GEORG-AUGUST-UNIVERSITÄT
GÖTTINGEN

Blinded sample size reestimation in clinical trials with time-to-event outcomes based on flexible parametric models

20 weeks Master Thesis
In partial fulfillment of the requirements for the degree
Master of Science (M.Sc.) in Applied Statistics
at the University of Göttingen

Tim Mori
from Düsseldorf
21914136

1. Supervisor: Prof. Dr. Tim Friede
2. Supervisor: Dr. Thomas Asendorf

University of Göttingen
Department of Medical Statistics

Submitted: May 18$^{\text{th}}$, 2022

Acknowledgments

I would like to thank Prof. Tim Friede and Prof. Satoshi Hattori for their guidance and their support throughout the writing of this thesis. I would also like to thank Dr. Sho Komukai for his support and for his help with reviewing my R code. I cherished our regular German-Japanese meetings and I learned a lot from all the interesting discussions. Moreover, I want to thank Dr. Thomas Asendorf for taking on the role of a second academic advisor and for providing helpful feedback. Also, I want to thank Malte, Tian and Ao for the friendly atmosphere at the office and the nice lunch breaks. Finally, I want to thank Carsten, Niklas, Sophie, Joanna and Marah for the the wonderful past two years of studying statistics together.

Abstract

The validity of clinical trials depends on sufficiently large sample sizes to ensure a pre-specified level of statistical power. In the context of time-to-event studies, a sufficient number of events needs to be observed to achieve this. For event-driven designs, in which the trial finishes once a pre-specified number of events is observed, the objective is to complete a clinical trial within a given time frame. Blinded sample size reestimation (BSSR) methods use non-comparative blinded interim trial data to adjust the sample size if the planning assumptions are wrong. In the context of time-to-event trials, interim data available at the time point of the sample size review provide some information about survival. However, this information evidently does not extend to the time point of the final analysis. Therefore, the estimated survival function based on the interim data needs to be extrapolated to reestimate the sample size. The current practice is to fit standard parametric models (e.g. exponential or Weibull models) to extrapolate the survival times for the purpose of BSSR. While extrapolation based on such models is useful, it may not always be suitable. First, there might be a lack of external evidence to justify the choice of the parametric model, for example when a treatment for a novel disease is investigated. Second, patient outcomes may change compared to previous cohorts due to advances in medical care or existing treatment regimens, which could result in potentially different survival distributions. In such cases a more flexible and robust parametric model for BSSR would be useful. The aim of the current thesis was to propose a flexible parametric approach for BSSR in clinical trials with time-to-event outcomes and to compare it to existing parametric approaches. Specifically, we extended the parametric BSSR framework of Friede et al. (2019) by carrying out the extrapolation based on the Royston-Parmar spline model for time-to-event data. We carried out a simulation study for an event-driven design based on exponential, Weibull and Gompertz distributed data. We found that a 1-knot spline BSSR was unbiased in the exponential and Weibull setting and performed best in the Gompertz misspecification scenario. We also considered a practical application of spline methods to a Multiple Sclerosis (MS) study. Overall, the current thesis provides evidence for the usefulness of flexible spline methods for BSSR in clinical trial with time-to-event outcomes. We discuss the implications of our findings for spline-based survival extrapolation and BSSR and outline a number of extensions, including a method combining the Kaplan-Meier estimate with spline-based estimates for BSSR.

Keywords: blinded sample size reestimation, event-driven designs, extrapolation, flexible parametric models, splines

Contents

1 Introduction

The validity of clinical trials depends on sufficiently large sample sizes to ensure a pre-specified level of statistical power. In the context of time-to-event studies, a sufficient number of events (for example deaths) needs to be observed to achieve this (Collett, 2015, Ch. 15). For event-driven designs, in which the trial finishes once a pre-specified number of events is observed, the objective is to complete a clinical trial within a given time frame (Friede et al., 2019). That is, a sufficient number of events needs to occur in a given time frame so that the study finishes on time. Such designs are especially important in medical areas without pre-existing treatments, where effective treatments are urgently needed (Friede et al., 2019).

The sample size planning for such studies depends on valid estimates of the survival function, which indicates the probability that a patient survives beyond a specific time (Collett, 2015). However, sometimes such information is unavailable. This was the case, for example, with early COVID-19 trials during the start of the global pandemic in 2020 (Stallard et al., 2020). So-called adaptive designs can help to counteract this uncertainty. In adaptive designs planned interim analyses are included, which use accumulating trial data to potentially modify aspects of an ongoing trial (Stallard et al., 2020). For example, if survival has been underestimated at the planning stage, lower number of events might become evident at an interim analysis. Subsequently, recruitment numbers can be increased (for example by prolonging the recruitment phase) to ensure that the trial has sufficient statistical power and finishes on time.

Blinded sample size reestimation (BSSR) methods, which use non-comparative interim data pooled across treatment groups, have been shown to be particularly useful in adjusting the sample size (Friede and Kieser, 2001). They are advantageous from a regulatory viewpoint, since they do not break treatment code. (Kieser and Friede, 2003). In the context of time-to-event trials, interim data available at the time point of the sample size review provide some information about survival. However, this information evidently does not extend to the time point of the final analysis. Therefore, the estimated survival function based on the interim data needs to be extrapolated to reestimate the sample size. The current practice is to fit standard parametric models (e.g. exponential or Weibull models) to extrapolate the survival times for the purpose of BSSR (Friede et al., 2019).

While extrapolation based on such models is useful, it may not always be suitable.

First, there might be a lack of external evidence to justify the choice of the parametric model, for example when a treatment for a novel disease is investigated. Second, patient outcomes may change compared to previous cohorts due to advances in medical care or existing treatment regimens, which could result in potentially different survival distributions (Schumacher and Schulgen-Kristiansen, 2008, Ch. 2). Third, some real life datasets exhibit complex hazard functions that cannot be captured by simple parametric models. For example, immuno-therapy treatments in oncology may induce delayed treatment responses and long-term survival, which can results in complex hazard shapes (Ouwens et al., 2019). In all of these cases a flexible parametric model would be useful to make the BSSR procedure more robust. In the first two cases, the additional flexibility allows to model survival curves that might deviate from a (potentially previously assumed) exponential or Weibull model. In the latter case, the flexibility might be necessary to capture the complex hazard function, for example when it involves multiple turning points (Latimer and Adler, 2022).

The aim of the current thesis was to propose a flexible parametric approach for BSSR in clinical trials with time-to-event outcomes and to compare it to existing parametric approaches. Specifically, we extend the parametric BSSR framework of Friede et al. (2019) by carrying out the extrapolation based on the Royston-Parmar spline model for time-to-event data (Royston and Parmar, 2002). Section 2 begins by introducing the terminology and reviewing existing BSSR procedures for clinical trials with time-to-event outcomes. Section 3 reviews the current literature on survival extrapolation with a special focus on the Royston-Parmar model. In Section 4 we bring these methods together and present our proposed flexible BSSR method. The operating characteristics of our method are investigated in a simulation study in Section 5 and we apply the Royston-Parmar spline model to a real world dataset on multiple sclerosis in Section 6. Finally, we summarize our findings and discuss their implications and limitations in Section 7.

2 Background: Clinical trials with time-to-event outcomes

2.1 Basic concepts and setting

In clinical trials we are frequently interested in the time until an event of interest occurs (Schumacher and Schulgen-Kristiansen, 2008, Ch. 5). For example, we might want to investigate how long patients diagnosed with a certain type of cancer survive on average. In that case the event of interest would be death and it is frequently, but not exclusively, what is investigated in so called time-to-event studies (Collett, 2015, Ch. 1). Because of this, the field associated with such types of analyses is also called survival analysis (Collett, 2015, Ch. 1). However, the event of interest may for example also be remission (for example in cancer treatments) or onset of an infection.

Survival data are defined by the feature that they have a well-defined time origin (e.g. start of treatment) from which the time until the occurrence of our event of interest is measured (Collett, 2015, Ch. 1). In a clinical trial, the time origin is typically the time of randomization. Let the random variable for the time-to-event process, which can take any non-negative values, be denoted by T. The distribution function of T is

$$F(t) = \mathrm{P}(T < t) = \int_0^t f(u)\mathrm{d}u \tag{1}$$

where $f(t)$ is the underlying probability density function (Collett, 2015, Ch. 1). While this function indicates the probability that the survival time is less than t, the more commonly used function is the so-called survival function (or survivor function). It is defined as

$$S(t) = \mathrm{P}(T \geq t) = 1 - F(t) \tag{2}$$

and represents the probability that a patient survives beyond any given time (Collett, 2015, Ch. 1).

A fundamental difference between time-to-event data and other (longitudinal) data commonly observed in clinical trials is that censoring can occur (Collett, 2015, Ch. 1). Censoring means that the endpoint of interest has not yet been observed for a given individual (Collett, 2015, Ch. 1). For example, at the end of follow-up of a cancer trial, some patients may not have experienced remission. Censoring of this kind is referred to as administrative censoring and it occurs simultaneously for all patients left in the study who have not yet experienced the event at that point (Collett, 2015, Ch. 1).

Another type of censoring is loss to follow up, where individual patients become censored, because no updated data on their event status can be obtained (Collett, 2015, Ch. 1). For example, a patient may move to another country or another clinic and may no longer be reachable by the investigators. In survival analysis we typically assume that the time-to-censoring is independent of the time-to-event process (Collett, 2015, Ch. 1). That is, the mechanism that causes censoring (e.g. loss to follow up) does not depend on the mechanism that influences survival. This assumption is referred to as independent or non-informative censoring (Collett, 2015, Ch. 1). It would be violated, if a patient would, for example, withdraw from a study as a consequence of a deteriorating health status, which has been caused by an inferior treatment.

Now that the notions of time-to-event and time-to-censoring processes have been introduced, we can present the relevant notation. The scenario of interest is a randomized clinical trial with two treatment groups. The treatment allocation is denoted by $Z = 0, 1$ and the probability of being assigned to the treatment group is denoted by $P(Z = 1) = \pi$. Furthermore, let T_{ij} and C_{ij} denote the time-to-event and time-to-censoring for patient j in group i. The observation time of a patient is the minimum of these two and the data is summarized as the observation time plus an event indicator, δ_{ij}. Here, $\delta_{ij} = 0$ indicates censoring and $\delta_{ij} = 0$ indicates an event at the observed follow-up time. Let f and g denotes the density functions and S and G the survival functions for the time-to-event and time-to-censoring process, respectively.

