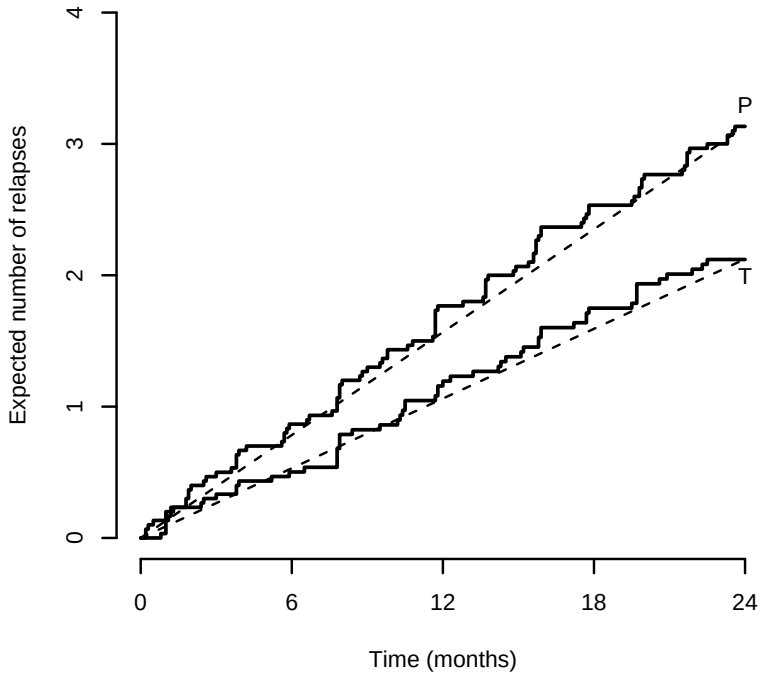
Figure 2: Relapse data for RRMS patients randomized to test (T) treatment or placebo (P). Each horizontal line corresponds to one patient with relapses indicated by dots. Two patients on T were lost to follow-up after four and six months, respectively. One patient on T discontinued treatment after nine months, but was followed-up for two years (dashed line).



For the time-to-first-relapse estimand, a Cox proportional hazards model is used to estimate the HR. A Wald test based on this model may be used for significance testing.

For the number-of-relapses estimand, NB is used for estimation and inference. The model includes log(duration on treatment) as offset variable. NB provides a consistent estimate for this estimand if the model assumptions (e.g. constant relapse rates) are appropriate (Appendix A.2.2.4). A Wald test based on NB may be used for significance testing. LWYY could be used as a sensitivity analysis as it also targets the number-of-relapses estimand, but under different assumptions. Other methods such as WLW or PWP can be used for supplementary analyses but would not be appropriate as sensitivity analyses, as these are not targeting the number-of-relapses estimand.

Figure 3: Expected number of relapses for test (T) treatment and placebo (P) against follow-up time: non-parametric Nelson-Aalen estimate (solid lines), and estimate assuming constant relapse rates (dashed lines).



4.1.3 Results

The relapse data of the acyclovir trial were manually extracted from Figure 1 in Lycke et al. (1996), and may slightly differ from the original data. Figure 2 shows relapse times for each patient in the clinical trial. Most patients (57 of 60) stayed on their randomized treatment for the planned duration of two years. Two patients on test treatment discontinued treatment after four and six months, respectively, and were not followed-up (censored). One patient on test treatment discontinued treatment after nine months, but was followed for two years. For this patient, relapse data after nine months were removed for the analysis (censored), in alignment with the hypothetical estimands.

Figure 3 shows the Nelson-Aalen estimate (Cook and Lawless, 2007) of the expected number of relapses against follow-up time, suggesting roughly constant relapse rates over time. Hence, both NB and LWYY seem appropriate

Table 2: Summary of analysis methods (RR: hazard or rate ratio; LCIL: lower 95% confidence interval limit; UCIL: upper 95% confidence interval limit). For the time-to-first-relapse estimand, Cox is the main analysis. For the number-of-relapses estimand, NB is the main analysis, with LWYY as a sensitivity analysis. WLW or PWP can be used as supplementary analyses.

Method	Estimand	RR	LCIL	UCIL	p-value
Cox	time-to-first-relapse	0.90	0.51	1.58	0.705
NB	number-of-relapses	0.67	0.43	1.05	0.082
LWYY	number-of-relapses	0.68	0.45	1.02	0.060
WLW	-	0.65	0.44	0.96	0.030
PWP	-	0.72	0.53	0.97	0.034

in terms of model assumptions (LWYY would also allow for non-constant relapse rates).

Table 2 summarizes the analysis results. For the time-to-first-relapse estimand, the Cox model estimates the hazard ratio as 0.90, which is far from being statistically significant. For the number-of-relapses estimand, NB estimates the relapse rate ratio as 0.67. This is very similar to the estimate of 0.68 obtained by LWYY. The method of moments estimate of the relapse rates within a treatment group is obtained by dividing the total number of relapses by the total follow-up time. Thus, in the placebo group, the 30 patients had 94 relapses with a total follow-up time of $60 = 2 \times 30$ years so that ARR $= 94/60 = 1.57$, i.e. patients have on average 1.57 relapses per year. In the test treatment group, 59 relapses were observed in a total follow-up of $2 \times 27 + (4 + 6 + 9)/12 = 55.6$ years (ARR $= 59/55.6 = 1.06$). The ratio of these ARR is 0.68 and coincides with the LWYY estimate. Although a relapse rate ratio of less than 0.70 could be clinically relevant, the treatment effect estimate is quite uncertain. The treatment effect in both NB and LWYY are not statistically significant.

As seen in Table 2, we obtain a much smaller treatment effect estimate for the time-to-first-relapse estimand than for the number-of-relapses estimand. This may be expected and is due to a selection effect; see (17) in Appendix A.2.2 and the corresponding discussions in Appendices A.1.5 and C.

4.1.4 Discussion

In this case study, the number-of-relapses estimand seems more appropriate to reflect the disease burden and also more sensitive to assess the treatment effect compared to the time-to-first-relapse estimand. In contrast, the time-to-first-relapse estimand ignores valuable information, and also has conceptual drawbacks (Appendix C).

We considered only hypothetical estimands in this case study. An alternative would be to use the treatment policy estimand (Section 3.1.1.1) where the effect regardless of treatment discontinuation is of interest. This estimand would require follow-up of all patients for two years regardless of treatment discontinuation. Patients who are lost to follow-up create a missing data problem and subsequent statistical analyses need to make untestable assumptions. These assumptions should be in line with the treatment policy estimand (NRC, 2010; Carpenter et al., 2013).

4.2 Chronic heart failure

The second case study is the ValHeft randomized trial of the angiotensin-receptor blocker valsartan in CHF (Cohn et al., 2001). This was a parallel group, placebo-controlled, double blind clinical trial with 5010 patients suffering from CHF of New York Heart Association (NYHA) class II, III or IV being randomly assigned to receive test treatment valsartan or placebo in a 1:1 ratio. The trial was designed with two primary endpoints: time to all-cause mortality and time to a combined endpoint of mortality and morbidity, defined as the incidence of cardiac arrest with resuscitation, HHF or receipt of intravenous inotropic or vasodilator therapy for at least four hours.

The trial results showed that overall mortality was similar in the two groups. The risk of the combined endpoint, however, was 13.2% lower with test treatment than with placebo (HR = 0.87; 97.5% confidence interval $[0.77, 0.97]$; p-value: 0.009), predominantly because of a lower number of patients hospitalized for HF: 455 (18.2%) on placebo and 346 (13.8%) on test treatment (p-value < 0.001). The comparison of both endpoints between test treatment and placebo was performed using a log-rank test.

All original analyses ignored the recurrent HHF occurring after the first event. When reanalyzing this example in the following, we assess different estimands that incorporate information on the recurrent HHF.

4.2.1 Estimands

For the definition of estimands that incorporate information on recurrent HHF we need to account for three intercurrent events: CVD, non-CVD and treatment discontinuation. We consider two estimands which differ with respect to the variable of interest (attribute B):

- Estimand 1 (HHF): Number of HHF while the patient is alive;
- Estimand 2 (HHF+CVD): Number of unfavorable events, i.e. number of HHF or CVD, up to and including the time of death.

These two estimands differ in that Estimand 2 counts CVD as an additional event for the variable of interest; see also Section 3.2.1.6.1.

For both estimands, the population (attribute A) is defined by the inclusion/exclusion criteria given in Cohn et al. (2001). In terms of attribute (C), we are interested in the treatment effect regardless of treatment discontinuation (treatment policy strategy) and while patients are alive (while-alive strategy), i.e. they did not die from any cause. The summary measure (D) is the rate ratio

$$\frac{\text{expected number of events per unit time alive in test treatment}}{\text{expected number of events per unit time alive in placebo}}.$$

4.2.2 Analysis

Various estimators are available for both estimands, dependent on the adjustment for the competing event of death which stops the recurrent event data process. LWYY (Appendix A.2.3.1) targets the two estimands of interest and may be used as the main analysis. NB with termination at the time of death could be considered as well. Conceptually, NB attempts to weight patients in both arms based on their likelihood to die. That is, patients with a larger chance to die early are upweighted relative to patients with a low probability to die during the trial. The weighting is based on both baseline covariates and information on the patient's recurrent event process prior to death. This is different from LWYY where the weighting is based only on baseline covariates. Finally, joint frailty models (JFM, Appendix A.2.2.4) could be considered as they model the recurrent morbidity and the mortality events simultaneously while accounting for the association between these

Table 3: Summary of number of HHF and CVD.

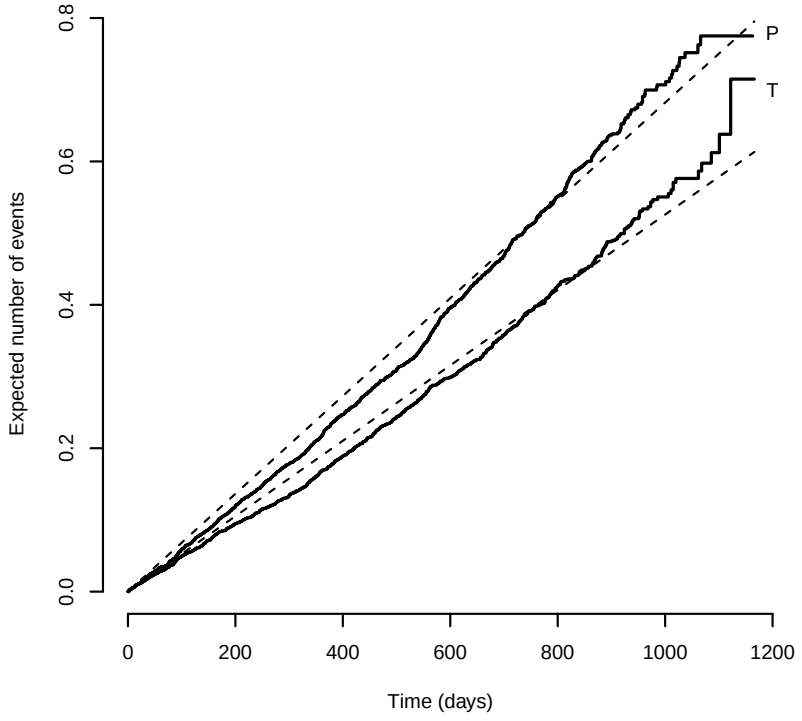
Number of HHF	Placebo $N_{\mathrm{P}} = 2499$	Test Treatment $N_{\mathrm{T}} = 2511$	Total $N_{\mathrm{TOT}} = 5010$
0	1878 (75.15%)	1974 (78.61%)	3852 (76.89 %)
1	344 (13.77%)	317 (12.62%)	661 (13.19 %)
2	146 (5.84%)	130 (5.18%)	276 (5.51 %)
3	56 (2.24%)	51 (2.03%)	107 (2.14 %)
4	36 (1.44%)	19 (0.76%)	55 (1.10 %)
5	21 (0.84%)	13 (0.52%)	34 (0.68 %)
6	5 (0.20%)	3 (0.12%)	8 (0.16 %)
7	6 (0.24%)	1 (0.04%)	7 (0.14 %)
8	3 (0.12%)	2 (0.08%)	5 (0.10 %)
9	2 (0.08%)	0 (0.00%)	2 (0.04 %)
10	1 (0.04%)	1 (0.04%)	2 (0.04 %)
12	1 (0.04%)	0 (0.00%)	1 (0.02 %)
Number of HHF	1189	922	2111
Number of CVD	419	427	846
Number of HHF or CVD	1608	1349	2957

two event processes. By linking the two processes, assumptions such as 'the higher your risk of experiencing morbidity events the higher your risk of dying' can be captured.

In the following we apply LWYY and NB to both estimands. In practice, LWYY may be chosen as the main analysis while NB could be a sensitivity analysis, or vice versa. We also use different JFM for Estimand 1, which in practice could serve as sensitivity analyses. The JFM differ in their assumptions with regard to

- the distribution of the random effect linking the recurrent event data and the survival processes (gamma versus lognormal distribution);
- the parametric form of the link between both processes. Either the same frailty term Z is used for both processes, or the processes use modified versions of the frailty term, i.e. we include Z for the recurrent event data process and Z^{α} for the death process; the parameter α is also estimated.

Figure 4: Expected number of HHF events for test (T) treatment and placebo (P) against follow-up time: non-parametric Nelson-Aalen estimate (solid lines); dashed lines are added as a visual aid to judge deviations from linearity.



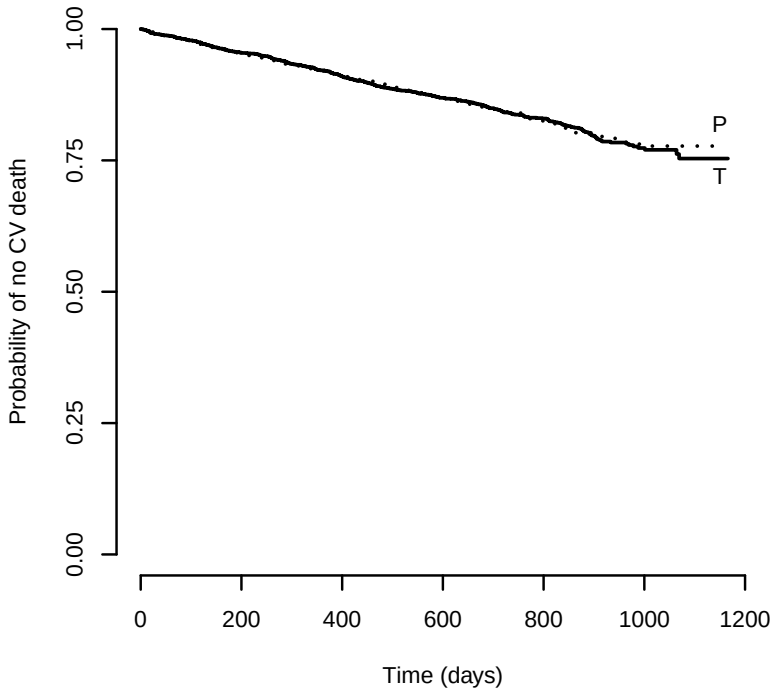
4.2.3 Results

We start with descriptive statistics and the traditional analyses for this type of data before moving to the estimation of Estimand 1 and Estimand 2.

A tabulation of the number of HHF observed in the ValHeft trial is shown in Table 3. From this table we can see that about 10% of patients had more than one HHF with one patient suffering 12 HHF. The total number of HHF is 2111 while the number of unfavorable events, i.e. HHF and CVD, is 2957. Counting only the number of first composite event of HHF and CVD amounts to 1618, so about half of the number of unfavorable events. The overall mean duration of follow-up in the ValHeft trial was 23 months with follow-up times ranging from 0 to 38 months.

In Figure 4 we display the mean cumulative function of HHF. We can see that

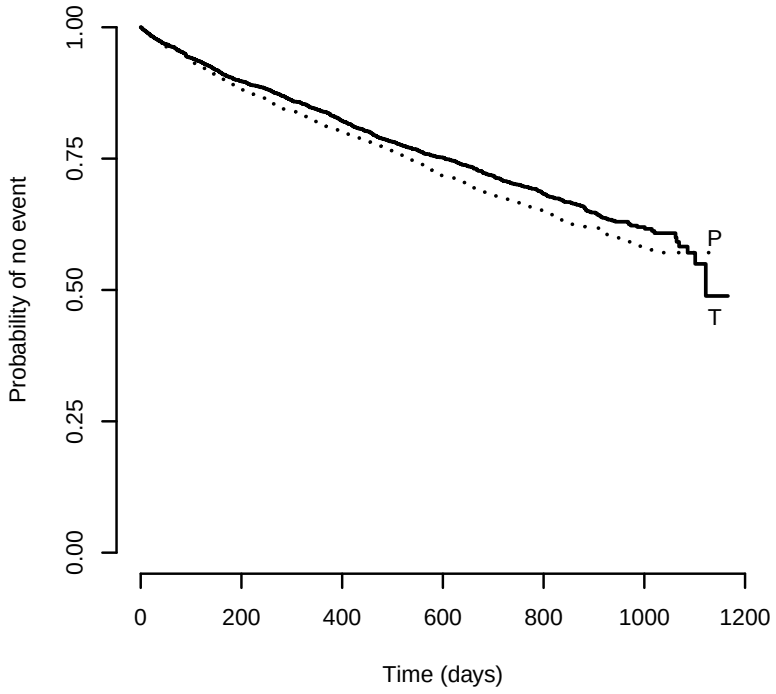
Figure 5: Kaplan-Meier estimate for time-to-CVD, for test (T) treatment and placebo (P).



more HHF occurred in the placebo group and that the rate of recurrent HHF appears to be roughly linear over time for both treatments. After 600 days of follow-up we observe an average of approximately 0.3 and 0.4 recurrent HHF per patient for test treatment and placebo, respectively. While these descriptive analyses suggest a reduction in recurrent HHF for the test treatment, as compared to placebo, this is not the case when focusing on CVD. Figure 5 shows the estimated survival functions for test treatment and placebo, respectively. The Kaplan-Meier curves overlap and likewise the corresponding log-rank test fails to show a significant difference (p-value: 0.8565). The Kaplan-Meier curves for the time to the first composite event of HHF and CVD shown in Figure 6, however, reveal a separation which is also confirmed by the log-rank test (p-value: 0.0233). Note that Kaplan-Meier curves are difficult to interpret in the context of competing risks (Appendix A.3.5).

Next we estimate the estimands laid out in Section 4.2.1. The results are shown in Table 4. All analyses considered for Estimand 1 reveal that test

Figure 6: Kaplan-Meier estimate for the time to the first composite event of HHF and CVD, for test (T) treatment and placebo (P).



treatment is superior to placebo in reducing the expected number of HHF per unit time while alive. The rate ratio is estimated to be approximately 0.77, i.e. the expected number of HHF per unit time in the test treatment group is reduced by 23% compared to the HHF rate per unit time in the placebo group. The different assumptions on the competing event process of death and the different assumptions on the relation between the recurrent HHF and death processes have negligible impact on the estimated effect size and the inference. Turning to Estimand 2, we observe a dilution of the treatment effect resulting in an event rate reduction of only 17%. This dilution was to be expected as no treatment effect was observed for all cause mortality. Including CVD in a composite variable as done for Estimand 2 thus leads to a dilution of the effect seen on HHF alone.

In addition, we also apply the unmatched win ratio approach from Dong et al. (2016). This approach compares each patient on test treatment with each patient on placebo based on hierarchically ordered endpoints:

Table 4: Summary of analysis methods (RR: rate ratio; LCIL: lower 95% confidence interval limit; UCIL: upper 95% confidence interval limit) for Estimand 1 (HHF) and Estimand 2 (HHF+CVD). Maximum likelihood estimates for α were 1 for JFM2 and JFM4

Method	Estimand	RR	LCIL	UCIL	p-value
LWYY	1	0.771	0.68	0.88	0.0001
NB	1	0.763	0.65	0.88	0.0003
JFM 1 (gamma frailty Z)	1	0.770	0.66	0.88	0.0004
JFM 2 (gamma frailty Z^α)	1	0.770	0.66	0.88	0.0005
JFM 3 (lognormal frailty Z)	1	0.771	0.66	0.88	0.0006
JFM 4 (lognormal frailty Z^α)	1	0.765	0.65	0.88	0.0007
LWYY	2	0.834	0.75	0.93	0.0016
NB	2	0.834	0.72	0.95	0.0084

1. CVD: the patient who lives longer wins;

2. HHF: If tied on CVD, then compare the rate of HHF, i.e. the number of HHF divided by the time on trial (until death or censoring due to trial end). The patient with the smaller rate wins.

Applying this approach to the ValHeft trial results in a win ratio of 1.13, 95% confidence interval $[1.03, 1.24]$ and a p-value of 0.0101 in favor of the test treatment.

4.2.4 Discussion

Traditional endpoints widely used in CHF trials do not include all relevant information on recurrent HHF and CVD. The missed opportunities with such approaches were discussed in Section 2. For the ValHeft trial, we presented descriptive statistics for the recurrent HHF and the CVD data. In addition, we looked at two estimands which differ in their variable definition. In terms of the intercurrent events, we focused on the treatment policy strategy for treatment discontinuations and on a while-alive strategy for all causes of death. We estimated these estimands using both LWYY and NB. Other analysis methods (JFM, win ratio) could be considered as sensitivity analyses. Alternative estimands as described in Section 3.2 could also be of value.

5 Efficiency comparison of recurrent event and time-to-first-event estimands

In this section we compare the efficiency of time-to-first-event and recurrent event analyses for commonly used estimands described in Section 3. We also assess the relative performance of various statistical methods for recurrent event data and discuss advantages and limitations on their use in practice.

We report the results of a comprehensive simulation study covering a wide range of practical scenarios to allow a direct quantitative comparison of the described methods, under the same conditions and using the same performance metrics. First, we consider settings where terminal events such as death are rare, e.g. in RRMS trials. Second, we investigate settings where terminal events are more common, e.g. in CHF trials with death as a terminal event.

We describe the simulation studies for each of these two settings (without and with terminal events) following the same outline: a) design of the simulation study, including its assumptions and scenarios; b) performance metrics (mean treatment effect, type I error, power) used to evaluate the statistical operating characteristics of each method; c) summary of the statistical performance of the methods, based on the simulation results; and d) conclusions.

Appendix C summarizes relevant published literature. Appendices D and E provide further technical details and results from additional simulations.

5.1 Settings without terminal event

In this section we consider clinical settings with recurrent event endpoints and where terminal events such as death are rare. The following simulations are motivated by clinical trials in patients with RRMS, where a reduction of relapses is of interest. More specifically, we simulate clinical trials to compare test against control treatment, where patients are followed for two years. Patients may discontinue their study treatment during the trial, and the intercurrent event of treatment discontinuation may be either independent of the recurrent events (non-informative treatment discontinuation) or dependent on the recurrent events (informative treatment discontinuation); see also Appendix A.1.4.

Estimands In Section 3.1 we discussed two commonly used estimands for recurrent event endpoints without terminal events, namely the treatment policy estimand and the hypothetical estimand. Considering these two estimands for each of the two types of study treatment discontinuation described above, we investigate the following four scenarios:

- Scenario 1: Hypothetical estimand; non-informative discontinuation,
- Scenario 2: Hypothetical estimand; informative discontinuation,
- Scenario 3: Treatment policy estimand; non-informative discontinuation,
- Scenario 4: Treatment policy estimand; informative discontinuation.

Analysis methods Analysis methods should ideally be chosen such that they target the estimand of interest. In the following we investigate four commonly used statistical analysis methods for recurrent event data (NB, LWYY, WLW, PWP), together with a time-to-first-event analysis (Cox model). We discuss in particular which estimand each analysis method is targeting and evaluate their operating characteristics in Section 5.1.3.

5.1.1 Design of simulation study

5.1.1.1 Primary endpoint, treatment effect and sample size The average number of relapses per year (i.e. the ARR) is a frequently used primary endpoint in RRMS trials. Following typical rates seen in RRMS trials, we set the baseline recurrent event rate $\lambda_0 = 0.5, 1.5$, corresponding to an active and a highly active disease population, respectively. We simulate recurrent events for a two-armed randomized controlled clinical trial with a planned fixed follow-up time of $T = 2$ years for each patient. We evaluate the performance of various analysis methods under varied sample size (50 to 250 patients per group, by 50) with fixed treatment effect size RR = 0.65, where the treatment effect RR is defined as the ratio of ARR in the test treatment over the control treatment.

5.1.1.2 Event-generating process For each patient, we generate recurrent event data under a homogeneous Poisson process. In RRMS, relapse

rates tend to be different between patients. We account for this overdispersion by including a patient-specific frailty factor which varies according to a gamma distribution with shape parameter $1/\theta$ and rate parameter $1/\theta$, having mean 1 and variance θ. The dispersion parameter θ measures the extent of heterogeneity in event rates among patients. We set $\theta = 0.25, 0.5, 1$, where larger values of θ correspond to larger between-patient variations with respect to relapse rates. The mixing of a homogeneous Poisson process with a gamma frailty gives a NB process which is often used to model overdispersed recurrent events. In additional simulations, we also generate recurrent event data under a non-homogeneous Poisson process. The inclusion of these simulations is motivated by time trends observed in clinical trials (Nicholas et al., 2011) and we choose a log-linear baseline intensity function to model the relapses. More details are given in Appendix D.1.

5.1.1.3 Treatment discontinuation process For the scenarios with non-informative treatment discontinuation, time to treatment discontinuation can be simulated independent of the recurrent event process. For scenarios with informative treatment discontinuation, a JFM is used to link treatment discontinuation with the recurrent event process such that patients with higher event rates are more likely to discontinue treatment. After treatment discontinuation, we continue to follow-up all patients. Each patient is observed until the trial end so that we do not assume any missing data in the simulation study. The event rate of the control treatment is assumed to be the same before and after discontinuation. However, the event rate for the test treatment is assumed to change to the control rate after treatment discontinuation. More details are given in Appendix D.2.

In total, we cover 480 different settings in the simulations, corresponding to the factorial combinations of baseline event rates ($\lambda_0 = 0.5, 1.5$), treatment effect size (RR $= 0.65, 1$), sample size per group ($n = 50$ to 250 by 50), frailty (dispersion parameter $\theta = 0.25, 0.5, 1$), event-generation process (homogeneous and non-homogeneous Poisson), treatment discontinuation (non-informative, informative) and estimand (hypothetical, treatment policy).

5.1.2 Measuring performance of methods

The performance of the various statistical methods is evaluated based on 10'000 simulated clinical trials. The following metrics are used.

1. *Mean of estimated treatment effects.* The mean estimate for the Cox model is the hazard ratio. The mean estimate for NB and LWYY is the rate ratio. For WLW and PWP, we compute both the event-specific estimates up to event 4 and the overall estimates of the treatment effect. The event-specific treatment effects are estimated by fitting the marginal and the conditional stratified Cox model for WLW and PWP, respectively. The overall treatment effect is estimated by fitting the stratified Cox model with the treatment parameter constrained to be equal across strata (Therneau and Grambsch, 2000). We denote the mean estimates by 'RR' for all five approaches when reporting the simulation results below.

2. *Type I error of the two-sided Wald test at a significance level of* $\alpha = 5\%$. The (strict) null hypothesis corresponds to an identical data generation process for both treatment and control group.

3. *Power of the Wald test to show a significant treatment effect.* The alternative is implicitly defined by the specific setting under consideration.

5.1.3 Simulation results

In the following we summarize the results of the simulation study using the performance metrics of Section 5.1.2. Because of the large number of simulations results, we show here only the results for the base case settings covering the homogeneous Poisson process with 50, 150, 250 patients per group and dispersion parameter $\theta = 0.25$. We include tables and plots for these base case settings to illustrate the key findings. The simulations results for the other settings are generally in line with the ones shown here. The complete output can be found in a separate document (Akacha et al., 2017).

5.1.3.1 Mean estimate of treatment effects

5.1.3.1.1 Estimand value For each of the four Scenarios $1 - 4$ introduced above, the estimand has a true numerical value. This value is unknown in actual trials, but can be calculated analytically for our simulation study; see Appendix D.3 for the derivations. The second column in Table 5 shows the four numerical estimand values, thus providing a reference for comparison. We can make the following observations.

Table 5: Settings without terminal event (Estimand vs Estimate): Numerical values of hypothetical estimand and treatment policy estimand under four scenarios. The ratio of the target of estimation (Estimate) for each of the five analysis methods over the corresponding estimand value (Estimand) is also shown. 'Estimand' values are calculated analytically, 'Estimate' values are calculated based on a simulated data set with 100'000 patients with $RR = 0.65$, $\theta = 0.25$, and $\lambda_0 = 0.5, 1.5$. Estimate/Estimand values larger (smaller) than 1 correspond to overestimation (underestimation).

	Estimand value	Estimate/Estimand		
		Method	$\lambda_0 = 0.5$	$\lambda_0 = 1.5$
Scenario 1: Non-informative (Hypothetical)	0.65	Cox	1.023	1.055
		NB	0.995	0.994
		LWYY	0.995	0.994
		WLW	0.886	0.895
		PWP	1.032	1.075
Scenario 2: Informative (Hypothetical)	0.65	Cox	1.043	1.071
		NB	1.017	1.009
		LWYY	1.020	1.014
		WLW	0.922	0.912
		PWP	1.051	1.082
Scenario 3: Non-informative (Treatment policy)	0.685	Cox	1.013	1.029
		NB	0.996	0.993
		LWYY	0.999	1.000
		WLW	0.892	0.893
		PWP	1.032	1.067
Scenario 4: Informative (Treatment policy)	0.7002	Cox	1.000	1.007
		NB	1.001	0.995
		LWYY	1.005	1.014
		WLW	0.894	0.887
		PWP	1.034	1.055

- The numerical value for the treatment policy estimand with informative treatment discontinuation is closer to 1 than with non-informative treatment discontinuation. This is expected because patients with higher frailty and hence more events have their treatment stopped earlier due to the dependence. This reduces the apparent treatment effect, moving it closer to 1. This is also seen from the analytic formula for Scenario 4 (Appendix D.3) where larger values of the dispersion parameter θ lead to estimand values closer to 1.

- The numerical value for the hypothetical estimands is 0.65 regardless of the type of treatment discontinuation because these estimands measure the effect as if the treatment had continued.

- The numerical values for the treatment policy estimands are closer to 1 than for the hypothetical estimands because the effect is diluted under treatment policy. More specifically, all patients are followed until the end of trial and some of the patients on test treatment, who discontinue their medication, then behave like patients on control.

5.1.3.1.2 Target of estimation for analysis methods In our simulation study we consider one time-to-first-event analysis method (Cox) and four recurrent event analysis methods (NB, LWYY, WLW, PWP). The target of estimation for these methods is the treatment effect estimate, i.e. the hazard ratio (Cox) or the rate ratio (NB, LWYY, WLW, PWP), obtained from a very large (infinite) number of patients. We approximate this value by simulating a single trial with 100'000 patients in total. The third column in Table 5 shows the ratio of the target of estimation ('Estimate') over the corresponding estimand value ('Estimand') for each of the four scenarios and five analysis methods. A ratio Estimate/Estimand close to 1 suggests that the analysis method targets that estimand. A ratio Estimate/Estimand larger (smaller) than 1 suggests that the analysis method underestimates (overestimates) the treatment effect. The results in Table 5 can be interpreted as follows.

- NB and LWYY give consistent mean effects for the treatment policy estimand (Scenarios 3 and 4) since they exactly target the treatment effect defined as ratio of mean event rate for test treatment over control treatment under a fixed follow-up time. Thus, both NB and LWYY can be considered as suitable analysis methods for the treatment policy estimand.

- For the hypothetical estimand with non-informative treatment discontinuation (Scenario 1), both NB and LWYY give again consistent mean effects and therefore can be considered as suitable analysis methods.

- LWYY is misspecified under informative treatment discontinuation with the hypothetical estimand (Scenario 2) since its model assumption of independent censoring is violated. Hence it gives an inconsistent estimate under Scenario 2 and the difference from the numerical estimand value will increase as the variance of the frailty term increases; see also the complete results in Akacha et al. (2017). NB is also not a correct

model under this scenario, and its performance depends on the informative treatment discontinuation process. However, the difference from the numerical estimand value will typically be smaller than for LWYY under a Poisson-gamma process, especially when the variance of the frailty term is large. One may argue that the treatment effect is typically diluted under plausible informative censoring mechanisms (as seen in the simulations), so that both NB and LWYY could be used as conservative analysis methods.

- WLW and PWP are not appropriate since their target values are different from the estimand values, i.e. they give inconsistent estimates under all scenarios. PWP systematically underestimates the treatment effect, while WLW systematically overestimates it.

- The Cox model underestimates the treatment effect for Scenarios 1, 2 and 3, but not for Scenario 4.

5.1.3.1.3 Mean estimates for typical sample sizes Table 6 presents the treatment effect estimates (i.e. hazard ratios or rate ratios) based on 10'000 simulated trials for the five statistical approaches under the four Scenarios $1 - 4$, with varying sample sizes per group and all other parameters being the same as in Table 5. In particular, the true estimand values under the four scenarios are the same as given in Table 5. For WLW and PWP, only the overall estimates are presented; see Appendix D.4 for the event-specific treatment effect estimates. The Monte Carlo standard error of the 10'000 simulations is about 0.0046. Thus, the asymptotic 95% confidence interval for e.g. RR = 0.65 is (0.641, 0.659).

Differences in means seen in Table 6 are mainly driven by differences between the underlying estimands as given in Table 5. Hence, analysis methods targeting different estimands should be compared with caution. Note that the true estimand value rather than the treatment effect parameter value used in the simulations (RR = 0.65) is the appropriate reference. It seems that NB and LWYY converge to the true estimand value for large n (e.g. $n \geq 150$), but none of the other methods, which again verifies the findings in Table 5.

Table 7 presents the treatment effect estimates when there is no treatment effect (RR = 1) for a baseline recurrent event rate $\lambda_0 = 0.5$ and dispersion parameter $\theta = 0.25$. Again, only the overall estimates are presented here for

Table 6: Settings without terminal event: Mean treatment effect estimates under four scenarios based on 10'000 clinical trial simulations, $RR = 0.65$, $\theta = 0.25$, $\lambda_0 = 0.5, 1.5$.

		$\lambda_0 = 0.5$			$\lambda_0 = 1.5$		
	Method	$n = 50$	$n = 150$	$n = 250$	$n = 50$	$n = 150$	$n = 250$
Scenario 1: Non-informative	Cox	0.7	0.68	0.675	0.705	0.694	0.692
(Hypothetical)	NB	0.672	0.656	0.653	0.657	0.652	0.652
Estimand value: 0.65	LWYY	0.671	0.656	0.653	0.657	0.652	0.652
	WLW	0.615	0.591	0.586	0.602	0.591	0.59
	PWP	0.69	0.678	0.676	0.704	0.701	0.702
Scenario 2: Informative	Cox	0.705	0.687	0.681	0.709	0.698	0.696
(Hypothetical)	NB	0.679	0.666	0.661	0.665	0.659	0.658
Estimand value: 0.65	LWYY	0.681	0.668	0.663	0.668	0.663	0.661
	WLW	0.628	0.607	0.599	0.609	0.597	0.594
	PWP	0.697	0.687	0.682	0.709	0.706	0.705
Scenario 3: Non-informative	Cox	0.726	0.708	0.703	0.723	0.713	0.711
(Treatment policy)	NB	0.705	0.691	0.688	0.692	0.687	0.686
Estimand value: 0.685	LWYY	0.706	0.692	0.689	0.695	0.69	0.691
	WLW	0.646	0.624	0.619	0.631	0.62	0.619
	PWP	0.724	0.713	0.711	0.736	0.733	0.734
Scenario 4: Informative	Cox	0.729	0.713	0.709	0.724	0.714	0.712
(Treatment policy)	NB	0.718	0.706	0.702	0.707	0.702	0.701
Estimand value: 0.7002	LWYY	0.721	0.709	0.706	0.717	0.714	0.714
	WLW	0.658	0.638	0.633	0.64	0.63	0.627
	PWP	0.737	0.729	0.726	0.746	0.744	0.743

WLW and PWP. All five approaches give hazard and rate ratio estimates close to 1 under all scenarios (actually the estimates are slightly above 1 in all cases being investigated).

5.1.3.2 Type I error rate Table 7 also includes the type I error rates when there is no treatment effect (RR = 1). All methods control the type I error rate at the significance level $\alpha = 0.05$ within the simulation error for moderate to large sample size under all four scenarios. For small sample sizes ($n = 50$) we note a moderate type I error rate inflation when using the two-sided Wald test, except for the Cox model which seems to control the type I error rate throughout.

5.1.3.3 Power Figure 7 presents the power of the five statistical approaches for different sample sizes under the same four scenarios and parameter configurations as in Tables 5 and 6.

Table 7: Settings without terminal event: Mean treatment effect estimates and type I error rate under four scenarios based on 10'000 clinical trial simulations, $RR = 1$, $\theta = 0.25$, $\lambda_0 = 0.5$.

		$n = 50$		$n = 150$		$n = 250$	
	Method	RR	Type I error	RR	Type I error	RR	Type I error
Scenario 1: Non-informative	Cox	1.036	0.047	1.013	0.048	1.007	0.047
(Hypothetical)	NB	1.028	0.054	1.008	0.053	1.005	0.049
	LWYY	1.029	0.058	1.008	0.053	1.005	0.049
	WLW	1.051	0.056	1.016	0.052	1.009	0.05
	PWP	1.024	0.055	1.007	0.053	1.004	0.049
Scenario 2: Informative	Cox	1.052	0.047	1.009	0.061	1.007	0.045
(Hypothetical)	NB	1.043	0.067	1.008	0.054	1.005	0.051
	LWYY	1.043	0.069	1.008	0.056	1.005	0.052
	WLW	1.073	0.066	1.014	0.057	1.009	0.046
	PWP	1.036	0.066	1.006	0.058	1.004	0.051
Scenario 3: Non-informative	Cox	1.032	0.048	1.012	0.05	1.006	0.046
(Treatment policy)	NB	1.026	0.053	1.008	0.056	1.004	0.048
	LWYY	1.026	0.055	1.008	0.056	1.004	0.047
	WLW	1.046	0.054	1.015	0.051	1.008	0.048
	PWP	1.022	0.054	1.006	0.055	1.003	0.048
Scenario 4: Informative	Cox	1.032	0.05	1.011	0.052	1.006	0.05
(Treatment policy)	NB	1.025	0.056	1.008	0.053	1.003	0.05
	LWYY	1.025	0.058	1.008	0.053	1.003	0.049
	WLW	1.045	0.057	1.015	0.053	1.007	0.051
	PWP	1.021	0.057	1.007	0.053	1.002	0.048

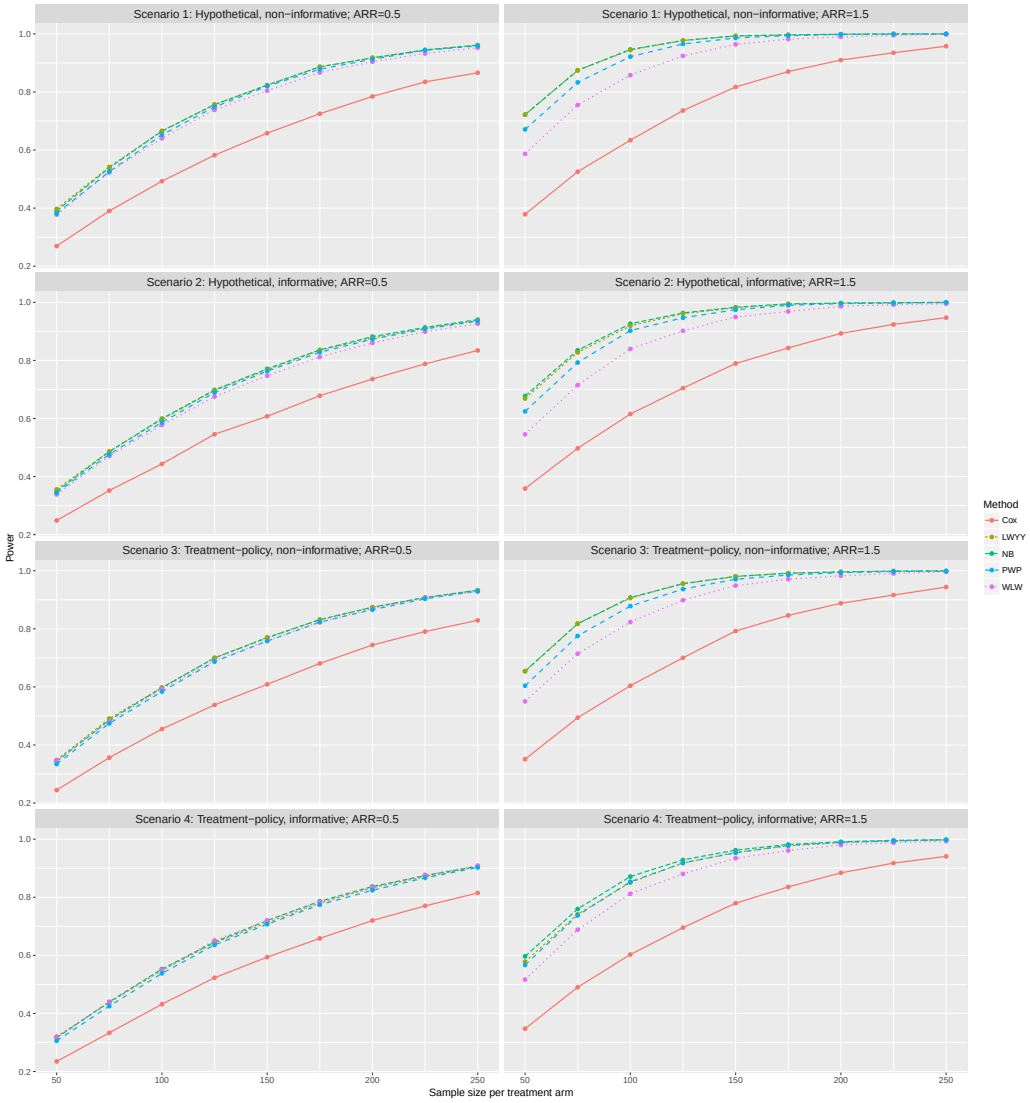
Clearly, the time-to-first-event analysis (Cox model) has considerably less power than any of the recurrent event analysis methods for all scenarios considered. Additionally, the power loss for the Cox regression is more pronounced in settings with high baseline event rates (e.g. high relapse rate in RRMS trials).

With respect to the four statistical approaches for recurrent event data, power is similar under all four scenarios in settings with low event rates ($\lambda_0 = 0.5$). In settings with high event rates ($\lambda_0 = 1.5$), NB and LWYY perform similarly, and better than both WLW and PWP.

5.1.4 Conclusions

The simulations show that a time-to-first-event analysis typically provides considerably less power than recurrent event analyses. Thus, the recurrent event methods are shown to be more efficient than a time-to-event analysis.

Figure 7: Setting without terminal event: Statistical power at varied sample size under four scenarios based on 10'000 clinical trial simulations, $RR = 0.65$, $\theta = 0.25$, $\lambda_0 = 0.5, 1.5$.



Overall, NB seems to perform best in the simulations and can be considered as a suitable main estimator. NB targets both the hypothetical and the treatment policy estimand, and also performs well under informative treatment discontinuation. LWYY could be used for a sensitivity analysis (or for the main analysis) since it targets the same estimands but makes different assumptions. WLW and PWP do not target the estimands of interest, and hence should be used with caution, possibly for a supplementary analysis.

The investigations emphasize the importance of first specifying the estimand of interest before selecting an appropriate statistical approach. Comparing analysis methods targeting different estimands should be done with great care.

The simulations were motivated by trials in RRMS. However, the assumptions and scenarios considered in the simulations, as well as the performance metrics used to summarize the results apply more broadly. The results and conclusions shown here may thus be extended to a wider range of therapeutic areas with recurrent event endpoints, where the rate of terminal events such as death is low, e.g. asthma and COPD (recurrent exacerbations), or epilepsy (recurrent seizures).

5.2 Settings with terminal event

We now investigate clinical settings where terminal events (e.g. death) are common, motivated by clinical trials in HF with preserved ejection fraction (HFpEF). More specifically, we simulate clinical trials to compare test versus control treatment with a total duration of five years and with patients being recruited uniformly over a period of three years. Hence, the minimal follow-up time is two years and the maximum follow-up time is five years.

Estimands The following three intercurrent events can be expected in typical HFpEF trials: CVD, non-CVD, and treatment discontinuation. Treatment discontinuation may be either unrelated (non-informative treatment discontinuation) or related to the recurrent HHF (informative treatment discontinuation). As discussed in Section 3.2.1.6.1, there are two main estimands for recurrent event endpoints with terminal events: Estimand 1 (HHF) which focuses only on recurrent HHF and Estimand 2 (HHF+CVD) which includes CVD as an additional event. These estimands handle the intercurrent events as follows:

- CVD: while-alive strategy for Estimand 1, composite strategy (part of the variable definition) for Estimand 2;
- Non-CVD: while-alive strategy for both estimands;
- Treatment discontinuation: treatment policy strategy for both estimands, as often done in long-term outcome trials.

Considering these two estimands for each of the two types of study treatment discontinuation described above, we investigate the following four scenarios:

- Scenario 1: Estimand 1 (HHF), non-informative discontinuation.
- Scenario 2: Estimand 1 (HHF), informative discontinuation.
- Scenario 3: Estimand 2 (HHF+CVD), non-informative discontinuation.
- Scenario 4: Estimand 2 (HHF+CVD), informative discontinuation.

Analysis methods The same five statistical methods considered for the setting without terminal event are of interest here as well: one time-to-first-event analysis (Cox) and four recurrent event analyses (NB, LWYY, WLW, PWP). We also considered the inclusion of a JFM as implemented in the function *frailtyPenal* of the R package *frailtypack* (Rondeau et al., 2012). Unfortunately, this implementation led to a number of computational problems so that results for this model could not be included; see Appendix E.1 for more details.

5.2.1 Design of simulation study

5.2.1.1 Primary endpoint and event rates Two types of endpoints are considered in the two estimands respectively: a recurrent endpoint that focuses only on the recurrent HHF in Estimand 1 and a recurrent composite endpoint that includes CVD as an additional event in Estimand 2.

We choose the event rate λ_{CV} for CVD such that an observed annualized event rate of 4% is obtained, motivated by the value of 3.9% observed in both the CHARM-Preserved trial and the BNP stratum in the TOPCAT trial. Also, we choose the event rate λ_{HHF} for repeated hospitalizations such

that an observed annualized control event rate of first composite event of 9% (events per patient-year) is obtained, similar to what has been observed in the CHARM-Preserved trial (9.1%) and in the BNP Stratum of the TOPCAT trial (8.5%).

5.2.1.2 Treatment effect and sample size We investigated in detail several base case situations similar to a typical HFpEF trial. For these we vary the treatment effect on recurrent HHF (rate ratio RR_{HHF}= 0.6, 0.7, 0.8, 0.9, 1.0) and on CVD (hazard ratio HR_{CV}= 0.6, 0.7, 0.8, 0.9, 1.0), while keeping the total sample size fixed at $N = 4'350$, i.e. $2'175$ patients per arm. We choose this sample size as it gives approximately 90% power to show a treatment effect for the recurrent composite endpoint with LWYY for $RR_{HHF} = 0.7$ and $HR_{CV} = 0.8$. We also vary the sample size ($N = 1'500, 2'000, \ldots, 5'000$) while keeping $RR_{HHF} = 0.7$ and $HR_{CV} = 0.8$ fixed.

5.2.1.3 Event-generating process Similar to the setting without terminal event, patient-specific frailties Z_i for the rate of recurrent hospitalizations are assumed to follow a gamma distribution with mean 1 and variance θ. To determine patient-specific frailties U_i for the rate of CVD, a joint frailty model (Rogers et al., 2016) is used, assuming $U_i = Z_i^{\alpha}$. The frailties are then correlated, but not identical, which seems clinically plausible. We choose $\alpha = 0.75$, a value similar to what has been observed when applying the joint frailty model to previous HF trials (Rogers et al., 2016). This leads to frailties for CVD having smaller influence than the ones for HHF, which seems plausible. We set the variance $\theta = 5.7$, as this leads to an observed ratio of number of total events to number of first events around 1.8. Similar ratios have been observed across a number of previous HF trials (Anker and McMurray, 2012). Conditional on the patient-specific frailty Z_i, time-to-next-hospitalization is exponentially distributed with rate λ_{HHF} and conditional on U_i, time-to-CVD is exponentially distributed with rate λ_{CV}. Time-to-non-CVD is independently simulated as an exponential process without patient-specific frailty. The event rate λ_{NCV} is chosen such that the proportion of non-CVD of all deaths is around 30%.

5.2.1.4 Treatment discontinuation process In the simulations, we vary whether treatment discontinuation is independent of both treatment

and HHF or depends on HHF, and thus, through the effect on hospitalizations, also indirectly on treatment. In the independent case it is simulated as an exponential process without patient-specific frailty. The treatment discontinuation event rate λ_{TD} is chosen such that that the rate of annual treatment discontinuation is 5 %. In the case of treatment discontinuation depending on hospitalizations, it is assumed that treatment is only discontinued directly after a hospitalization event. This is clinically plausible, as patients and their doctors might take recurrent HHF as a non-response to treatment. The probability of discontinuing after each HHF is chosen as 0%, 5%, 10%, 15% and 20%. Both for informative (dependent) and non-informative treatment discontinuation patients are still followed up for events after discontinuation, and the event rate is the same as the control event rate after stopping treatment. Non-CVD is treated as a censoring event. More details on the simulation set-up, including exact values of the parameters, are provided in Appendix E.2.

5.2.1.5 Variations of base case situations We also consider variations of the base case situations (for non-informative treatment discontinuation only). These correspond to alternative settings that have some clinical plausibility for the HF indication. In these settings, only one aspect is varied at a time. However, where necessary, the control event rates λ_{HHF} and λ_{CV} as well as the frailty variance θ are adapted so that the observed control annualized CVD rate is 4%, the observed control annualized rate of first composite event is 9% and the observed ratio of the number of all events and the number of first events is 1.8 (as for base case situations). The following aspects are investigated in these variations of the base case.

- *Inter-event Weibull:* The time-to-next-hospitalization as well as time-to-CVD are assumed to follow a Weibull distribution instead of an exponential distribution. The Weibull shape parameter is chosen as $\gamma = 0.75$, which leads to an increased hazard shortly after a hospitalization and stabilizes after some time, reflecting that a patient might still be in a vulnerable state shortly after an event.

- *Autoregressive event rate:* It is assumed that the rate of further HHF and CVD is multiplied after each HHF by an additional factor. This would reflect some permanent deterioration in the patient's health after

each hospitalization. The multiplicative factor is chosen as 1.1 and 1.2, respectively.

- *Detrimental CVD effect:* Treatment is assumed to have a positive effect on HHF, but a detrimental effect on CVD, i.e. we consider settings with $RR_{HHF} < 1$ and $HR_{CV} > 1$.

- *Frailty correlation:* In the relation of the frailty terms ($U = Z^{\alpha}$), α is set to 0.5 (1) instead of 0.75, leading to a lower (higher) correlation between HHF and CVD.

5.2.2 Measuring performance of methods

We evaluate the performance of the various statistical methods based on 10'000 simulated clinical trials, using the same three metrics as in Section 5.1.2: mean of estimated treatment effects, type I error rate, and power. For the type I error rate evaluations, the (strict) null hypothesis corresponding to an identical data generation process for both treatment and control groups is again of main interest. For Estimand 1 we also consider the null hypothesis that there is no treatment effect on HHF ($RR_{HHF} = 1$), but possibly a treatment effect on CVD ($HR_{CV} \neq 1$).

5.2.3 Simulation results

We only present a subset of tables and figures in this section to illustrate key findings. The simulations results for the other settings are generally in line with the ones shown here. The complete output can be found in Appendix E.5.

5.2.3.1 Mean estimate of treatment effects

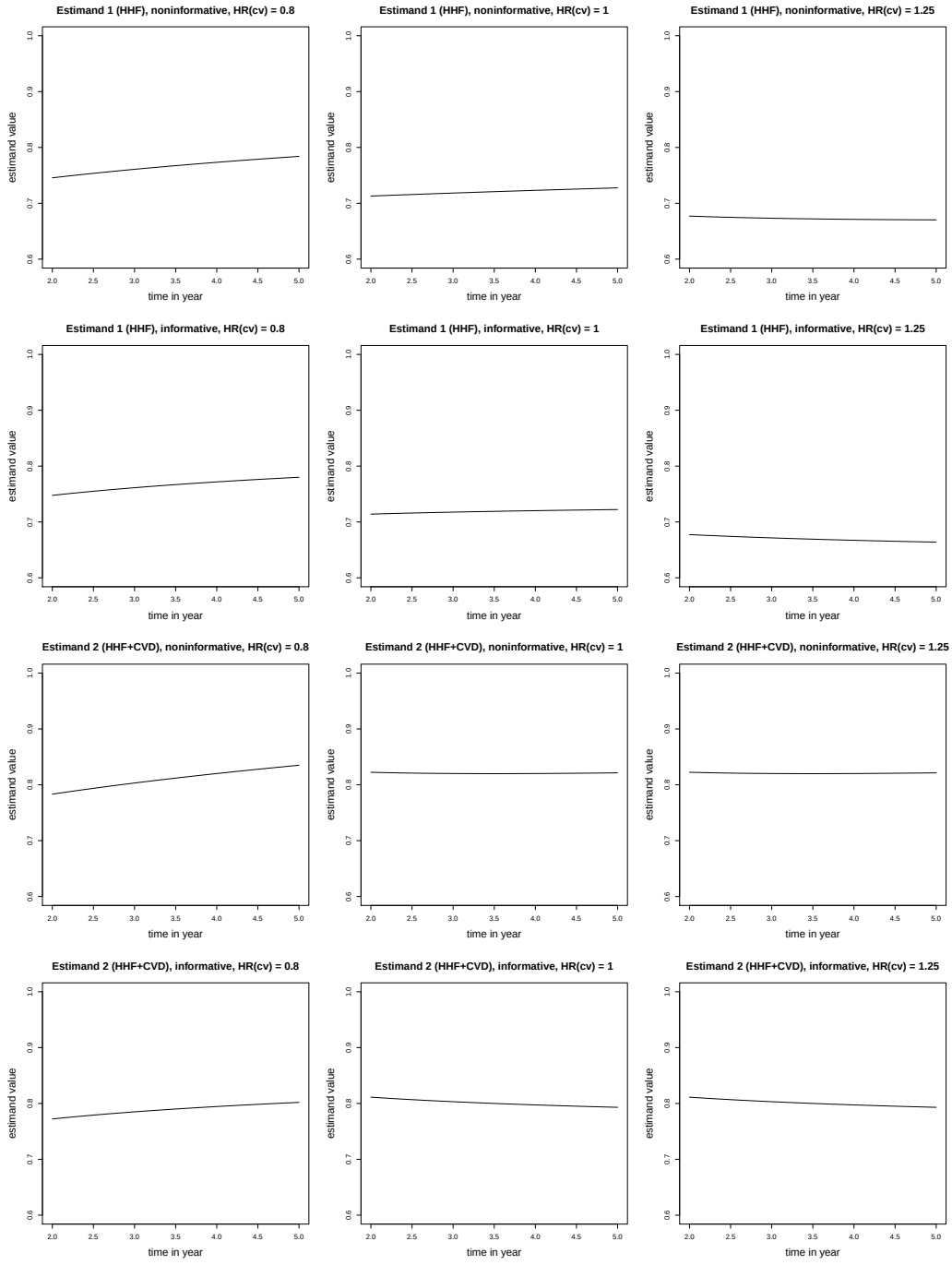
5.2.3.1.1 Estimand value The true estimand value is of interest for each of the four scenarios, that are the combination of Estimands 1 and 2 with non-informative and informative treatment discontinuation. Analytical derivation of the true estimand value was not feasible for flexible follow-up times and arbitrary correlation between the frailty of HHF and CVD. Hence,

Table 8: Settings with terminal event (Estimand vs Estimate): True estimand values under four scenarios, as well as the treatment effects estimates from five approaches. Simulated data for 100'000 patients are generated with $RR_{HHF} = 0.7$, $HR_{CV} = 0.8; 1.0; 1.25$.

	Estimand value			Method	Estimates		
HR_{CV}	0.8	1.0	1.25		0.8	1.0	1.25
Scenario 1: Non-informative				Cox	0.841	0.799	0.782
Estimand 1 (HHF)	0.783	0.722	0.688	NB	0.752	0.700	0.684
				LWYY	0.784	0.722	0.687
				WLW	0.789	0.731	0.702
				PWP	0.849	0.811	0.791
Scenario 2: Informative				Cox	0.822	0.789	0.769
Estimand 1 (HHF)	0.770	0.728	0.686	NB	0.741	0.704	0.679
				LWYY	0.771	0.727	0.684
				WLW	0.774	0.731	0.692
				PWP	0.843	0.817	0.787
Scenario 3: Non-informative				Cox	0.875	0.898	0.935
Estimand 2 (HHF+CVD)	0.809	0.806	0.822	NB	0.766	0.814	0.885
				LWYY	0.809	0.806	0.821
				WLW	0.817	0.818	0.839
				PWP	0.878	0.907	0.944
Scenario 4: Informative				Cox	0.859	0.881	0.929
Estimand 2 (HHF+CVD)	0.800	0.800	0.820	NB	0.767	0.797	0.889
				LWYY	0.801	0.800	0.819
				WLW	0.807	0.806	0.831
				PWP	0.879	0.900	0.944

we consider here only the setting where all the patients have a fixed follow-up time of 3.5 years and use a correlation between the frailty of HHF and CVD of 1. The analytical estimand values are derived in Appendix E.3. The estimand values are shown in Table 8 and are consistent with the values obtained analytically (Appendix E.3). Table 8 shows that for Estimand 1, the treatment effect is stronger as HR_{CV} increases, since the worst patient outcome of CVD precludes all future HHF for that patient, while patients in a less serious condition may remain on trial and experience many HHF, which means this estimand favors the treatment with the higher CVD rate. In contrast, the treatment effect is fairly constant for Estimand 2. The estimand values for informative and non-informative treatment discontinuation are comparable. As mentioned above, the true estimand value was calculated for a fixed follow-up time of 3.5 years. Values would change if instead a fixed follow-up time of 2 or 4 years would have been used, as illustrated by Figure 8.

Figure 8: True estimand value for varying fixed follow-up time.



5.2.3.1.2 Target of estimation for analysis methods The target of estimation for the time-to-first-event analysis method (Cox) and the four recurrent event analysis methods (NB, LWYY, WLW, PWP) is the estimate obtained from a very large (infinite) number of patients. This value can approximately be computed from one simulated dataset containing 100000 patients, with fixed follow-up time of 3.5 years. Table 8 shows the target of estimation for each scenario and each analysis method. These results can be interpreted as follows.

- LWYY targets the estimand of interest for all four scenarios. LWYY is based on the principle of averaging across patients first and then comparing between treatments, which is aligned with the estimands; see Section 3.2.1.6.1. Therefore LWYY can be considered as the main analysis.
- Cox, NB, WLW and PWP seem not appropriate as their target values are different from the estimand values for all scenarios.

5.2.3.1.3 Mean estimates of analysis methods for typical sample sizes For understanding the target of estimation of the five analysis methods, an idealized situation was considered with extremely large sample sizes, and fixed follow-up time. Here we investigate the treatment effect estimates for the different methods with realistic sample sizes ($N = 4350$), a more realistic correlation for the frailty ($\gamma = 0.75$), and we also assume that the total duration of the trial is 5 years with patients being recruited uniformly over a period of 3 years (flexible follow-up time). In such cases where the follow-up time varies among patients, estimators essentially target an average of different estimands (corresponding to different fixed follow-up times). As illustrated above (Figure 8), the impact of different averages due to different trial recruitment characteristics (e.g. fast or slow recruitment) would typically be small.

Table 9 shows the mean treatment effect estimates for the five analysis methods, for Estimands 1 and 2, and the case of non-informative treatment discontinuation. Results for the event-specific estimates for WLW and PWP can be found in Appendix E.4. When comparing the means, a similar pattern as in Table 8 is seen. Some additional findings are:

Table 9: Settings with terminal event: Mean treatment effect estimates for Estimands 1 and 2 with non-informative treatment discontinuation based on 10'000 clinical trial simulations, sample size $N = 4350$.

Endpoint	RR_{HHF}	Method	$HR_{CV} = 0.6$	$HR_{CV} = 0.8$	$HR_{CV} = 1.0$
Estimand 1 (HHF)	0.6	Cox	0.780	0.755	0.731
		NB	0.659	0.631	0.607
		LWYY	0.704	0.664	0.628
		WLW	0.719	0.680	0.647
		PWP	0.793	0.767	0.744
	0.8	Cox	0.928	0.902	0.878
		NB	0.866	0.834	0.805
		LWYY	0.914	0.863	0.817
		WLW	0.916	0.872	0.831
		PWP	0.931	0.907	0.883
	1.0	Cox	1.055	1.030	1.004
		NB	1.075	1.040	1.006
		LWYY	1.124	1.062	1.006
		WLW	1.101	1.051	1.005
		PWP	1.050	1.025	1.002
Estimand 2 (HHF+CVD)	0.6	Cox	0.770	0.811	0.851
		NB	0.624	0.676	0.730
		LWYY	0.700	0.714	0.728
		WLW	0.712	0.730	0.748
		PWP	0.782	0.819	0.855
	0.8	Cox	0.859	0.896	0.932
		NB	0.759	0.813	0.868
		LWYY	0.853	0.859	0.866
		WLW	0.853	0.867	0.880
		PWP	0.867	0.901	0.933
	1.0	Cox	0.936	0.971	1.003
		NB	0.894	0.950	1.005
		LWYY	1.006	1.005	1.004
		WLW	0.985	0.995	1.004
		PWP	0.941	0.971	1.001

- All methods provide estimates around 1 under the global null hypothesis ($HR_{CV} = RR_{HHF} = 1$), in line with our expectations.
- For Estimand 1, the null hypothesis of no treatment effect on HHF ($RR_{HHF} = 1$) but a treatment effect on CVD ($HR_{CV} \neq 1$) is also of interest. The treatment effect estimates are monotonically increasing with increasing effects on CVD (smaller HR_{CV}). Hence if a treatment reduces CVD, especially in severely ill patients who subsequently experience many hospitalizations, the treatment appears to be less effective.

Table 10: Settings with terminal event: Mean treatment effect estimates for Estimands 1 and 2 with informative treatment discontinuation based on 10'000 clinical trial simulations, $RR_{HHF} = 0.7$ and sample size N = 4350 (Trt.Disc. = Probability of treatment discontinuation).

Endpoint	HR_{CV}	Method	Trt.Disc. = 0 %	Trt.Disc. = 10 %	Trt.Disc. = 20 %
Estimand 1 (HHF)	0.6	Cox	0.843	0.843	0.843
		NB	0.741	0.763	0.781
		LWYY	0.789	0.804	0.817
		WLW	0.800	0.808	0.816
		PWP	0.848	0.861	0.873
	0.8	Cox	0.819	0.818	0.819
		NB	0.713	0.735	0.753
		LWYY	0.743	0.762	0.778
		WLW	0.759	0.769	0.779
		PWP	0.825	0.838	0.851
	1.0	Cox	0.796	0.795	0.796
		NB	0.688	0.709	0.728
		LWYY	0.704	0.726	0.744
		WLW	0.723	0.735	0.746
		PWP	0.802	0.817	0.830
Estimand 2 (HHF+CVD)	0.6	Cox	0.800	0.800	0.800
		NB	0.669	0.690	0.708
		LWYY	0.754	0.770	0.784
		WLW	0.763	0.772	0.781
		PWP	0.807	0.822	0.835
	0.8	Cox	0.843	0.843	0.843
		NB	0.727	0.745	0.761
		LWYY	0.768	0.784	0.798
		WLW	0.783	0.791	0.799
		PWP	0.847	0.859	0.870
	1.0	Cox	0.884	0.884	0.884
		NB	0.785	0.800	0.813
		LWYY	0.782	0.798	0.811
		WLW	0.802	0.809	0.817
		PWP	0.886	0.895	0.904

Table 10 gives the mean treatment effect estimates of the five analysis methods for Estimands 1 and 2 for the case of informative treatment discontinuation. One sees that the mean estimates of treatment effect for the recurrent event methods are getting closer to 1 with an increasing rate of treatment discontinuation, while the mean estimates for the Cox model are unaffected. The other patterns observed for the mean treatment effect in case of non-informative treatment discontinuation are also observed for the informative treatment discontinuation case.

Table 11: Settings with terminal event: Mean treatment effect estimates and type I error rates for Estimands 1 and 2 with non-informative treatment discontinuation based on 10'000 clinical trial simulations, $RR_{HHF} = 1$ and sample size $N = 4350$.

Endpoint	HR_{CV}	Method	Estimate	Type I error
Estimand 1 (HHF)	0.6	Cox	1.055	0.115
		NB	1.075	0.120
		LWYY	1.124	0.254
		WLW	1.101	0.207
		PWP	1.050	0.142
	0.8	Cox	1.030	0.066
		NB	1.040	0.066
		LWYY	1.062	0.098
		WLW	1.051	0.088
		PWP	1.025	0.071
	1.0	Cox	1.004	0.048
		NB	1.006	0.050
		LWYY	1.006	0.046
		WLW	1.005	0.049
		PWP	1.002	0.050
Estimand 2 (HHF+CVD)	1.0	Cox	1.003	0.046
		NB	1.005	0.046
		LWYY	1.004	0.046
		WLW	1.004	0.050
		PWP	1.001	0.049

5.2.3.2 Type I error rate For non-informative treatment discontinuation, Table 11 shows type I error rates and mean treatment effect estimates under both the global null hypothesis ($HR_{CV} = RR_{HHF} = 1$) and the local null hypothesis ($RR_{HHF} = 1$, $HR_{CV} \neq 1$)).

- All methods provide control of the type I error rate under the global null hypothesis, with point estimates very close to 1.
- For Estimand 1 and a larger treatment effect on CVD (smaller HR_{CV}), the type I error rates of all considered methods increase and exceed the desired two-sided 5% significance level. The type I error inflation is largest for LWYY, followed by WLW, PWP, NB and Cox. The main reason for the type I error inflation is the fact that the point estimates become larger than 1 with decreasing HR_{CV}. As we use two-sided tests this then leads to an increased number of false rejections of the null hypothesis, but favoring the control treatment with no effect on CVD.

Table 12: Settings with terminal event: Mean treatment effect estimates and type I error rates for Estimands 1 and 2 with informative treatment discontinuation based on 10'000 clinical trial simulations, $RR_{HHF} = 1$ and sample size $N = 4350$ (Trt.Disc. = Probability of treatment discontinuation).