To compare the survival outcomes of two treatment groups we typically consider the log-rank test (Collett, 2015, Ch. 2). We are interested in testing the null hypothesis that the survival outcomes of individuals in the two groups do not differ. That is, we wish to test $H_0 : \theta = 1$, where θ is the hazard ratio for the treatment. Assume that there are r unique death times across the two groups, which are denoted as $t_{(1)} < t_{(2)} < ... < t_{(r)}$. Let d_{ij} denote the number of deaths in group i at time $t_{(j)}$. Also, assume that n_{ij} patients in group i are at risk of death just before time $t_{(j)}$. Let the respective quantities pooled across both treatment groups be denoted by $d_j = d_{1j} + d_{2j}$ and $n_j = n_{1j} + n_{2j}$. Then, under the null hypothesis the expected number of deaths can be estimated as $e_{ij} = n_{ij} d_j / n_j$. We can sum up the differences between the observed and expected number of deaths at each unique time point as $U_L = \sum_{j=1}^{r} (d_{1i} - e_{1j})$. The variance of this statistic is $\text{var}(U_L) = \sum_{j=1}^{r} v_{1j}$, where v_{1j} is defined as $v_{1j} = \frac{n_1 n_{2j} d_j (n_j - d_j)}{n_j^2 (n_j - 1)}$. Then, the log-rank test-statistics is computed as

$$W_L = \frac{U_L^2}{\mathrm{var}(U_L)} \tag{3}$$

and it asymptotically follows a chi-squared distribution with one degree of freedom (Collett, 2015, Ch. 2).

If we wish to account for explanatory variables in our analysis, such as demographic variables (e.g. age, sex) or physiological variables (e.g. heart rate, diet), we can extend this approach to the Cox model (Collett, 2015, Ch. 3). In the Cox model we directly model the hazard function, $h(t)$, which expresses the hazard or risk of experiencing an event at some time t (Collett, 2015, Ch. 3). Formally, the hazard function is defined as

$$h(t) = \lim_{t\Delta \to 0} \frac{\mathrm{P}(t \leq T < t + t\Delta | T \geq t)}{t\Delta}. \tag{4}$$

Based on an estimated of the hazard function we can also obtain an estimate of the survival function, since these two functions are related as $S(t) = \exp[-\int_0^t h(x)\mathrm{d}x]$.

The hazard for an individual j is specified by the Cox model as

$$h_j(t) = h_0(t)e^{X_j\beta}, \tag{5}$$

where, X_j is a $1 \times p$ vector of covariate values and β is $p \times 1$ vector of coefficients (Therneau and Grambsch, 2000, Ch. 3). h_0 is an unspecified non-negative function, which is referred to as the baseline hazard (Therneau and Grambsch, 2000, Ch. 3). The Cox model is a proportional hazards model, since the hazard ratio between two individuals with fixed covariate vectors is constant over time (Therneau and Grambsch, 2000, Ch. 3).

The Cox model is flexible in that no distributional assumptions are made about the survival times (Collett, 2015, Ch. 3). Since it requires the proportional hazards assumptions, it is often referred to as a semiparametric method (Royston et al., 2011, Ch. 1). While this is a desirable feature, there may be occasions in which a parametric model for the survival times is preferred. If we can safely assume a certain probability distribution for our data, then we can choose to fit the corresponding parametric survival model (e.g. an exponential or Weibull model). This will allow for more precise inferences, since the standard errors of estimates (e.g. for the median survival times) will be smaller under correct model specification (Collett, 2015, Ch. 5). Moreover, estimates of the baseline hazard function based on a Cox model can be highly erratic (Royston and Parmar, 2002). Obtaining a smooth estimate based on a parametric model may be useful, since the behavior of the hazard function provides important insights into the time-course of an illness

(Royston and Parmar, 2002). Another important aspect is that parametric models easily allow for extrapolation of the hazard and survival function (Royston et al., 2011, Ch. 1). This is relevant for out of sample predictions or for anticipating future survival outcomes based on limited trial data. This notion is important in the context of blinded sample size reestimation and will be discussed in Section 2.3. However, we first start by introducing the basic concepts of sample size planning for clinical trials with time-to-event outcomes.

2.2 Sample size planning

When planning a clinical trial with time-to-event outcomes we need to carry out two steps: First need to determine how many events are necessary for our trial to have sufficient statistical power. Second, we have to consider how many patients we have to recruit and how long our follow up period will be, in order to ensure that the necessary number of events is observed by the end of our study (Therneau and Grambsch, 2000, Ch. 3).

The power of a statistical test, $1-\beta$, is defined as the probability of rejecting a false null hypothesis (Collett, 2015, Ch. 15). In time-to-event studies the power depends among other things on the number of events and the magnitude of treatment effect that is under investigation (Collett, 2015, Ch. 15). Therefore, in the first step of the sample size calculation we consider how many events are necessary to ensure a pre-specified level of statistical power (e.g. 90%) when we assume a certain treatment effect (Collett, 2015, Ch. 15). The assumed treatment effect can be specified as a hazard ratio, denoted by θ^*. For example, suppose that a current treatment results in a five-year survival 30% and we expect our new treatment to improve survival to 45% at five years (Therneau and Grambsch, 2000, Ch. 3). Then the hazard ratio can be computed as $\theta^* = \log(0.45)/\log(0.30) = 0.663$. The required number of events is such that the probability of finding that the observed hazard ratio is significantly different from 1 (e.g. using a log-rank test or cox regression) is $1-\beta$, when the true, unknown hazard ratio is indeed θ^* (Collett, 2015, Ch. 15). The required number of events, d, can be computed as

$$d = \frac{(1+k)^2}{k}\frac{(z_{1-\alpha}+z_{1-\beta})^2}{log\,(\theta^*)^2} \tag{6}$$

for a k:1 treatment allocation and a one-sided significance level of α (Schoenfeld, 1983). If k is equal to one (this corresponds to a 1:1 treatment allocation), this reduces to the commonly seen formula $d = 4\frac{(z_{1-\alpha}+z_{1-\beta})^2}{log\,(\theta^*)^2}$ (Collett, 2015, Ch. 15). If a two-sided test is carried out, then the above quantile of the normal distribution should be replaced by

$z_{1-\alpha/2}$ (Friede et al., 2019).

Once we have computed the necessary number of events, the second question is how many patients we should enroll and how long we should follow up on them. Since time-to-event studies include a recruitment period and a follow-up period, the trial design is determined by three quantities: the length of the recruitment period, the recruitment rate and the length of follow up. It is common to consider this on a monthly time scale, in which case we we have to decide on the number of recruitment months, the monthly the recruitment rate and the number of follow-up months. These three quantities need to be chosen in such a way that we can expect the necessary number of events to occur by the planned end of our trial. Note that a constant recruitment rate is often assumed, but it could also vary over time (e.g. when trial recruitment is anticipated to start off slowly and increase with time).

An important distinction needs to be made here between fixed and flexible follow-up designs (Friede et al., 2019). In a fixed follow-up design, each patient is followed-up for the same time period (e.g. 24 months) or until the event of interest occurs, whatever comes first. In a flexible follow-up design, the study closes at a specific time point (e.g. once a specified number of events has been reached) and all patients are followed-up until that point or until the event of interest occurs. Since patients are recruited at different time points during the recruitment period, this results in variable individual follow-up times for the flexible follow-up design (Friede et al., 2019).

The fixed follow-up design is simpler in terms of the sample size calculations. Let $S_E(t)$ and $S_C(t)$ denote the survival functions in the treatment and control group, respectively. Since every patient is followed up for exactly the same time (say, M months) only anticipated values at the end of follow up, $S_E(M)$ and $S_C(M)$ need to be available (Whitehead, 2001). With a known treatment allocation, the pooled survival function $\bar{S}(M)$ can then easily be calculated. Based on this, we can compute the probability of experiencing an event during the trial as $\mathrm{P}(\text{event}) = 1 - \bar{S}(M)$. Using that, the expected number of events, $\mathrm{E}(D)$, can be computed as

$$\mathrm{E}(D) = n \times \mathrm{P}(\text{event}). \tag{7}$$

With a constant monthly recruitment rate, r, the number of patients simplifies to $n = rR$, where R is number of recruitment months. Based on this the trial team can juggle the values of r, R and M to ensure that the number of events is large enough to ensure the pre-specified level of statistical power.

In a flexible follow-up design, a different approach must be used. Here, we have a follow-up period of F months, which begins once the recruitment period of R months has finished. Only the last recruited patient (right before recruitment closure) is followed up for exactly F months (Whitehead, 2001). In contrast, the first recruited patient will be followed up for $R + F$ months (or until an event) and patients recruited in between will have follow up times between F and $R + F$ (Whitehead, 2001). In such cases we need to approximate the probability that a patient experiences an event over the duration of the trial (Collett, 2015, Ch. 15). When we assume constant accrual, one approach is to use the Simpson's rule, based on which the probability of experiencing an event can be approximated as

$$\mathrm{P(event)} = 1 - \frac{1}{6}\{\bar{S}(F) + 4\bar{S}(0.5R + F) + \bar{S}(R + F)\}. \tag{8}$$

Another approach is to consider recruitment months separately (Therneau and Grambsch, 2000, Ch. 3). For example, Todd et al. (2012) have suggested to approximate the expected number of events (for a 1:1 treatment allocation) as

$$\mathrm{E}(D) = r\{R - 1/2 \sum_{C,E} \sum_{k=F}^{R+F-1} S_i(k)\}. \tag{9}$$

As can be seen, the survival function is evaluated at a series of discrete time points, k (e.g. months), which allows to directly account for the variable follow-up times (Todd et al., 2012). With such an approach, estimates of the survival functions $S_C(t)$ and $S_E(t)$ for $F < t < R + F$ need to be available. Whitehead (2001) suggests that a form for $S_C(t)$ can be anticipated and a corresponding form for $S_E(t)$ can be derived under the proportional hazards model with an assumed reference improvement. They propose that the form of $S_C(t)$ can be expressed as a parametric survival function or as a step function. Whitehead (2001) derived formulas for the exponential and piecewise exponential case and provided them in the appendices of his paper. Note that above formulas do not account for the time-to-censoring process. A formula for the expected number of events that also accounts for censoring will be introduced in Subsection 2.3.2.