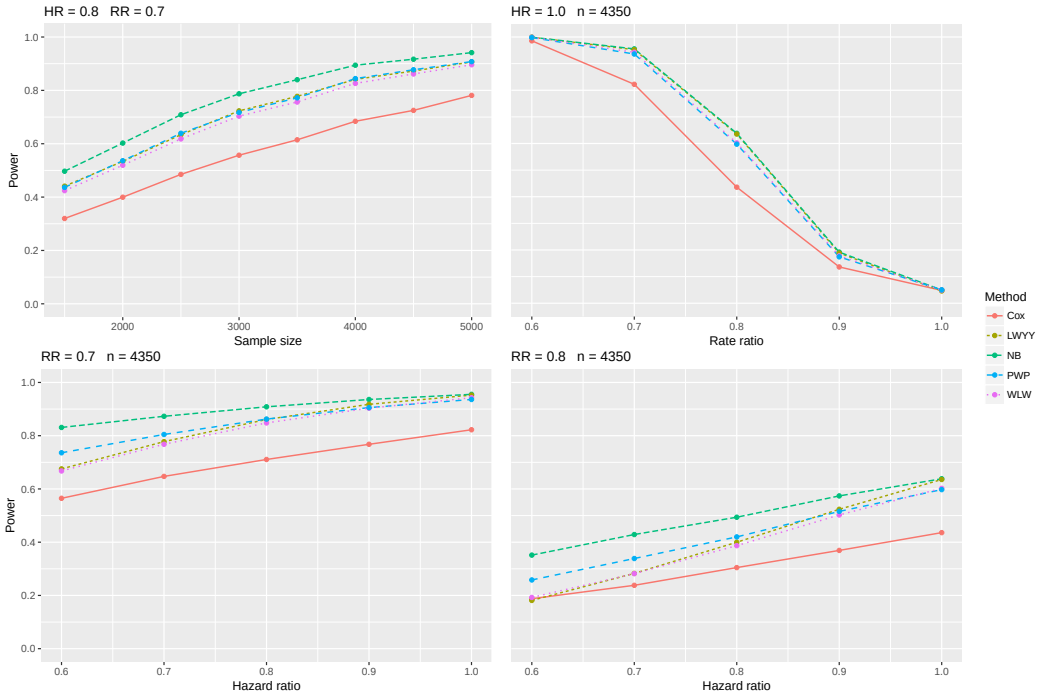
Endpoint	HR_{CV}	Method	Trt.Disc. = 0 %		Trt.Disc. = 10 %		Trt.Disc. = 20 %	
			Estimate	Type I error	Estimate	Type I error	Estimate	Type I error
Estimand 1 (HHF)	0.6	Cox	1.055	0.115	1.055	0.116	1.055	0.116
		NB	1.075	0.118	1.073	0.112	1.073	0.112
		LWYY	1.127	0.267	1.110	0.216	1.110	0.182
		WLW	1.102	0.214	1.095	0.191	1.095	0.174
		PWP	1.049	0.138	1.048	0.137	1.048	0.132
	0.8	Cox	1.030	0.066	1.029	0.066	1.029	0.067
		NB	1.040	0.066	1.038	0.066	1.038	0.063
		LWYY	1.063	0.100	1.055	0.089	1.055	0.080
		WLW	1.052	0.089	1.048	0.084	1.048	0.079
		PWP	1.025	0.071	1.024	0.070	1.024	0.070
	1.0	Cox	1.004	0.048	1.004	0.048	1.004	0.048
		NB	1.006	0.050	1.006	0.050	1.006	0.050
		LWYY	1.006	0.046	1.006	0.046	1.006	0.046
		WLW	1.005	0.049	1.005	0.049	1.005	0.049
		PWP	1.002	0.050	1.002	0.050	1.002	0.050
Estimand 2 (HHF+CVD)	1.0	Cox	1.003	0.046	1.003	0.046	1.003	0.046
		NB	1.005	0.046	1.005	0.046	1.005	0.046
		LWYY	1.004	0.046	1.004	0.046	1.004	0.046
		WLW	1.004	0.050	1.004	0.050	1.004	0.050
		PWP	1.001	0.049	1.001	0.049	1.001	0.049

For informative treatment discontinuation, the corresponding summaries of the simulations are shown in Table 12, and lead to the same conclusions.

5.2.3.3 Power Figures 9 and 10 show the power of Estimands 1 and 2 for selected scenarios with non-informative treatment discontinuation. The main observation is that the recurrent event methods generally provide larger power than the standard time-to-first-event model, with few exceptions. It can also be seen that

- For the scenario $HR_{CV} = 0.8$ and $RR_{HHF} = 0.7$ the power is ordered as follows: NB > LWYY > WLW > PWP > Cox. For Estimand 1 the difference between NB and the other recurrent event methods is higher than for Estimand 2.
- In case there is no treatment effect on CVD, i.e. $HR_{CV} = 1$, LWYY and

Figure 9: Setting with terminal event: Statistical power for Estimand 1 with non-informative treatment discontinuation based on 10'000 clinical trial simulations, sample size $N = 4350$.

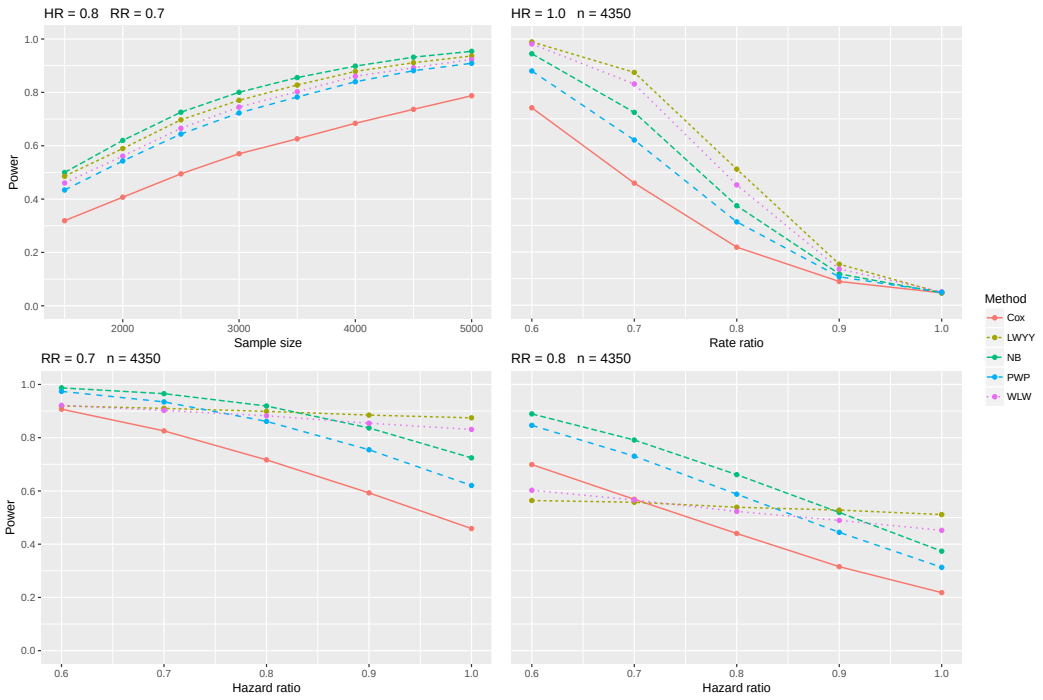


WLW turn out to be more powerful than NB and PWP for Estimand 2. The two graphs at the bottom of Figure 10 also indicate that LWYY and WLW are only more powerful than NB in case the treatment effect on CVD is low. For Estimand 1 NB provides the highest power in all considered scenarios.

- It can also be seen in Figure 10 that for Estimand 2 LWYY and WLW are almost uninfluenced by changes in HR_{CV}. That gives them a higher power than other methods in case of only a small or no treatment effect on HR_{CV}, but a lower power for a larger effect on CVD.

- The two graphs at the bottom of Figure 9 show that for Estimand 1 the power of all methods increases with decreasing effect on CVD (HR_{CV} closer to 1), which is an undesirable behavior. It is due to the fact

Figure 10: Setting with terminal event: Statistical power for Estimand 2 with non-informative treatment discontinuation based on 10'000 clinical trial simulations, sample size $N = 4350$.



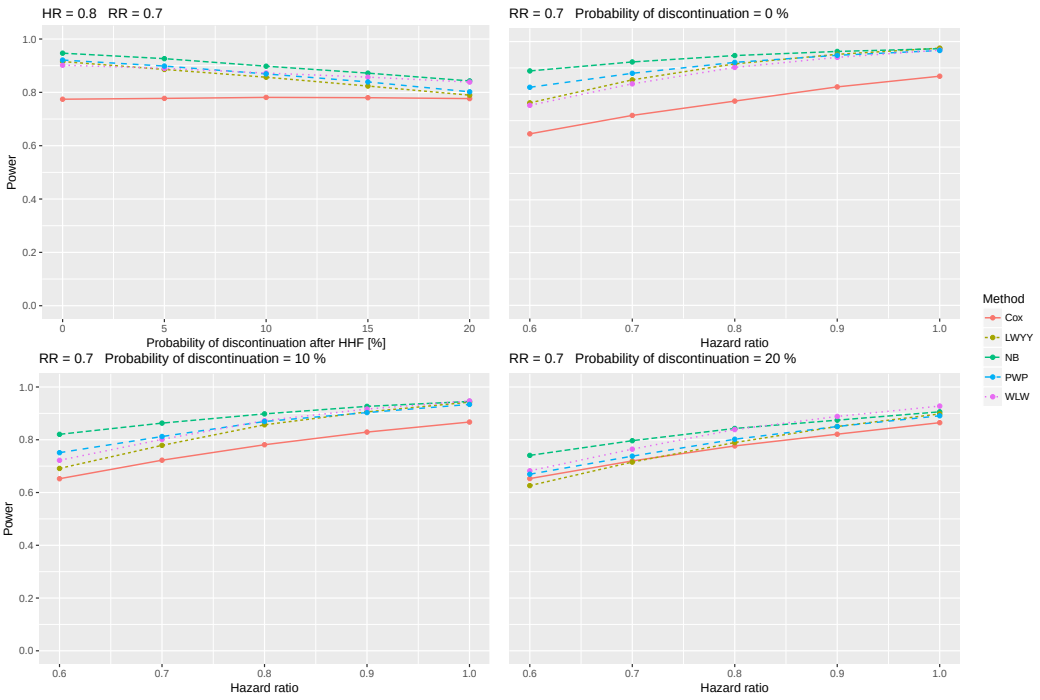
that for large treatment effect on CVD, i.e. small HR_{CV}, the point estimates of all models overestimate the true RR_{HF} (see Table 9) because of dependent censoring and survivor bias.

- Comparing the power of Estimand 1 with Estimand 2, the latter has a higher power if there is a large effect on CVD (HR_{CV} is low), while Estimand 1 has a higher power for HR_{CV} close to 1.

Figures 11 and 12 show the power of Estimands 1 and 2 for selected scenarios with informative treatment discontinuation.

- The results are generally similar to the results for non-informative treatment discontinuation. As expected, the power of all recurrent event methods decreases with a higher discontinuation probability, but in al-

Figure 11: Setting with terminal event: Statistical power for Estimand 1 with informative treatment discontinuation based on 10'000 clinical trial simulations, sample size $N = 4350$.

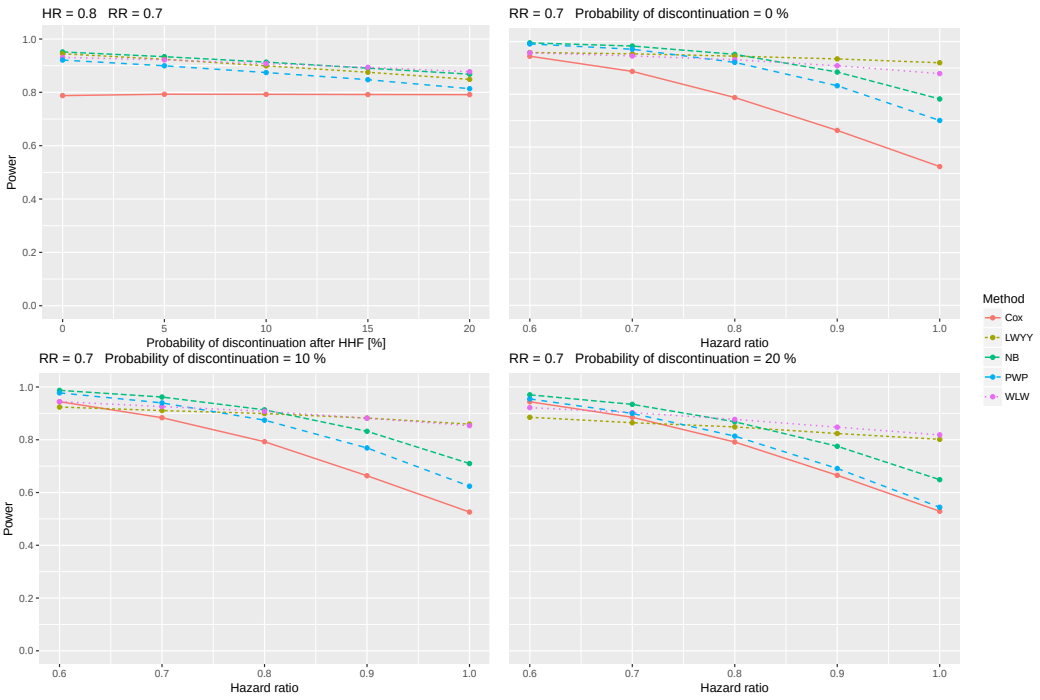


most all cases it is still higher than the one for the Cox model, even for 20% discontinuation probability.

- PWP seems to be most affected by a higher discontinuation rate, while WLW is least affected. This seems plausible, as the treatment effect would be, e.g., more diluted by discontinuations after the first HHF in a time from first to second event analysis than in the analysis of time from treatment start to second event.

5.2.3.4 Further simulation results Additional simulations were done by varying the base case situations, and detailed results can be found in Appendix E.5. No major discrepancies have been observed compared to the results of the base case situations, with one exception. A type I error inflation

Figure 12: Setting with terminal event: Statistical power for Estimand 2 with informative treatment discontinuation based on 10'000 clinical trial simulations, sample size $N = 4350$.



was seen for NB when inter-event times followed a Weibull distribution, and for an autoregressive event rate process. This is in line with expectations as these scenarios deviate from the assumptions of NB, namely the constant baseline rate.

5.2.4 Conclusions

The conclusion of the simulation study for the setting with terminal events is essentially the same as for the one without terminal event: the recurrent event methods were shown to be more efficient than a time-to-first-event analysis, as they provided a higher power in almost all considered scenarios.

For Estimand 1 the higher power was accompanied by sometimes considerably inflated type I error rates, if there was either a positive or negative

treatment effect on CVD. This is due to Estimand 1 favoring the treatment with a worse effect on CVD. At least with the investigated methods, the use of Estimand 1 seems therefore not appropriate, unless it is reasonable to assume that there is no or only a very small treatment effect on CVD. In this case, analysis methods targeting Estimand 1 may be appropriate.

In contrast, no type I error rate increase was seen for Estimand 2. As for the simulation without terminal event, NB and LWYY provide better interpretable treatment effect estimates than WLW and PWP. When patients have the same follow-up time, LWYY directly estimates a meaningful estimand. LWYY also had the highest power for a small or no CVD effect, while NB had the highest power for a larger treatment effect on CVD. Note that NB can have a moderately inflated type I error if the assumption of a constant baseline rate is violated. Knowledge on the magnitude of the treatment effect on CVD might thus be helpful to choose among the different methods. For an uncertain effect on CVD, it seems that LWYY is a good choice.

Although our simulations were motivated by HFpEF trials, similar results may be expected in other indications with a non-negligible terminal event.

6 Conclusions

Chronic diseases are often characterized through the repeated occurrence of events like relapses in RRMS or hospitalizations in CHF. Treatments for such diseases are then expected to impact the first as well as subsequent recurrent events. Hence, their effect is best characterized by using a recurrent event endpoint rather than a time-to-first-event endpoint.

In many diseases where recurrent events reflect disease activity, deaths during a clinical trial are rare. Examples include RRMS (recurrent relapses), asthma or COPD (recurrent exacerbations), migraine (recurrent headache attacks), and epilepsy (recurrent epileptic seizures); see Section 2.2. For such diseases, rate ratios are well-established treatment effect measures (estimands) based on recurrent event endpoints (Section 3.1). These estimands are clinically relevant and easy to interpret, as illustrated with a RRMS case study (Section 4.1). Statistical analysis methods (estimators) targeting these estimands include the commonly used NB and LWYY models (Section 3.1). In contrast, estimands based on a time-to-first-event endpoint are rarely considered in such diseases, as they do not fully capture the clinically relevant

information (Section 2.2). The simulation results show that recurrent event methods are more efficient than a time-to-event analysis (Cox model) as the latter provides considerably less power than the former (Section 5.1). Hence, more precise inference on the treatment effect can be achieved when using recurrent event endpoints, or the same precision can be obtained with less patients enrolled into a clinical trial.

For chronic diseases, where deaths during a clinical trial are more common, the situation is far more complex. A prime example is CHF, where recurrent HHF characterize disease burden but the risk of death is not negligible (Section 2.1). Although the clinical meaningfulness of recurrent HHF has been recognized in the recent CHMP (2017) guideline on the clinical investigation of medicinal products for the treatment of CHF, it is acknowledged that experience in this setting is limited. There are two main challenges. First, the definition of a clinically interpretable treatment effect measure needs careful attention. Various estimand proposals are described and discussed that capture potential treatment effects on both HHF and CV death (Section 3.2). While all these estimands have limitations, those using a while-alive strategy seem often to be appropriate. Second, finding suitable analysis methods (estimators) that target the estimand of interest is challenging. When targeting e.g. the while-alive estimand, NB and LWYY seem adequate; see Sections 3.2 and 5.2 for a detailed discussion, and Section 4.2 for a CHF case study to illustrate the concepts. Extensive simulations show that estimators including a recurrent event endpoint, such as NB and LWYY, are typically more efficient than a Cox model based on a time-to-first-event endpoint (Section 5.2).

The results described in this request support the claim that treatment effect measures can be defined based on recurrent event endpoints that are clinically interpretable and allow for efficient statistical analyses.

Acknowledgments

We thank Richard Cook, Dieter Häring, Björn Holzhauer, Günther Müller-Velten, Tobias Mütze and Jennifer Rogers for helpful and constructive comments on draft versions of this document. We also thank Alexander Seipp for help with writing and extending some of the simulation programs in R. Finally, we thank our regulatory affairs colleagues Denis Burkhalter, Mireille Muller, and Henrik Tang Vestergaard for their continuous support.

References

Aalen, Cook, and Rysland (2015). Does Cox analysis of a randomized survival study yield a causal treatment effect? *Lifetime Data Analysis 21*, 579–593.

Abraham et al. (2011). Wireless pulmonary artery haemodynamic monitoring in chronic heart failure: a randomised controlled trial. *Lancet 377*(9766), 658–666.

Akacha et al. (2017). *Simulations for efficacy comparisons of time-to-event with recurrent event analyses*. Technical Report. Available at `https://www.biostat.uni-hannover.de/fileadmin/institut/pdf/complete.pdf`.

Allignol et al. (2011). Understanding competing risks: a simulation point of view. *BMC Med. Res. Methodol. 11*(1), 86.

Andersen et al. (1993). *Statistical Models Based on Counting Processes*. Springer.

Andersen and Gill (1982). Cox's regression model for counting processes: A large sample study. *Ann. Stat. 10*, 1100–1120.

Anker et al. (2016). Traditional and new composite endpoints in heart failure clinical trials: facilitating comprehensive efficacy assessments and improving trial efficiency. *European Journal of Heart Failure 18*, 482–489.

Anker and McMurray (2012). Time to move on from time-to-first: should all events be included in the analysis of clinical trials? *Eur. Heart J. 33*(22), 2764–2765.

Belot et al. (2014). A joint frailty model to estimate the recurrence process and the disease-specific mortality process without needing the cause of death. *Stat. Med. 33*(18), 3147–3166.

Bender, Augustin, and Blettner (2005). Generating survival times to simulate Cox proportional hazards models. *Stat. Med. 24*(11), 1713–23.

Bernardo and Harrington (2001). Sample size calculations for the two-sample problem using the multiplicative intensity model. *Stat. Med. 20*(4), 557–79.

Beyersmann et al. (2009). Simulating competing risks data in survival analysis. *Stat. Med. 28*(6), 956–971.

Calabresi et al. (2014a). Pegylated interferon beta-1a for relapsing-remitting multiple sclerosis (ADVANCE): a randomised, phase 3, double-blind study. *Lancet Neurol. 13*(7), 657–665.

Calabresi et al. (2014b). Safety and efficacy of fingolimod in patients with relapsing-remitting multiple sclerosis (FREEDOMS II): a double-blind, randomised, placebo-controlled, phase 3 trial. *Lancet Neurol. 13*(6), 545–556.

Carpenter, Roger, and Kenward (2013). Analysis of longitudinal trials with protocol deviation: A framework for relevant, accessible assumptions, and inference via multiple imputation. *J. Biopharm. Stat. 23*(6), 1352–1371.

CHAMPIONS Study Group (2006). IM interferon β-1a delays definite multiple sclerosis 5 years after a first demyelinating event. *Neurology 66*(5), 678–684.

Cheung et al. (2010). Estimation of intervention effects using first or multiple episodes in clinical trials: The Andersen-Gill model re-examined. *Stat. Med. 29*(3), 328–36.

CHMP (2007). *Guideline on clinical investigation of medicinal products for the treatment of migraine.* EMA.

CHMP (2010a). *Guideline on clinical investigation of medicinal products for the treatment of asthma.* EMA.

CHMP (2010b). *Guideline on clinical investigation of medicinal products in the treatment of epileptic disorders.* EMA.

CHMP (2012a). *Guideline on clinical investigation of medicinal products in the treatment of chronic obstructive pulmonary disease (COPD).* EMA.

CHMP (2012b). *Guideline on clinical investigation of medicinal products in the treatment or prevention of diabetes mellitus.* EMA.

CHMP (2015). *Guideline on clinical investigation of medicinal products for the treatment of Multiple Sclerosis.* EMA.

CHMP (2016a). *Concept paper on the need for revision of the guideline on clinical investigation of medicinal product for the treatment of migraine.* EMA.

CHMP (2016b). *Concept paper on the need for revision of the guideline on clinical investigation of medicinal product in the treatment of epileptic disorders.* EMA.

CHMP (2017). *Guideline on clinical investigation of medicinal products for the treatment of chronic heart failure.* EMA.

Claggett et al. (2014). Treatment selections using risk-benefit profiles based on data from comparative randomized clinical trials with multiple endpoints. *Biostatistics 16*(1), 60–72.

Claggett, Wei, and Pfeffer (2013). Moving beyond our comfort zone. *Eur. Heart J. 34*, 869–871.

Cleland (2002). How to assess new treatments for the management of heart failure: composite scoring systems to assess the patients clinical journey. *European Journal of Heart Failure 4*, 243–247.

Cohen et al. (2010). Oral fingolimod or intramuscular interferon for relapsing multiple sclerosis. *New Engl. J. Med. 362*(5), 402–415.

Cohn et al. (2001). A randomized trial of the angiotensin-receptor blocker valsartan in chronic heart failure. *New Engl. J. Med. 345*, 1667–1675.

Collins et al. (2013). Is hospital admission for heart failure really necessary? The role of the emergency department and observation unit in preventing hospital-

ization and rehospitalization. *J. Am. Coll. Cardiol. 61*(2), 121–126.

Comi et al. (2009). Effect of glatiramer acetate on conversion to clinically definite multiple sclerosis in patients with clinically isolated syndrome (PreCISe study): a randomised, double-blind, placebo-controlled trial. *Lancet 374*(9700), 1503–1511.

Confavreux et al. (2014). Oral teriflunomide for patients with relapsing multiple sclerosis (tower): a randomised, double-blind, placebo-controlled, phase 3 trial. *Lancet Neurol. 13*(3), 247–256.

Cook and Lawless (1997). Discussion of Wei and Glidden. *Stat. Med. 16*, 841–3.

Cook and Lawless (2007). *The statistical analysis of recurrent events*. Springer.

Cowling, Hutton, and Shaw (2006). Joint modeling of event counts and survival times. *J. Royal Stat. Soc. C 55(1)*, 31–39.

Cumming, Kelsey, and Nevitt (1990). Methodologic issues in the study of frequent and recurrent health problems falls in the elderly. *Ann. Epidemiol. 1*(1), 49–56.

Dong et al. (2016). A generalized analytic solution to the win ratio to analyze a composite endpoint considering the clinical importance order among components. *Pharm. Stat. 15(5)*, 430–437.

D'Souza, Kappos, and Czaplinski (2008). Reconsidering clinical outcomes in multiple sclerosis: relapses, impairment, disability and beyond. *J. Neurol. Sci. 274*(1), 76–79.

Duchateau et al. (2003). Evolution of Recurrent Asthma Event Rate over Time in Frailty Models. *J. Royal Stat. Soc. C 52*(3), 355–363.

Felker, Anstrom, and Rogers (2008). A global ranking approach to end points in trials of mechanical circulatory support devices. *J. Card. Fail. 14(5)*, 368–372.

Fine and Gray (1999). A proportional hazards model for the subdistribution of a competing risk. *J. Am. Stat. Assoc. 94*(446), 496–509.

Fox et al. (2012). Placebo-controlled phase 3 study of oral bg-12 or glatiramer in multiple sclerosis. *New Engl. J. Med. 367*(12), 1087–1097.

Frangakis and Rubin (2002). Principal stratification in causal inference. *Biometrics 58*(1), 21–29.

Ghosh and Lin (2000). Nonparametric analysis of recurrent events and death. *Biometrics 56*, 554–562.

Ghosh and Lin (2002). Marginal regression models for recurrent and terminal events. *Statistica Sinica 12*, 663–688.

GINA (2017). *Global strategy for asthma management and prevention*. Global Initiative for Asthma. Available from `http://ginasthma.org/`.

Glynn, R. J. and J. E. Buring (1996). Ways of measuring rates of recurrent events. *BMJ 312*(7027), 364.

Gold et al. (2012). Placebo-controlled phase 3 study of oral bg-12 for relapsing multiple sclerosis. *New Engl. J. Med. 367*(12), 1098–1107.

Hauser et al. (2017). Ocrelizumab versus interferon beta-1a in relapsing multiple sclerosis. *New Engl. J. Med. 376*(3), 221–234.

Hengelbrock et al. (2016). Safety data from randomized controlled trials: applying models for recurrent events. *Pharm. Stat. 15*(4), 315–323.

Hernan and Robins (2018). *Causal Inference*. Chapman and Hall/CRC.

Hougaard (2000). *Analysis of multivariate survival data*. Springer.

ICH (1998). *Statistical principles for clinical trials E9*. ICH.

ICH (2017). *Draft ICH E9(R1) addendum on estimands and sensitivity analysis in clinical trials to the guideline on statistical principles for clinical trials*. ICH.

Imbens and Rubin (2015). *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press.

Ip et al. (2015). Comparison of risks of cardiovascular events in the elderly using standard survival analysis and multiple-events and recurrent-events methods. *BMC Med. Res. Methodol. 15*(1), 15.

Jahn-Eimermacher (2008). Comparison of the Andersen-Gill model with Poisson and negative binomial regression on recurrent event data. *Computational Statistics and Data Analysis 52*(11), 4989–4997.

Jahn-Eimermacher et al. (2015). Simulating recurrent event data with hazard functions defined on a total time scale. *BMC Med. Res. Methodol. 15*, 16.

Jahn-Eimermacher et al. (2017). A DAG-based comparison of interventional effect underestimation between composite endpoint and multi-state analysis in cardiovascular trials. *BMC Med. Res. Methodol. 17*, 92.

Kalbfleisch and Prentice (2002). *The statistical analysis of failure time data*. Wiley.

Kappos et al. (2006). Treatment with interferon beta-1b delays conversion to clinically definite and McDonald MS in patients with clinically isolated syndromes. *Neurology 67*(7), 1242–1249.

Kappos et al. (2010). A placebo-controlled trial of oral fingolimod in relapsing multiple sclerosis. *New Engl. J. Med. 362*(5), 387–401.

Kappos et al. (2015). Daclizumab hyp versus interferon beta-1a in relapsing multiple sclerosis. *New Engl. J. Med. 373*(15), 1418–1428.

Keene et al. (2008a). Statistical analysis of COPD exacerbations. *Eur. Respir. J. 32*(5), 1421–1422.

Keene et al. (2008b). Statistical analysis of exacerbation rates in COPD: TRISTAN and ISOLDE revisited. *Eur. Respir. J. 32*(1), 17–24.

Kelly and Lim (2000). Survival analysis for recurrent event data: an application to childhood infectious diseases. *Stat. Med. 19*(1), 13–33.

Klein JP, G. P. (1992). *Survival analysis: State of the art*. Kluwer.

Kuramoto, Sobolev, and Donaldson (2008). On reporting results from randomized controlled trials with recurrent events. *BMC Med. Res. Methodol. 8 (35)*.

Lavery, Verhey, and Waldman (2014). Outcome measures in relapsing-remitting multiple sclerosis: capturing disability and disease progression in clinical trials. *Multiple Sclerosis International*. Article ID 262350.

Law et al. (2017). Misspecification of at-risk periods and distributional assumptions in estimating COPD exacerbation rates: The resultant bias in treatment effect estimation. *Pharm. Stat. 16*, 201–209.

Lewis and Shedler (1976). Simulation of nonhomogeneous Poisson processes with log linear rate function. *Biometrika 63*(3), 501–505.

Lewis and Shedler (1979). Simulation of nonhomogeneous Poisson processes by thinning. *Naval Research Logistics Quartely 26*, 403–13.

Lin et al. (2000). Semiparametric regression for the mean and rate functions of recurrent events. *J. R. Statist. Soc. B 62*, 711–730.

Lin and Wei (1989). The robust inference for the Cox proportional hazards model. *J. Am. Statist. Assoc. 84*, 1074–1078.

Little and Rubin (2000). Causal effects in clinical and epidemiological studies via potential outcomes: concepts and analytical approaches. *Annual Review of Public Health 21*(1), 121–145.

Liu and Huang (2008). The use of Gaussian quadrature for estimation in frailty proportional hazards models. *Stat. Med. 27*, 2665–2683.

Liu, Wolfe, and Huang (2004). Shared frailty models for recurrent events and terminal event. *Biometrics 60*, 219–238.

Lycke et al. (1996). Acyclovir treatment of relapsing-remitting multiple sclerosis - A randomized, placebo-controlled, double-blind study. *J. Neurol. 243*, 214–224.

Mahé and Chevret (2001). Analysis of recurrent failure times data: should the baseline hazard be stratfied? *Stat. Med. 20*(24), 3807–3815.

Mahler and Criner (2007). Assessment tools for chronic obstructive pulmonary disease: do newer metrics allow for disease modification? *Proceedings of the American Thoracic Society 4*(7), 507–511.

Mao and Lin (2016). Semiparametric regression for the weighted composite endpoint of recurrent and terminal events. *Biostatistics 17(2)*, 390–403.

Martinussen, T. and T. H. Scheike (2006). *Dynamic Regression Models for Survival Data*. Springer.

Mazroui et al. (2012). General joint frailty model for recurrent event data with a dependent terminal event: Application to follicular lymphoma data. *Stat. Med. 31*, 1162–1176.

Metcalfe and Thompson (2006). The importance of varying the event generation process in simulation studies of statistical methods for recurrent events. *Stat. Med. 25*(1), 165–179.

Metcalfe and Thompson (2007). Wei, Lin and Weissfeld's marginal analysis of multivariate failure time data: should it be applied to a recurrent events outcome? *Statistical Methods in Medical Research 16*, 103–122.

Mikol et al. (2008). Comparison of subcutaneous interferon beta-1a with glatiramer acetate in patients with relapsing multiple sclerosis (the REbif vs Glatiramer Acetate in Relapsing MS Disease [REGARD] study): a multicentre, randomised, parallel, open-label trial. *Lancet Neurol. 7*(10), 903–914.

Nicholas et al. (2011). Trends in annualized relapse rates in relapsing-remitting multiple sclerosis and consequences for clinical trial design. *Multiple Sclerosis Journal 17*(1), 1211–1217.

Nicholas et al. (2012). Time-patterns of annualized relapse rates in randomized placebocontrolled clinical trials in relapsing multiple sclerosis: a systematic review and meta-analysis. *Multiple Sclerosis Journal 18*(9), 1290–1296.

NRC (2010). *The Prevention and Treatment of Missing Data in Clinical Trials*. Washington, DC: The National Academies Press.

O'Connor et al. (2009). 250 μg or 500 μg interferon beta-1b versus 20 mg glatiramer acetate in relapsing-remitting multiple sclerosis: a prospective, randomised, multicentre study. *Lancet Neurol. 8*(10), 889–897.

O'Connor et al. (2011). Randomized trial of oral teriflunomide for relapsing multiple sclerosis. *New Engl. J. Med. 365*(14), 1293–1303.

Packer (2001). Proposal for a new clinical end point to evaluate the efficacy of drugs and devices in the treatment of chronic heart failure. *J. Card. Fail. 7*, 176–82.

Packer et al. (2013). Effect of levosimendan on the short-term clinical course of patients with acutely decompensated heart failure. *JACC: Heart Failure 1*, 103–111.

Paik, Tsai, and Ottman (1994). Multivariate survival analysis using piecewise gamma frailties. *Biometrics 50*, 975–988.

Pak, Uno, and Kim (2017). Interpretability of cancer clinical trial results using

restricted mean survival time as an alternative to the hazard ratio. *JAMA Oncol. 3*(12), 1692–1696.

Panitch et al. (2005). Benefits of high-dose, high-frequency interferon beta-1a in relapsing–remitting multiple sclerosis are sustained to 16 months: final comparative results of the evidence trial. *J. Neurol. Sci. 239*(1), 67–74.

Pfeffer et al. (2015). Regional variation in patients and outcomes in the treatment of preserved cardiac function heart failure with an aldosterone antagonist (TOPCAT) trial. *Circulation 131*, 34–42.

Pocock et al. (2011). The win ratio: a new approach to the analysis of composite endpoints in clinical trials based on clinical priorities. *Eur. Heart J. 33*(2), 176–182.

Polman et al. (2006). A randomized, placebo-controlled trial of natalizumab for relapsing multiple sclerosis. *New Engl. J. Med. 354*(9), 899–910.

Prentice, Williams, and Peterson (1981). On the regression analysis of multivariate failure time data. *Biometrika 68*, 373–379.

Rogers et al. (2012). Eplerenone in patients with systolic heart failure and mild symptoms: Analysis of repeat hospitalizations. *Circulation 126*(19), 2317–2323.

Rogers et al. (2014a). Analysing recurrent hospitalizations in heart failure: a review of statistical methodology, with application to CHARM-Preserved. *Eur. J. Heart Fail. 16*(1), 33–40.

Rogers et al. (2014b). Effect of Rosuvastatin on Repeat Heart Failure Hospitalizations. *JACC: Heart Failure 2*(3), 289–297.

Rogers et al. (2016). Analysis of recurrent events with an associated informative dropout time: Application of the joint frailty model. *Stat. Med. 35*(13), 2195–205.

Rondeau, Mazroui, and Gonzales (2012). An R package for the analysis of correlated survival data with frailty models using penalized likelihood estimation or parametric estimation. *Journal of Statistical Software 47*, 1–28.

Royston and Parmar (2011). The use of restricted mean survival time to estimate the treatment effect in randomized clinical trials when the proportional hazards assumption is in doubt. *Stat. Med. 30*(19), 2409–2421.

Rudick et al. (2006). Natalizumab plus interferon beta-1a for relapsing multiple sclerosis. *New Engl. J. Med. 354*(9), 911–923.

Rufibach (2017). *Treatment Effect Quantification for Time-to-event Endpoints – Estimands, Analysis Strategies, and beyond.* arXiv. Available at `https://arxiv.org/pdf/1711.07518.pdf`.

Sampson et al. (2010). Composite outcomes: weighting component events accord-

ing to severity assisted interpretation but reduced statistical power. *J. Clin. Epidemiol. 63*, 1156–1158.

Sato et al. (2017). Evaluating the efficacy, safety, and tolerability of serelaxin when added to standard therapy in asian patients with acute heart failure: the design and rationale of relax-ahf-asia trial. *J. Card. Fail. 23(1)*, 63–71.

Schoenfeld (1983). Sample-size formula for the proportional-hazards regression model. *Biometrics 39*(2), 499–503.

Solomon et al. (2017). Angiotensin receptor neprilysin inhibition in heart failure with preserved ejection fraction: rationale and design of the paragon-hf trial. *JACC: Heart Failure 5*(7), 471–482.

Subherwal et al. (2012). Use of alternative methodologies for evaluation of composite end points in trials of therapies for critical limb ischemia. *American Heart Journal 164*, 277–284.

Taylor et al. (2004). Combination of isosorbide dinitrate and hydralazine in blacks with heart failure. *New Engl. J. Med. 351*, 2049–2057.

Therneau and Grambsch (2000). *Modeling Survival Data: Extending the Cox Model*. Springer.

Tuli et al. (2000). Risk factors for repeated cerebrospinal shunt failures in pediatric patients with hydrocephalus. *Journal of Neurosurgery 92*, 31–8.

Uno et al. (2014). Moving beyond the hazard ratio in quantifying the between-group difference in survival analysis. *J. Clin. Oncol. 32*(22), 2380–2385.

Uno et al. (2015). Alternatives to hazard ratios for comparing the efficacy or safety of therapies in noninferiority studiesalternatives to hazard ratios. *Ann. Intern. Med. 163*(2), 127–134.

van Munster and Uitdehaag (2017). Outcome measures in clinical trials for multiple sclerosis. *CNS drugs*, 1–20.

Villegas, Julià, and Ocaña (2013). Empirical study of correlated survival times for recurrent events with proportional hazards margins and the effect of correlation and censoring. *BMC Med. Res. Methodol. 13*(1), 95.

Wang et al. (2009). Statistical methods for the analysis of relapse data in MS clinical trials. *J. Neurol. Sci. 285*(1), 206–211.

Wei, Lin, and Weissfeld (1989). Regression analysis of multivariate incomplete failure time data by modeling marginal distributions. *J. Am. Stat. Assoc. 84*, 1065–1073.

Zannad et al. (2013). Clinical outcome endpoints in heart failure trials: a european society of cardiology heart failure association consensus document. *Eur. J. Heart Fail. 15*(10), 1082–1094.

Appendix

Contents

A Statistical methodology

A.1 Overview of statistical methods for recurrent events and time-to-first-event

A.1.1 Introduction

In this section, we discuss statistical considerations for the analysis of recurrent event data. As a special case we will touch upon the case where interest lies primarily on the first event.

With regard to statistical considerations for recurrent event data, the following characteristics of the data require attention and will be discussed in this section or in related appendices:

- Dependence among repeated events on the same patient;
- Unexplained heterogeneity between patients;
- Early discontinuation from the trial, also called censoring, for various reasons ranging from administrative reasons to lack of efficacy;
- Early termination of the recurrent event process due to death or other so-called terminal events.

In Section A.1.2, we start by discussing potential trial designs and data collection methods. Then follows an exploration of models for recurrent event data, when the rate of terminal events is low, in Section A.1.3. Details for the models are provided in the Appendix A.2. Consequences of censoring and terminal events are discussed in Section A.1.4 and A.1.5, respectively. Additional considerations, including the special case when focus only lies on the first event, are given in Section A.1.6.

General references that discuss statistical considerations for recurrent event data are Cook and Lawless (2007), Hougaard (2000) and Therneau and Grambsch (2000).

A.1.2 Study designs

The setup considered is a clinical trial following a set of patients during some time interval. One typical trial design is that for each patient there is a predefined observation period, e.g., one year. Ideally, all patients are followed for this period with allowance for scheduling the final visit after one year plus or minus a short period, say one or two weeks. However, due to early discontinuation from the trial the actual period with event information may be shorter than the intended period.

Patients can discontinue early for reasons that may or may not be related to the occurrence of events. Also, patients may die, especially in long-term studies involving patients suffering from a serious disease. There may also be other events occurring, such as initiation of rescue medication, which may influence the chance of future occurrences of the recurrent events. Further related aspects will be discussed in Section A.1.4 and A.1.5, respectively.

Another possible trial design is to enroll patients over a time period and follow them until terminating the trial. Similar to the first case, not all patients will be followed to the end. The advantage of this second design is that the patients enrolled first may be followed for relatively longer time giving information on long-term drug use without delaying the end of trial. Generally, the second design will have larger variation in length of follow-up for the patients, but in most cases, the same analysis techniques can be used.

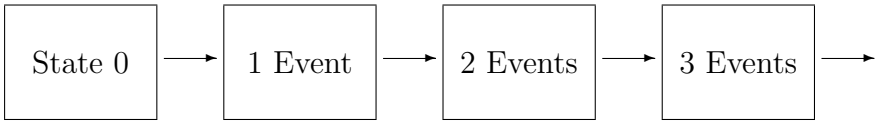
The recurrent events may be recorded in different levels of detail:

- Immediate recording of each event. This is the typical way of handling the most severe events. Either the patient is admitted to the hospital or required to call in to the clinical center and report that an event has happened. This includes day and time of day as well as other relevant information.

- Diary kept by the patient. The patient has to record the events in a diary that is presented to the investigator at the next trial visit.

In the case of joint consideration of recurrent events and a terminal event, it is most appropriate to have immediate recording of the events and therefore, this is the case considered in this request.

Figure 13: Recurrent events considered as a multi-state model.



A.1.3 The multi-state setup without terminal events

In Figure 13, we illustrate the basic setup of using a multi-state model to describe the recurrent event data process over time for a single patient. The process is assumed to start in 'State 0' at time 0.

An epidemiological trial of the lifetime risk of heart attacks could consider time 0 as the time of birth of the patient, whereas the typical drug clinical trial will define time 0 as enrolment of the patient in the trial (referring to randomization or the first dose of the drug). A consequence of the latter definition is that the outcome refers to events after start of treatment and thus events happening before the trial are either neglected or included only through covariates that in some sense reflect the event history before the trial.

A statistical model for the multi-state set-up depicted in Figure 13 models the transition hazards between the states; as illustrated by the arrows in the figure. As intuitively clear from the illustration, the transition hazard can depend on the number of events (j) that have already occurred for the relevant patient before time t. In general, this transition hazard can even depend on any aspect of the history of the process before time t for the patient, meaning t, as well as $T_1, ..., T_j$.

Dependent on the specific assumptions on the transition hazards, different regression models for the recurrent event data process can be formulated. More specifically, fully parametric models such as the Poisson model (Appendix A.2.2.1) and the NB model (Appendix A.2.2.4) or semi-parametric models such as the Andersen-Gill (AG) model (Appendix A.2.2.2) and the Prentice-Williams-Peterson (PWP) model (Appendix A.2.2.3) could be considered. All these models base on the proportional hazards assumption, i.e. the transition hazards are assumed to be proportional for any two covariate sets x and $x^{'}$.

In the context of (semi-) parametric models different distributional assumption can be made for the event process of interest and some of these models account for overdispersion while others do not, see Appendix A.2.2. Similarly, some models make stronger or weaker assumptions with regard to censoring, see Section A.1.4.

The event process over time can also be investigated in a non-parametric fashion using the Nelson-Aalen estimator for the mean cumulative function (MCF) over time, see Cook and Lawless (2007). At the point in time t, the MCF shows the expected number of events per patient by time t. Examples are provided in Section 4.

In this section, we focused on the multi-state setup and models based on the transition hazards. Such models attempt to fully specify the counting process depicted in Figure 13. Alternative frameworks that require fewer assumptions focus only on marginal features, e.g., the expected number of events by time t. Models which fall into this class of approaches include the Lin-Wei-Yang-Ying (LWYY) model and the Wei-Lin-Weissfeld (WLW) model, see also Appendix A.2.3.

A.1.4 Implications of censoring

Censoring means that the development for the patient is no longer followed and always implies a loss of information, because we do not know what happens to the patient after censoring. Censoring, however, should not be confused with mortality and other terminal events, which will be separately discussed in Section A.1.5.

When dealing with censoring in statistical analyses, the first priority is to avoid bias and the second priority is to keep the unavoidable loss of precision as small as possible. The aim is to appropriately estimate relevant characteristics of the recurrent event process, e.g. the transition hazards or the expected number of events by a certain time t, for the population of interest based on incomplete, i.e. censored, data.

Some assumptions along the lines of the *missing at random* for continuous longitudinal data are required:

1. The completely observed population should be well-defined;

2. Censoring should not leave us with a biased sample;

3. Censoring should be non-informative.

The first requirement states that censoring should not prevent the possibility of experiencing the event of interest. This implies that we need to distinguish between situations with and without terminal events. Situations with terminal events will be discussed in Section A.1.5.

The second requirement is that of independent censoring. This means that patients censored at any given time t should not be a biased sample of those who are at risk at time t. In other words, the extra information that the patient is uncensored at time t does not change the transition hazard. Independent censoring should be thought of as 'conditional on given covariates'. This means that censoring may depend on covariates as long as these covariates are accounted for in the statistical model of interest. As an example of violating this assumption, one could consider censoring if patients are admitted to a hospital. One could imagine that patients admitted to a hospital are more seriously ill and thus have higher risk of many event types. Censoring patients at admission would then lead to an underestimation of the risk of events.

The third requirement states that there are no shared parameters between the recurrent event process and the censoring process, see also Cook and Lawless (2007).

Other types of incomplete observations including

- left censoring: event only known to have happen before a certain time;
- interval censoring: event only known to lie in an interval; and
- left truncation: patients only observed from a later entry time

will not be discussed further here.

Turning to the models which were mentioned in Section A.1.3, we note that they provide valid inference under different assumptions for the censoring mechanism. Some methods, e.g., Poisson and LWYY, allow for the censoring to depend on the covariates in the model. Thus, the independence assumption means that the probability of events after censoring, say at time u, is the same as for uncensored patients continuing after time u with the same value of the covariates.

Some methods, e.g. NB and PWP, also allow for the censoring to depend on the event history. The idea is that an event can give a burden that may make a patient want to go out of the trial. For such a patient going out of the trial at time u after having experienced j events, the independence assumption implies that the chance of future events is the same as for patients continuing after time u, which at time u have experienced exactly j events. As an extreme case, one could decide to automatically censor patients after having experienced a certain number of events, say 10. In this case, it is meaningless to compare censored patients with continuing patients and indeed the relevant hazard functions cannot be identified without making untestable assumptions specifying relationships to the hazard functions for patients with fewer than 10 events. To make a more complex example, consider the NB model or another frailty model (Appendix A.2.2.4), where the frailty is assumed to refer to an unobserved disease severity leading to high or low risk of events. Censoring dependent on the unobserved frailty leads to dependent censoring, but censoring dependent on the actual number of events (which indirectly depends on the frailty) leads to independent censoring.

A.1.5 Implications of terminal events

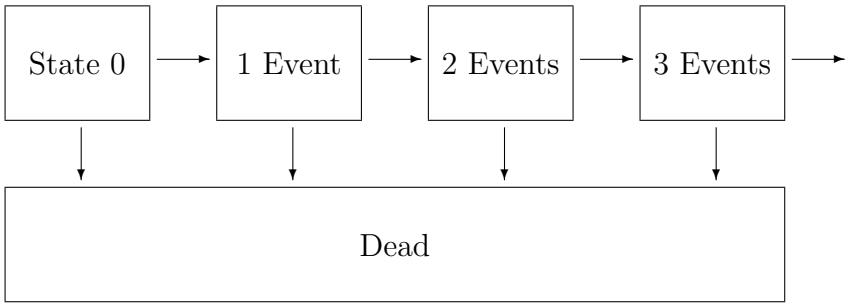
A consequence of the definition of censoring is that death is not covered by censoring. Death is often referred to as terminal event or competing event.

After death no events will occur and therefore *following the patient* is logically the same as *not following the patient*. In other words, we do not lose any information on the patient by not following him after death, because there is no information that can be lost. However, the presence and frequency of mortality has important conceptual consequences for deciding on the scientific question of interest, the estimand (see Section 3), the analysis method as well as for the interpretation of the results.

Other events than death can lead to similar implications and conceptual challenges, e.g., intake of rescue medication may make it irrelevant to follow the disease process further. In that case, one may consider the use of rescue medication as a terminal event.

In terms of a modeling framework, the multi-state setup is also suitable in the presence of terminal events such as mortality. One needs to add a state, 'Dead', corresponding to the occurrence of death or another terminal event, i.e. with zero probability to return from the dead, see Figure 14.

Figure 14: Recurrent events considered as a multi-state model with a terminal event.



Each of the original states will then have a transition into the 'Dead' state. These transition hazards as well as those between the original states will get an extra condition stating that the patient is alive. To make an example, the hazard of experiencing an event if the patient has previously experienced three events, is then changed to describe the hazard of experiencing an event if the patient is alive and has previously experienced three events. In the case of rescue medication as terminal event, the extra condition is that the patient has not initiated rescue medication.

Following general multi-state model principles, the hazard is conditional on the history of the patient until the relevant time point. This can be illustrated by the transition hazard from going from the '2 Events' state to the '3 Events' state, which is the hazard of experiencing an event at time t given that the patient has experienced exactly two events before time t and is alive immediately before time t. The condition that the patient is alive immediately before time t is added compared to the setup of Section A.1.3. In such a multi-state framework, it is possible to estimate the various hazard functions and study their potential dependence on the number of events (j) as well as treatment and other covariates (x). The general multi-state setup considered here can also be reduced to fewer parameters by assuming a frailty model, see Appendix A.2.4 for further discussion.

With terminal events the key issue is how to handle the termination of the event process - not only in a statistical sense but also in a conceptual sense. For example, the interpretation of a single event count, say N_t as discussed

in Section A.1.3 becomes more difficult. The event count may be low for two very different reasons, either because the risk of experiencing the event is low or because the patient has died early and therefore not experienced many events. The same applies to the corresponding population mean.

Different strategies to account for the termination of the recurrent event process due to terminal events (here: death) are discussed in Section 3 but we also touch upon some of them here, see also Chapter 17, Hernan and Robins (2018):

- Consider the terminal event to be a form of censoring and try to adjust for the selection bias that may be introduced through the early discontinuation of the patients. If successful, this approach effectively simulates a population in which death is either abolished or independent of the risk factors for the recurrent events. In either case, the resulting estimates are hard to interpret and may not correspond to a meaningful quantity. This is particularly true if the death is disease related, see also the discussion on hypothetical estimands in Section 3 and related aspects presented in Appendix A.3.5.1.

- Do not consider the terminal event as a form of censoring and deterministically set the time to the next event to infinity. That is, dead patients are considered to have probability zero to have an additional recurrent event between death and the administrative end of follow-up. As mentioned before, this quantity may also not be a good reflection of the treatment effect on the disease burden. The event count may be low for two very different reasons, either because the risk of experiencing the event is low or because the patient has died early and therefore not experienced many events.

- If mortality and recurrent events are thought to reflect the same disease process, one can define a composite endpoint that counts death just like a new case of the recurring events, see discussion on composite estimand in Section 3 and the CHF case study in Section 4.

- One could of course also study mortality irrespectively of the recurrent events process. In case, this analysis shows a significant effect, whether positive or negative, it is important information for the interpretation of the relevance of the treatment. In case the effect is non-significant, it is still relevant to investigate the treatment effect on the recurrent

event process. For the latter investigation, we again need to decide how to account for the terminal event.

A.1.6 Additional considerations

Time-to-first-event data: While recurrent event data are often collected, sometimes events after the first event are ignored. A comparison of test and control treatment is then based on a time-to-event variable, e.g. 'time-to-first-relapse' for RRMS or 'time-to-disease-related-death' for a serious disease. Standard methods from the survival time literature, e.g. Therneau and Grambsch (2000), can then be applied and are briefly summarized in Appendix A.3. In terms of regression models for time-to-first-event data, the Cox proportional hazards model which is a semi-parametric model is often used. The treatment effect is then usually summarized using the hazard ratio, see Appendix A.3.4 for a more detailed discussion. Note that the first coordinates of WLW and PWP correspond to traditional time-to-event analyses, see also Appendix A.2.3.2 and A.2.2.3.

It is worth noting that the challenges discussed in the presence of terminal events also occur when interest lies in the first event only. Related discussions and models usually run under the header competing risk and are presented in Appendix A.3.5.

Sensitivity analysis: All models presented in this section base on some assumptions, e.g., parametric assumptions for the counting process and the frailty terms or assumptions for the censoring mechanism. To assess the robustness across a range of plausible assumptions it is advisable to perform a sensitivity analysis, see also Section 3.

Consider the example of frailty models. We focus on models that use a constant frailty which follows a gamma distribution. If one is in doubt as to these assumptions, one can extend the model. If one is concerned about the assumption of a gamma distribution, there are alternative frailty distributions that can be used instead. Keeping the general variation fixed, such alternatives will have different right and left tails. Observing differences between the various models is most easy if the number of events is relatively large. The doubt on assumption of constant frailty can be addressed by time-dependent frailties. Most of the literature suggestions have considered piecewise constant frailty, e.g., Paik et al. (1994).

Duration of events: In this request, we consider the duration of events to be so short that it makes no difference whether the duration is accounted for or not. Some events may, however, have a duration that is so long that it makes sense to account for it in some way. For example, if the event is admission to a hospital, the duration of the hospitalization may be so long that it could influence the results. One could modify the calculations recognizing that patients are not at risk of being admitted to a hospital, if they already are at the hospital. For most diseases, the duration of a hospitalization is so short that this does not make a major issue. This problem has been considered by Law et al. (2017).

A.2 Recurrent event methods

A.2.1 Notation

Patients will be indexed by i, but wherever possible, this index will be neglected. Thus, the following will address only a single patient.

Time t is measured since the patient started in the trial (randomization or first dose of drug). The cumulative number of events that the patient has realized at time t will be denoted N_t. It is convenient to have a notation for the event times, so the times of recurrent events will be denoted $0 < T_1 < T_2 < ....$ To this, we add the convention that $T_0 = 0$ and $T_j = \infty$, if the patient is not observed to experience $j - th$ events. With the convention, we can state that $N_t = j$ means $T_j \leq t < T_{j+1}$.

The notation might give the impression that there will be infinitely many events but this is not the case. First, no events can happen after death, meaning that death will stop the development of the process. We will use W to denote the time of death. Second, there may be other reasons that the process stops. For example, a person can be immune to the events, implying that no events will ever occur for that person, or become immune, implying that after some time, there will not be new events. Third, something may happen that make consideration of future events irrelevant for the purpose at hand. This could, e.g., refer to use of rescue medication or in some cases, withdrawal due to adverse events. Fourth, events may continue but are not observed within the setting of the clinical trial. This case is denoted censoring.

The patient is followed for events from time 0 to time V, which can be the

time of censoring or refer to a terminal event (death), in which case $V = W$. The total number of recurrent events observed for the patient will be denoted k, meaning $N_V = k$.

To keep control of the observation period, the at-risk function $R(t)$ is introduced being 1 when the patient can experience events (whether recurrent events or death). Informally, this can be phrased as $R(t) = 1\{V \geq t\}$ (where $1\{\text{condition}\}$ refers to the indicator function being 1 when the condition is satisfied and 0 when it is not) but formally the expression could be misunderstood, because the value of $R(t)$ must be known at time t and it is expressed using V, which is not known at time t as we allow for censoring being a consequence of a random process developing in real time. A more detailed consideration of the implications of mortality and censoring is presented in Sections A.1.4 and A.1.5.

There is a p-dimensional covariate $x_1, ..., x_p$, which can also be expressed as a vector x. The values are individual but following the convention, the subscript i is not written in the formula. One or more of the covariates will reflect the treatment group and the effect of this variable is the quantity of key interest in the clinical trial. One of the covariates may be 1 in order to have an intercept in the model. The hazard will typically be assumed to depend on the covariates through the linear score $\beta' x = \beta_1 x_1 + ... + \beta_p x_p$, where $\beta_1, ..., \beta_p$ are the regression coefficients corresponding to the covariates. As phrased here, the covariates are independent of time but in some cases, the setup works even when the covariates are time-dependent, reflecting either an external process or the history of the actual process, meaning reflecting the events that have taken place before the current time t. To discuss and compare the various models in details, some models will use other symbols for the regression coefficients.

Some models may consider *gap times* (time between events) instead (e.g. the PWP model used in Section 4 and Section 5). These can be derived from the basic time observations as $\Delta_j = T_j - T_{j-1}$. While this makes a simple unified formula, it should be mentioned that it is only for $j \geq 2$ that this is indeed a gap time. As there is no requirement that there is an event happening at time 0 (T_0), time Δ_1 is not a gap time but refers to time since trial start.

A.2.2 Methods referring to the multi-state setup

The statistical model describes the transition hazards between the states and is illustrated by the arrows in Figure 13. The hazard for state j is defined as

$$\lambda_j(t) = \frac{Pr\{N_{t+dt} = j+1 \mid N_{t^-} = j, N_v(0 < v < t)\}}{dt}, \tag{1}$$

where the notation N_{t^-} refers to the left limit of N_t and dt refers to an infinitesimal small time interval.

This somewhat technical point is to make the quantities mathematically precisely defined, so that N_t has jumps of size 1, and is continuous from the right. Technically, $\lambda_j(t)$ is defined for all t but for a single patient, the expression is only used between event times number j and $j+1$. So the risk set is patients ongoing in the trial and with exactly j events at the relevant time point. For a single patient, there may be a jump from $\lambda_j(t)$ to $\lambda_{j+1}(t)$, when an event happens. There may, however, also be jumps at other times, exemplified by a piecewise constant hazards model. In any case, the hazard function has to be continuous from the left.

If the transition hazard only depends on the history through the accumulated number of events (j), the process is a Markov process. The Cox proportional hazards Markov model has transition hazard function

$$\lambda_j(t; x) = \lambda_j(t) \exp(\beta' x). \tag{2}$$

As described below, it is also possible to introduce patient-level random effects.

An alternative expression using the random variables is

$$\lambda_j(t) = \frac{Pr\{t \leq T_{j+1} < t + dt \mid T_j < t; T_{j+1} \geq t; T_1, ..., T_j\}}{dt}. \tag{3}$$

In this expression, the term $T_{j+1} \geq t$ should not be interpreted as T_{j+1} being known, only that it is known that the $j+1$-th event has not occurred before time t.

The statistical model for the time-to-first-event can immediately be read off this model. The distribution of this time follows automatically from the first transition hazard (that is, by inserting $j = 0$ in Equation (2)). For the first event, this gives the survivor function (meaning the probability of not having

experienced any events in the interval $(0, t]$)

$$S_0(t; x) = \exp\{-\int_0^t \lambda_0(v) \exp(\beta' x) dv\}. \quad (4)$$

Expressions referring to other number of events are dependent on the actual model and will therefore be presented later.

A.2.2.1 The Poisson model The Poisson model is the statistical model, where the transition hazards do not depend on the history of the process; that is, the hazard of experiencing an event is independent of j and $T_1, ..., T_j$. It can depend on covariates as shown below for the loglinear model

$$\lambda_j(t; x) = \lambda(t) \exp(\beta' x). \quad (5)$$

As this is the hazard function for any event, it is, in particular, also the hazard for the first event, so studying the time-to-first-event is the same as studying the distribution with the hazard function described in the formula.

The special case of a homogeneous Poisson process is obtained by assuming constant hazard $\lambda(t) = \lambda$. Otherwise $\lambda(t)$ can be a parametric or non-parametric function.

The number of events experienced by a patient over a time period from 0 to t follows a Poisson distribution with mean given as the integral of the hazard function, that is,

$$EN_t = \exp(\beta' x) \int_0^t \lambda(v) dv = \exp(\beta' x) \Lambda(t). \quad (6)$$

When the hazard is constant, this expression simplifies to $EN_t = \exp(\beta' x) \lambda t$. The probability distribution of the number of events happening in the interval $(0, t]$ is

$$p(k) = Pr(N_t = k) = \rho^k e^{-\rho} / k!, \quad (7)$$

where $\rho = EN_t$.

The Poisson model is a classical model, but generally, it is insufficient for application to recurrent events in clinical trials. It is based on the simple assumption that all events occur completely independent of each other and thus assumes that there are no patient differences and that occurrence of an event does not change future risk. A consequence of these assumptions is that variation is completely determined as $Var(N_t) = EN_t$.

A.2.2.2 The Andersen-Gill model The model considered by Andersen and Gill (1982), is the semi-parametric Poisson model of Equation (5). However, the model allowed for time-dependent covariates, either reflecting external variables or variables describing the history of the process, thus allowing for dependence between the events. The standard practice when using this model is to use covariates independent of time and estimate the regression parameters in the model by means of the corresponding Cox partial likelihood. The distribution of the number of events is covered by Equations (6) and (7).

To account for potential dependence and overdispersion, one does not use the original variance estimate but instead estimate the uncertainty based on a robust variance estimate, as suggested by Lin and Wei (1989) and Lin et al. (2000). So this is still heavily based on the Poisson model, but the method recognizes that the true variation is higher than suggested by the Poisson model.

A.2.2.3 The Prentice-Williams-Peterson model This model, which is often called the PWP model, was suggested by Prentice et al. (1981) and is the model defined in Equation (2). Compared to the Poisson/AG model, it presents the same hazard for the first event but builds event dependence into the model by allowing the semi-parametric hazard to change each time an event occurs. Based on general thinking, in most cases the hazard is suggested to increase with j, meaning that if you have had events before, your risk of future events is increased. The treatment effect in terms of relative risk is the same whatever the number of events. This means that if two patients with covariate vectors x_1, respectively x_2, have experienced the same number of events (say j) at time t, the ratio of their hazards is $\exp(\beta'(x_1 - x_2))$. The model does not directly consider the ratio of hazards in case the two patients have not experienced the same number of events.

It becomes more difficult to express the distribution of number of events than in the Poisson model, so this is formulated without covariates. The first probabilities follow from the formulas

$$p(0) = \exp\{-\Lambda_0(t)\}. \tag{8}$$

$$p(1) = \int_0^t \lambda_0(t_1) \exp[-\Lambda_0(t_1) - \{(\Lambda_1(t) - \Lambda_1(t_1)\}]dt_1, \tag{9}$$

where t_1 refers to the time of the first event.

$$p(2) = \int_0^t \int_{t_1}^t \lambda_0(t_1)\lambda_1(t_2) \exp[-\Lambda_0(t_1) - \{\Lambda_1(t_2) - \Lambda_1(t_1)\} - \{\Lambda_2(t) - \Lambda_2(t_2)\}] dt_2 dt_1, \tag{10}$$

where t_1 and t_2 refer to the time of the first and second event, respectively. In general, there will be as many integrals as there are events happening.

There is no simple relation between the treatment effect in the defining model and the treatment effect in the event count distribution which limits its practical use in clinical trials.

A.2.2.4 The Negative Binomial Model and other frailty models

The frailty model, which essentially dates back to Greeenwood and Yule (1920), extends the Poisson model by a patient-level random effect.

This type of model suggests that patients have different risks of events but these differences cannot (or can only partially) be explained by the measured covariates. This creates an over-dispersion compared to the Poisson model. The random effect, denoted Z_i for the i-th patient, is called the frailty and has a multiplicative effect on the hazard, which is formulated conditionally on Z_i. The description below refers to a single patient, but with the subscript i omitted.

The hazard can depend on covariates as shown below for the loglinear model

$$\mu(t; x \mid Z) = Z\mu(t) \exp(\omega' x). \tag{11}$$

The notation for the baseline hazard function is changed from $\lambda(t)$ to $\mu(t)$ to emphasize that it is a conditional hazard. For the same reason the conditional regression coefficients are denoted ω.

While in principle many different distribution families can be used for Z in this expression, see Hougaard (2000), (Chapter 9), the classical gamma model is sufficient for this request, so the density of z is assumed to be

$$f(z) = \theta^\delta z^{\delta-1} \exp(-\theta z)/\Gamma(\delta), \tag{12}$$

where δ is the shape parameter and θ, the inverse scale parameter. In particular, this distribution has mean $EZ = \delta/\theta$. One obtains the same model

by using this bivariate parameter as by restricting the mean to one (by requesting $\delta = \theta$) and then including a constant as one of the elements of the covariate vector.

As the frailty is unobserved, it is natural to derive the distribution of the observed quantities by integrating out the frailty. The distribution of the number of events in an interval is conditionally Poisson, similar to Equation (6), but after integration in the gamma frailty case, this becomes a NB distribution with mean number of events

$$EN_t = \delta \exp(\omega' x) M(t)/\theta, \tag{13}$$

where $M(t) = \int_0^t \mu(v)dv$.
This gives the event distribution

$$p(k) = \frac{\{\theta/\exp(\omega' x)M(t)\}^{\delta}\Gamma(\delta + k)}{\{1 + \theta/\exp(\omega' x)M(t)\}^{(\delta+k)}\Gamma(\delta)k!}. \tag{14}$$

One can similarly derive the hazard functions in the multi-state model. For any frailty distribution, the multi-state model is of the Markov type and for gamma frailty, the expressions are further simplified. The hazard of experiencing a first event is

$$\lambda_0(t; x) = \mu(t) \exp(\omega' x)\delta/\{\theta + M(t) \exp(\omega' x)\}, \tag{15}$$

Interestingly, the transition hazard to experience the $j + 1$-th event is

$$\lambda_j(t; x) = \lambda_0(t; x)(\delta + j)/\delta. \tag{16}$$

A consequence of this result is that the gamma frailty model shows proportional hazards over the accumulated number of events. It also clearly demonstrates that the more events you have experienced before, the more events you are predicted to experience in the future.

However, the model does not present proportional hazards across covariates. To assess the treatment effect, one can compare two different covariate values, say x_1 and x_2, corresponding to, e.g., the case with and without treatment. First, one can consider the ratio of the hazards in Equation (11), which is popularly known as a *within patient comparison* meaning the ratio for a specific patient if he tried the two different treatments. This ratio is $\exp(\omega'(x_1 - x_2))$, which might be as expected. Second, one can consider

the ratio of means in Equation (13), which is popularly known as a *population comparison* meaning the ratio of the mean cumulative events in the population after having averaged over patients if they tried the two different treatments. Interestingly, this ratio is also $\exp(\omega'(x_1 - x_2))$. Third, one can consider the ratio of hazards for the first event (Equation (15)), which is

$$\exp(\omega'(x_1 - x_2))\frac{\theta + M(t)\exp(\omega' x_2)}{\theta + M(t)\exp(\omega' x_1)}, \tag{17}$$

which at time 0 equals $\exp(\omega'(x_1 - x_2))$, but monotonically goes to 1 as time increases. This shows a quantitative conflict between considering the first event only and considering all events. This result has important consequences as it implies that the treatment effect evaluated in a first event hazard will be smaller (relative rates closer to 1) than the treatment effect in a multiple event evaluation due to selection effect (meaning that high risk patients are quickly removed from this first risk set). So this makes a theoretical explanation that even though a treatment effect is present at all times, when assessed in a recurrent events frame, it can disappear over time when considering only the time to the first event. In this frame, the recurrent events data reflect the disease better than time to the first event.

Obviously, this model can be extended, e.g., one can substitute the constant frailty with a time-dependent frailty. However, this makes the whole setup more complicated and implies a more complex interpretation, so this is not considered here.

A.2.2.5 Estimation of models based on the multi-state setup Using standard survival data methods, see e.g. Therneau and Grambsch (2000), there are no technical problems in calculating the estimates and other quantities for the Poisson, AG and PWP models, whether the hazard has a parametric, semi-parametric form or is allowed to be non-parametric.

The only real requirement is that the relevant risk sets are non-empty. Exactly which risk sets are relevant depend on the more detailed model assumptions. For the non-parametric versions of the Poisson and AG models, the overall risk set needs to be non-empty at all times from 0 to t. For the non-parametric PWP model, the hazard $\lambda_j(t)$ is only identifiable, when there are patients at risk, who have experienced exactly j events earlier. As all patients start with 0 events, this implies that for each $j > 0$ there is an earliest time, where $\lambda_j(t)$ can be identified. As the treatment effect is assumed

shared over j and t, this does not really create problems for the treatment effect estimate as it has contributions from all values of j and t, where the j-th risk set includes patients from two or more treatment groups.

Also for the non-parametric versions of the frailty model, the overall risk set needs to be non-empty at all times from 0 to t. To identify the frailty distribution parameters, it is also important that at least some patients experience two or more events. The frailty model describes over-dispersion, so if the actual data display under-dispersion, the estimate will correspond to the boundary model of no frailty effect (degenerate frailty distribution). Even in the case of no patients experiencing two or more events, the gamma frailty model may be identifiable because Equation (17) shows converging hazard ratio. This implies that for the first event time, the frailty describes non-proportional hazards rather than over-dispersion. In practice, these points do not create major problems. The risk set problem is handled by knowing in a clinical trial of duration eight weeks, say, one will not attempt at concluding how the event risk will develop later than eight weeks. Whether there are patients experiencing two or more events will also be clear from the descriptive results.

Parametric models will partially or fully handle the above issues by extrapolating the relevant hazards into the areas with empty risk sets.

Data is most conveniently coded in the so-called counting process notation, which has one record for each event and one record for a potential final time period ending with censoring. The PWP approach will then have separate strata for each event number.

The gamma (as well as lognormal) frailty model with non-parametric hazard is covered by the basic R procedure coxph. The SAS procedure can similarly handle non-parametric hazard, but allow only for the lognormal frailty. The SAS procedure nlmixed can be instructed to handle the parametric case.