Clearly, much consideration has to go into the second step of the sample size calculation, where the necessary number of events is translated into a trial design. Ultimately, however, this step remains an informed guess, which relies on a number of assumptions. This includes assumptions about recruitment, the incidence of the event of interest and the treatment effect (Therneau and Grambsch, 2000, Ch. 3). In practice, both the incidence of the event of interest and the recruitment may be lower than anticipated.

This phenomenon has been referred to as Lasagna's law (Therneau and Grambsch, 2000, Ch. 3). If the clinical trial is then carried out based on the planning assumptions, this can have severe consequences. In a time-driven design, in which the study ends at the planned end of follow-up, this can result in an underpowered trial, since the necessary number of events has not yet been reached (Peace, 2009). In an event-driven design, where the trial only finishes once the necessary number of events have been observed, this can lead to a severely increased trial duration (Friede et al., 2019). To prevent such problematic outcomes, a blinded sample size reestimation (BSSR) procedure may be carried out while recruitment is still ongoing (Friede et al., 2019). We discuss such procedures in detail in the next section.

2.3 Blinded sample size reestimation

Blinded sample size reestimation (BSSR) methods use non-comparative blinded data available at an interim analysis to assess whether the trial design needs to be modified (Friede et al., 2019). In clinical trials with time-to-event outcomes the BSSR takes place during the recruitment period and we analyse the preliminary data of the patients that have been recruited so far. For example, we can update the observed recruitment pattern and gather information on current and future recruitment (Friede et al., 2019). Moreover, observed censoring trends can be accounted for when estimating the expected number of events (Friede et al., 2019). Importantly, we can use the updated information on survival to re-assess how many events we expect by the planned end of the trial. An inherent issue, however, is that information on survival does not extend to the time point of the final analysis. This means that survival information based on the interim data needs to be extrapolated in order to reestimate the expected number of events (Friede et al., 2019). To address this issue, different BSSR approaches have been proposed, which can be categorized into non-parametric (Whitehead et al., 2001; Whitehead, 2001; Todd et al., 2012) and parametric approaches (Hade et al., 2010; McClure et al., 2012; Friede et al., 2019).

2.3.1 Non-parametric approaches

Whitehead et al. (2001) proposed a non-parametric extrapolation approach, which relies on the Kaplan-Meier estimate based on the available follow-up times. To obtain an estimate of the entire survival function, they first estimate the survival function for the available follow-up times. Then, the average deviation of these Kaplan-Meier estimates from the planning estimates (e.g. based on previous data) is computed on the complementary log-log scale. This difference is subsequently used to obtain extrapolated values of the survival function by shifting the planning estimates, again on the complementary log-log scale, by this observed average deviation.

Since only blinded data are used, these calculations are carried out for the pooled interim time-to-event data (Whitehead et al., 2001). For their sample-size reestimation calculations, they introduce the idea of splitting the obtained pooled survival curve into treatment and control survival curves based on an assumed reference improvement. The reference improvement is quantified as a log hazard ratio, θ_R. It can be calculated at the planning stage based on the assumed survival at end of follow up in the treatment and control group. For example, if the fixed follow-up time was 36 months, it can be computed as $\theta_R = -\log(-\log(S_E(36))) + \log(-\log(S_c(36)))$, where $S_E(36)$ and $S_C(26)$ are the assumed survival probabilities at 36 months in the treatment and control group, respectively. Using this reference improvement and the pooled estimates of the survival function, $\bar{S}(t_i)$, they propose to iteratively solve the equation $e^{\theta_R} = \frac{\log S_C(t_i)}{\log[2\bar{S}(t_i) - S_C(t_i)]}$ to obtain estimates of $S_C(t_i)$. Then, these can in turn be used to obtain estimates of $S_E(t_i)$ as $S_E(t_i) = 2\bar{S}(t_i) - S_C(t_i)$.

A variation of above method was investigated by Todd et al. (2012). Rather than shifting the planning estimates based on the average deviation observed so far, they suggest computing the deviation at the last observed time point. Again, the computation of the deviation and the shifting are carried out on the complementary log-log scale. Their rationale for the modification was that the most recently observed deviation might be a better predictor for future survival than the average deviation observed over the entire time course (Todd et al., 2012). In a simulation study they found that both approaches had similar performances. In particular, they found that, even in misspecification scenarios, both methods maintained the type I error rate and the desired power (Todd et al., 2012).

2.3.2 Parametric approaches

Parametric models are a convenient tool for BSSR, since they can be estimated based on the limited data available at the time of interim analysis and naturally allow for an extrapolation of survival into the future (Hade et al., 2010; Friede et al., 2019). Hade et al. (2010) carried out a blinded sample size reestimation for a breast cancer trial and they extrapolated the interim survival function by means of a parametric Weibull model. As an extra step, they generated bootstrap intervals for their sample size estimates to assess the uncertainty in their reestimation procedure (Hade et al., 2010). They also carried out some simulations assuming Weibull distributed data and found that their parametric approach maintained the type I error rate and correctly increased sample size in mis-specification scenarios. When the planning assumptions were correct the sample size was only rarely increased (Hade et al., 2010).

In another clinical trial, which was concerned with the secondary prevention of small subcortical strokes, McClure et al. (2012) carried out an unplanned sample size reestimation based on an exponential model. While the trial was on-going, new randomized trial results indicated an event rate that was lower than anticipated by the trial team at the design stage (McClure et al., 2012). They estimated the exponential overall event rate from their interim data, including the 95% confidence interval limits, and assessed the impact of potential design modifications by means of simulations. They found that in many simulation scenarios, the original design would be underpowered and recommended to recruit additional patients.

While these two applications studies illustrate the potential usefulness of parametric BSSR approaches in time-to-events settings, technical details and the methodological framework were not discussed. To close this gap, Friede et al. (2019) developed a methodological framework for parametric BSSR procedures for clinical trials with time-to-event outcomes. Moreover, they were the first to specifically consider BSSR in the context of event-driven designs.

For event-driven designs, in which the trial finishes once a pre-specified number of events is observed, the main goal is to complete the clinical trial within a given time frame (Friede et al., 2019). Study closure is defined by reaching a certain number of events, which ensures a pre-specified level of statistical power under a given alternative. Event-driven designs are commonly used in clinical trials in therapeutic areas such as oncology and cardiovascular diseases. As Friede et al. (2019) point out, study duration is not controlled with such a procedure. That is, when planning assumptions on event rates were too high, it may take a significantly longer time than anticipated to reach the necessary number of events. This can be particularly problematic in medical areas where no other treatment options are currently available (Friede et al., 2019). Moreover, this could be an issue for trials funded by a grant for a certain period of time. In this context, an interim analysis can recognize deviations from the planning stage in terms of recruitment patterns, survival and censoring. This knowledge can then be used in a BSSR procedure to modify the design in such a way that the study finishes on time.

Let L denote the desired end of the study, which is some time point after the end of the recruitment $R < L$. Furthermore, let D_i denote he number of events in treatment group i by the end of the trial. Friede et al. (2019) provide the general parametric notation for the expected number of events in group i as

$$\mathrm{E}(D_i) = \sum_{l=1}^{R} r_{il}\mathrm{P}(T_{ij} < C_{ij}, T_{ij} < L) = \sum_{l=1}^{R} r_{il} \int_0^{L-l} f_i(t)G_i(t)\mathrm{d}t. \tag{10}$$

As can be seen in the formula, the expected number of events depends on the length of the recruitment period R, monthly recruitment per group r_{il} and the length of the follow-up period L. Moreover, it depends on the probability density function of the time-to-event process $f(t)$ and the survival function of the time-to-censoring process $G(t)$. This formula is a more general version of Equation (9), because it accounts for both a variable accrual process and the time-to-censoring process. Moreover, the two treatment groups are considered separately here.

The time-to-event and time-to-censoring processes can be modeled using an appropriate parametric function. Friede et al. (2019) propose using an exponential model, Weibull model or piecewise exponential model, which can be fit to the interim data. If we assume independent group specific exponential event times, $f_i(t) = \lambda_i e^{-\lambda_i t}$, and common exponential censoring times, $g(t) = \gamma e^{-\gamma t}$, there exists and explicit solution to the integral $\int_0^{L-l} f_i(t)G(t)\mathrm{d}t$. Friede et al. (2019) show that the expected number of events can then be written as $\mathrm{E}(D_i) = \sum_{l=1}^{R} r_{il} \frac{\lambda_i}{\lambda_i+\gamma}\left(1-e^{-(\lambda_i+\gamma)(L-l)}\right)$. Here λ_i is the event rate of the group specific time-to-event process and γ is the event rate of the common time-to-censoring process. If the time-to-event process follows a Weibull distribution, no analytical solution to the integral exists and numerical integration methods need to be resorted to.

Since blindness needs to be maintained at the interim analysis, the group-specific parametric functions cannot be directly estimated from the data. Friede et al. (2019) assume that the time-to-censoring process is identical across treatment groups and thus estimating it from the pooled sample is sufficient. For the time-to-event process, they split the pooled estimates based on the assumed hazard ratio, θ^*. Specifically, when $\bar{\lambda}$ is the estimated pooled event rate, the event rate of the treatment group can be calculated as $\lambda_2 = \frac{1+k}{1+k/\theta^*}\bar{\lambda}$. Then, the event rate of the control group can be obtained as $\lambda_1 = \lambda_2/\theta^*$.

Based on the formulas above, Friede et al. (2019) propose that the BSSR can be carried out in the following steps:

1. Fit parametric distributions for the time-to-event and time-to-censoring processes using the blinded interim data and gather information on observed and predicted recruitment.

2. Based on the interim estimates, re-evaluate the trial design with regard to the expected number of events given the originally planned sample size.

3. If the expected number is at least as large as d, the design does not need to be changed.

4. If the expected number is smaller than d, the design needs to be changed to maintain the desired trial length of L months. Specifically, the sample size needs to be increased to ensure that that at least d events are observed within L months. Friede et al. (2019) proposed to increase the sample size by prolonging the recruitment period, R. The necessary number of additional recruitment months was determined by the following iterative process: Prolong recruitment by 1 month and re-evaluate the expected number of events. If it is still smaller than d, repeat this step iteratively. Once the expected number of events is at least as large as d we have found the required number of additional recruitment months and the design is modified accordingly.