A.2.3 Methods not referring to the multi-state setup

A.2.3.1 LWYY model The LWYY model named after Lin and Wei (1989) and Lin et al. (2000) aims to directly estimate EN_t without relying on the multi-state model. This means that one assumes the formula

$$EN_t = \exp(\omega' x)H(t), \tag{18}$$

where $H(t)$ is an increasing non-parametric function. This expression is essentially the same formula as those in Equations (6) and (13) but without considering that the derivation of those formulas was made under specific model assumptions, respectively an independence model and a frailty model.

This model will consider events without relating to patient history. One consequence is that the event number j has no role to play in the calculations. The advantage of this approach is that fewer assumptions are implemented but this comes at a price of not fully considering the intra-patient dependence and thus the results may be inefficient and/or biased in some way. Another consequence is that censoring has to be independent of the accumulated number of events.

A.2.3.2 The Wei-Lin-Weissfeld marginal model An approach in a marginal model frame is the so-called WLW model, named after Wei et al. (1989). The idea is to first make an analysis of T_1 in a survival data setting; that is, allowing for censoring. This is indeed the classical way of analyzing the time to the first event. This is technically the same as the PWP model of Section A.2.2.3.

The result is an estimate of the treatment effect. The next step is to make a similar analysis of T_2. This is, however, controversial because the analysis does not account for the fact that $T_1 < T_2$. In other words, patients are considered as being at risk for their second event also before having experienced their first event. Similar analyses are then performed for T_3, T_4 and so on, until the number of events experienced becomes too low to make the treatment estimate informative. Each analysis (strata) leads to one treatment effect estimate and these estimates are then pooled to make an overall effect estimate. The standard errors used are the robust standard errors accounting for the dependence.

An alternative approach is to estimate the overall treatment effect under an assumption that the treatment regression coefficients across strata are the same and again using robust standard errors for accounting for the dependence.

There are two major problems with this approach. One problem is that it is unclear what the treatment effect really estimates. Section A.2.2.4 showed that the first event analysis estimate is different than the frailty model estimate. The first coordinate of the WLW estimate corresponds to the first

event analysis. However, the other coordinates are estimating markedly different aspects of the recurrent events process making interpretation difficult. The other problem is that censoring is not independent, e.g., the first event has to happen before the second event is at risk.

A.2.4 Methods referring to the multi-state setup with a terminal event

As explained in Section A.1.5, considering terminal events implies a switch from the setup in Figure 13 to that of Figure 14 and this implies that the interpretation of the hazard functions changes. However, the formula expressions for the hazard of the recurrent events may be chosen as in the case without terminal events, such as in Equation (2) without a frailty term and Equation (11) in the presence of frailty.

For the death hazard, an expression is needed. The death hazard from state j could be expressed as

$$\xi_j(t; x) = \xi_j(t) \exp(\varphi' x).$$

The simplest case is the non-differential mortality case, where this is independent of j and x, that is, of the form $\xi(t)$. This gives the simplest interpretation of the event hazards, $\lambda_j(t; x)$. However, based on general principles, we have to consider the possibility that mortality depends on both treatment and the number of events that have occurred.

Being a multi-state model, it is, at least in principle, possible to derive the transition probabilities starting from the hazard functions. This can give, e.g., the probability that a patient is alive at time t and has experienced exactly j events before time t. By adding over the corresponding dead and alive states, one can also find the probability of having experienced exactly j events before time t. Due to the possibility of death, the number of events experienced will be smaller than that in the similar model without terminal events. The problem with just looking at the number of events implies that a treatment with high mortality will appear to be better than a treatment with low mortality.

In the multi-state framework depicted in Figure 14, it is possible to estimate the various hazard functions and study their potential dependence on the number of events (j) as well as treatment and other covariates (x). The

treatment may have different effect on the hazard for recurrent events and death. This general setup can also be reduced to fewer parameters by assuming a frailty model. In this section, we will first discuss models without frailty, then models with frailty and finally move to considering the treatment effect.

A.2.4.1 Models without a frailty term For the recurrent event hazard, there are two classical models to consider. First, the Poisson model, where the event hazard is independent of the events already occurred, corresponding to the expression in Equation (5). As this model does not address the possibility of patient differences that are not described by covariates, it will underestimate the variability in many cases.

Second, a more general model, corresponding to the expression in Equation (2), where the hazard of future events may depend on the currently experienced number of events. This corresponds to the PWP model using a time scale of "time since trial start". This has higher flexibility than the Poisson model, but it becomes more difficult to quantify the treatment effect, in the sense that the overall number of events in the treatment groups may develop differently than described by the hazard treatment effect (β) as the patient population is a mixture of patients with different values of the accumulated number of events (j) and this population changes over time (t).

Regarding the mortality, the thinking is that the recurrent events carry a risk of death in a short as well as long time perspective. This suggests a death hazard model of the form $\xi_j(t; x)$, where it may be convenient to describe the dependence on j by a parametric relationship.

A.2.4.2 Models with a frailty term In the case without terminal events, it was shown how the general multi-state model as in Equation (1) could be reduced to fewer parameters by assuming a frailty model of the form in Equation (11). The frailty is interpreted as a patient-level random effect in the hazard of experiencing recurrent events. A similar simplification can be done in the case with terminal events.

The simplest possible model is to combine a hazard of recurrent events, as described in Equation (11) with an assumption of non-differential death hazards, defined as $\xi(t)$ above. However, this assumption seems unrealistically simplistic. In general terms, one would expect mortality to increase with

disease severity and/or disease course.

Starting with the disease course, the thinking is that the recurrent events carry a risk of death in a short as well as long time perspective, as described above. This suggests a death hazard model of the form $\xi_j(t;x)$, where it may be convenient to describe the dependence on j by a parametric relationship.

An alternative approach based on a disease severity justification is to suggest a frailty model that assumes the same frailty, say Z, for the mortality as for the recurrent events, as suggested by Rogers et al. (2014a). This model assumes that the recurrent events and mortality share parameters. Expressed in popular terms, this model says that patients that have a double hazard of recurrent events will also have double hazard of death. To make a more flexible model, a correlated bivariate frailty (Z, U) could be suggested so that Z is the value applicable for recurrent events and U is the value for death risk, corresponding to a mortality hazard of the form

$$U\kappa(t)\exp(\gamma' x).$$

In this modelling setup, Z is well-defined because each patient can experience the event several times. However, results regarding U are more sensitive to the specific model used, because each patient cannot experience death more than once. Therefore a compromise model may suggested, stating a relation between Z and U, more precisely $U = Z^\alpha$, as suggested by Rogers et al. (2016). The advantage of this model is that it uses all data, and for the recurrent events it has a random effects interpretation, which in the multi-state model without frailties leads to a hazard that increases with event count (j), conceptually like the expression in Equation (16). The relation described by the bivariate frailty implies that the death hazard conditional on the accumulated number of events (analogous to $\xi_j(t)$) also increases with j but not necessarily to the same extent.

A.2.4.3 Modeling and estimating treatment effects For the more detailed models, a key question is how to model the treatment effect.

Considering the general clinical trial viewpoint that the results should be adequately described by a single primary endpoint estimate, there are really two choices. One choice is to have separate treatment effect parameters for each type of events, above phrased by parameters β/ω for the recurrent events and φ/γ for the death risk, in the model without/with a frailty term. The

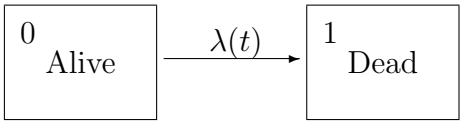
primary analysis should then refer to the recurrent event parameter, β/ω, whereas the death hazard parameter φ/γ, is only considered in a secondary analysis. The other choice is to assume the same treatment effect across endpoints in order to cover all treatment effects simultaneously. This implies $\beta = \varphi$, respectively $\omega = \gamma$.

To put the treatment effect in perspective, it may be desirable to quantify the number of events experienced. In practice, this refers to EN_τ for a specified time point τ. In the case without terminal events, this is expressed in Equation (6) for the Poisson model and Equation (13) for the frailty model. In the PWP model, the similar quantity requires a sum of multi-dimensional integrals and it may be convenient to only count events up to some limit.

In the presence of terminal events, one can evaluate the expected number in the full multi-state model (that is, corresponding to the left side of Equation (6), and Equation (13)), meaning calculating all transition probabilities and use these as input to a mean value calculation. This alternative is less attractive because it gives an advantage to the treatment with the highest mortality.

A better approach is to use the expression of the right hand side of Equation (6) and Equation (13), but in that case, the interpretation as a mean is lost. Popularly, one would denote this as studying recurrent events 'conditional on survival' but this expression is not mathematically precise because the expression is an integral over time. Survival until time τ is not required but the hazard contribution at time t ($0 < t \leq \tau$) requires survival until time t. The advantage of this approach is that a treatment with high mortality will not present with a reduction in the measure of recurrent event risk. With a frailty model for the recurrent events and a non-differential mortality (where the death hazard does not depend on the number of events or the frailty but potentially on the treatment), the mean frailty, EZ, will stay constant over time and the event hazard marginalized over Z will develop similar to the hazards conditional on the frailty. In the joint frailty models suggested by Rogers et al. (2014a) and Rogers et al. (2016), the mean frailty among survivors will decrease over time due to the relation between frailty and death hazard. The estimation method automatically accounts for this effect so that the event hazard will refer to a person with fixed frailty ($Z = 1$). Calculating the integrated hazard for each treatment can then be used to assess the treatment effect on the risk of experiencing the recurrent events.

Figure 15: Survival data illustrated as a two-state model with $\lambda(t)$ as transition rate



A.3 Time-to-first-event methods

A.3.1 Notation

Survival data are realizations of nonnegative random variables, and the object of a survival analysis is to describe and understand the distribution of these random variables or survival times T. This distribution can be described by the survival function with $t \geq 0$

$$S(t) = P(T > t)$$

or equivalently by the rate or hazard function

$$\lambda(t) = -\frac{d}{dt}\log(S(t)),$$

which has the attractive interpretation

$$\lambda(t)dt \approx P(T < t + dt \mid T \geq t),$$

i.e. the conditional probability of dying in the next small time interval ($[t, t + dt)$) given alive immediately before the beginning of the interval. Here, dt refers to an infinitesimal time interval. Statistical models for continuous survival data are most often formulated in terms of the hazard function.

Figure 15 illustrates survival data as a two-state model starting in state 0 (alive) at time $t = 0$ with the hazard function as the transition rate from state 0 to state 1 (dead).

The integrated or cumulative hazard function is defined as

$$\Lambda(t) = \int_0^t \lambda(u)du.$$

The survival function $S(t)$ is the fraction of patients having survived until time t and the hazard function $\lambda(t)$ describes the instantaneous risk per time unit of failing 'now' given alive at time t. We have the following important one-to-one relationship between survival probability and rate

$$S(t) = \exp(-\Lambda(t)). \tag{19}$$

For $i = 1, \ldots, n$ patients, let $t_1, \ldots, t_n$ be failure or censoring times and $d_1, \ldots d_n$ the indicator (0 or 1) of a failure observed at those times. Let $N(t)$ be the counting process counting the number of failures observed before or at time t

$$N(t) = \#\{i : t_i \leq t, d_i = 1\}$$

The number of patients at risk just before time t (t^-) is denoted by

$$Y(t) = \#R(t),$$

where

$$R(t) = \{i : t_i \geq t\} \tag{20}$$

is the risk set at time t identifying the patients still at risk in the trial.

To make the mathematical theory work, it is required that the event count $N(t)$ is continuous from the right, whereas the risk set $R(t)$ as well as the hazard functions (such as $\lambda(t)$) are continuous from the left, so this needs introduction of the time symbol t^- referring to the time immediately before t, so that $N(t^-)$ refers to the left limit of $N(t)$.

Individual covariates are given as $x_i = x_{i1}, \ldots, x_{ip}$, which is a p-dimensional vector.

A.3.2 Non-parametric methods

The survival distribution can be estimated by the Kaplan-Meier estimator

$$\widehat{S}(t) = \prod_{t_i \leq t} \left(1 - \frac{\Delta N(t_i)}{Y(t_i)}\right), \tag{21}$$

where $\Delta N(t) = N(t) - N(t^-)$ is the number of failures at time t. The cumulative hazard function can similarly be estimated by the Nelson-Aalen estimator

$$\widehat{\Lambda}(t) = \sum_{t_i \leq t} \frac{\Delta N(t_i)}{Y(t_i)}. \tag{22}$$

Non-parametric comparison of the survival distributions in groups of patients can be compared by e.g. logrank tests, Kalbfleisch and Prentice (2002).

A.3.3 Parametric models

The exponential distribution is the simplest lifetime distribution assuming a constant hazard function

$$\lambda(t) = \lambda, \quad \Lambda(t) = \lambda t \quad \text{and} \quad S(t) = \exp(-\lambda t).$$

for all $t \geq 0$. The maximum likelihood estimate for λ_0 is

$$\widehat{\lambda} = \frac{\sum d_i}{\sum t_i},$$

also known as the occurrence/exposure rate.

An extension of the exponential distribution is obtained by assuming piecewise constant rates on a number (say Q) of pre-specified time intervals,

$$\lambda(t) = \lambda_q \quad \text{for} \quad c_{q-1} < t \leq c_q, \quad q = 1, \ldots, Q, \quad c_0 = 0.$$

This leads to interval-specific occurrence/exposure rates and provides the basis for further analysis (e.g. Poisson regression). The piecewise constant rate model provides a sensible and flexible summary of many phenomena and is often used in epidemiology and large register studies.

Another extension of the exponential distribution is the Weibull model, which provides a fairly flexible class of distributions

$$\lambda(t) = \lambda\rho(\lambda t)^{\rho-1} \quad \text{and} \quad S(t) = \exp(-(\lambda t)^{\rho}),$$

where $\lambda > 0$ is the inverse scale parameter and $\rho > 0$ is the shape parameter. The exponential distribution is obtained for $\rho = 1$.

Under the assumption of independent censoring, the likelihood function for the models becomes

$$L(\theta) \propto \prod_{i=1}^{n} \lambda(t_i;\theta)^{d_i} S(t_i;\theta), \tag{23}$$

where θ is a vector of the unknown parameters to be estimated. Standard inference via score function and observed information is available, see e.g. Andersen et al. (1993).

A.3.4 Regression models

As previously noted, statistical models for continuous survival data are most often formulated in terms of the hazard function. In particular, when studying how survival time depends on covariates, like treatment and prognostic variables. Let $x_i = x_{i1}, \ldots, x_{ip}$ be a p-dimensional baseline covariate vector for each patient. The hazard is typically assumed to depend on the covariates through the linear score

$$\beta' x_i = \beta_1 x_{i1} + \cdots + \beta_p x_{ip}, \tag{24}$$

where β is the vector of regression parameters.

The Cox proportional hazards model is a semi-parametric model defined by

$$\lambda_i(t) = \lambda_0(t) \exp(\beta' x_i), \tag{25}$$

where the baseline hazard function $\lambda_0(t)$ is an unspecified function of time. The model assumes that the effects of covariates are additive and linear on the log-rate scale and $\lambda_i(t)/\lambda_v(t)$ (where i and v refer to two different patients) does not depend on time. The latter is an assumption of proportional hazards. In particular, for a binary treatment covariate x_1 with two categories treated ($x_1 = 1$) and untreated ($x_1 = 0$), $\exp(\beta_1)$ is the hazard ratio between treated and untreated.

The statistical analysis of the Cox model is based on the partial likelihood function, which in the case of no ties of survival times is given by

$$L(\beta) = \prod_{i=1}^{n} \left(\frac{\exp(\beta' x_i)}{\sum_{j \in R(t_i)} \exp(\beta' x_j)} \right)^{d_i}, \tag{26}$$

where $R(t_i)$ is the risk set defined previously (20). The partial likelihood function may be obtained from the general likelihood function (23) by profiling out the baseline hazard function $\lambda_0(t)$.

The cumulative baseline hazard function $\Lambda_0(t) = \int_0^t \lambda_0(u)du$ from the Cox model can be estimated by the Breslow estimator

$$\widehat{\Lambda}_0(t) = \sum_{t_i \leq t} \frac{d_i}{\sum_{j \in R(t_i)} \exp(\widehat{\beta}' x_j)},$$

where $\widehat{\beta}$ is the maximum likelihood estimate of β. In case of no covariates, the Breslow estimator equals the Nelson-Aalen estimator.

The stratified Cox model can be used to include more than one baseline hazard function, so that for a patient in stratum s, the hazard is

$$\lambda_i(t) = \lambda_{0s}(t) \exp(\beta' x_i), \tag{27}$$

where $\lambda_{0s}(t)$ for $s = 1, \ldots, S$ is the baseline hazard function in each of S strata. This is useful if the proportional hazards assumptions is questionable for some categorical covariate.

The survival function at time t is:

$$S(t \mid x) = \exp(-\Lambda_0(t))^{\exp(\beta' x)}$$

$$\log(-\log(S(t \mid x))) = \log(\Lambda_0(t)) + \beta' x. \tag{28}$$

Importantly, the linear score (24) may be extended to depend on time t by including time-dependent covariates. These covariates need to be left continuous and be known in real time (meaning that at time t, the covariate value $x(t)$ needs to be known). The Cox model becomes

$$\lambda_i(t) = \lambda_0(t) \exp(\beta' x_i^*(t)).$$

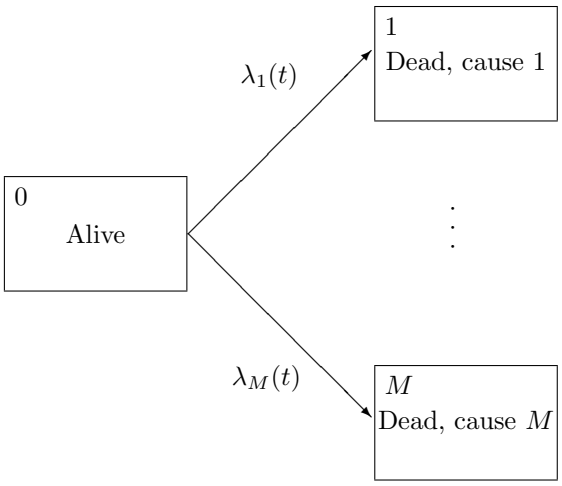
Here $x_i^*(t)$ is some summary of the covariate history $(x(u); u < t)$. Time-dependent covariates can be combined with stratified model and strata may also be time-dependent.

Parametric proportional hazards regression models are obtained by replacing the unspecified baseline hazard function $\lambda_0(t)$ by a parametric function, e.g. one of those reviewed in Section A.3.3. The statistical inference is usually based on a likelihood function analogous to (23). Other regression models include Aalen's additive hazard rate model and the accelerated failure time regression model, see e.g. Martinussen and Scheike (2006) and Kalbfleisch and Prentice (2002).

A.3.5 Implications of terminal events and associated models

A simple extension of the two-state model (Figure 15) is the competing risks model where the terminal event state 'Dead' is split into, say M, exclusive causes of death as illustrated in Figure 16.

Figure 16: Competing risks model illustrated by causes of death.



A.3.5.1 Cause-specific hazard model In the competing risks model there is a cause-specific hazard function, $m = 1, \ldots, M$, from state 'alive' to each of the causes of death

$$\lambda_m(t)dt \approx \text{Prob}(\text{state } m \text{ time } t + dt \mid \text{state 0 time } t^-).$$

The state occupation probabilities include the overall survival function

$$S(t) = P(\text{alive at time } t) = \exp\left(-\sum_{m=1}^{M}\int_0^t \lambda_m(u)du\right) = \exp\left(-\sum_{m=1}^{M}\Lambda_m(t)\right)$$

and the *cumulative incidences* $m = 1, \ldots, M$

$$F_m(t) = P(\text{state } m \text{ at time } t) = P(T \leq t, D = m) = \int_0^t S(u-)\lambda_m(u)du, \tag{29}$$

where D is an indicator for cause of death. The overall risk of dying becomes a sum of all the cumulative incidences

$$F(t) = 1 - S(t) = \sum_{m=1}^{M} F_m(t).$$

As $S(t) + \sum_{m=1}^{M} F_m(t) = 1$ the cumulative incidences are sometimes called sub-distribution functions to underline that they are not true distribution functions.

The cumulative cause-specific hazard can be estimated from the Nelson-Aalen estimator (22) using only failures from the relevant cause

$$\widehat{\Lambda}_j(t) = \sum_{t_i \leq t} \frac{\Delta N(t_i) I(d_i = j)}{Y(t_i)},$$

which is an increasing step function with steps at each observed time of failure from cause j. This formula is only correct when there are no ties among death times, but can be extended to cover ties as well. The overall survival function can be estimated by the Kaplan-Meier estimator (21) and the cumulative incidences can then be estimated by plugging-in estimates

$$\widehat{F}_j(t) = \sum_{t_i \leq t} \widehat{S}(t_i^-) \frac{\Delta N(t_i) I(d_i = j)}{Y(t_i)}, \tag{30}$$

often called the Aalen-Johansen estimator.

Alternatively, it may be tempting to estimate the cumulative incidences by calculating the Kaplan-Meier estimator for each of the transitions by applying censoring for other transitions, which corresponds to assume that all other cause-specific hazards continue after death with the same values, that is, assuming that the hazard for all causes are the same after death as they were before, which is obviously not meaningful. So this calculation fails because, the requirement for the target population to be well-defined is not full-filled, because we attempt to make inference for a potentially completely observed population, where patients can survive also after having died from other causes of death. Such a population is hypothetical. The risk will always be overestimated if using '1 - Kaplan-Meier' instead of the Aalen-Johansen estimator (30).

Another way to describe this point is that while the one-to-one relationship (19) holds for the total mortality, it does not hold for the cause-specific incidences in the competing risks model because all cause-specific intensities are needed, when computing each of the cumulative incidences, $F_m(t), m = 1 \ldots, M$, as shown in Equation (29). Thus, $\exp\{-\Lambda_m(t)\}$ cannot be interpreted as a probability.

In contrast to estimation of the cumulative incidences, inference for the cause-specific hazards can be done using the standard hazard-based models for survival data. Thus, semi-parametric and parametric regression models for the cause-specific hazard function can be applied, e.g. a Cox model for cause j (and omitting patient index i)

$$\lambda_j(t \mid x) = \lambda_{0j}(t) \exp(\beta_j' x),$$

with separate baseline hazard functions and separate regression coefficients for each cause. Differences between the hazard functions can be assessed and by testing $\beta_j = 0$, one can evaluate whether the covariate x_j has an influence on the hazard. This is particularly relevant for studying a treatment effect.

It is technically possible to fit Cox models for cause-specific hazards with identical or proportional baselines for some causes and regression coefficients that are shared between several causes. These features may be more relevant for other multi-state models like recurrent events than the competing risks model. Having models for the cause-specific hazards it is possible to estimate the cumulative incidences by 'plugging-in'

$$\widehat{F}_j(t \mid x) = \int_0^t \widehat{S}(u- \mid x) d\widehat{\Lambda}_j(u \mid x),$$

where

$$\widehat{\Lambda}_j(u \mid x) = \widehat{\Lambda}_{0j}(u \mid x) \exp(\widehat{\beta}_j' x)$$

is the cumulative cause-j-hazard estimate from the Cox model and $\widehat{S}(u \mid x)$ the Cox model based estimator for the overall survival function, e.g.

$$\widehat{S}(u \mid x) = \exp\left(- \sum_j \widehat{\Lambda}_j(u \mid x) \right).$$

The way in which a covariate affects a rate can be different from the way in which it affects the corresponding probability, as this will depend on how the covariate affects the rates also for the competing causes. In conclusion, comparing cause-specific hazards by hypothesis testing is immediate, but for judging the clinical relevance, it is useful to consider both the integrated cause-specific hazards and the cumulative incidence functions.

A.3.5.2 The Fine-Gray model The involvement of all causes into the formula (Equation (29)) for the cumulative incidence for a single cause has led to development of direct regression models for the cumulative incidences of which the Fine-Gray model is the most widely used, Fine and Gray (1999). Like the Cox model is a model for all-cause mortality (28), the Fine-Gray model is a model for cumulative incidences

$$\log(-\log(1 - F_j(t \mid x))) = \log(\widetilde{\Lambda}_{0j}(t)) + \widetilde{\beta}'x.$$

i.e. for

$$\widetilde{\lambda}_j(t) = -\frac{d}{dt}\log(1 - F_j(t \mid x)).$$

That is, the transformation which for all-cause mortality takes us from cumulative risk to hazard is used for a cumulative incidence in a competing risks model. The resulting $\widetilde{\lambda}_j(t)$ is denoted the sub-distribution hazard and the Fine-Gray model is thus a proportional sub-distribution hazards model. However, while a 'sub-distribution hazard' sounds like a hazard, it is not, and the resulting parameters $\exp(\widetilde{\beta})$ in the Fine-Gray model have an indirect interpretation as 'sub-distribution hazard ratios'.

B Estimands for time-to-first-event endpoints with competing terminal events

In this section, we will consider the case where interest lies in time-to-disease-related-death, e.g. CVD, and where the intercurrent event of disease-unrelated death is a competing terminal event. Very similar considerations would also apply to the case where time-to-first-morbidity-event is of main interest and subject to a terminal event, e.g. disease-related or unrelated death.

B.1 Treatment policy estimand

As discussed in Section 3.2.1, a treatment policy estimand is not suitable for terminal intercurrent events such as disease-unrelated death.

B.2 Composite estimand

The composite strategy includes the intercurrent event of disease-unrelated death in the variable definition. Using the four estimand attributes, the composite estimand may be described as follows:

(A) The population is defined through appropriate inclusion/exclusion criteria to reflect the targeted patient population for approval;

(B) The variable of interest is the time to disease-related or disease-unrelated death up to two years;

(C) The intercurrent event of disease-unrelated death is captured through the variable definition;

(D) The same summary measures as discussed in Section 3.1.2 can be considered.

By defining a composite endpoint this estimand assesses treatment effects on any cause of death. Investigational treatments are usually not expected to delay disease-unrelated deaths and the use of this strategy may thus result in a somewhat unspecific treatment effect measure. However, this estimand choice may be relevant when a treatment is expected to improve time to disease-related death while resulting in disease-unrelated deaths due to, e.g., adverse reactions. More generally, this estimand choice may be suitable whenever we want to acknowledge that disease-unrelated death is an unfavorable outcome that ought to be attributed to the treatments under investigation.

A design that targets this estimand is a randomised parallel group design where patients are followed up for two years or until death.

The analysis considerations for this estimand are the same as for the treatment policy estimand in the absence of competing terminal events, see Section 3.1.2.

B.3 Hypothetical estimand

This estimands shares the same estimand attributes (A) and (D) as the composite estimand, but differs in the other attributes.

(B) The variable of interest is the time to disease-related death up to two years;

(C) Here we consider a hypothetical setting/world where death due to disease-unrelated reasons was abolished.

Dependent on the specific setting at hand, this estimand may not be clinically meaningful and relevant.

The design and analysis considerations for this estimand are the same as for the hypothetical estimand in the absence of competing terminal events, see Section 3.1.2.

B.4 Principal stratum estimand

The principal stratum estimand (Frangakis and Rubin, 2002) shares the same attributes (B) and (D) as the hypothetical estimand. The population (principal stratum) is defined as follows:

(A) *Population:* Defined through patients who would not die due to a disease-unrelated cause over a period of two years, regardless of treatment assignment, within the targeted population defined by inclusion/exclusion criteria.

As disease-unrelated deaths do not occur for this principal stratum population, attribute (C) becomes

(C) *Intercurrent events:* The intercurrent event of disease-unrelated death is captured through the population definition.

The principal stratum estimand has a causal interpretation as it refers to the treatment effect in a subgroup properly defined by intercurrent events.

Considerations in terms of trial design and statistical analysis discussed for the principal strata estimand and recurrent event endpoints in Section 3.2.1 also apply to this estimand.

B.5 While-alive estimand

The while-alive or while-on-treatment estimand focuses on the treatment effect on disease-related death while patients are at risk of experiencing this

event. Again, the definition of this estimand is similar to that for recurrent event endpoints presented in Section 3.2.1, just the variable definition and the summary measure need to be adjusted. Considering an effect at a certain time point (e.g. at two years) is no longer of main interest as the variable of interest is defined while a patient is alive - this can be shorter or longer than two years.

In terms of statistical analysis, the Kaplan-Meier approach to obtain survival probabilities at different time points should not be relied on when interest lies in this estimand. Kaplan-Meier estimates are obtained by censoring patients when they die from the competing event, assuming that they are still at risk of dying from the event of interest even after they are censored, when in fact they are at zero risk of dying twice. This leads to biased estimates as has been discussed in various statistical publications; see, e.g., Hougaard (2000).

A more suitable summary measure can be based on the hazard function (at different time points t) which is defined as the proportion of patients who experience the event of interest at time t among those who are still alive at time t. Weighted hazard ratios across relevant time windows can then be used as summary measures. Cause-specific hazard models can be used to estimate the (weighted) hazard ratios, see Appendix A.3.5.1. Although (weighted) hazard ratios are problematic and may not offer a causal interpretation, they may serve as intermediate step for the estimation of survival probabilities and risks, see Hernan and Robins (2018).

C Published literature related to the simulations

C.1 Simulation methods

Bender et al. (2005) proposed a general method for the simulation of univariate survival data by applying inverse sampling. For simulating recurrent event time data following a gap time model, this approach can be applied to simulate the inter-event times. However, to simulate data following a calendar-time model, the method by Bender *et al.* must be adapted; and corresponding simulation strategies have been proposed in Jahn-Eimermacher et al. (2015). Alternative strategies (e.g. thinning homogeneous Poisson

processes) are provided in Lewis and Shedler (1976, 1979). When a potential terminal event has to be considered in addition, methods for simulating competing risk data (Beyersmann et al., 2009; Allignol et al., 2011) can be combined with methods for simulating recurrent event data.

C.2 Selection effects

A trial population might be heterogeneous in the patients' risk for events, even conditional on the intervention and further covariates. In these situations, the mean event rate at time t of those patients that have been free of any event until t (that is the at-risk-set in the Cox model) is changing non-proportionally between both intervention groups. The violation of the proportional hazards assumption causes the intervention effect estimates as derived from a Cox model to be biased (selection bias). However, in a recurrent event setting, patients remain at risk after experiencing a first event, and therefore no selection in the at-risk-set takes place. For this reason, the AG model and its parametric counterparts, the NB model and the Poisson model, can also provide unbiased intervention effect estimates in situations with unmodeled heterogeneity. Selection bias in the Cox model has been - in addition to others - analytically derived in Aalen et al. (2015) and has been demonstrated in several simulation studies with data following homogeneous or non-homogeneous mixed Poisson processes, that have been analyzed by Cox, AG, NB, and Poisson models (Metcalfe and Thompson, 2006; Hengelbrock et al., 2016; Cheung et al., 2010; Jahn-Eimermacher et al., 2017). Results from clinical trial data (Rogers et al., 2014a; Ip et al., 2015; Mahé and Chevret, 2001) further support the findings obtained in simulation studies. In addition to Aalen et al. (2015), accessible explanations of selection bias can also be found in Jahn-Eimermacher et al. (2017) and Hengelbrock et al. (2016). Whereas the aforementioned unstratified models for recurrent event data can prevent selection bias, it will be reintroduced by stratifying the analysis model by the event number as is done in the PWP modeling approach (Metcalfe and Thompson, 2007; Kelly and Lim, 2000; Hengelbrock et al., 2016; Therneau and Grambsch, 2000). For univariate data, frailty models have been proposed for preventing selection bias. In the context of HF studies, a comparison of the Cox model for time to death with results derived from a frailty model indeed reveals differences between these estimates (Rogers et al., 2016). Hengelbrock et al. (2016) further demonstrated that

adding a frailty term to the PWP model will also remove selection bias.

C.3 Total intervention effects and carry-over effects

In stratified models, also called autoregressive models by some authors, the risk of further events is changing after each event, thus violating the Poisson assumption of the (in these situations misspecified) AG, NB and Poisson models. The intervention effect estimates from such misspecified models refer to a total intervention effect, as has analytically been derived for the AG model in Cheung et al. (2010). In contrast, the stratified version (PWP) estimates the direct intervention effect that is corresponding to the model parameter. When risks increase with each event, preventing or delaying an event also prevents or delays a patient from being at an increased risk for further events. This so-called indirect intervention effect contributes to a larger total effect as compared to an intervention's direct effect on the risk rates. For similar reasons, the total intervention effect decreases when risks decrease with each event. Besides the analytical results in Cheung et al. (2010), the estimates of total effects have been derived in many simulation studies (Jahn-Eimermacher, 2008; Metcalfe and Thompson, 2006, 2007; Kelly and Lim, 2000; Hengelbrock et al., 2016; Villegas et al., 2013; Therneau and Grambsch, 2000). Furthermore, Cheung et al. (2010) showed in simulation studies that the 95% confidence interval as derived from a misspecified AG model keeps the 95% coverage probability for the total intervention effect. Next to Cheung et al. (2010), accessible explanations of total versus direct intervention effects can also be found in Jahn-Eimermacher et al. (2017) and (Hengelbrock et al., 2016). Some clinical trial results further support these findings (Ip et al., 2015; Mahé and Chevret, 2001). Also, the Wei-Lin-Weissfeld model is in general misspecified, as the proportional hazards assumption is violated within all but the first stratum. As a consequence, the mean intervention effect estimates differ from the model parameters. The so-called carry-over-effects(a delay of a first event causes a delay of all further events) contribute to larger mean intervention effect estimates as compared to the model parameters, as has been demonstrated in several simulation studies (Kelly and Lim, 2000; Metcalfe and Thompson, 2007; Villegas et al., 2013; Therneau and Grambsch, 2000). These studies also confirm that estimates derived from a PWP model are not affected by carry-over-effects, as here a different definition of the at-risk-set is applied.