To assess the operating characteristics of their proposed BSSR method, Friede et al. (2019) carried out a simulation study of event-driven trials, in which their method was compared to a fixed sample size design. In the fixed design, no sample size reestimation is carried out and the study is implemented according to the decisions made at the design stage (Friede et al., 2019). In their simulation they considered exponentially distributed event and censoring times and modeled the design of the simulation after a recent clinical trial in secondary progressive multiple sclerosis (SPMS) (Kappos et al., 2018). In the planning stage of the simulation studies they assumed that the probability of disease progression was 30% after 24 months individual observation time. Then, data were simulated to violate that assumption by having probabilities of disease progression ranging from 20%, 21% ..., 30% (Friede et al., 2019).

In a nutshell, they found that their proposed BSSR procedure (based on an exponential model) correctly recognized that too few events would be observed in the misspecification scenarios and consequently recruited additional patients, which resulted in the study generally finishing on time (Friede et al., 2019). However, the number of additional patients to be recruited was capped at a maximum of 3 or 6 additional months of recruitment. This meant that for scenarios with significantly lower disease progression probabilities the desired study duration could not be maintained in spite of the maximum number of patients having been added (Friede et al., 2019). Moreover, they found that due to sampling errors, the BSSR procedure sometimes recruited additional patients even when this was not necessary, since the disease progression probabilities were correctly specified (Friede et al., 2019). Overall though, the BSSR procedure was successful in maintaining the desired trial duration at the costs of recruiting additional patients. Importantly, Friede et al. (2019) reported that the BSSR procedure did not result in an increase in the type I error rate. In the simulated null hypothesis scenario the rejection probabilities were found to be close to the nominal significance of 0.05 (Friede et al., 2019). Besides, they also investigated the method's robustness to misspecification of the assumed hazard rate, which is used to split up the pooled estimates. They found that the operating characteristics were still favorable for the BSSR procedures compared to the fixed sample size design in spite of this misspecification. In particular, the BSSR designs more frequently finished on time, again at the costs of recruiting additional patients (Friede et al., 2019).

Evidently, parametric approaches for BSSR in time-to-event trials are useful tools for improving the trial design when the planning assumptions are not correct. A problem with parametric modeling, however, is that it relies on distributional assumptions for the data. Such assumptions should be justified based on some external data (Friede et al., 2019). If the distributional assumptions of a chosen model do not hold, survival extrapolation and hence estimates of the expected number of events may be biased. In a BSSR setting this could lead to large additional costs due to delayed trial closure and delayed admission of a potentially effective new treatment. Consequently, trial statisticians should closely consider how they choose to extrapolate survival based on available interim data. To obtain a better understanding of the available methods, we carried out a literature review on survival extrapolation methods, which we present in the next section.

3 Literature review: Methods for survival extrapolation

As has become evident from the BSSR approaches presented above, different methods of survival extrapolation are available to trial statisticians carrying out a blinded review. Indeed, survival extrapolation has been an issue of interest in medical statistics from different perspectives. Prominently, both health economists and trial statisticians have been studying extrapolation methods, albeit with different goals. For the former, the interest lies in long-term extrapolation of survival for a given treatment (beyond currently available trial data) to assess costs and effectiveness on an economic scale. For trial statisticians, in contrast, the interest typically lies in predicting the analysis time or the expected number of events within a given trial. The following sections provides a brief overview of the (overlapping) extrapolation methods commonly used in both of these settings. Two specific extrapolation methods were of particular interest to the current research project. The first are so-called hybrid methods, which combine a non-parametric Kaplan-Meier estimate with some parametric extrapolation for the tail of the survival curve. The second is the Royston-Parmar model, which uses restricted cubic splines to model the survival function. These methods will be covered in more detail in Subsections 3.2 and 3.3.

3.1 Extrapolation in health economics and clinical trial literature

A comprehensive and widely cited review of survival extrapolation in health economics has been presented by Latimer (2013). He analyzed the empirical use of extrapolation methods in UK based health technology assessment studies concerned with advanced or metastatic cancer. Health technology assessment (HTA) studies assess the short- and long-term consequences of health technologies (e.g. pharmaceutical treatments) in order to support policy decision making (Draborg et al., 2005). In interventions that affect survival such economic evaluations typically consider a lifetime horizon and therefore extrapolation of currently available trial data is necessary (Latimer, 2013). The goal of Latimer (2013) was to analyse what extrapolation methods are commonly used in practice and how these methods were justified. Their sample consisted of 45 HTAs that had been completed by December 2009.

The majority of the studies considered carried out some sort of extrapolation (Latimer, 2013). Parametric models were the most common extrapolation method, being used in 71% of the studies. In most cases Weibull or Exponential models were fit (72% and 63% of extrapolations), while other parametric models - for example Gompertz, log-normal, log-logistic or generalized gamma models - were used only occasionally (Latimer, 2013). Complex parametric models (e.g. spline models or piecewise models) were found not to have been used in practice at that point (Latimer, 2013). In very few studies, external registry data was considered or hybrid methods (which combine a Kaplan-Meier estimate with a parametric tail) were used. In almost all cases the parametric models were fit considering all available data, but in few cases a restricted data set was used (Latimer, 2013). For example, one study only considered data up to a certain time point. Apart from that, proportional hazards were assumed in the majority of the studies (59%) (Latimer, 2013). Parametric models were often fit to the control group and survival curves for the experimental group were obtained by applying the observed hazard ratio. Latimer (2013) notes that few studies made explicit assumptions about the duration of the treatment effect, which implies that the hazard ratio observed in the trial would be assumed to last for the entire extrapolation period.

As for the model selection and justification, Latimer (2013) observed that 69% of studies reported some justification for why the given extrapolation method was used. However, they note that these justifications were often very brief and model selection had not been considered extensively. For example, they mention that only 37% of studies fit more than one parametric model (Latimer, 2013). To assess the model fit, several studies carried out a visual inspection comparing the model fit to the Kaplan-Meier curve for the observed part of the data. Moreover, some studies used information criteria such as AIC and BIC to compare different models. Occasionally, log-cumulative hazards plots were analysed to assess the fit of a model (Latimer, 2013). Clinical validity and external data were rarely considered when assessing the extrapolation of a chosen model (Latimer, 2013).

Based on his observations, Latimer (2013) concludes that HTA studies did not pay sufficient attention to how they extrapolated survival curves and how they justified their model selection. Therefore, he proposes an algorithmic procedure for model selection, which revolves around a systematic assessment of a variety of candidate models. He emphasizes the importance of justifying the model selection in terms of both internal validity (fit to observed data) and external validity (plausibility of extrapolated tail of the survival curve) (Latimer, 2013). Moreover, he suggests that complex models - e.g. spline models - might be a useful addition to standard parametric models and should be considered more in the future (Latimer, 2013).

An updated review of extrapolation used methods in UK based HTAs of cancer treatments has recently been carried out by Bell Gorrod et al. (2019). They considered HTAs completed between 2011 and 2017, which resulted in a sample of 58 studies (Bell Gorrod et al., 2019). They report that - similar as before - in the majority of studies (91%), standard parametric models were used. However, while hybrid methods and piecewise constant models had previously only very rarely been observed, they had now become increasingly popular (17% and 23%, respectively) (Bell Gorrod et al., 2019). Moreover, HTA researchers had begun to occasionally use complex parametric models to account for complex hazard functions observed in their datasets (Bell Gorrod et al., 2019). In particular, two HTA studies used flexible parametric spline models. Another review by Grumberg et al. (2022) of French HTAs concerning immune-checkpoint inhibitors generally reported similar results regarding the use of extrapolation methods. They found that standard parametric and hybrid methods were commonly used. Moreover, the use of spline-based extra-polation methods had also been observed (Grumberg et al., 2022). Since such flexible approaches have slowly become more popular, a comprehensive review of flexible extrapolation approaches in HTAs has recently been provided by Rutherford et al. (2020) for the UK-based National Institute for Health and Care Excellence (NICE). We will discuss flexible parametric spline models in detail in Section 3.3 of this thesis.

In the clinical trial literature, extrapolation of survival curves has been investigated in order to predict when a certain number of events can be expected. In general, similar approaches have been suggested as in the health economics literature. For example, Bagiella and Heitjan (2001) developed a method to predicting analysis times by extrapolating survival times based on an exponential model. Ying and Heitjan (2008) extended this approach by extrapolating based on a more flexible Weibull model. Methods for blinded data have been investigated by Donovan et al. (2006), who propose a Bayesian mixture model approach. Besides, a number of hybrid approaches have also been proposed, which combine a non-parametric Kaplan-Meier estimate with some parametric extrapolation for tail of the survival curve (Moeschberger and Klein, 1985; Gelber et al., 1993; Ying et al., 2004; Fang and Su, 2011). These will be presented in more detail the next section.

Recently, some novel approaches have been suggested. For example, Chen (2016) proposed a parametric mixture cure rate model to predict analysis times in immuno-oncology studies, which can account for delayed and long-term survival effects in a proportion of patients. As a flexible alternative to specific parametric models, Lan and Heitjan (2018) developed an adaptive Bayesian model selection procedure, which selects among cure or non-cure models with an exponential or Weibull distribution. Rather than focusing on the selection of a specific model, Ou et al. (2019) developed a Bayesian prediction synthesis method, which combines predictions from different models. A detailed consideration of these novel methods, however, is beyond the scope of this thesis.