C.4 Parametric vs semi-parametric models

The NB and Poisson are parametric models. When all patient have the same follow-up time, these may still be appropriate even with time-changing event rates. When follow-up times are not the same in all patients, this is handled through the inclusion of the log follow-up time as an offset, and essentially relies on exponentially distributed inter-event-times (eventually conditional on a random frailty term). The AG model is the semi-parametric counterpart without a distributional restriction on inter-event-times. Several simulation studies and real data examples show comparable results when derived from semi-parametric and parametric models even when the exponential distributional assumption does not hold (Rogers et al., 2014a; Metcalfe and Thompson, 2006; Jahn-Eimermacher, 2008; Duchateau et al., 2003)

C.5 Competing terminal event

In HF studies, CVD is a competing terminal event. When analyzing HHF and CVD within a composite endpoint and intervention effects differ for the two components, a mixed intervention effect is estimated. This has been observed in some clinical trials (Rogers et al., 2012, 2014b). When the hospitalisations are analysed by applying the AG model and thus handling CVD as an independent censoring event, the terminal event reduces the at-risk-sets over time to the survivors only. For this reason, the resulting intervention effect estimates are prone to selection bias in contrast to the situations without a competing terminal event (Jahn-Eimermacher et al., 2017). The joint frailty model can prevent selection bias. However, in clinical trial data, no substantial differences between the results derived from the joint frailty and marginal models (AG and Poisson model) have been observed (Rogers et al., 2014b). Some simulation studies further compare different joint frailty models (Mazroui et al., 2012; Belot et al., 2014).

C.6 Time scale

Most of the considered statistical models can apply either a calendar time scale or a gap time scale (Kelly and Lim, 2000). Only for exponentially distributed inter-event-times do both time scales refer to the same model as underlies the Poisson and NB model (eventually conditional on the random

term). Under more general distributional assumptions, statistical results depend on the specification of the time scale as has been observed in simulation studies (Metcalfe and Thompson, 2006; Villegas et al., 2013) and real data (Duchateau et al., 2003).

C.7 Power comparisons

Comparing the power of statistical methods for rejecting a null hypothesis $H_0 = \{\beta = 0\}$ is only well interpretable if the compared methods asymptotically provide the same β estimates. Otherwise, the differences in asymptotic effect estimators will cause differences in power, which therefore no longer relate to efficiency only. As most of the statistical methods described so far differ in the treatment effect they are estimating, we focus here on the power comparison between the Cox and AG model only. Power and sample size formulas have been derived for data following a Poisson process (Schoenfeld, 1983; Bernardo and Harrington, 2001). For both methods (Cox, AG), the number of observed events required to obtain a power of $1 - \gamma$ for rejecting the null hypothesis $H_0 = \{\beta = 0\}$ at the two-sided significance level of α when comparing two equally sized groups is given as

$$L = 4 \cdot \left(\frac{z_{1-\alpha/2} + z_{1-\gamma}}{\beta} \right)^2$$

As the AG model incorporates recurrent events and thus uses more events than the Cox model, the AG approach will always be more efficient under Poisson processes. Simulation results that show standard error estimates support this finding (Kelly and Lim, 2000; Metcalfe and Thompson, 2006). In these papers, the simulation results given for the first event under the PWP method coincide with results that would be obtained from a Cox model. The results further indicate, that the use of robust standard errors will not affect the power under a Poisson process as they hardly differ from the naive ones.

When data follow a mixed Poisson process, the Cox but not the AG model is prone to selection bias. Selection bias will cause a decrease in the power. The power of the AG model also decreases as the robust standard error estimates are increasing with increasing variance of the random effect. Some simulation studies indicate, that the AG model is still more efficient under the particularly investigated data generation processes (Kelly and Lim, 2000; Metcalfe and Thompson, 2006).

D Details for simulation studies in settings without terminal events

D.1 Event-generating process

Let T_{ij} be the waiting time between patients $(j-1)th$ and its jth event for $j > 1$ and the time elapsed from starting point 0 to its first event for $j = 1$, we generate recurrent event data under both homogeneous Poisson process and non-homogeneous Poisson process.

1. Under homogeneous Poisson process that assumes constant control ARR over time, we generate for each patient i, conditional on $Z_i = z_i$, the inter-event times T_{ij} from independent realizations of an exponential distribution with scale parameter $\lambda = \lambda_0 z_i \exp(x_i\beta)$.

2. The non-homogeneous Poisson process accounts for a considerable variation of the control ARR over the last years (Nicholas et al., 2011). We choose a log-linear baseline intensity function $\lambda(t) = \exp(\alpha_0 + \alpha_1 t) = \lambda_0 \exp(\alpha_1 t)$ to model the relapse counts. To get an impression of the sizes of α_0 and α_1 we visit the meta-analysis published by (Nicholas et al., 2012), where they reported a decreasing trend of control relapse rates over time within randomized clinical trials in relapsing MS, the overall rate ratio of first year versus second year of the ARR from 1.1 to 1.6 correspond to a range of α_1 from -0.1 to -0.5. Additionally, we define α_0 that leads to a baseline rate of $\lambda_0 = \exp(\alpha_0) = 0.5$. The inter-event times T_{ij} are generated based on the algorithm in (Lewis and Shedler, 1976).

D.2 Treatment discontinuation

The planned follow-up time for each patient is $T = 2$ years, treatment discontinuation is treated as non-informative or informative.

1. Non-informative treatment discontinuation is simulated by randomly assigning each patient a follow-up of length $T(1 - p_i l_i)$ with $(p_i)_{i=1,\ldots,n}$ be independent realizations of binomial distribution with success probability 0.2, and $(l_i)_{i=1,\ldots,n}$ be independent realizations of uniform distribution.

2. Informative treatment discontinuation is simulated in two ways. First, we model the recurrent event process and the treatment discontinuation process jointly by a joint frailty model, so that times to informative treatment discontinuation are generated from exponential distribution with scale parameter $r = r_0 z_i \exp(x_i \gamma)$, where $r_0 = 0.2$ is the baseline treatment discontinuation rate, and γ is the coefficient of treatment effect on treatment discontinuation, which is not necessary equal to the coefficient of treatment effect on recurrent event β. Second, the treatment discontinuation process is generated conditional on the number of recurrent events, the larger the number of recurrent events, the more likely the patient being censored. This is achieved by generating a realization of Bernoulli distribution with success probability 0.3 each time an event occurs, and the patient discontinues right after the event in case we get a realization of 1.

D.3 Numeric estimand values

In this section, we compute the numeric estimand value under four scenarios. Here we only consider the homogeneous Poisson process, and the informative treatment discontinuation where it is modeled jointly with recurrent event process by a joint frailty model (see more details about treatment discontinuation in the Appendix D.2).

Suppose we are interested in the treatment effect defined as ratio of mean event rate in treatment group over the control group. In order to derive the numeric estimand values under four scenarios, first we denote the treatment effect on recurrent event by $e^{\beta} = 0.65$, treatment discontinuation rate is $\gamma_0 = 0.2$, follow-up time is $T = 2$, dispersion parameter is $\theta = 0.25$, and the treatment effect on discontinuation is $e^{\gamma} = 0.65$.

For hypothetical estimands (scenario 1 and 2), the numeric estimand value is 0.65 no matter the treatment discontinuation is informative or non-informative. This is because this estimand represents the values as if the treatment had continued.

For treatment-policy estimand, non-informative treatment discontinuation (scenario 3), the analytic form of estimand is derived as $(1 - \gamma_0/2)e^{\beta} + \gamma_0/2$, by plugging in the parameters, we get 0.685. For treatment-policy estimand, informative treatment discontinuation (scenario 4), the analytic

form of estimand is derived as $1 - \frac{(1-e^{\beta})}{\gamma_0 T e^{\gamma}} \left[1 - \left(\frac{1}{1+\theta\gamma_0 e^{\gamma} T}\right)^{\frac{1}{\theta}}\right]$, by plugging in the parameters, we get 0.7002. Please see below for more detailed derivations.

D.3.1 Informative treatment discontinuation

The treatment effect on the recurrent event rate is e^{β} with rates $z\lambda_0$ in control group and $z\lambda_0 e^{\beta}$ in treatment group up until treatment discontinuation and the discontinuation rate is $z\gamma_0$ in control group and $z\gamma_0 e^{\gamma}$ in treatment group for a patient with gamma random effect z.

The time spent at rate $\lambda_0 z e^{\beta}$ is limited by the smaller of time to treatment discontinuation and the length of trial T. Thus, the expected number of events of treatment group given z is

$$\begin{aligned}
&\int_0^T z\lambda_0 e^{\beta} e^{-z\gamma_0 e^{\gamma} t} dt + \int_0^T z\lambda_0 \left(1 - e^{-z\gamma_0 e^{\gamma} t}\right) dt \\
&= z\lambda_0 \left\{T + (e^{\beta} - 1) \int_0^T e^{-z\gamma_0 e^{\gamma} t} dt\right\} \\
&= z\lambda_0 \left\{T + (e^{\beta} - 1)\left(1 - e^{-z\gamma_0 e^{\gamma} T}\right) / z\gamma_0 e^{\gamma}\right\}
\end{aligned}$$

Now we need to integrate out the frailty z, which follows a gamma distribution with shape parameter $1/\theta$ and rate parameter $1/\theta$, having mean 1 and variance θ

$$\begin{aligned}
&\int_0^{\infty} \lambda_0 \left\{Tz + \frac{(e^{\beta} - 1)}{\gamma_0 e^{\gamma}}\left(1 - e^{-z\gamma_0 e^{\gamma} T}\right)\right\} \frac{1}{\Gamma(1/\theta)} \left(\frac{z}{\theta}\right)^{\frac{1}{\theta}-1} e^{-z/\theta} \frac{dz}{\theta} \\
&= \lambda_0 T \left\{1 + \frac{(e^{\beta} - 1)}{T\gamma_0 e^{\gamma}} \left[1 - \int_0^{\infty} \frac{1}{\Gamma(1/\theta)} \left(\frac{z}{\theta}\right)^{\frac{1}{\theta}-1} e^{-z(\gamma_0 e^{\gamma} T + 1/\theta)} \frac{dz}{\theta}\right]\right\} \\
&= \lambda_0 T \left\{1 + \frac{(e^{\beta} - 1)}{T\gamma_0 e^{\gamma}} \left[1 - \left(\frac{1}{\theta}\right)^{\frac{1}{\theta}} (\gamma_0 e^{\gamma} T + 1/\theta)^{-\frac{1}{\theta}}\right]\right\} \\
&= \lambda_0 T \left\{1 + \frac{(e^{\beta} - 1)}{T\gamma_0 e^{\gamma}} \left[1 - (1 + \theta\gamma_0 e^{\gamma} T)^{-\frac{1}{\theta}}\right]\right\}
\end{aligned}$$

So when we divide through by $\lambda_0 T$, the expected number of events in control group, we end up with

$$1 - \frac{(1-e^{\beta})}{\gamma_0 T e^{\gamma}} \left[1 - \left(\frac{1}{1+\theta\gamma_0 e^{\gamma} T} \right)^{\frac{1}{\theta}} \right]$$

D.3.2 Non-informative treatment discontinuation

The treatment effect on the recurrent event rate is e^{β} with rates $z\lambda_0$ in control group and $z\lambda_0 e^{\beta}$ in treatment group up until treatment discontinuation and the discontinuation rate is γ_0 in control group and γ_0 in treatment group for a patient with gamma random effect z. We assume 10% of the time patients in the active arm are back on control (20% are censored, and each spends on average half their time before and half after censoring).

The time spent at rate $\lambda_0 z e^{\beta}$ is limited by the smaller of time to treatment discontinuation and the length of trial T. Thus, the expected number of events of treatment group given z is

$$\begin{aligned} &\int_0^T z\lambda_0 e^{\beta}(1-\gamma_0/2)dt + \int_0^T z\lambda_0\gamma_0/2 dt \\ &= z\lambda_0 T \left\{ (1-\gamma_0/2)e^{\beta} + \gamma_0/2 \right\} \end{aligned}$$

Now we need to integrate out the frailty z, which follows a gamma distribution with mean 1 and variance θ

$$\begin{aligned} &\int_0^{\infty} z\lambda_0 T \left\{ (1-\gamma_0/2)e^{\beta} + \gamma_0/2 \right\} dz \\ &= \lambda_0 T \left\{ (1-\gamma_0/2)e^{\beta} + \gamma_0/2 \right\} \end{aligned}$$

So when we divide through by $\lambda_0 T$, the expected number of events in control group, we end up with

$$(1-\gamma_0/2)e^{\beta} + \gamma_0/2$$

D.4 Event specific estimates for WLW and PWP models

Table 13 presents the event specific treatment effect estimates for WLW and PWP under four scenarios as above. This table just shows the result when

baseline recurrent event rate $\lambda_{HHF} = 0.5$, dispersion parameter $\theta = 0.25$. For these two models, non-convergence can happen when both sample size and the baseline event rate are small, so non-convergence percentage is also presented. To compare WLW and PWP in terms of event specific treatment effect estimates, we have the following findings:

- The treatment effect estimates of first event from PWP and WLW correspond to Cox model; on each of the events beyond the first, WLW gives smaller HR compared to PWP. All four scenarios have the same pattern, but scenarios 2, 3 and 4 give larger event specific HR estimates than scenario 1, which is consistent with the overall estimates.

- The HR estimates from the PWP are of markedly greater than 1 for events beyond the first, increasingly so for successive events, since PWP provides an indication of the direct effect of treatment upon each ordered event. However, PWP estimates are not based on comparisons of full randomized groups, and so the resulting conclusions must be suitably cautious.

- The distinctive treatment effect estimates obtained by WLW are a direct result of the risk set definition in WLW. By allowing patients to be at risk of the kth event before they have undergone the $(k-1)th$ event, WLW has been seen as failing to accommodate the ordered nature of recurrent events (Klein JP, 1992), (Cook and Lawless, 1997), (Tuli et al., 2000). An alternative view of this issue is that the treatment effect on event k will be "carried over" to subsequent events, so "biasing" the estimates of the treatment effects for those later events.

- When the baseline recurrent event rate is 0.5, the treatment effect estimates for event 3 and 4 are not reliable for small sample size, e.g. $n < 150$ per arm, since the non-convergence percentage is large.

Table 13: Without terminal event: event specific mean treatment effects estimate and non-convergence percentage of WLW and PWP under four scenarios, with $HR = 0.65$, $\lambda_0 = 0.5$, and $\theta = 0.25$.

			n=50		n=150		n=250	
	Method	Event	HR	non-converge(%)	HR	non-converge(%)	HR	non-converge(%)
Scenario 1: Non-informative (Hypothetical)	WLW	1	0.7	0	0.68	0	0.675	0
		2	0.769	0.18	0.739	0	0.738	0
		3	0.789	18.98	0.556	0.57	0.538	0.01
		4	1.22	74.81	0.772	31.27	0.571	13.42
	PWP	1	0.7	0	0.68	0	0.675	0
		2	1.181	0.18	1.057	0	1.045	0
		3	1.891	21.25	1.158	0.6	1.087	0.01
		4	2.668	83.06	2.205	34.37	1.631	14.26
Scenario 2: Informative (Hypothetical)	WLW	1	0.705	0	0.687	0	0.681	0
		2	0.79	0.19	0.752	0	0.745	0
		3	0.855	23.89	0.596	1.1	0.561	0.1
		4	1.147	81.87	0.851	39.6	0.643	19.88
	PWP	1	0.705	0	0.687	0	0.681	0
		2	1.212	0.19	1.072	0	1.049	0
		3	2.019	27.04	1.207	1.19	1.115	0.1
		4	2.337	89.41	2.343	44.58	1.809	21.43
Scenario 3: Non-informative (Treatment-policy)	WLW	1	0.726	0	0.708	0	0.703	0
		2	0.796	0.04	0.773	0	0.771	0
		3	0.784	10.91	0.608	0.14	0.591	0
		4	1.171	63.92	0.758	18.14	0.594	5.58
	PWP	1	0.726	0	0.708	0	0.703	0
		2	1.184	0.04	1.078	0	1.066	0
		3	1.725	12.02	1.18	0.12	1.124	0
		4	2.589	72.9	1.941	19.98	1.484	5.88
Scenario 4: Informative (Treatment-policy)	WLW	1	0.729	0	0.713	0	0.709	0
		2	0.818	0	0.793	0	0.789	0
		3	0.839	9.47	0.664	0.1	0.644	0
		4	1.217	60.07	0.826	13.98	0.67	3.88
	PWP	1	0.729	0	0.713	0	0.709	0
		2	1.22	0	1.107	0	1.091	0
		3	1.797	10.13	1.247	0.08	1.192	0
		4	2.49	68.9	1.949	15.48	1.509	4.04

E Details for simulation studies in settings with terminal events

E.1 Issues with implementation of the joint frailty model

Computational issues were encountered with the function frailtyPenal of the R package frailtypack (Rondeau et al., 2012). This included biased estimates, long run times of more than an hour for a single data set and massive memory occupation. The problems did not occur for every data set, but for a not too small proportion of about 20%. For many individual data sets it was possible to prevent these issues by choosing different starting values or fine-tuning parameters like the number of knots or smoothing parameters when using splines to model the baseline hazards. But no general setting could be found that would have allowed running a simulation study with many iterations. In a regulatory context it is also questionable to apply a model whose application requires manual tuning of parameters and would not allow pre-specification of detailed settings. We also tried the implementation of the joint frailty as described in Liu and Huang (2008) with PROC NLMIXED in SAS, which uses piecewise constant baseline hazard functions for HHF and CVD. The procedure generally converged relatively quickly to meaningful values. At least when a normal random effect was assumed, for the gamma frailty the convergence took about 30 minutes. But PROC NLMIXED sometimes stopped before reaching the global maximum of the log-likelihood. This could be prevented by either restarting the algorithm with the 'final' values or by using a time scale in years instead of months. The latter increased the value of the parameters and prevented numerical issues in the fitting process. Since the rest of the simulation study was done on a parallel server with R, the joint frailty model in SAS is not included in the results, as transferring the simulated data to SAS and running them there would have required a long time.

In summary, the joint frailty model remains an attractive model in a setting with recurrent events and a terminal event. The NLMIXED implementation seems to work ok, but care must be taken that convergence to the global maximum of the log-likelihood function was achieved. Further investigations of this model and possibly improved software implementations are of interest.

E.2 Event-generating process

As a first step the enrollment timepoints were assigned equally-spaced across the 3 years of enrollment. Treatment groups were randomly assigned. In case no terminal event such as CVD or non-CVD occurred, patients were administratively censored at the end of the trial (5 years overall trial duration).

For the base case scenario the time from enrollment to CVD and to the next HHF for patient $i, i = 1, ..., n$, were generated using a joint frailty model, where the inter-event times are exponentially distributed conditional on the gamma distributed frailties z_i (mean 1 and variance θ) with rates given as

$$\lambda^*_{CV} = \lambda_{CV}\ z_i^{\alpha}\ \exp(x_i\beta_{CV}), \tag{31}$$

$$\lambda^*_{HHF} = \lambda_{HHF}\ z_i\ \exp(x_i\beta_{HHF}), \tag{32}$$

where λ_{CV} and λ_{HHF} are the respective rates in the control group, α defines the correlation between the two processes, x_i is the individual treatment identifier ($x_i = 1$ for the active treatment group, $x_i = 0$ for the control group) and $\exp(\beta_{CV}) = HR_{CV}$, $\exp(\beta_{HHF}) = RR_{HHF}$ are the treatment effects on CVD and HHFs, respectively. Depending on the respective scenario the frailty correlation was set to $\alpha = 0.5, 0.75$ or 1.

The control rates λ_{CV} and λ_{HHF} as well as the frailty variance θ were chosen so that the observed annualized control CVD rate is 4% (number of events per patient-year at risk), the observed annualized control rate of first composite event is 9% and the overall observed ratio of all to first events is 1.8. These parameters were adapted accordingly for the variations of the base case in order to observe the above mentioned annualized rates and ratio of all to first events. Table 14 gives the respective parameters for each considered scenario.

Time from enrollment to non-CVD was independently simulated according to an exponential distribution, where the rate λ_{NCV} (same in both treatment groups) was chosen such that the proportion of non-CVD of all deaths is around 30%. For all considered scenarios this was achieved with an annual rate of $\lambda_{NCV} = 0.01716$.

Time from enrollment to treatment discontinuation was varied to be either independent of the joint frailty process for CVD and HHFs or to depend on the hospitalization events (and indirectly on treatment). In the independent case it was simulated as an exponential process with rate $\lambda_{TD} = -\log(0.95) = 0.05129$ so that the proportion of treatment discontinuation

Table 14: Exact parameter values for the annual control rates $\lambda_{0,CV}$ and $\lambda_{0,HHF}$ as well as the frailty variance θ for all considered scenarios to obtain an observed annualized control CVD rate of 4%, an observed annualized control rate of first composite event of 9% and an overall observed ratio of all to first events of 1.8.

Scenario	λ_{CV}	λ_{HHF}	θ
Base case	0.07032	0.15444	5.7
Inter-event Weibull	0.168	0.3228	5.1
Autoregressive $\kappa = 1.1$	0.0678	0.1386	5.2
Autoregressive $\kappa = 1.2$	0.06492	0.1254	4.7
Frailty correlation $\alpha = 0.5$	0.07752	0.13956	6.0
Frailty correlation $\alpha = 1$	0.0612	0.1692	5.7
Numerical estimand	0.06036	0.16788	5.7

after 1 year is 5%. In the case of treatment discontinuation depending on hospitalizations, it was assumed that treatment is only discontinued directly after a hospitalization event, i.e. the higher the number of recurrent events the more likely treatment is discontinued. This was achieved by generating a realization of a Bernoulli distribution with success probability 0%, 5%, 10%, 15% or 20% each time an event occurs. After treatment discontinuation it was assumed that active treated patients 'jump' to the respective rate of the control group.

As one variation of the base case the inter-event times for patient i was assumed to be Weibull distributed conditional on the gamma distributed frailties with shape $\gamma = 0.75$ for both treatment groups and scale parameters equal to the rate parameters for the exponential case given in (31) and (32). The parametrization of the Weibull distribution used is such that with scale parameter λ and shape parameter γ the cumulative hazard function is given as $\Lambda(t) = \lambda t^{\gamma}$. As mentioned above the respective control rates λ_{CV} and λ_{HHF} as well as the frailty variance θ were adapted accordingly in order to obtain the required annualized rates and ratio of all to first events.

For the scenario with autoregressive event rate the same joint frailty process as for the base case scenario was chosen, except for the rates of the conditional

exponential process for patient i which were defined as

$$\lambda^*_{CV}(t) = \lambda_{CV}\ \kappa^{N_i(t)}\ z_i^{\alpha}\ \exp(x_i\beta_{CV}), \tag{33}$$

$$\lambda^*_{HHF}(t) = \lambda_{HHF}\ \kappa^{N_i(t)}\ z_i\ \exp(x_i\beta_{HHF}). \tag{34}$$

Here, $N_i(t)$ defines the number of hospitalizations that have already occurred for patient i until time t and κ defines the increase of the rates after each hospitalization which was set to $\kappa = 1.1, 1.2$. That means, the rates for both the time to CVD and the time to the next HHF are increasing after each hospitalization.

The exact parameter values for the determination of the numerical estimand with fixed follow-up time of 3.5 years and frailty correlation $\alpha = 1$ that resulted in the required annualized control rates and ratio of all to first events are also listed in Table 14.

E.3 Numeric estimand values

In this section, we compute the numeric estimand values under four scenarios. We consider here only the setting where all the patient have a fixed follow-up time of 3.5 years and use a correlation between the frailty of HHF and CVD of 1.

E.3.1 Assumptions and notations

- $N_1(t)$: number of HHF a patient has experienced by time t in treatment group
- $N_0(t)$: number of HHF a patient has experienced by time t in control group
- $M_1(t)$: number of HHF and CVD a patient has experienced by time t in treatment group
- $M_0(t)$: number of HHF and CVD a patient has experienced by time t in control group
- Z: frailty for recurrent event, has a gamma distribution with mean 1 and dispersion parameter $\theta = 5.7$, so the probability density function

is

$$f_Z(Z) = \frac{1}{\Gamma(1/\theta)} \left(\frac{Z}{\theta}\right)^{\frac{1}{\theta}-1} e^{-Z/\theta} \frac{1}{\theta}$$

- $U = Z^\alpha$: frailty for CVD, $\alpha = 1$
- $\lambda_{HHF} = 0.16788$: baseline recurrent event (hospitalization) rate
- $\lambda_{CV} = 0.06036$: baseline CVD rate
- $\lambda_{NCV} = 0.01716$: baseline non-CVD rate
- $\lambda_{TD} = 0.05129$: treatment discontinuation rate
- $RR_{HHF} = e^\beta = 0.7$: rate ratio of recurrent event (hospitalization)
- $HR_{CV} = e^\gamma = 0.8, 1.0, 1.25$: hazard ratio of CVD
- $E\{dN_1^*(t)/dt\} = \lambda_{HHF} Z e^\beta$: expected event rate for patients in treatment group
- $E\{dN_0^*(t)/dt\} = \lambda_{HHF} Z$: expected event rate for patients in control group
- $S_1(t)$: overall survival in treatment group
- $S_0(t)$: overall survival in control group
- $S_{CV1}(t, Z) = \exp(-\lambda_{CV} Z^\alpha e^\gamma t)$: survival function for CVD in treatment group
- $S_{CV0}(t, Z) = \exp(-\lambda_{CV} Z^\alpha t)$: survival function for CVD in control group
- $S_{NCV}(t) = \exp(-\lambda_{NCV} t)$: survival function for non CVD, $S_{NCV}(t)$ is independent of Z
- $S_{TD1}(t) = \exp(-\lambda_{TD} t)$: survival function for non-informative treatment discontinuation
- $S_{TD21}(t, Z) = E\{\prod_{s \le t}(1 - q dN_1^*(s))|Z\} = \prod_{s \le t}(1 - q\lambda_{HHF} Z e^\beta ds) = \exp(-q\lambda_{HHF} Z e^\beta t)$: survival function for informative treatment discontinuation in treatment group, $q = 10\%$

- $S_{TD20}(t,Z) = E\{\prod_{s\le t}(1-qdN_0^*(s))|Z\} = \prod_{s\le t}(1-q\lambda_{HHF}Zds) = \exp(-q\lambda_{HHF}Zt)$: survival function for informative treatment discontinuation in control group, $q = 10\%$

E.3.2 Analytic formula of summary measures

The summary measure for the two estimands are:

- Estimand 1 (HHF): $RR1 = \frac{E\{N_1(T)\}/\int_0^T S_1(t)dt}{E\{N_0(T)\}/\int_0^T S_0(t)dt}$. This could also be written as the product of two estimands: a). effect on recurrent events: $\frac{E\{N_1(T)\}}{E\{N_0(T)\}}$; b). effect on mortality: $\frac{\int_0^T S_0(t)dt}{\int_0^T S_1(t)dt} = \frac{RMST_0(T)}{RMST_1(T)}$, where $RMST = \int_0^T S(t)dt$ is the restricted mean survival time.
- Estimand 2 (HHF+CVD): $RR2 = \frac{E\{M_1(T)\}/\int_0^T S_1(t)dt}{E\{M_0(T)\}/\int_0^T S_0(t)dt}$.

In the following we will derive the analytic formula of summary measures for the two estimands under two types of treatment discontinuation.

E.3.2.1 Estimand 1 (HHF)

The summary measure is $RR1 = \frac{E\{N_1(T)\}/\int_0^T S_1(t)dt}{E\{N_0(T)\}/\int_0^T S_0(t)dt}$.

E.3.2.1.1 Non-informative treatment discontinuation

$$\begin{aligned}
\frac{E\{N_1(T)\}}{\int_0^T S_1(t)dt} &= \frac{\int_0^T\int_0^\infty \{S_{CV1}(t,Z)S_{NCV}(t)S_{TD1}(t)\lambda_{HHF}Ze^\beta + S_{CV1}(t,Z)S_{NCV}(t)[1-S_{TD1}(t)]\lambda_{HHF}Z\}f_Z(Z)dZdt}{\int_0^T\int_0^\infty \{S_{CV1}(t,Z)S_{NCV}(t)\}f_Z(Z)dZdt} \\
&= \frac{\int_0^T\int_0^\infty \{\lambda_{HHF}ZS_{CV1}(t,Z)S_{NCV}(t)[S_{TD1}(t)(e^\beta-1)+1]\}dZdt}{\int_0^T\int_0^\infty \{S_{CV1}(t,Z)S_{NCV}(t)\}f_Z(Z)dZdt} \\
&= \frac{\int_0^T\int_0^\infty \{\lambda_{HHF}Z\exp(-\lambda_{CV}Z^\alpha e^\gamma t)\exp(-\lambda_{NCV}t)[\exp(-\lambda_{TD}t)(e^\beta-1)+1]\}f_Z(Z)dZdt}{\int_0^T\int_0^\infty \{\exp(-\lambda_{CV}Z^\alpha e^\gamma t)\exp(-\lambda_{NCV}t)\}f_Z(Z)dZdt} \\
\frac{E\{N_0(T)\}}{\int_0^T S_0(t)dt} &= \frac{\int_0^T\int_0^\infty \{S_{CV0}(t,Z)S_{NCV}(t)S_{TD1}(t)\lambda_{HHF}Z + S_{CV0}(t,Z)S_{NCV}(t)[1-S_{TD1}(t)]\lambda_{HHF}Z\}f_Z(Z)dZdt}{\int_0^T\int_0^\infty \{S_{CV0}(t,Z)S_{NCV}(t)\}f_Z(Z)dZdt} \\
&= \frac{\int_0^T\int_0^\infty \{\lambda_{HHF}ZS_{CV0}(t,Z)S_{NCV}(t)\}f_Z(Z)dZdt}{\int_0^T\int_0^\infty \{S_{CV0}(t,Z)S_{NCV}(t)\}f_Z(Z)dZdt} \\
&= \frac{\int_0^T\int_0^\infty \{\lambda_{HHF}Z\exp(-\lambda_{CV}Z^\alpha t)\exp(-\lambda_{NCV}t)\}f_Z(Z)dZdt}{\int_0^T\int_0^\infty \{\exp(-\lambda_{CV}Z^\alpha t)\exp(-\lambda_{NCV}t)\}f_Z(Z)dZdt}
\end{aligned}$$

Now integrate out Z, it is only integratable when $\alpha = 1$. Let

$$
\begin{aligned}
A &= \int_0^\infty Z\exp(-\lambda_{CV}Ze^{\gamma}t)f_Z(Z)dZ \\
&= \int_0^\infty \frac{(\frac{1}{\theta})^{\frac{1}{\theta}}}{\Gamma(1/\theta)}Z^{\frac{1}{\theta}}\exp(-\frac{1}{\theta}Z)\exp(-\lambda_{CV}e^{\gamma}tZ)dZ \\
&= \frac{(\frac{1}{\theta})^{\frac{1}{\theta}}}{\Gamma(1/\theta)} \times \frac{\Gamma(1+1/\theta)}{(\lambda_{CV}e^{\gamma}t+\frac{1}{\theta})^{\frac{1}{\theta}+1}}\int_0^\infty \frac{(\lambda_{CV}e^{\gamma}t+\frac{1}{\theta})^{\frac{1}{\theta}+1}}{\Gamma(1+1/\theta)}Z^{\frac{1}{\theta}}\exp\{-(\frac{1}{\theta}+\lambda_{CV}e^{\gamma}t)Z\}dZ \\
&= (\frac{1}{\lambda_{CV}e^{\gamma}\theta t+1})^{\frac{1}{\theta}+1} \\
B &= \int_0^\infty Z\exp(-\lambda_{CV}Zt)f_Z(Z)dZ = (\frac{1}{\lambda_{CV}\theta t+1})^{\frac{1}{\theta}+1} \\
C &= \int_0^\infty \exp(-\lambda_{CV}Ze^{\gamma}t)f_Z(Z)dZ = (\frac{1}{\lambda_{CV}e^{\gamma}\theta t+1})^{\frac{1}{\theta}} \\
D &= \int_0^\infty \exp(-\lambda_{CV}Zt)f_Z(Z)dZ = (\frac{1}{\lambda_{CV}\theta t+1})^{\frac{1}{\theta}}
\end{aligned}
$$

Therefore

$$
\begin{aligned}
\frac{E\{N_1(T)\}}{\int_0^T S_1(t)dt} &= \frac{\int_0^T A\exp(-\lambda_{NCV}t)[\exp(-\lambda_{TD}t)(e^{\beta}-1)+1]dt}{\int_0^T C\exp(-\lambda_{NCV}t)dt} \\
&= \frac{\int_0^T (\frac{1}{\lambda_{CV}e^{\gamma}\theta t+1})^{\frac{1}{\theta}+1}\exp(-\lambda_{NCV}t)[\exp(-\lambda_{TD}t)(e^{\beta}-1)+1]dt}{\int_0^T (\frac{1}{\lambda_{CV}e^{\gamma}\theta t+1})^{\frac{1}{\theta}}\exp(-\lambda_{NCV}t)dt} \\
\frac{E\{N_0(T)\}}{\int_0^T S_0(t)dt} &= \frac{\int_0^T B\exp(-\lambda_{NCV}t)dt}{\int_0^T D\exp(-\lambda_{NCV}t)dt} = \frac{\int_0^T (\frac{1}{\lambda_{CV}\theta t+1})^{\frac{1}{\theta}+1}\exp(-\lambda_{NCV}t)dt}{\int_0^T (\frac{1}{\lambda_{CV}\theta t+1})^{\frac{1}{\theta}}\exp(-\lambda_{NCV}t)dt}
\end{aligned}
$$

E.3.2.1.2 Informative treatment discontinuation

$$
\begin{aligned}
\frac{E\{N_1(T)\}}{\int_0^T S_1(t)dt} &= \frac{\int_0^T\int_0^\infty\{S_{CV1}(t,Z)S_{NCV}(t)S_{TD21}(t,Z)\lambda_{HHF}Ze^{\beta}+S_{CV1}(t,Z)S_{NCV}(t)[1-S_{TD21}(t,Z)]\lambda_{HHF}Z\}f_Z(Z)dZdt}{\int_0^T\int_0^\infty\{S_{CV1}(t,Z)S_{NCV}(t)\}f_Z(Z)dZdt} \\
&= \frac{\int_0^T\int_0^\infty\{\lambda_{HHF}ZS_{CV1}(t,Z)S_{NCV}(t)[S_{TD21}(t)(e^{\beta}-1)+1]\}f_Z(Z)dZdt}{\int_0^T\int_0^\infty\{S_{CV1}(t,Z)S_{NCV}(t)\}f_Z(Z)dZdt} \\
&= \frac{\int_0^T\int_0^\infty\{\lambda_{HHF}Z\exp(-\lambda_{CV}Z^{\alpha}e^{\gamma}t)\exp(-\lambda_{NCV}t)[\exp(-q\lambda_{HHF}Ze^{\beta}t)(e^{\beta}-1)+1]\}f_Z(Z)dZdt}{\int_0^T\int_0^\infty\{\exp(-\lambda_{CV}Z^{\alpha}e^{\gamma}t)\exp(-\lambda_{NCV}t)\}f_Z(Z)dZdt} \\
\frac{E\{N_0(T)\}}{\int_0^T S_0(t)dt} &= \frac{\int_0^T\int_0^\infty\{S_{CV0}(t,Z)S_{NCV}(t)S_{TD20}(t,Z)\lambda_{HHF}Z+S_{CV0}(t,Z)S_{NCV}(t)[1-S_{TD20}(t,Z)]\lambda_{HHF}Z\}f_Z(Z)dZdt}{\int_0^T\int_0^\infty\{S_{CV0}(t,Z)S_{NCV}(t)\}f_Z(Z)dZdt} \\
&= \frac{\int_0^T\int_0^\infty\{\lambda_{HHF}ZS_{CV0}(t,Z)S_{NCV}(t)\}f_Z(Z)dZdt}{\int_0^T\int_0^\infty\{S_{CV0}(t,Z)S_{NCV}(t)\}f_Z(Z)dZdt} \\
&= \frac{\int_0^T\int_0^\infty\{\lambda_{HHF}Z\exp(-\lambda_{CV}Z^{\alpha}t)\exp(-\lambda_{NCV}t)\}f_Z(Z)dZdt}{\int_0^T\int_0^\infty\{\exp(-\lambda_{CV}Z^{\alpha}t)\exp(-\lambda_{NCV}t)\}f_Z(Z)dZdt}
\end{aligned}
$$

Now integrate out Z, it is only integratable when $\alpha = 1$. Let

$$G = \int_0^\infty Z \exp(-\lambda_{CV} Z e^\gamma t) \exp(-q\lambda_{HHF} Z e^\beta t) f_Z(Z) dZ = (\frac{1}{(\lambda_{CV} e^\gamma + q\lambda_{HHF} e^\beta)\theta t + 1})^{\frac{1}{\theta}+1}$$

Therefore, we have

$$\begin{aligned}
\frac{E\{N_1(T)\}}{\int_0^T S_1(t)dt} &= \frac{\int_0^T \exp(-\lambda_{NCV} t)[G(e^\beta - 1) + A]dt}{\int_0^T C \exp(-\lambda_{NCV} t)dt} \\
&= \frac{\int_0^T (\exp(-\lambda_{NCV} t)[(\frac{1}{(\lambda_{CV} e^\gamma + q\lambda_{HHF} e^\beta)\theta t+1})^{\frac{1}{\theta}+1}(e^\beta - 1) + (\frac{1}{\lambda_{CV} e^\gamma \theta t+1})^{\frac{1}{\theta}+1}]dt}{\int_0^T (\frac{1}{\lambda_{CV} e^\gamma \theta t+1})^{\frac{1}{\theta}} \exp(-\lambda_{NCV} t)dt} \\
\frac{E\{N_0(T)\}}{\int_0^T S_0(t)dt} &= \frac{\int_0^T B \exp(-\lambda_{NCV} t)dt}{\int_0^T D \exp(-\lambda_{NCV} t)dt} = \frac{\int_0^T (\frac{1}{\lambda_{CV} \theta t+1})^{\frac{1}{\theta}+1} \exp(-\lambda_{NCV} t)dt}{\int_0^T (\frac{1}{\lambda_{CV} \theta t+1})^{\frac{1}{\theta}} \exp(-\lambda_{NCV} t)dt}
\end{aligned}$$

E.3.2.2 Estimand 2 (HHF+CVD) The summary measure is $RR2 = \frac{E\{M_1(T)\}/\int_0^T S_1(t)dt}{E\{M_0(T)\}/\int_0^T S_0(t)dt}$.