3.2 Hybrid methods

Hybrid methods combine the non-parametric Kaplan-Meier estimator with some parametric model that is used for extrapolation. This is intuitively appealing, because they make use of the fact that an unbiased, non-parametric estimator of the survival function exists for the available follow-up time. Already in 1985, Moeschberger and Klein considered using the Kaplan-Meier estimator until the last observed event and further extrapolating the survival curve based on an appropriate parametric model, such as a Weibull model. Their goal was to extrapolate the survival curve in order to compute mean survival times or survival percentiles. They suggested that the parametric model should be fit based on all available observations using the maximum likelihood method (Moeschberger and Klein, 1985). As an alternative option, they considered a restricted maximum likelihood estimation, where the extrapolated Weibull tail was tied to the tail of the Kaplan-Meier estimator at the last observed event (Moeschberger and Klein, 1985). Based on this idea and the work of Bagiella and Heitjan (2001), Ying et al. (2004) extended this hybrid approach to the context of predicting event times in clinical trials. Moreover, they extended the method to include prediction intervals based on Bayesian bootstrapping (Ying et al., 2004). In a simulation study they demonstrated that the hybrid method is superior to the parametric approach of Bagiella and Heitjan (2001) when the assumptions of the parametric model are wrong.

Gelber et al. (1993) proposed a slightly different hybrid approach, where the Kaplan-Meier estimator is not used until the last observed event, but only until a specific time point such as the median follow-up time. Then, the parametric model for the extrapolation is fit only based on the observations beyond this time point. Their rationale was that in some cases it may be more useful if the extrapolation model is based on the latter part of the data rather than all all data. As an example they name the healthy entrant phenomenon, where patients selected for a trial are initially healthy (since they are able to enter the trial), but subsequently experience relapses and death (Gelber et al., 1993). In such a case, they argue, it would be more useful to extrapolate based on the tail of the Kaplan-Meier, because it is more representative of the hazards that are to be expected (Gelber et al., 1993). The choice of a cutpoint, however, is not trivial. While the median follow-up time may be useful as a measure of maturity of the trial, enough patients should be available after the cutpoint to ensure a stable model fit (Gelber et al., 1993). Moreover, the choice of the cutpoint and the fit of the parametric model may be assessed using a probability plot and, if available, should be justified based on additional evidence or external data (Gelber et al., 1993).

The family of hybrid method was further expanded by Fang and Su (2011), who developed a two-step approach to account for potential change points in the survival function. They point out that the approach by Ying et al. (2004) assumes a smooth survival function without changepoints, since the parametric tail is estimated based on all available observations. However, as they point out, survival functions may have one or multiple change points after which the hazard function may become significantly different. They suggest a change-point detection procedure, which formalizes Gelber at al.'s (1993) idea that extrapolation should be potentially based on a latter, more representative part of the data. Specifically, the suggest the following two-step approach: First, a piecewise exponential model is fit to the interim data to detect potential change points in the survival function (Fang and Su, 2011). Second, if one or several change points are detected, the Kaplan-Meier estimator is only used until the last change point. Then, extrapolation is done based on an exponential model with the hazards observed for the last interval of the piecewise exponential model (Fang and Su, 2011). If no change points are detected, they suggest that either Bagiella and Heitjans' (2001) parametric approach or Ying et al.'s (2004) hybrid method can be used. They also briefly introduce a generalization of this two-step approach based on a piecewise Weibull model, rather than a piecewise exponential model (Fang and Su, 2011). In a simulation study, Fang and Su (2011) found that their proposed method outperformed the exponential model, piecewise exponential model and Ying et al.'s (2004) hybrid method in a misspecification scenario with piecewise

Weibull survival data. Fang and Su's (2011) hybrid method with change point detection has been implemented in the R software package "eventTrack" (Rufibach, 2021). A large limitation of this method (as well as the R package) is, however, that time-to-censoring has not been considered in the prediction of the expected number of events, which limits its practical applicability.

3.3 The Royston-Parmar spline model

A modeling approach that has recently received a lot of attention by health economists in the context of survival extrapolation (Kearns et al., 2021; Gray et al., 2021; Cooper et al., 2022) is the Royston-Parmar spline model (Royston and Parmar, 2002). A comprehensive review of the method has been provided in a textbook by Royston et al. (2011) and an accessible summary is available in Chapter 9 of Collett (2015). Royston and Parmar (2002) developed their spline method as a parametric alternative to the flexible Cox model commonly used in survival analysis. Rather than treating the baseline hazard as a nuisance parameter, their aim was to explicitly model it in order to gain a better understanding of the time-course of an illness (Royston and Parmar, 2002). Importantly, if the baseline hazard is explicitly modeled using a parametric model, this model can be used to extrapolate and make out of sample predictions (Royston et al., 2011, Ch. 1). Naturally, standard parametric models (e.g. exponential, Weibull or Gompertz models) can be used for this purpose, but they may fail to capture complex hazard shapes observed in clinical practice.

For example, it is not uncommon for real-life datasets to exhibit multiple turning points, for example in cancer research (Latimer and Adler, 2022). Rutherford et al. (2020) discuss that patients may be relatively healthy upon recruitment, but probably will experience an increase in mortality due to their disease. However, the composition of the cohort is expected to change over time, due to sicker patients dying and treatment responders surviving, which would result in a decreasing hazard (Rutherford et al., 2020). Following that, however, the effectiveness of the investigated treatment may reduce over time or progression of the disease may take place, resulting in a new increase in hazards (Rutherford et al., 2020). Such potentially complex hazards cannot be captured by standard parametric models. Instead, more flexible parametric models are necessary in such situations (Latimer and Adler, 2022).

3.3.1 Model formulation

The Royston-Parmar model can be seen as a generalization of the Weibull model and thus it can naturally be introduced in this manner (Royston and Parmar, 2002). Assume our survival times follow a Weibull distribution $S(t) = \exp(-\lambda t^{\gamma})$, with scale parameter λ and shape parameter γ. If we apply the log-log transformation, we obtain the log cumulative hazard function

$$\begin{aligned}\ln\{H(t)\} &= \ln[-\ln\{S(t)\}] \\ &= \ln[-\ln\{\exp(-\lambda t^{\gamma})\}] \\ &= \ln(\lambda) + \gamma \ln(t) \qquad (11)\end{aligned}$$

As we can see, the log cumulative hazard function is linear in $\ln(t)$. However, if the distribution of the survival times departs from a Weibull distribution, then $\ln\{H(t)\}$ will no longer be linear in $\ln(t)$ (Royston and Parmar, 2002). Instead, these two quantities will be related by some non-linear function. The idea of Royston and Parmar (2002) was to model such a non-linear relationship by using so-called restricted cubic splines and thereby creating a class of more flexible parametric models.

Splines are a class of flexible functions that are commonly used to model non-linear relationships (Lambert and Royston, 2009). Spline functions are defined by piecewise polynomials that are fitted separately over a specified number of intervals (Royston et al., 2011, Ch. 4). In practice, cubic polynomials are most commonly used, in which case the function is referred to as a cubic spline (Royston et al., 2011, Ch. 4). The points that define the intervals of the piecewise polynomials are called knots (Royston et al., 2011, Ch. 4). To ensure that the overall spline function is smooth so-called continuity restrictions are imposed (Royston et al., 2011, Ch. 4). First, the cubic polynomials within each interval are forced to join at the knot locations to ensure a continuous function. Second, the first and second derivative of the spline function are forced to be continuous, which leads to a smooth function. Finally, restricted cubic splines are defined by the additional feature, that they are not cubic but linear beyond their boundary knots (Royston et al., 2011, Ch. 4). Imposing this linearity is useful, because it often leads to a more sensible functional forms in the tails of the spline function, where often data is sparse (Royston et al., 2011, Ch. 4).

Royston and Parmar (2002) propose that the smallest and largest uncensored log survival times are used as the boundary knots. The number of piecewise polynomials being used depends on the number of internal knots, which is specified before fitting the model (Royston et al., 2011, Ch. 4). The higher the number of internal knots, the more flexible the spline model becomes.

To illustrate the mathematical formulation of the restricted cubic splines model, we can first introduce the notation for a spline model with a single internal knot (Collett, 2015, Ch. 6). Let $y = \ln(t)$ denote the uncensored log survival times. Note that no index for the patients is included here, since we are modeling survival without the inclusion of covariates. The smallest value k_{min} is the lower boundary knot and the largest value k_{max} is the upper boundary knot. The internal knot k_1 is set to be the median value of y and splits the log survival times into two halves. Now, a cubic expression in y is defined for each of the two intervals $y \in (k_{min}, k_1)$ and $y \in (k_1, k_{max})$ (Collett, 2015, Ch. 6). The restricted cubic spline is defined by the property that we assume a linear term in y for $y < k_{min}$ and $y > k_{max}$. The mathematical formulation for this 1-knot spline model (Collett, 2015, Ch. 1) is

$$\ln\{H(t)\} = \gamma_0 + \gamma_1 y + \gamma_2 \nu_1(y), \tag{12}$$

where

$$\nu_1(y) = (y - k_1)^3_+ - \lambda_1 (y - k_{min})^3_+ - (1 - \lambda_1)(y - k_{max})^3_+,$$

with

$$(y - a)^3_+ = \max\{0, (y - a)^3\},$$

for any value a, and

$$\lambda_1 = \frac{k_{max} - k_1}{k_{max} - k_{min}}.$$

Royston and Parmar (2002) derived this formulation for the restricted cubic splines based on the linearity constraint by using the knowledge that the 2nd and 3rd derivative of the function must be 0 beyond the boundary knot. Their derivation can be found in Appendix B of Royston and Parmar (2002). Collett (2015, Ch. 6) illustrates that for $y > k_{max}$ the 1-knot spline function reduces to $\gamma_0 + \gamma_1 y + \gamma_2 (k_{max} - k_1)(k_{min} - k_1)(3y - k_{min} - k_1 - k_{max})$ and thus, indeed, becomes linear in log-time beyond the boundary knot.

The more general restricted cubic spline with m internal knots $k_1 < ... < k_m$ is the Royston-Parmar model (RP model) (Collett, 2015, Ch. 6). The general model contains m non-linear terms $\nu_1(y), \nu_2(y), ..., \nu_m(y)$, one for each internal knot (Collett, 2015, Ch. 6). These non-linear terms are referred to as basis functions by Royston and Parmar (2002).