E.3.2.2.1 Non-informative treatment discontinuation There are four terms in the following formula, the first two on line 1 represent HHF, the second two on line 2 represent CVD.

- First term, the probability of patients who did not die from CVD or non-CVD times the probability of patients did not discontinue at time t, times the event rate of hospitalization in treatment group.
- Second term, the probability of patients who did not die from CVD or non-CVD times the probability of patients discontinued at time t, times the event rate of hospitalization in control group.
- Third term, the probability of patients who did not die from non-CVD times the probability of patients did not discontinue at time t, times the probability of CVD in treatment group.
- Fourth term, the probability of patients who did not die from non-CVD times the probability of patients discontinued at time t, times the probability of CVD in control group.

$$
\begin{aligned}
E\{M_1(T)\} &= \int_0^T \int_0^\infty \{S_{CV1}(t,Z)S_{NCV}(t)S_{TD1}(t)\lambda_{HHF}Ze^{\beta} + S_{CV1}(t,Z)S_{NCV}(t)[1-S_{TD1}(t)]\lambda_{HHF}Z\}f_Z(Z)dZdt \\
&+ \int_0^\infty \int_0^T \{S_{NCV}(t)S_{TD1}(t)d(-S_{CV1}(t,Z))/dt + S_{NCV}(t)[1-S_{TD1}(t)]d(-S_{CV0}(t,Z))/dt\}f_Z(Z)dZ \\
&= \int_0^T \int_0^\infty \{S_{CV1}(t,Z)S_{NCV}(t)S_{TD1}(t)\lambda_{HHF}Ze^{\beta} + S_{CV1}(t,Z)S_{NCV}(t)[1-S_{TD1}(t)]\lambda_{HHF}Z \\
&+ S_{CV1}(t,Z)S_{NCV}(t)S_{TD1}(t)\lambda_{CV}Z^{\alpha}e^{\gamma} + S_{CV0}(t,Z)S_{NCV}(t)[1-S_{TD1}(t)]\lambda_{CV}Z^{\alpha}\}f_Z(Z)dZdt \\
&= \int_0^T \exp(-\lambda_{NCV}t)\int_0^\infty \{\lambda_{HHF}Z\exp(-\lambda_{CV}Z^{\alpha}e^{\gamma}t)[\exp(-\lambda_{TD}t)(e^{\beta}-1)+1] \\
&+ \exp(-\lambda_{TD}t)\lambda_{CV}Z^{\alpha}e^{\gamma}\exp(-\lambda_{CV}Z^{\alpha}e^{\gamma}t) + (1-\exp(-\lambda_{TD}t))\lambda_{CV}Z^{\alpha}\exp(-\lambda_{CV}Z^{\alpha}t)\}f_Z(Z)dZdt \\
E\{M_0(T)\} &= \int_0^T \int_0^\infty \{S_{CV0}(t,Z)S_{NCV}(t)S_{TD1}(t)\lambda_{HHF}Z + S_{CV0}(t,Z)S_{NCV}(t)[1-S_{TD1}(t)]\lambda_{HHF}Z\}f_Z(Z)dZdt \\
&+ \int_0^\infty \int_0^T \{S_{NCV}(t)S_{TD1}(t)d(-S_{CV0}(t,Z))/dt + S_{NCV}(t)[1-S_{TD1}(t)]d(-S_{CV0}(t,Z))/dt\}f_Z(Z)dZ \\
&= \int_0^T \int_0^\infty \{S_{CV0}(t,Z)S_{NCV}(t)\lambda_{HHF}Z + S_{CV0}(t,Z)S_{NCV}(t)\lambda_{CV}Z^{\alpha}\}f_Z(Z)dZdt \\
&= \int_0^T \exp(-\lambda_{NCV}t)\int_0^\infty \{\lambda_{HHF}Z\exp(-\lambda_{CV}Z^{\alpha}t) + \lambda_{CV}Z^{\alpha}\exp(-\lambda_{CV}Z^{\alpha}t)\}f_Z(Z)dZdt
\end{aligned}
$$

Now integrate out Z, it is only integratable when $\alpha = 1$.

$$
\begin{aligned}
E\{M_1(T)\} &= \int_0^T \exp(-\lambda_{NCV}t)\int_0^\infty \{Z\exp(-\lambda_{CV}Ze^{\gamma}t)[\lambda_{HHF}(\exp(-\lambda_{TD}t)(e^{\beta}-1)+1) + \exp(-\lambda_{TD}t)\lambda_{CV}e^{\gamma})] \\
&+ (1-\exp(-\lambda_{TD}t))\lambda_{CV}Z\exp(-\lambda_{CV}Zt)\}dZdt \\
&= \int_0^T \exp(-\lambda_{NCV}t)\{A[\lambda_{HHF}\exp(-\lambda_{TD}t)(e^{\beta}-1)+1) + \exp(-\lambda_{TD}t)\lambda_{CV}e^{\gamma})] + (1-\exp(-\lambda_{TD}t))\lambda_{CV}B\}dt \\
E\{M_0(T)\} &= \int_0^T \exp(-\lambda_{NCV}t)\int_0^\infty \{Z\exp(-\lambda_{CV}Zt)(\lambda_{HHF}+\lambda_{CV})\} \\
&= \int_0^T \exp(-\lambda_{NCV}t)\{B(\lambda_{HHF}+\lambda_{CV})\}dt
\end{aligned}
$$

Therefore, we have

$$
\begin{aligned}
\frac{E\{M_1(T)\}}{\int_0^T S_1(t)dt} &= \frac{\int_0^T \exp(-\lambda_{NCV}t)\{A[\lambda_{HHF}(\exp(-\lambda_{TD}t)(e^{\beta}-1)+1) + \exp(-\lambda_{TD}t)\lambda_{CV}e^{\gamma})] + (1-\exp(-\lambda_{TD}t))\lambda_{CV}B\}dt}{\int_0^T C\exp(-\lambda_{NCV}t)dt} \\
\frac{E\{M_0(T)\}}{\int_0^T S_0(t)dt} &= \frac{\int_0^T \exp(-\lambda_{NCV}t)\{B(\lambda_{HHF}+\lambda_{CV})\}dt}{\int_0^T D\exp(-\lambda_{NCV}t)dt}
\end{aligned}
$$

E.3.2.2.2 Informative treatment discontinuation

$$
\begin{aligned}
E\{M_1(T)\} &= \int_0^T \int_0^\infty \{S_{CV1}(t,Z)S_{NCV}(t)S_{TD21}(t,Z)\lambda_{HHF}Ze^{\beta} + S_{CV1}(t,Z)S_{NCV}(t)[1-S_{TD21}(t,Z)]\lambda_{HHF}Z\}f_Z(Z)dZdt \\
&+ \int_0^\infty \int_0^T \{S_{NCV}(t)S_{TD21}(t,Z)d(-S_{CV1}(t,Z))/dt + S_{NCV}(t)[1-S_{TD21}(t,Z)]d(-S_{CV0}(t,Z))/dt\}f_Z(Z)dZ \\
&= \int_0^T \int_0^\infty \{S_{CV1}(t,Z)S_{NCV}(t)S_{TD21}(t,Z)\lambda_{HHF}Ze^{\beta} + S_{CV1}(t,Z)S_{NCV}(t)[1-S_{TD21}(t,Z)]\lambda_{HHF}Z \\
&+ S_{CV1}(t,Z)S_{NCV}(t)S_{TD21}(t,Z)\lambda_{CV}Z^{\alpha}e^{\gamma} + S_{CV0}(t,Z)S_{NCV}(t)[1-S_{TD21}(t,Z)]\lambda_{CV}Z^{\alpha}\}f_Z(Z)dZdt \\
&= \int_0^T \exp(-\lambda_{NCV}t)\int_0^\infty \{\lambda_{HHF}Z\exp(-\lambda_{CV}Z^{\alpha}e^{\gamma}t)[\exp(-q\lambda_{HHF}Ze^{\beta}t)(e^{\beta}-1)+1] \\
&+ \exp(-q\lambda_{HHF}Ze^{\beta}t)\lambda_{CV}Z^{\alpha}e^{\gamma}\exp(-\lambda_{CV}Z^{\alpha}e^{\gamma}t) + (1-\exp(-q\lambda_{HHF}Ze^{\beta}t))\lambda_{CV}Z^{\alpha}\exp(-\lambda_{CV}Z^{\alpha}t)\}f_Z(Z)dZdt \\
E\{M_0(T)\} &= \int_0^T \int_0^\infty \{S_{CV0}(t,Z)S_{NCV}(t)S_{TD20}(t,Z)\lambda_{HHF}Z + S_{CV0}(t,Z)S_{NCV}(t)[1-S_{TD20}(t,Z)]\lambda_{HHF}Z\}f_Z(Z)dZdt \\
&+ \int_0^\infty \int_0^T \{S_{NCV}(t)S_{TD20}(t,Z)d(-S_{CV0}(t,Z))/dt + S_{NCV}(t)[1-S_{TD20}(t,Z)]d(-S_{CV0}(t,Z))/dt\}f_Z(Z)dZ \\
&= \int_0^T \int_0^\infty \{S_{CV0}(t,Z)S_{NCV}(t)\lambda_{HHF}Z + S_{CV0}(t,Z)S_{NCV}(t)\lambda_{CV}Z^{\alpha}\}f_Z(Z)dZdt \\
&= \int_0^T \exp(-\lambda_{NCV}t)\int_0^\infty \{\lambda_{HHF}Z\exp(-\lambda_{CV}Z^{\alpha}t) + \lambda_{CV}Z^{\alpha}\exp(-\lambda_{CV}Z^{\alpha}t)\}f_Z(Z)dZdt
\end{aligned}
$$

Now integrate out Z, it is only integratable when $\alpha = 1$. Let

$$
H = \int_0^\infty Z\exp(-\lambda_{CV}Zt)\exp(-q\lambda_{HHF}Ze^{\beta}t)f_Z(Z)dZ = \left(\frac{1}{(\lambda_{CV}+q\lambda_{HHF}e^{\beta})\theta t+1}\right)^{\frac{1}{\theta}+1}
$$

Therefore, we have

$$
\begin{aligned}
\frac{E\{M_1(T)\}}{\int_0^T S_1(t)dt} &= \frac{\int_0^T \exp(-\lambda_{NCV}t)\{\lambda_{HHF}[G(e^{\beta}-1)+A] + \lambda_{CV}e^{\gamma}G + \lambda_{CV}(B-H)\}dt}{\int_0^T C\exp(-\lambda_{NCV}t)dt} \\
\frac{E\{M_0(T)\}}{\int_0^T S_0(t)dt} &= \frac{\int_0^T \exp(-\lambda_{NCV}t)B(\lambda_{HHF}+\lambda_{CV})dt}{\int_0^T D\exp(-\lambda_{NCV}t)dt}
\end{aligned}
\tag{35}
$$

E.3.3 Numeric values

After numerical integration on t and plugging in all the parameters in the summary measure of the two estimands, we get Table 15.

Table 15: Numerical estimand values for two estimands with two types of treatment discontinuation. Data is generated with $RR_{HHF} = 0.7$, $HR_{CV} = 0.8, 1.0, 1.25$

	Estimand value		
HR_{CV}	0.8	1.0	1.25
Scenario 1: Estimand 1 (HHF), non-informative	0.767	0.721	0.672
Scenario 2: Estimand 1 (HHF), informative	0.767	0.719	0.669
Scenario 3: Estimand 2 (HHF+CVD), non-informative	0.812	0.815	0.820
Scenario 4: Estimand 2 (HHF+CVD), informative	0.790	0.793	0.800

E.4 Event specific estimates for WLW and PWP models - terminal event

The event specific mean treatment effects of WLW and PWP for the base case are shown in Table 16 for Estimand 1 (HHF) and Estimand 2 (HHF+CVD).

- For the PWP event specific estimates it can be observed that with increasing event number the event specific estimates are increasing (this applies to both Estimand 1 (HHF) and Estimand 2 (HHF+CVD) and all considered scenarios).
- For the WLW event specific estimates are decreasing with increasing event numbers in most considered scenarios. However, in case the treatment effect on CVD is high compared with the treatment effect on HHFs ($HR_{CV} < RR_{HHF}$), it is the other way around.
- In addition, the WLW effect estimates for the 4th event are decreasing when HR_{CV} converges to 1. The PWP estimates for the 4th event are generally very close to 1.
- The reason for the differences between the event specific estimates of the WLW model is that 2nd, 3rd and 4th events are highly influenced by dependent censoring.
- Unlike for the scenario without terminal event no cases of non-convergence were observed here, presumably because of the larger sample size.

Table 16: Event specific mean treatment effect estimates of WLW and PWP for the composite endpoint (CVD & HHFs) and the recurrent endpoint (only HHFs) in the base case scenario with sample size $N = 4350$.

Endpoint	RR_{HHF}	Method	Event	$HR_{CV} = 0.6$	$HR_{CV} = 0.8$	$HR_{CV} = 1.0$
Estimand 1 (HHF)	0.6	WLW	1	0.780	0.755	0.731
			2	0.684	0.637	0.595
			3	0.541	0.467	0.407
			4	0.397	0.307	0.243
		PWP	1	0.780	0.755	0.731
			2	0.811	0.785	0.763
			3	0.865	0.844	0.828
			4	0.959	0.945	0.940
	0.8	WLW	1	0.928	0.902	0.878
			2	0.910	0.857	0.807
			3	0.891	0.788	0.694
			4	0.884	0.714	0.572
		PWP	1	0.928	0.902	0.878
			2	0.939	0.917	0.893
			3	0.966	0.950	0.926
			4	1.009	1.002	0.990
	1.0	WLW	1	1.055	1.030	1.004
			2	1.121	1.062	1.008
			3	1.277	1.142	1.023
			4	1.578	1.295	1.071
		PWP	1	1.055	1.030	1.004
			2	1.051	1.026	1.005
			3	1.055	1.035	1.017
			4	1.068	1.054	1.045
Estimand 2 (HHF+CVD)	0.6	WLW	1	0.770	0.811	0.851
			2	0.675	0.674	0.673
			3	0.531	0.481	0.439
			4	0.386	0.305	0.245
		PWP	1	0.770	0.811	0.851
			2	0.800	0.831	0.863
			3	0.851	0.870	0.892
			4	0.936	0.942	0.952
	0.8	WLW	1	0.859	0.896	0.932
			2	0.851	0.850	0.847
			3	0.844	0.778	0.715
			4	0.847	0.698	0.572
		PWP	1	0.859	0.896	0.932
			2	0.878	0.910	0.938
			3	0.916	0.938	0.954
			4	0.972	0.984	0.993
	1.0	WLW	1	0.936	0.971	1.003
			2	1.011	1.009	1.005
			3	1.179	1.095	1.018
			4	1.494	1.251	1.053
		PWP	1	0.936	0.971	1.003
			2	0.949	0.975	1.003
			3	0.975	0.993	1.012
			4	1.014	1.021	1.031

E.5 Simulation results for variations of the base case

The most interesting findings for the variations are summarized below, we only consider non-informative treatment discontinuation for all variations of the base case. Further details of the results are displayed by each variation.

- **Inter-event Weibull**: Regarding power the relative behavior of the methods is similar as in the base case. There is a somewhat higher power for small HR_{CV}, especially for NB but to a lesser extent also for PWP and Cox. The mean treatment effect estimates tend to be further away from 1 than in the base case for all methods. This is especially pronounced for NB in the case of small HR_{CV}, e.g., mean $\widehat{RR}_{HHF} = 0.649$ for Estimand 2 (HHF+CVD) with $HR_{CV} = RR_{HHF} = 0.7$. There is an increased type I error of 0.066 for the composite and of 0.064 for Estimand 1 (HHF) for NB in the case $RR_{HHF} = HR_{CV} = 1$ which is not seen for the other methods. Reason for this probably is the deviation from the distributional assumptions of the NB model (mixed Poisson-gamma), which doesn't influence the other semi-parametric models as the inter-event Weibull scenario fulfills the proportional hazards assumption.

- **Autoregressive event rate**: Overall similar behavior as for the base case. For Estimand 2 (HHF+CVD) there is a higher relative power for NB for HRs closer to 1, and higher relative power for Cox, WLW and PWP for HRs further away from 1. For Estimand 1 (HHF) there is a lower relative power for LWYY. For both Estimand 1 (HHF) and Estimand 2 (HHF+CVD) the mean treatment effect tends to be a bit further away from 1, most pronounced for NB. Small increase in type I error up to 0.056 for Estimand 2 (HHF+CVD) in the case $RR_{HHF} = HR_{CV} = 1$ for LWYY, WLW and PWP with the multiplicative factor 1.2, a bit higher increase in type I error of 0.065 for NB. For Estimand 1 (HHF) the type I error inflation of NB is even higher, e.g. 0.079 for $RR_{HHF} = HR_{CV} = 1$. This is probably due to the deviation from the constant baseline rate assumption of the NB model.

- **Detrimental CVD effect**: Trends seen for positive to neutral CVD effect in the base case continue for a detrimental CVD effect. In the case of power, for HR_{CV} above 1 there are further increases for Estimand 1 (HHF). For Estimand 2 (HHF+CVD), there is only a small decrease

in power for LWYY and WLW and a strong decrease for the other methods.

- **Frailty correlation**: For Estimand 2 (HHF+CVD) there is generally higher power for smaller exponent α. Relative to the other methods NB and PWP seem to have higher power for small α. For fixed RR_{HHF}, LWYY and WLW have slightly decreasing power as HR_{CV} approaches 1 for $\alpha = 0.5$ and roughly constant power for $\alpha = 1.0$. There are only minor differences to the base case for Estimand 1 (HHF). For Estimand 2 (HHF+CVD) treatment effect estimates are further away from 1 for $\alpha = 0.5$ and small HR_{CV}, while for $\alpha = 1.0$ they are closer to 1. This is similar for Estimand 1 (HHF), but less pronounced. For smaller correlation ($\alpha = 0.5$) all recurrent event methods tend to be more conservative than for larger correlation ($\alpha = 1.0$). In particular for Estimand 1 (HHF) with small HR_{CV} the type I error inflation becomes significantly smaller with smaller correlation ($\alpha = 0.5$). This observation is most likely caused by a decreased influence of dependent censoring for smaller correlation between CVD and HHFs.

E.5.1 Inter-event Weibull

Table 17: Mean treatment effect estimates for Estimand 1 (HHF) and Estimand 2 (HHF+CVD) for inter-event Weibull with sample size $N = 4350$, with non-informative treatment discontinuation.

Endpoint	RR_{HHF}	Method	$HR_{CV} = 0.6$	$HR_{CV} = 0.8$	$HR_{CV} = 1.0$
Estimand 1 (HHF)	0.6	Cox	0.760	0.735	0.712
		NB	0.602	0.580	0.560
		LWYY	0.683	0.640	0.603
		WLW	0.699	0.661	0.626
		PWP	0.776	0.752	0.728
	0.8	Cox	0.917	0.892	0.867
		NB	0.827	0.800	0.775
		LWYY	0.911	0.852	0.801
		WLW	0.909	0.861	0.818
		PWP	0.924	0.899	0.875
	1.0	Cox	1.053	1.027	1.002
		NB	1.065	1.035	1.006
		LWYY	1.146	1.069	1.004
		WLW	1.106	1.052	1.003
		PWP	1.050	1.025	1.001
Estimand 2 (HHF+CVD)	0.6	Cox	0.750	0.798	0.843
		NB	0.551	0.625	0.701
		LWYY	0.678	0.694	0.710
		WLW	0.692	0.715	0.736
		PWP	0.765	0.809	0.850
	0.8	Cox	0.842	0.886	0.927
		NB	0.691	0.769	0.851
		LWYY	0.843	0.848	0.855
		WLW	0.839	0.856	0.873
		PWP	0.854	0.892	0.931
	1.0	Cox	0.922	0.963	1.002
		NB	0.834	0.919	1.006
		LWYY	1.012	1.006	1.003
		WLW	0.977	0.990	1.002
		PWP	0.930	0.966	1.001

Figure 17: Statistical power for Estimand 1 (HHF) for inter-event Weibull with sample size $N = 4350$, with non-informative treatment discontinuation.

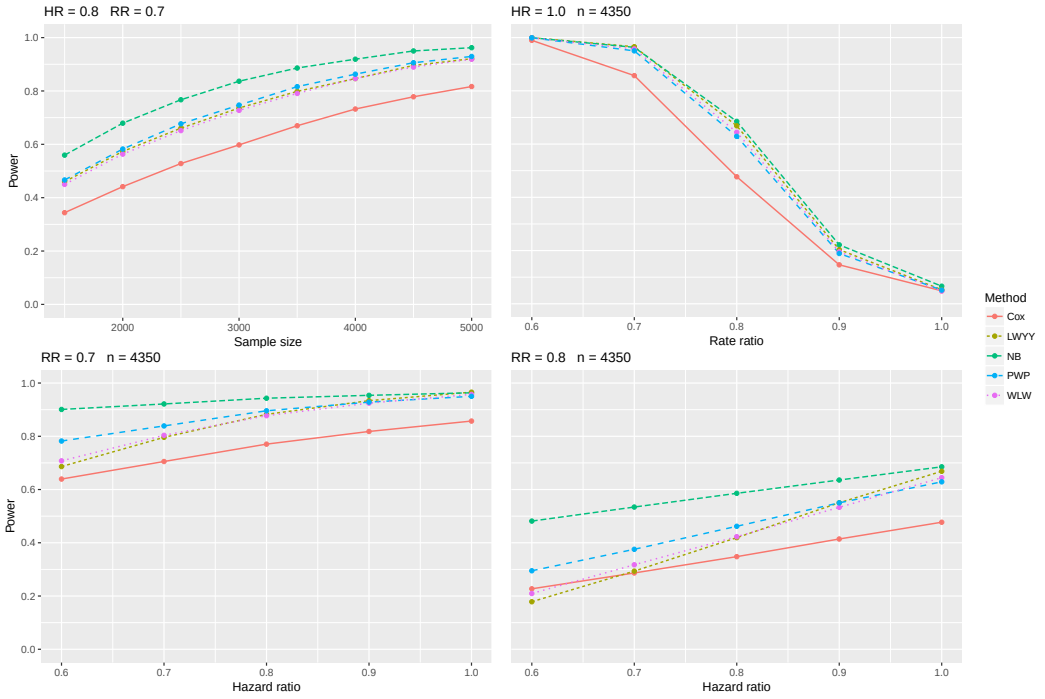


Figure 18: Statistical power for Estimand 2 (HHF+CVD) for inter-event Weibull with sample size $N = 4350$, with non-informative treatment discontinuation.

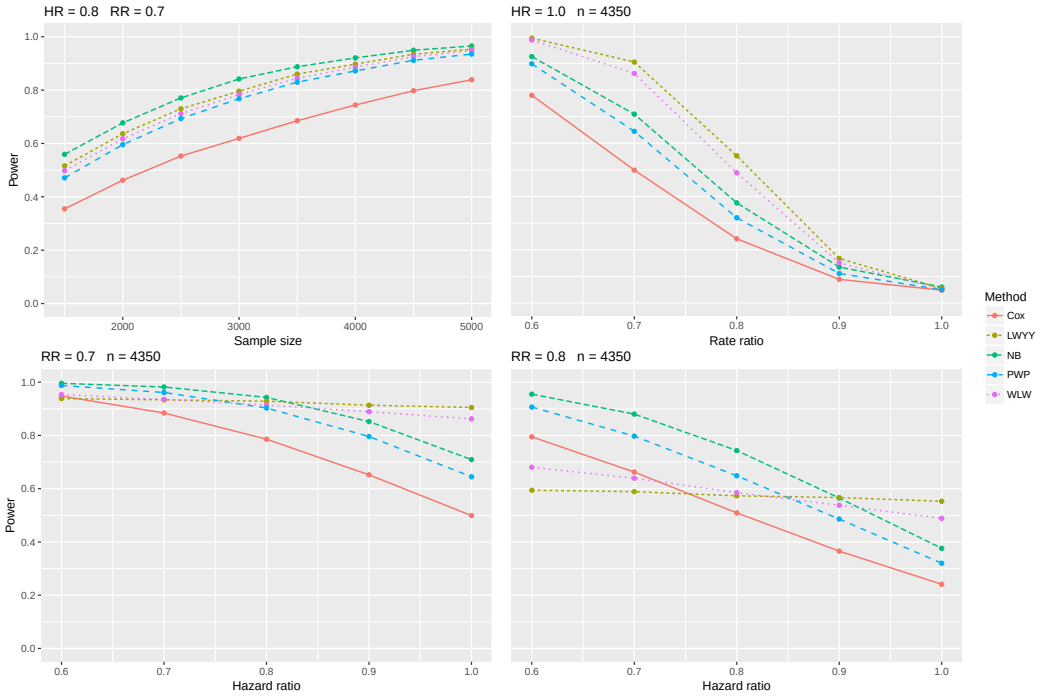


Table 18: Mean treatment effect estimates and type I error rates for Estimand 1 (HHF) and Estimand 2 (HHF+CVD) with $RR_{HHF} = 1$ for inter-event Weibull and sample size $N = 4350$, with non-informative treatment discontinuation.

Endpoint	HR_{CV}	Method	Estimate	Type I error
Estimand 1 (HHF)	0.6	Cox	1.053	0.099
		NB	1.065	0.104
		LWYY	1.146	0.285
		WLW	1.106	0.210
		PWP	1.050	0.138
	0.8	Cox	1.027	0.063
		NB	1.035	0.078
		LWYY	1.069	0.110
		WLW	1.052	0.088
		PWP	1.025	0.076
	1.0	Cox	1.002	0.050
		NB	1.006	0.065
		LWYY	1.004	0.053
		WLW	1.003	0.050
		PWP	1.001	0.050
Estimand 2 (HHF+CVD)	1.0	Cox	1.002	0.051
		NB	1.006	0.064
		LWYY	1.003	0.051
		WLW	1.002	0.050
		PWP	1.001	0.051

E.5.2 Autoregressive event rate

Table 19: Mean treatment effect estimates for Estimand 1 (HHF) and Estimand 2 (HHF+CVD) for autoregressive event rate (factor 1.1) with sample size N = 4350, with non-informative treatment discontinuation.

Endpoint	RR_{HHF}	Method	$HR_{CV} = 0.6$	$HR_{CV} = 0.8$	$HR_{CV} = 1.0$
Estimand 1 (HHF)	0.6	Cox	0.769	0.744	0.721
		NB	0.640	0.610	0.585
		LWYY	0.692	0.648	0.612
		WLW	0.708	0.670	0.637
		PWP	0.784	0.759	0.736
	0.8	Cox	0.921	0.896	0.872
		NB	0.861	0.823	0.790
		LWYY	0.914	0.856	0.806
		WLW	0.911	0.866	0.825
		PWP	0.928	0.903	0.880
	1.0	Cox	1.055	1.028	1.003
		NB	1.099	1.049	1.006
		LWYY	1.148	1.069	1.005
		WLW	1.103	1.051	1.004
		PWP	1.051	1.026	1.002
Estimand 2 (HHF+CVD)	0.6	Cox	0.759	0.804	0.846
		NB	0.606	0.658	0.712
		LWYY	0.687	0.701	0.716
		WLW	0.701	0.722	0.743
		PWP	0.773	0.814	0.853
	0.8	Cox	0.849	0.890	0.929
		NB	0.752	0.804	0.856
		LWYY	0.850	0.853	0.858
		WLW	0.845	0.861	0.876
		PWP	0.860	0.897	0.932
	1.0	Cox	0.930	0.966	1.002
		NB	0.906	0.955	1.006
		LWYY	1.021	1.009	1.004
		WLW	0.980	0.992	1.003
		PWP	0.936	0.969	1.002

Figure 19: Statistical power for Estimand 1 (HHF) for autoregressive event rate (factor 1.1) with sample size $N = 4350$, with non-informative treatment discontinuation.

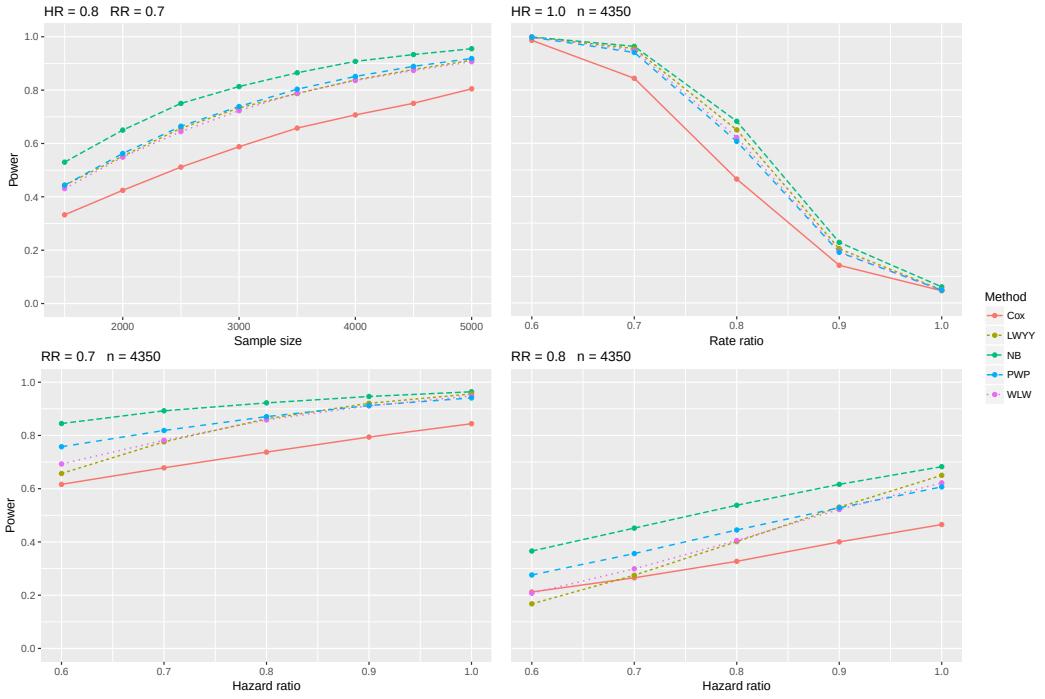


Figure 20: Statistical power for Estimand 2 (HHF+CVD) for autoregressive event rate (factor 1.1) with sample size $N = 4350$, with non-informative treatment discontinuation.