The Royston-Parmar model is then defined as

$$\ln\{H(t)\} = \gamma_0 + \gamma_1 y + \gamma_2 \nu_1(y) + ... + \gamma_{m+1}\nu_m(y), \tag{13}$$

with the basis function for the jth knot at k_j with $j = 1, 2, ..., m$ being defined as

$$\nu_j(y) = (y - k_j)^3_+ - \lambda_j(y - k_{min})^3_+ - (1 - \lambda_j)(y - k_{max})^3_+,$$

and

$$\lambda_j = \frac{k_{max} - k_j}{k_{max} - k_{min}}.$$

The Royston-Parmar model is fully parametric and can thus be fitted using standard maximum likelihood methods (Royston and Parmar, 2002). Asymptotic standard errors for the parameters are readily available, which allows for the construction of point-wise confidence intervals of the log cumulative hazard function (Royston and Parmar, 2002). These can then be transformed to also obtain confidence intervals for the survival function and the cumulative hazard function (Royston and Parmar, 2002). Assume the Royston-Parmar model for the log cumulative hazard has been fit and denote the spline function as s. Then, the corresponding survival, density and hazard functions can be obtained as (Royston and Parmar, 2002)

$$S(t) = \exp(-\exp s) \tag{14}$$

$$f(t) = \frac{\mathrm{d}s}{\mathrm{d}t}(s - \exp s) \tag{15}$$

$$h(t) = \frac{\mathrm{d}s}{\mathrm{d}t}\exp(s) \tag{16}$$

Note that, in particular, estimates of the survival function are very straightforward to obtain.

When several internal knots are used the issue of the knot placement needs to be re-considered. For the 1-knot spline model we briefly mentioned that the median of the uncensored log-survival times could be used. This was a special case and in general Royston and Parmar (2002) suggest that as a default internal knots should be spaced equally between percentiles of the distribution of the uncensored log survival times (Royston and Parmar, 2002). For example, in a 2-knot model, they knots would be placed at the 33% and 67% percentiles and in a 3-knot model at the 25%, 50% and 75% percentiles. Royston et al. (2011, Ch. 4) have pointed out that optimal knot positioning does not seem to be crucial for a good model fit and so robustness of the method can be expected. Note that when a 0-knot model is fit, the model reduces to a linear function of the log cumulative hazard function in log-time (Royston and Parmar, 2002). As we have shown in Equation (11) such linearity holds for a Weibull distribution and thus a 0-knot spline model reduces to a Weibull model.

We want to briefly note here that a similar flexible spline-based approach to that of Royston and Parmar (2002) for modeling survival has been proposed by Crowther and Lambert (2014). They also use restricted cubic splines, but they model the log hazard function rather than the log cumulative hazard function (Crowther and Lambert, 2014). Such a model formulation is useful when there are multiple time-dependent effects (Crowther and Lambert, 2014). When modeling without covariates for the purpose of extrapolation, however, the formulation of Royston and Parmar (2002) should be preferred. First, the log cumulative hazard function is much more stable than the log hazard function and it is therefore easier to capture the shape of the function (Rutherford et al., 2015). Second, if the log cumulative hazard function is modeled, the cumulative hazard function and thus the survival function can be obtained without numerical integration (Rutherford et al., 2015). Therefore, we deemed the Royston-Parmar model more practical for our application.

3.3.2 Number of knots and model selection

When fitting Royston-Parmar models (or any spline model for that matter), an important question is not only where, but how many internal knots should be fitted. In their original publication Royston and Parmar (2002) recommend one internal knot as a reasonable initial choice. They point out that often a significant improvement over a simple Weibull model is obtained by adding one knot, but adding further knots does not always further improve the model fit (Royston and Parmar, 2002). Moreover, they indicate that models with more than three internal knots might be unstable (Royston and Parmar, 2002). However, Royston et al. (2011, Ch. 5) have pointed out that with larger datasets more complex models with up to 5 or 6 internal knots might be useful. Beyond the initial recommendation of using one internal knot, Royston and Parmar (2002) recommend that other models should be checked as a sensitivity analysis. In particular, they suggest to informally look at the Akaike Information Criterion (AIC) of the models. The AIC is defined as $AIC = -2\log\hat{L} + 2q$, where $\hat{L}$ denotes the likelihood and q denotes the number of parameters of the model (Akaike, 1974). However, Royston and Parmar (2002) warn of using information criteria mechanically for model selection. Similarly, Royston et al.

(2011, Ch. 5) point out that with very large datasets (>10,000 observations) formal inference using information criteria may not be helpful. They refer to instances where a supposedly better fitting model was almost indistinguishable from less complex spline models with fewer knots (Royston et al., 2011, Ch. 5). Therefore, they recommend that in practice model selection should be also be based on a "feel" of the model fit, rather than a theoretically optimal fit (Royston et al., 2011, Ch. 5).

To better understand the issue of the number of knots and model selection in Royston-Parmar models, Rutherford et al. (2015) have carried out extensive simulations. They considered both simple hazard scenarios (Weibull distributions) as well as various complex hazard scenarios (simulated through mixture Weibull distributions). They fit Royston-Parmar models with up to 9 internal knots and assessed the absolute area difference between the fitted and true function across 1,000 simulations (Rutherford et al., 2015). Moreover, to assess model selection, they computed the AIC and BIC for each model and registered which model was selected based on the respective information criterion in each simulation run (Rutherford et al., 2015). The BIC is defined as $BIC = -2\log\hat{L} + \log(n)q$, where n denotes the number of uncensored observations (Collett, 2015, Ch. 3). It typically penalizes complex models stronger than the AIC, since the $\log(n)$ in the penalty term will be larger than 2 for $n \geq 8$.

In summary, Rutherford et al. (2015) found in their simulation that using 2-4 internal knots was usually sufficient to capture even very complex hazard shapes (Rutherford et al., 2015). Moreover, depending on the scenario, even adding only one internal knot could lead to a significantly improved performance in complex hazards scenarios compared to a standard Weibull model (Rutherford et al., 2015). However, they also noted that overfitting could occur in spline models, especially as the model complexity increases (Rutherford et al., 2015). This seemed to be more prominent in small sample sizes compared to large sample sizes (Rutherford et al., 2015).

With regards to model selection, they generally found that in complex hazard scenarios the AIC tended to select a higher number of internal knots, with an average of 4-5 internal knots for the smallest sample size of $n = 300$ (Rutherford et al., 2015). With larger sample sizes ($n = 3,000$ or $n = 30,000$), the average number of internal knots selected by AIC went up and frequently a very large number of internal knots was selected (Rutherford et al., 2015). Interestingly, this also was the case for simulated Weibull data, where a 0-knot (Weibull) model would be sufficient to model the data generating process.

Specifically, the mean number of internal knots selected by the AIC were 1.76, 2.48 and 7.81 for sample sizes of 300, 3,000 and 30,000, respectively (Rutherford et al., 2015). This implies that as the sample size increases (and particularly with very large sample sizes) the AIC cannot be relied on to ensure a parsimonious model selection for Royston-Parmar models. This can be problematic, because highly flexible spline models with a large number of internal knots can be prone to overfitting (Rutherford et al., 2015). Indeed, Rutherford et al. (2015) found that for simulated Weibull data, highly flexible spline models had a poorer fit due to occasional over-fitting. That is, the spline model picked up local deviations from the Weibull model and modeled the observed data too closely (Rutherford et al., 2015). This seemed to be more problematic in small ($n = 300$) or modestly sized ($n = 3,000$) samples, where random variations in the data-generating process might be more pronounced (Rutherford et al., 2015). Fortunately though, in small sample sizes the AIC tends to recommend simpler models, so this might mitigate the issue of overfitting due to complex spline models to some extent.

In general, overfitting could potentially be prevented by using a more stringent information criteria like the BIC. For example, Rutherford et al. (2015) found that even for a sample size of $n = 30,000$ the average number of internal knots chosen by the BIC were only 1.44, when the data were generated based on a Weibull model. Of course, though, the larger penalty term employed by the BIC will also lead to less complex models being selected where it might be necessary. Indeed, Rutherford et al. (2015) report that in the complex hazards scenarios, the BIC typically selected 1-2 internal knots less than the AIC. Similarly as before though, the suggested number of internal knots increased with an increase in sample size (Rutherford et al., 2015).

To conclude, information criteria may provide a helpful guidance in terms of how many internal knots might be appropriate for a given dataset. However, care needs to be taken especially in larger sample sizes, because unnecessarily complex models might be selected (Rutherford et al., 2015). In practice, Rutherford et al. (2015) suggest to carry out a sensitivity analysis by plotting the hazard function for a range of values for the number of internal knots. If the hazard function hardly changes with a higher number of internal knots, this is a sign of overfitting and a simple model should be preferred (Rutherford et al., 2015).

Recently, penalized spline functions have been proposed as a generalization of the Royston-Parmar spline model (Liu et al., 2018). Rutherford et al. (2020) state that it is currently unclear how such a penalized approach would influence the extrapolation performance of spline models. However, considering such models is beyond the scope of the current thesis.

3.3.3 Extrapolation

Regarding extrapolation based on the Royston-Parmar model, the restricted cubic splines imply that the values beyond the boundary knot k_{max} follow a local Weibull distribution. This holds, because linearity of $\ln[-\ln\{S(t)\}]$ in $\ln t$ is assumed beyond the last available time point. It should be noted, however, that these local Weibull predictions will be different from those of a standard Weibull model fit to the entire data set. The linear trend in the restricted cubic splines (and thus the Weibull extrapolation) arises as a continuation of the last cubic polynomial given the continuity restrictions imposed by the model. The larger the number of knots (that is, the closer to the end of follow-up the last knot will be placed) the more the extrapolation may be based on the tail of the observed data (Rutherford et al., 2020). Since it is still a Weibull model, however, the hazard function will be monotonic beyond the right boundary knot.

A possible remedy to this limitation would be to employ another, less frequently used version of Royston-Parmar model, which was also introduced in their original article in 2002 (Royston and Parmar, 2002). While their generalization of the Weibull model represented a flexible proportional hazards model, they also introduced a flexible proportional odds model based on a generalization of the log-logistic model (Royston and Parmar, 2002). Rather than using restricted cubic splines to model the log cumulative hazard as a function of log-time, they proposed to alternatively model the log-odds of survival beyond t using the same approach (Royston and Parmar, 2002).

It is well known that if survival times follow a log-logistic distribution, the log-odds of survival beyond t are linear in $\ln(t)$ (Collett, 2015, Ch. 6). Specifically, assume that survival times follow a log-logistic distribution $S(t) = (1 + e^{\theta} t^{k})^{-1}$ (Collett, 2015, Ch. 6). Then the odds of survival beyond time t are $\frac{S(t)}{1-S(t)} = e^{-\theta} t^{-k}$. Consequently, the log-odds are linear in $\ln t$, since $\ln\{\frac{S(t)}{1-S(t)}\} = -\theta - k \ln t$ (Collett, 2015, Ch. 6). Given this knowledge, the same generalization can be carried out as previously noted for the Weibull model. If the survival times do not follow a log-logistic distribution, them the log-odds of surviving beyond t will be related to $\ln t$ by some non-linear function (Royston and Parmar, 2002). Again restricted cubic splines can be used as a flexible function to capture this relationship, resulting in a flexible class of proportional odds models (Royston and Parmar, 2002). Formulas to convert the model estimates into estimates of the survival, density and hazard function are provided in Royston and Parmar (2002). The authors have called these generalizations of the Weibull and log-logistic model, PH spline model and PO spline model, respectively (Royston and Parmar, 2002). A third alternative formulation of the model on the probit scale has been suggested by Royston and Parmar (2002), which can be seen as a generalization of the log-normal model. However, it has received little attention both in the original paper and in the literature that followed (for example, see Collett (2015, Ch. 6)), so it will not be presented here. The important point here is that the PO spline model provides a different extrapolation mechanism from the PH spline model, since the extrapolated part beyond the right boundary knot will now follow a local log-logistic distribution (Royston and Parmar, 2002). This holds, because beyond the boundary knot the restricted cubic spline models the relationship between the log-odds of surviving and $\ln t$ as linear.

Properties of survival extrapolation using Royston-Parmar models have been briefly investigated in Chapter 6 of Royston et al. (2011). There they considered an extrapolation scenario similar to the one encountered in blinded sample size reestimation scenarios, namely the simple modeling of the survival function without covariates. Specifically, they compared various Royston-Parmar spline models to standard parametric models (here: Weibull, log-logistic, log-normal models) by fitting them to the Rotterdam breast cancer dataset (Royston et al., 2011, Ch. 6). They found that that the standard parametric models had a poorer fit to the observed 10 years survival data as well as a poor extrapolation performance (Royston et al., 2011, Ch. 6). In contrast, the Royston-Parmar spline models predicted survival well for the extended follow-up times of 20 years (Royston et al., 2011, Ch. 6). They fit both the PH spline and PO spline model with 1-3 internal knots and found that all of those models extrapolated similarly well (Royston and Parmar, 2002).

To investigate the robustness of the extrapolation performance of the Royston-Parmar models, they considered several modifications. First, they changed the right-hand boundary knot in the Royston-Parmar models (beyond which a linear trend is assumed on the respective scale) and found that the extrapolated survival curves were hardly affected by this (Royston et al., 2011, Ch. 6). Second, they reduced the available follow-up times from 10 to 5 years. This significantly reduced the extrapolation accuracy of the standard parametric models, but had had little effect on the Royston-Parmar models (Royston et al., 2011, Ch. 6). These preliminary investigations suggest that these flexible spline models might be useful extrapolators, which may outperform the standard parametric models in some datasets.

3.3.4 Case studies and simulation studies

Due to the increased interest in using spline models for survival extrapolation, some researchers have undertaken systematic case studies and simulation studies to assess the extrapolation performance of the Royston-Parmar model (Rutherford et al., 2020; Kearns et al., 2021; Gray et al., 2021). Gray et al. (2021) compared the extrapolation performance of standard parametric models and the Royston-Parmar model by fitting both to artificially right-censored registry cohort data of advanced cancer. While these observational cancer registry data differ from clinical trial data in terms of cohort size and heterogeneity, they have the advantage of containing long-term follow-up times, which can be used to assess long-term extrapolation performance (Gray et al., 2021). Moreover, the large registry dataset could be used to generate various analysis cohorts. Specifically, Gray et al. (2021) considered five different cancer types (breast, colorectal, NSCLC, SCLC and pancreatic cancer) and 3 different age groups (18-59, 60-69, 70+). The purpose of this was to obtain datasets with different survival distributions and hazards shapes, which could provide a basis for a extensive comparison of the extrapolation methods of interest (Gray et al., 2021). The authors considered a time frame of 10 years and patients with follow up times >10 years were right censored. Then, three kinds of interim data were generated by artificially right censoring patients at the time points when either 50%, 35% or 20% of patients were still alive (Gray et al., 2021). The exact time of interim analysis and the length of the extrapolation period therefore varied greatly among the analysis cohorts, depending on the cancer type and the age group. For example, the early interim analysis (50% surviving) for the young cohort (18-59 years) was carried out after 26.03 months for breast cancer patients and after only 1.72 months for pancreatic cancer patients (Gray et al., 2021). The sample sizes of the different analysis cohorts ranged from around 5,000 to 30,000 (Gray et al., 2021).

Gray et al. (2021) fit six standard parametric models to the data: an Exponential, Weibull, Gompertz, log-logistic, log-normal and generalized gamma model. The Royston-Parmar model was fit it on all three available scales: proportional hazards, proportional odds and probit. Each spline model was fit with with 1-3 internal knots and the model with the lowest AIC was selected on each scale (Gray et al., 2021). In their simulated datasets, 3-knot spline models were selected the most frequently by the AIC (Gray et al., 2021). This is unsurprising, given the previous findings of Rutherford et al. (2015) that the AIC tends to select complex spline models in large datasets. As primary performance measures Gray et al. (2021) considered the predicted restricted mean survival time (RMST) and the prediction error in the percentage of patients surviving at 10 years.

To summarize, Gray et al. (2021) found for their datasets that when long follow-up data were available (20% survival) both the log-logistic and log-normal model, as well as the spline odds and spline probit models performed best in terms of the RMST and the prediction error. However, with shorter follow-up times the two spline models outperformed the log-logistic and log-normal models (Gray et al., 2021). Notably, the improvement in RMST was not only due to a better fit to the observed data, but also a more accurate survival extrapolation of the spline models (Gray et al., 2021).

Gray et al. (2021) illustrated the improved performance of the flexible spline models based on some example cohorts. They selected cohorts with various hazard shapes and showed that for example in a complex bimodal hazard distribution, only spline models were able to accurately capture the hazard shape (Gray et al., 2021). Other example cohorts displayed unimodal or simple monotonic hazards, all of which were well captured by at least one of the spline models (PH, PO or probit) (Gray et al., 2021). The authors conclude from this that spline models are useful for extrapolation, at least in the context of this large registry cohort they investigated.

While they investigated various scenarios by means of splitting up their dataset in terms of cancer type and age group, it is not certain whether these findings generalize to other datasets. Moreover, this observational dataset is different from clinical trial datasets, which would be encountered in a BSSR setting. Specifically, the registry data comprises large datasets with little censoring (Gray et al., 2021). In contrast, interim datasets of clinical trials may have smaller sample sizes and more uncertain tails due to few events and administrative censoring (Gray et al., 2021). Uncertainty in the tail region of interim data may be particularly problematic for spline models, since overfitting of the tail of the observed data might occur. Therefore, the authors suggest that spline-based extrapolation should be investigated in clinical trial settings in future simulation studies (Gray et al., 2021). Overall though, their study provides some encouraging preliminary evidence about the extrapolation performance of the Royston-Parmar spline model.

The insights of Gray et al.'s (2021) comprehensive case study have recently been complemented by a simulation study by Kearns et al. (2021). The aim of their study was to compare standard parametric models (referred to as "current practice" by them) and emerging practice models. The latter included the Royston-Parmar model as well as fractional polynomials (FPs), Generalised additive models (GAMs) and Dynamic survival models (DSMs) (Kearns et al., 2021). Here, we will focus on the results regarding the royston-parmar model. Kearns et al. (2021) considered PO spline and PH spline models, but did not consider probit spline models. They fit the spline models with up to 5 internal knots and selected the model with the lowest AIC (Kearns et al., 2021).

The data simulation was similar to that of Rutherford et al. (2015) in that a complex hazard scenario was simulated by means of a mixture Weibull distribution (Kearns et al., 2021). A single hazard distribution with two turning points was considered: the hazard function initially peaks at around 0.5 years and then slowly starts increasing again at about 1.75 years (Kearns et al., 2021). The goal was to simulate a complex, real world dataset, which would also account for the impact of aging on mortality in the long run. While the data-generating mechanism remained fixed, the simulation scenarios were defined by the sample size ($n = 100, 300$ or 600) and the different lengths of available follow-up (2, 3 or 4 years) (Kearns et al., 2021). This resulted in a total of 9 simulation scenarios, for each of which the simulations were carried out 200 times (Kearns et al., 2021). At all 3 follow-up points the second turning point in the hazard function had occurred, but the scenarios differed in terms of how much time had elapsed since then (Kearns et al., 2021). The performance measures used were the mean squared error (MSE) and the bias in the estimation of the logarithm of the time-varying hazard function (Kearns et al., 2021). Details on how the bias and MSE were computed for this quantity are presented in additional file 1 provided by Kearns et al. (2021). Since long-term extrapolation is common in health economic evaluations, extrapolation here was carried out until 20 years (Kearns et al., 2021).

In general, Kearns et al. (2021) found that the standard parametric models were unable to fit the observed portion of the hazard rate. The emerging models tended to better capture this, but often displayed a larger variability (Kearns et al., 2021). Regarding extrapolation performance, the long-term increase in hazards was only identified when the longest follow-up times (4 years) were available and only by two types of DSMs and by the GAM (Kearns et al., 2021). For these models the bias decreased with sample size and results were approximately unbiased for $n = 600$ (Kearns et al., 2021). However, these methods also tended to suffer from particularly large variability. This problem of large variance became even worse with shorter available follow-up times (2 or 3 years) as well as with smaller sample sizes, resulting in some extreme estimates that were far away from the true value of the hazard function (Kearns et al., 2021).