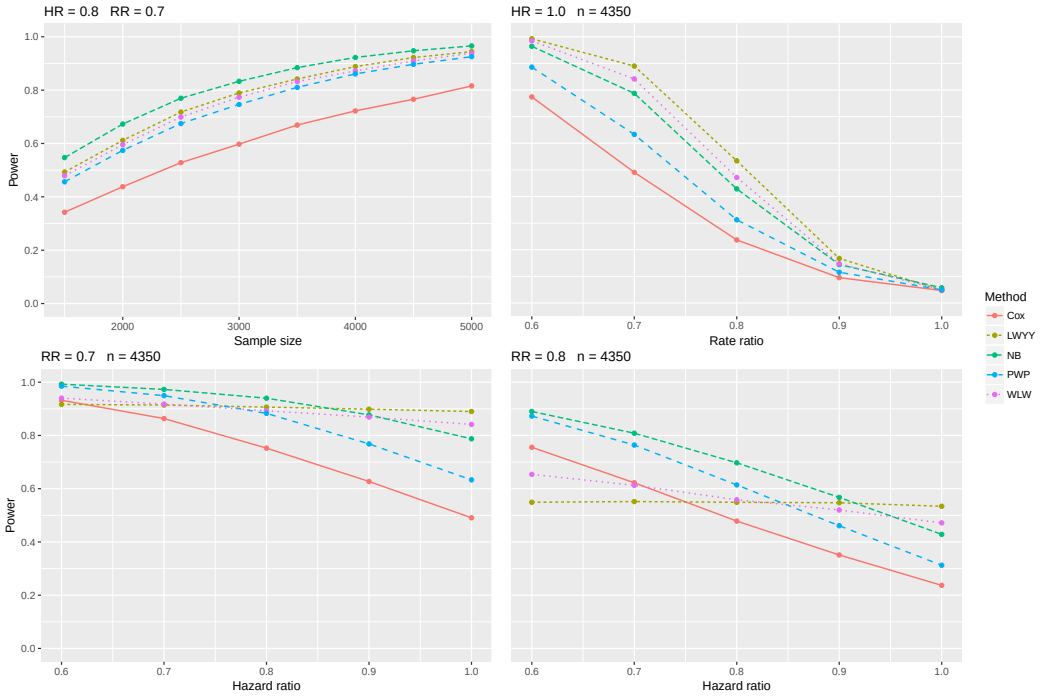


Table 20: Mean treatment effect estimates and type I error rates for Estimand 1 (HHF) and Estimand 2 (HHF+CVD) for autoregressive event rate (factor 1.1) with $RR_{HHF} = 1$ and sample size $N = 4350$, with non-informative treatment discontinuation

Endpoint	HR_{CV}	Method	Estimate	Type I error
Estimand 1 (HHF)	0.6	Cox	1.055	0.111
		NB	1.099	0.176
		LWYY	1.148	0.302
		WLW	1.103	0.206
		PWP	1.051	0.142
	0.8	Cox	1.028	0.063
		NB	1.049	0.088
		LWYY	1.069	0.102
		WLW	1.051	0.085
		PWP	1.026	0.070
	1.0	Cox	1.003	0.046
		NB	1.006	0.060
		LWYY	1.005	0.049
		WLW	1.004	0.048
		PWP	1.002	0.048
Estimand 2 (HHF+CVD)	1.0	Cox	1.002	0.047
		NB	1.006	0.057
		LWYY	1.004	0.048
		WLW	1.003	0.049
		PWP	1.002	0.049

Table 21: Mean treatment effect estimates for Estimand 1 (HHF) and Estimand 2 (HHF+CVD) for autoregressive event rate (factor 1.2) with sample size N = 4350, with non-informative treatment discontinuation.

Endpoint	RR_{HHF}	Method	$HR_{CV} = 0.6$	$HR_{CV} = 0.8$	$HR_{CV} = 1.0$
Estimand 1 (HHF)	0.6	Cox	0.758	0.734	0.713
		NB	0.622	0.591	0.565
		LWYY	0.680	0.635	0.599
		WLW	0.697	0.660	0.628
		PWP	0.775	0.750	0.728
	0.8	Cox	0.915	0.890	0.867
		NB	0.861	0.814	0.775
		LWYY	0.917	0.850	0.797
		WLW	0.906	0.860	0.819
		PWP	0.923	0.898	0.875
	1.0	Cox	1.052	1.027	1.003
		NB	1.126	1.059	1.006
		LWYY	1.168	1.076	1.005
		WLW	1.102	1.051	1.004
		PWP	1.050	1.025	1.002
Estimand 2 (HHF+CVD)	0.6	Cox	0.749	0.797	0.843
		NB	0.592	0.643	0.695
		LWYY	0.675	0.689	0.705
		WLW	0.691	0.715	0.737
		PWP	0.764	0.807	0.849
	0.8	Cox	0.841	0.885	0.927
		NB	0.750	0.796	0.844
		LWYY	0.848	0.847	0.851
		WLW	0.837	0.856	0.873
		PWP	0.853	0.892	0.930
	1.0	Cox	0.923	0.964	1.002
		NB	0.921	0.960	1.005
		LWYY	1.032	1.013	1.003
		WLW	0.974	0.989	1.003
		PWP	0.930	0.966	1.001

Figure 21: Statistical power for Estimand 1 (HHF) for autoregressive event rate (factor 1.2) with sample size $N = 4350$, with non-informative treatment discontinuation.

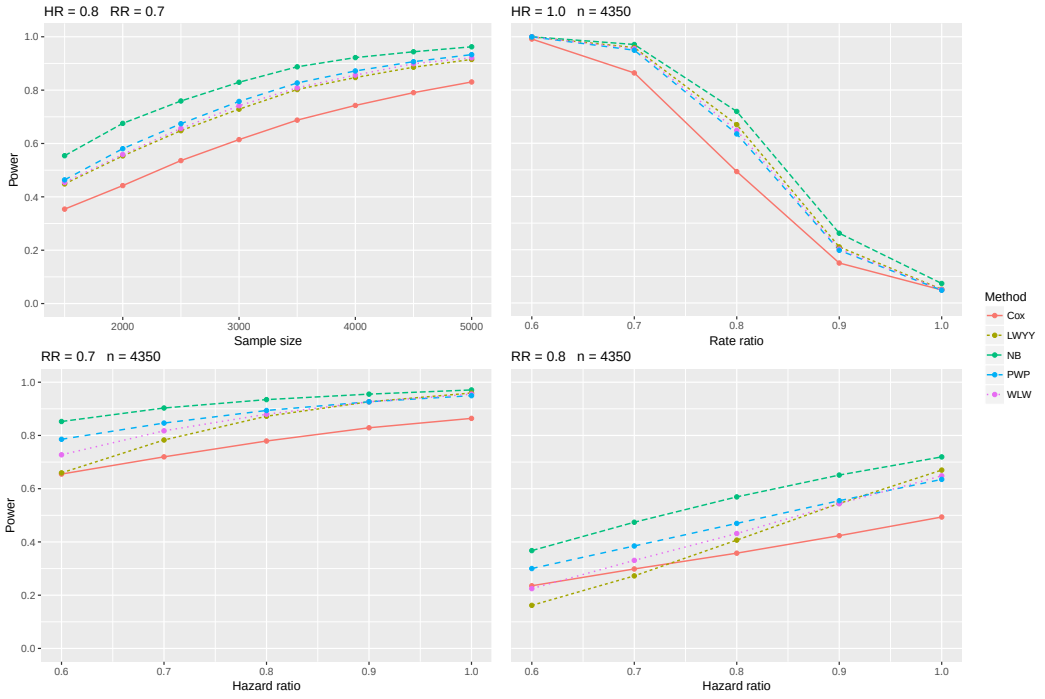


Figure 22: Statistical power for Estimand 2 (HHF+CVD) for autoregressive event rate (factor 1.2) with sample size $N = 4350$, with non-informative treatment discontinuation.

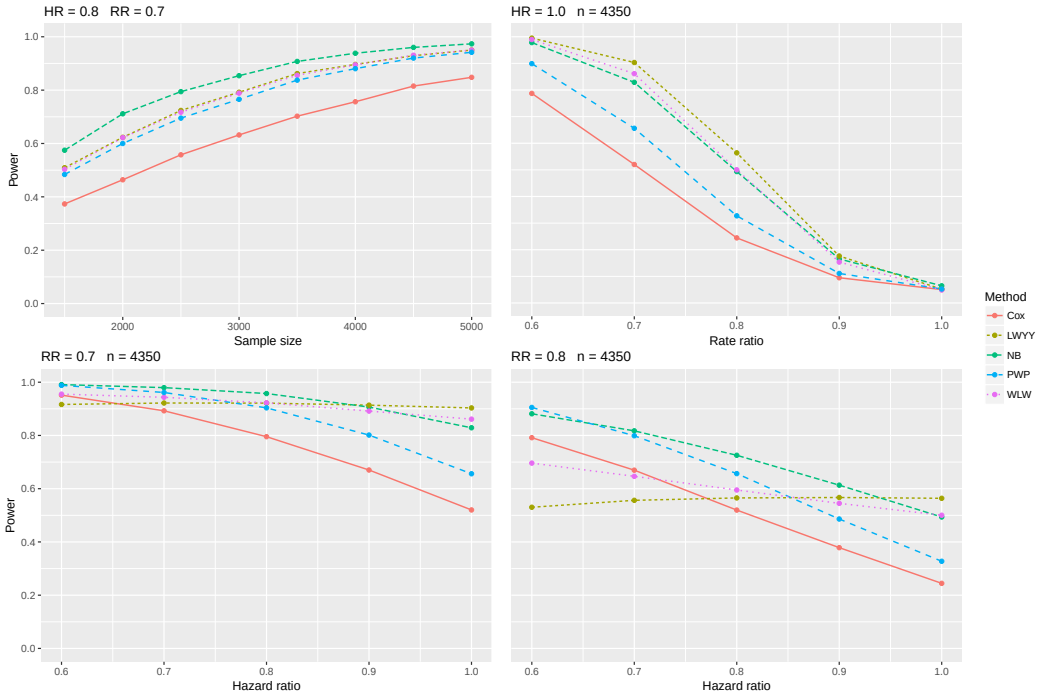


Table 22: Mean treatment effect estimates and type I error rates for Estimand 1 (HHF) and Estimand 2 (HHF+CVD) for autoregressive event rate (factor 1.2) with $RR_{HHF} = 1$ and sample size $N = 4350$, with non-informative treatment discontinuation.

Endpoint	HR_{CV}	Method	Estimate	Type I error
Estimand 1 (HHF)	0.6	Cox	1.052	0.104
		NB	1.126	0.243
		LWYY	1.168	0.335
		WLW	1.102	0.202
		PWP	1.050	0.138
	0.8	Cox	1.027	0.064
		NB	1.059	0.110
		LWYY	1.076	0.113
		WLW	1.051	0.084
		PWP	1.025	0.068
	1.0	Cox	1.003	0.049
		NB	1.006	0.073
		LWYY	1.005	0.052
		WLW	1.004	0.050
		PWP	1.002	0.048
Estimand 2 (HHF+CVD)	1.0	Cox	1.002	0.050
		NB	1.005	0.064
		LWYY	1.003	0.052
		WLW	1.003	0.049
		PWP	1.001	0.053

E.5.3 Detrimental CVD effect

Table 23: Mean treatment effect estimates for Estimand 1 (HHF) and Estimand 2 (HHF+CVD) for detrimental CVD effect with sample size $N = 4350$, with non-informative treatment discontinuation.

Endpoint	RR_{HHF}	Method	$HR_{CV} = 1/0.9$	$HR_{CV} = 1/0.8$	$HR_{CV} = 1/0.7$
Estimand 1 (HHF)	0.6	Cox	0.718	0.703	0.686
		NB	0.594	0.578	0.561
		LWYY	0.610	0.588	0.564
		WLW	0.629	0.608	0.584
		PWP	0.732	0.717	0.699
	0.8	Cox	0.865	0.849	0.829
		NB	0.790	0.772	0.749
		LWYY	0.794	0.767	0.735
		WLW	0.810	0.785	0.755
		PWP	0.871	0.856	0.837
	1.0	Cox	0.991	0.976	0.955
		NB	0.989	0.969	0.943
		LWYY	0.978	0.946	0.906
		WLW	0.982	0.954	0.919
		PWP	0.990	0.976	0.956
Estimand 2 (HHF+CVD)	0.6	Cox	0.872	0.898	0.930
		NB	0.759	0.796	0.844
		LWYY	0.735	0.745	0.757
		WLW	0.757	0.768	0.782
		PWP	0.874	0.898	0.928
	0.8	Cox	0.952	0.975	1.004
		NB	0.898	0.936	0.985
		LWYY	0.870	0.876	0.883
		WLW	0.887	0.895	0.906
		PWP	0.951	0.973	1.000
	1.0	Cox	1.021	1.043	1.069
		NB	1.036	1.075	1.124
		LWYY	1.005	1.007	1.008
		WLW	1.010	1.016	1.023
		PWP	1.018	1.038	1.063

Figure 23: Statistical power for Estimand 1 (HHF) for detrimental CVD effect with sample size $N = 4350$, with non-informative treatment discontinuation.

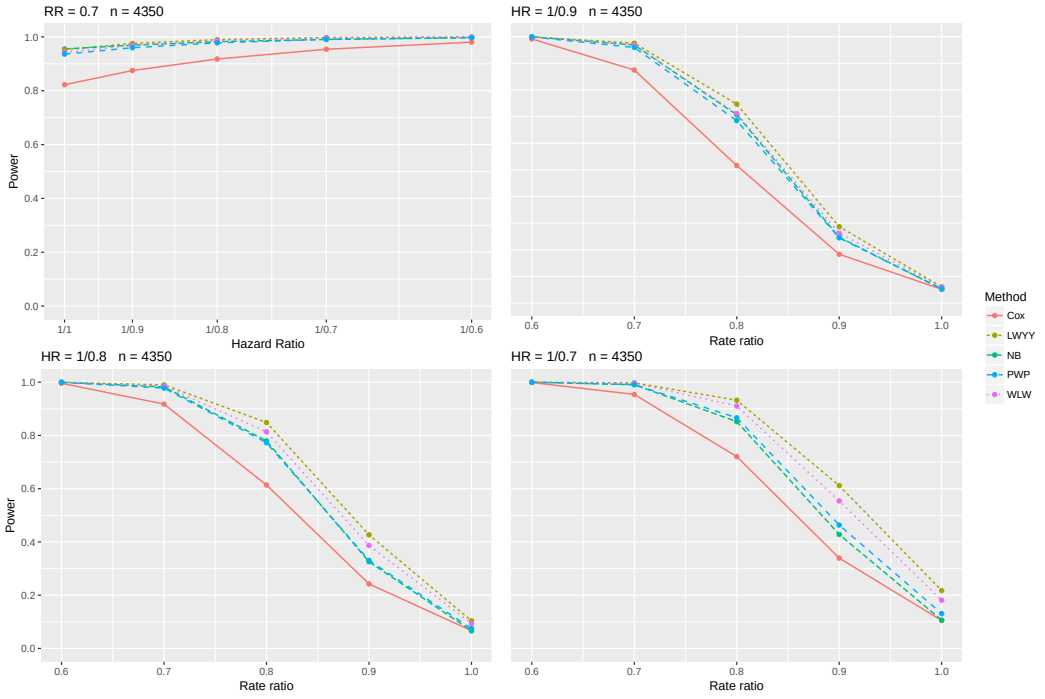


Figure 24: Statistical power for Estimand 2 (HHF+CVD) for detrimental CVD effect with sample size $N = 4350$, with non-informative treatment discontinuation.

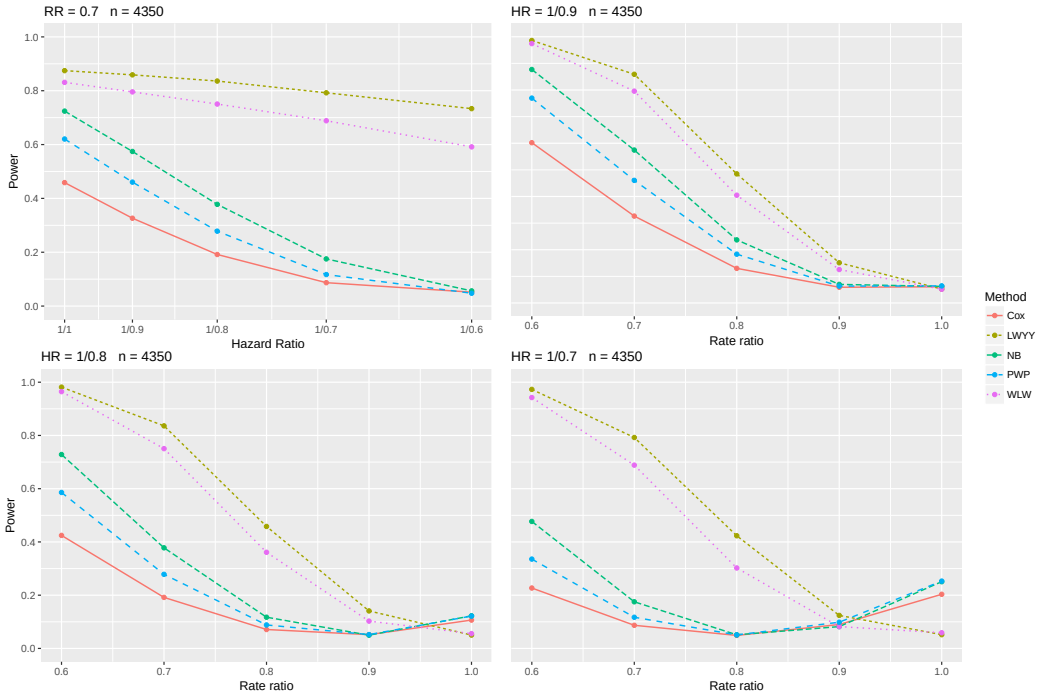


Table 24: Mean treatment effect estimates and type I error rates for Estimand 1 (HHF) and Estimand 2 (HHF+CVD) for detrimental CVD effect with $RR_{HHF} = 1$ and sample size $N = 4350$, with non-informative treatment discontinuation.

Endpoint	HR_{CV}	Method	Estimate	Type I error
Estimand 1 (HHF)	1/0.9	Cox	0.991	0.052
		NB	0.989	0.053
		LWYY	0.978	0.061
		WLW	0.982	0.059
		PWP	0.990	0.053
	1/0.8	Cox	0.976	0.067
		NB	0.969	0.066
		LWYY	0.946	0.104
		WLW	0.954	0.093
		PWP	0.976	0.075
	1/0.7	Cox	0.955	0.105
		NB	0.943	0.106
		LWYY	0.906	0.217
		WLW	0.919	0.181
		PWP	0.956	0.131
Estimand 2 (HHF+CVD)	1.0	Cox	1.003	0.046
		NB	1.005	0.046
		LWYY	1.004	0.046
		WLW	1.004	0.050
		PWP	1.001	0.049

E.5.4 Frailty correlation

Table 25: Mean treatment effect estimates for Estimand 1 (HHF) and Estimand 2 (HHF+CVD) for frailty correlation 0.5 with sample size $N = 4350$, with non-informative treatment discontinuation.

Endpoint	RR_{HHF}	Method	$HR_{CV} = 0.6$	$HR_{CV} = 0.8$	$HR_{CV} = 1.0$
Estimand 1 (HHF)	0.6	Cox	0.777	0.756	0.736
		NB	0.653	0.633	0.615
		LWYY	0.688	0.658	0.630
		WLW	0.710	0.681	0.655
		PWP	0.788	0.771	0.754
	0.8	Cox	0.922	0.901	0.880
		NB	0.854	0.831	0.810
		LWYY	0.892	0.854	0.819
		WLW	0.901	0.868	0.836
		PWP	0.922	0.905	0.888
	1.0	Cox	1.046	1.024	1.003
		NB	1.054	1.029	1.005
		LWYY	1.093	1.048	1.004
		WLW	1.077	1.040	1.004
		PWP	1.035	1.018	1.002
Estimand 2 (HHF+CVD)	0.6	Cox	0.755	0.806	0.856
		NB	0.617	0.675	0.735
		LWYY	0.681	0.706	0.730
		WLW	0.698	0.726	0.754
		PWP	0.767	0.811	0.855
	0.8	Cox	0.840	0.888	0.934
		NB	0.749	0.810	0.870
		LWYY	0.830	0.849	0.868
		WLW	0.834	0.860	0.884
		PWP	0.852	0.894	0.933
	1.0	Cox	0.912	0.958	1.002
		NB	0.879	0.941	1.004
		LWYY	0.977	0.990	1.003
		WLW	0.958	0.981	1.003
		PWP	0.924	0.963	1.001

Figure 25: Statistical power for Estimand 1 (HHF) for frailty correlation 0.5 with sample size $N = 4350$, with non-informative treatment discontinuation.

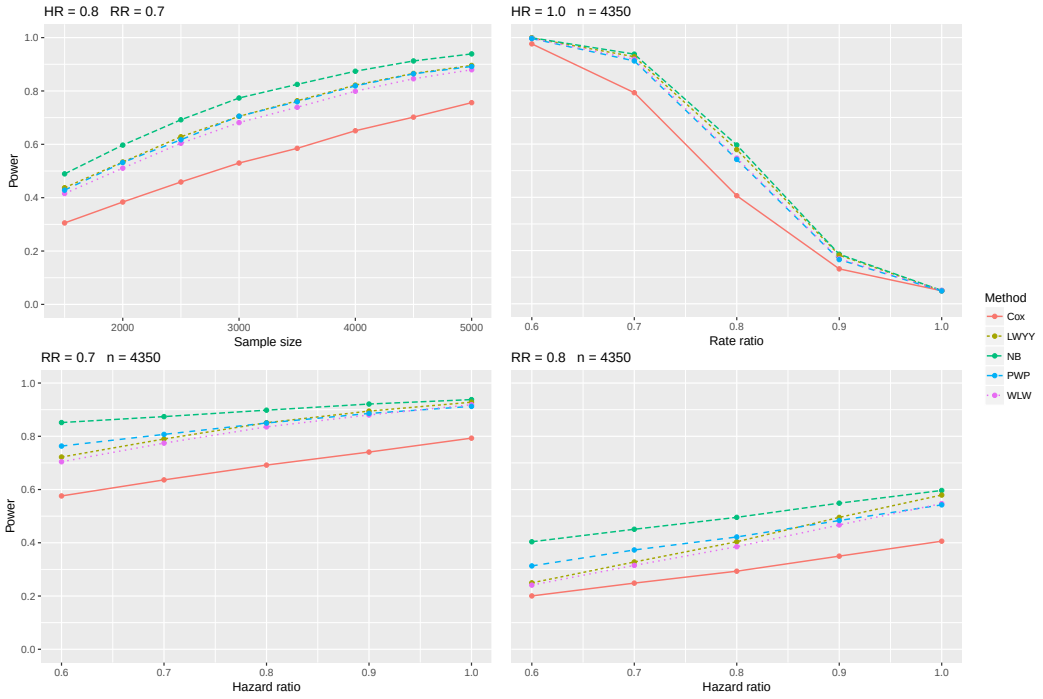


Figure 26: Statistical power for Estimand 2 (HHF+CVD) for frailty correlation 0.5 with sample size $N = 4350$, with non-informative treatment discontinuation.

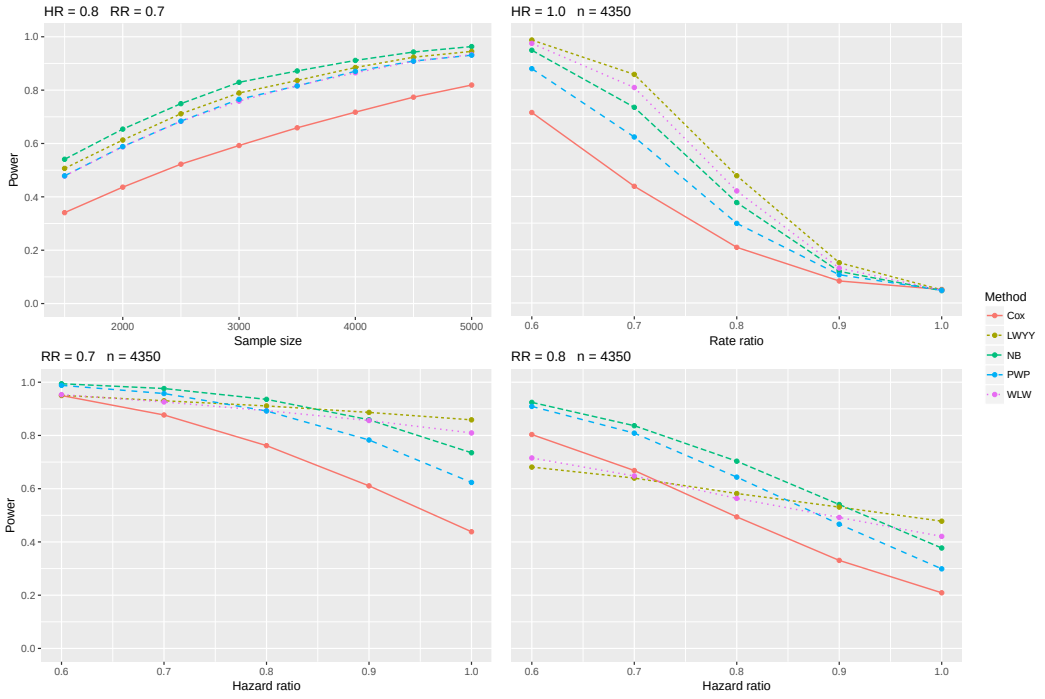


Table 26: Mean treatment effect estimates and type I error rates for for Estimand 1 (HHF) and Estimand 2 (HHF+CVD) for frailty correlation 0.5 with $RR_{HHF} = 1$ and sample size $N = 4350$, with non-informative treatment discontinuation.

Endpoint	HR_{CV}	Method	Estimate	Type I error
Estimand 1 (HHF)	0.6	Cox	1.046	0.088
		NB	1.054	0.081
		LWYY	1.093	0.152
		WLW	1.077	0.132
		PWP	1.035	0.087
	0.8	Cox	1.024	0.063
		NB	1.029	0.061
		LWYY	1.048	0.076
		WLW	1.040	0.070
		PWP	1.018	0.061
	1.0	Cox	1.003	0.049
		NB	1.005	0.050
		LWYY	1.004	0.049
		WLW	1.004	0.050
		PWP	1.002	0.048
Estimand 2 (HHF+CVD)	0.6	Cox	1.002	0.051
		NB	1.004	0.046
		LWYY	1.003	0.049
		WLW	1.003	0.048
		PWP	1.001	0.048

Table 27: Mean treatment effect estimates for Estimand 1 (HHF) and Estimand 2 (HHF+CVD) for frailty correlation 1.0 with sample size $N = 4350$, with non-informative treatment discontinuation.

Endpoint	RR_{HHF}	Method	$HR_{CV} = 0.6$	$HR_{CV} = 0.8$	$HR_{CV} = 1.0$
Estimand 1 (HHF)	0.6	Cox	0.783	0.753	0.727
		NB	0.664	0.629	0.600
		LWYY	0.717	0.667	0.627
		WLW	0.725	0.679	0.639
		PWP	0.796	0.764	0.736
	0.8	Cox	0.932	0.903	0.875
		NB	0.879	0.837	0.800
		LWYY	0.933	0.869	0.816
		WLW	0.929	0.874	0.826
		PWP	0.941	0.909	0.879
	1.0	Cox	1.061	1.032	1.003
		NB	1.096	1.049	1.006
		LWYY	1.148	1.071	1.005
		WLW	1.120	1.060	1.004
		PWP	1.063	1.032	1.002
Estimand 2 (HHF+CVD)	0.6	Cox	0.782	0.815	0.848
		NB	0.629	0.677	0.727
		LWYY	0.717	0.721	0.728
		WLW	0.725	0.733	0.743
		PWP	0.796	0.826	0.856
	0.8	Cox	0.872	0.902	0.930
		NB	0.768	0.817	0.866
		LWYY	0.874	0.869	0.866
		WLW	0.870	0.874	0.877
		PWP	0.881	0.908	0.934
	1.0	Cox	0.951	0.978	1.003
		NB	0.906	0.956	1.005
		LWYY	1.031	1.016	1.004
		WLW	1.005	1.005	1.003
		PWP	0.955	0.979	1.002

Figure 27: Statistical power for Estimand 1 (HHF) for frailty correlation 1.0 with sample size $N = 4350$, with non-informative treatment discontinuation.

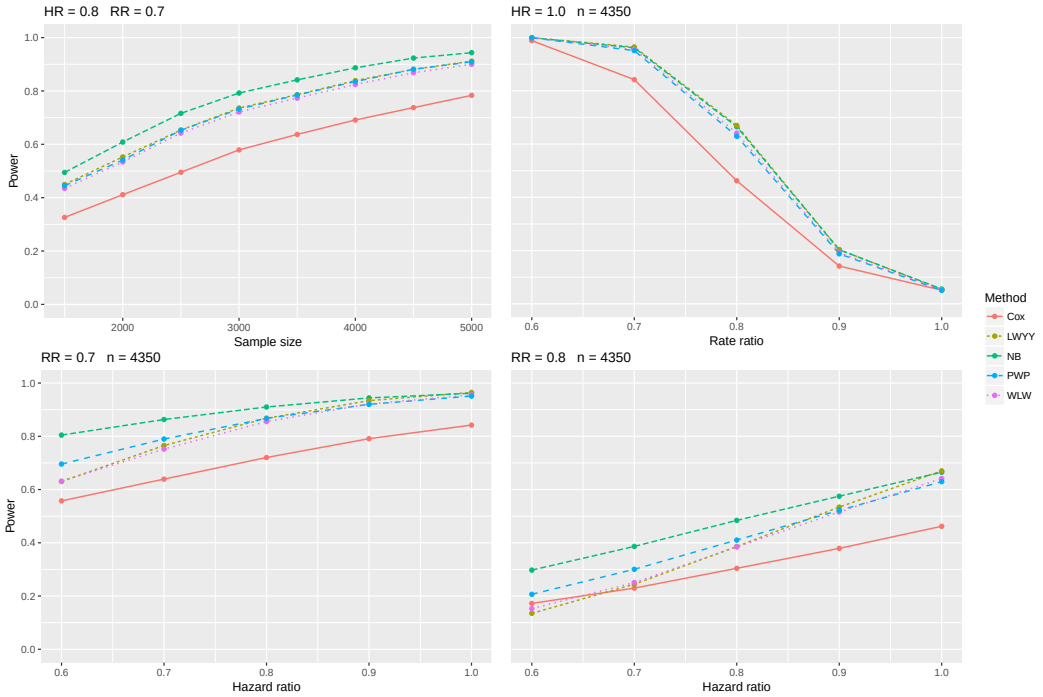


Figure 28: Statistical power for Estimand 2 (HHF+CVD) for frailty correlation 1.0 with sample size $N = 4350$, with non-informative treatment discontinuation.

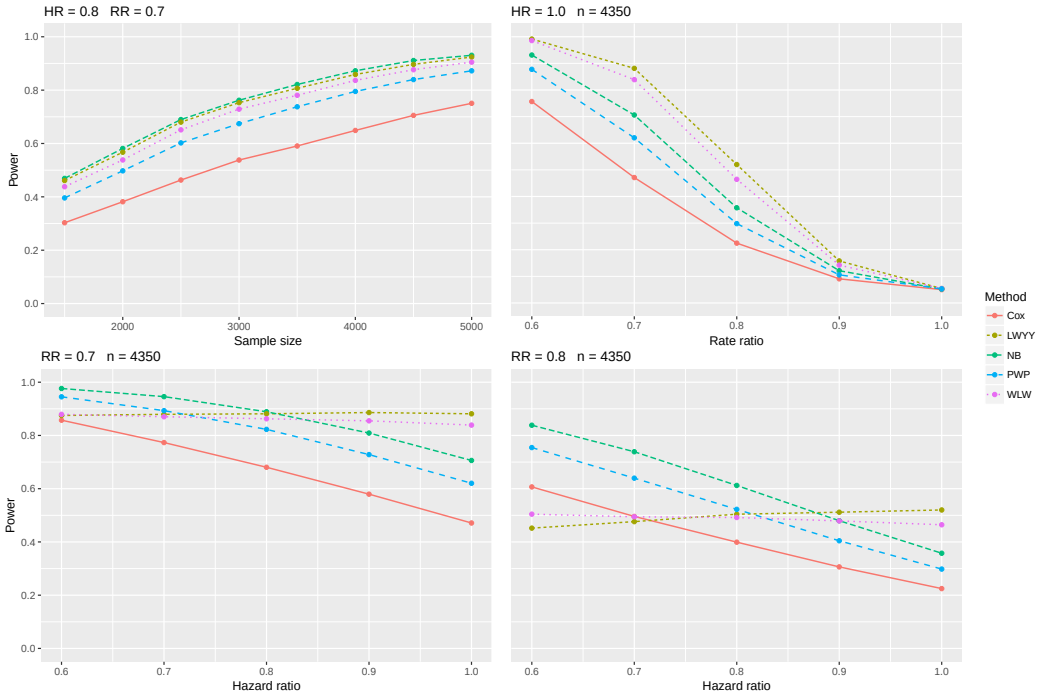


Table 28: Mean treatment effect estimates and type I error rates for for Estimand 1 (HHF) and Estimand 2 (HHF+CVD) for frailty correlation 1.0 with $RR_{HHF} = 1$ and sample size $N = 4350$, with non-informative treatment discontinuation.

Endpoint	HR_{CV}	Method	Estimate	Type I error
Estimand 1 (HHF)	0.6	Cox	1.061	0.130
		NB	1.096	0.164
		LWYY	1.148	0.364
		WLW	1.120	0.280
		PWP	1.063	0.206
	0.8	Cox	1.032	0.070
		NB	1.049	0.076
		LWYY	1.071	0.122
		WLW	1.060	0.103
		PWP	1.032	0.085
	1.0	Cox	1.003	0.051
		NB	1.006	0.056
		LWYY	1.005	0.054
		WLW	1.004	0.053
		PWP	1.002	0.050
Estimand 2 (HHF+CVD)	1.0	Cox	1.003	0.050
		NB	1.005	0.052
		LWYY	1.004	0.054
		WLW	1.003	0.054
		PWP	1.002	0.053