The Royston-Parmar model was not able to capture the long-term increase in the hazard rate (Kearns et al., 2021). That is, the good fit to the observed part of the data did not translate into an unbiased extrapolation. However, the Royston-Parmar models had a much lower variance than the other emerging practice models (Kearns et al., 2021). Therefore, while having some bias, there were not any extreme differences from the true hazard, as was observed for the DSMs and GAM (Kearns et al., 2021). As a result of this, the Royston-Parmar model frequently outperformed the other two in terms of MSE (Kearns et al., 2021). However, a DSM with a damped trend, while not capturing the long-term increase in hazards, had the lowest MSE (Kearns et al., 2021). Surprisingly, standard parametric models happened to have only a very small bias and MSE in this complex hazard simulation (Kearns et al., 2021). However, this was not because they accurately captured the hazard shape (they did not fit the observed part well), but because they extrapolated only a small decrease in hazards, which was not so far away from the true slow increase of hazard over time (Kearns et al., 2021). Therefore, this can be considered as an artifact of the simulated datasets at hand and this might not hold for other simulated datasets with complex hazards.

Overall, the conclusions that can be drawn from Kearn et al.'s (2021) simulation study are limited, since only one specific data-generating mechanism was considered based on a mixture Weibull distribution. However, we may draw some preliminary conclusions. Among the emerging practice models, the Royston-Parmar was not the best one in terms of detecting the hazard shape for the extrapolation part. However, the models that were able to capture this suffered from very large variability (as indicated by increased MSEs), which limits their practical applicability. The Royston-Parmar model had a decent MSE, which implies that the flexibility towards observed data together with only a moderate bias in the extrapolated tail offer a reasonable trade off in terms of flexibility and variance. This may be particularly true in scenarios with smaller sample sizes and limited follow up times, since these issues amplified the increased variance of the other emerging practice models. This alleviates, to some extent, Gray et al.'s (2021) concern that Royston-

Parmar models might suffer disproportionately from uncertainty in the data introduced by to censoring and smaller sample sizes. Apart from that, the study also highlights that selecting models based on within-sample fit (e.g. as measured by the AIC) does not guarantee an accurate extrapolation (Kearns et al., 2021). This indicates that integrating external evidence might be useful or even necessary for successful survival extrapolation (Kearns et al., 2021).

The inclusion of external information into spline-based extrapolation, among other things, has been investigated in a simulation study carried out by Rutherford et al. (2020). They considered 4 different simulation scenarios, which corresponded either to simple Weibull models or mixture Weibull models (Rutherford et al., 2020). Moreover, they included a scenario with a cure fraction. All scenarios were considered with either low or medium survival and low or high heterogeneity in individual survival functions (Rutherford et al., 2020). Apart from that, all models included background mortality to account for age related other causes of death (Rutherford et al., 2020). The follow-up time was 3 years. Extrapolation performance was assessed in terms of bias in the estimated mean overall survival times (Rutherford et al., 2020).

They found that spline models fit the observed data well in all scenarios as indicated by a very low bias in the restricted mean survival time (RMST) of the available follow up (Rutherford et al., 2020). As for the extrapolation performance, they found the the Royston-Parmar extrapolated very successfully in a Weibull scenario with increasing hazards, regardless of heterogeneity (Rutherford et al., 2020). However, in a Weibull scenario with decreasing hazards, they found that the Royston-Parmar had biased estimates of mean overall survival when survival was long and heterogeneity was large (Rutherford et al., 2020). This dependence on short survival and small heterogeneity was also observed for Weibull mixture distributed data (Rutherford et al., 2020). Finally, when the simulated data included a cure fraction, the Royston-Parmar model did not perform well and had bias in mean overall survival that was similar to that of standard parametric models (Rutherford et al., 2020). However, they found that including background mortality into the Royston-parmar model by means of a relative survival framework resulted in a marked reduction in bias (Rutherford et al., 2020). How the Royston-Parmar model can be extended into a relative survival framework has been described by Nelson et al. (2007).

Overall, these simulation and case studies provide some preliminary evidence of the strengths and limitation of using the Royston-Parmar model for survival extrapolation. It seems like spline models are useful in modeling complex hazards and can allow for more accurate extrapolation in such scenarios. And while they might not always remain unbiased, the seem to be relatively reliable in that their variance is not too inflated compared to other flexible modeling approaches. Of course, care must still be taken with regards to overfitting and model selection remains a non-trivial issue. All in all, however, spline models seem promising in terms of their extrapolation performance and they might be a useful additional tool to model survival in the context of BSSR. An extension of Friede's (2019) parametric BSSR method based on the Royston-Parmar spline model will therefore be presented in the next section.

A limitation of above case studies and simulation studies is that they have all been carried out to emulate the context of a health economic evaluation. Extrapolation in such a setting is different from that in a BSSR setting, since typically smaller sample sizes and shorter follow-ups will be available at an interim analysis. Moreover, extrapolation in an HTA covers a very long time range (often a lifetime horizon), whereas in an interim analysis we are interested in making predictions about the planned finish of the trial itself. Therefore, additional simulation studies are needed to assess the usefulness of the Royston-Parmar model for extrapolating survival in the context of a BSSR. Such a simulation study will be presented in Section 5, after the flexible BSSR method based on the Royston-Parmar model has been presented in the following section.

4 Proposed flexible blinded sample size reestimation procedure

4.1 Spline-based blinded sample size reestimation

As has become evident from the literature review, the Royston-Parmar model is a flexible alternative to standard parametric models, which might make parametric BSSR for clinical trials with time-to-event outcomes more robust to model misspecifications. This could be relevant, for example, if external data turns out to be unreliable (or unavailable). For example, a new patient cohort may differ from previous ones in that existing treatment recommendations might have changed in the meantime. This could lead to improved therapeutic outcomes, which could result in fewer events in a given time frame compared to previous cohorts. Moreover, a flexible model might be necessary if the observed hazard is more complex than a Weibull hazard, as has been hypothesized for example for immuno-oncology treatments (Ouwens et al., 2019). For our proposed BSSR method we assume, like Friede et al. (2019), that the censoring process is the same for both treatment groups. Then, Formula (10) simplifies to

$$E(D_i) = \sum_{l=1}^{R} r_{il} \int_0^{L-l} f_i(t)G(t)\mathrm{d}t. \tag{17}$$

In contrast to Friede et al. (2019), we propose that the expected number of events pooled across both treatment groups $E(D) = E(D_1) + E(D_2)$ can be estimated directly without considering both groups separately. Note that the marginal survival function $S(t)$ can be written as $S(t) = S_0(t)(1 - \pi) + S_1(t)\pi$. The treatment assignment probability, π, is known from the trial design. Based on the pooled recruitment numbers, we can then approximate the monthly recruitment numbers in each group as $r_{1l} \approx r_l\pi$ and $r_{0l} \approx r_l(1 - \pi)$. Then, using the Riemann-Stieltjes integral notation, it holds that

$$\begin{aligned}
E(D_0) + E(D_1) &= \sum_{l=1}^{R} r_{0l} \int_0^{L-l} f_0(t)G(t)\mathrm{d}t + \sum_{l=1}^{R} r_{1l} \int_0^{L-l} f_1(t)G(t)\mathrm{d}t \\
&= -\sum_{l=1}^{R} r_{0l} \int_0^{L-l} G(t)\mathrm{d}S_0(t) - \sum_{l=1}^{R} r_{1l} \int_0^{L-l} G(t)\mathrm{d}S_1(t) \\
&= -\sum_{l=1}^{R} r_l(1-\pi) \int_0^{L-l} G(t)\mathrm{d}S_0(t) - \sum_{l=1}^{R} r_l\pi \int_0^{L-l} G(t)\mathrm{d}S_1(t) \\
&= -\sum_{l=1}^{R} r_l \int_0^{L-l} G(t)\mathrm{d}\{S_0(t)(1-\pi) + S_1(t)\pi\} \\
&= -\sum_{l=1}^{R} r_l \int_0^{L-l} G(t)\mathrm{d}S(t) \qquad (18)
\end{aligned}$$

Thus, to estimate the total number of events, $E(D)$ we only need the pooled monthly recruitment numbers, r_l, and pooled estimates of the survival and censoring distribution, $G(t)$ and $S(t)$. This implies that a procedure for splitting up the survival curves - as has been proposed by Whitehead et al. (2001) and Friede et al. (2019) - might not be necessary to estimate the total expected number of events. Therefore, the model that is fit to the pooled survival data during an interim analysis can be used directly to extrapolate the survival times for the purpose of sample size reestimation.

We propose that a Royston-Parmar PH spline model is used to model the time-to-event and time-to-censoring processes, $S(t)$ and $G(t)$. The PH spline model is a generalization of the Weibull model and thus naturally extends the parametric PH models (exponential and Weibull) that have been proposed by Friede et al. (2019) in the context of BSSR. PO spline and probit spline models offer alternative extrapolation mechanisms, but considering these was beyond the scope of the current thesis. When a PH spline model is used to estimate $S(t)$ and $G(t)$, an analytic solution to the integral $\int_0^{L-l} G(t)\mathrm{d}S(t)$ was not apparent to us. Therefore, in our implementation we used numerical integration to compute the expected number of events.

The numerical integration was performed using the built-in R function *integrate()*, which performs adaptive quadrature (R Core Team, 2017). In adaptive quadrature, two different numerical integration methods are compared given some tolerance, which indicates the requested accuracy (Gander and Gautschi, 2000). If the relative difference between the two methods is larger than the specified tolerance, the interval to be integrated is divided and the two integration methods are applied to both integral parts (Gander and Gautschi, 2000). This division of the intervals is repeated recursively until the two integration methods agree within the specified tolerance or until another stopping criterion applies (Gander and Gautschi, 2000). The implementation of the built-in R function uses globally adaptive interval subdivision based on the Gauss-Kronrod approach (R Core Team, 2017). Details on this procedure are available in Section 4 of Gander and Gautschi (2000). For our numerical integration we used the default settings of the *integrate()* function, in which the maximum number of subintervals was set to 100 and the tolerance was set to 0.0001221 on our machine (R Core Team, 2017).

When the Royston-Parmar model is deemed relevant for an interim dataset at hand, the degree of flexibility of the model still has to be determined. For a detailed discussion of how many internal knots should be fit and how a spline model should be selected we refer to the corresponding paragraph in Section 3.3 of our literature review. In a nutshell, one internal knot may be a reasonable starting point and more than 3-4 knots will typically not be necessary (Rutherford et al., 2015). Note that a 0-knot model corresponds to a Weibull model (Royston and Parmar, 2002). The AIC may be used as a guidance, but a sensitivity analysis should be conducted by plotting the hazard function for a range of values for the number of internal knots (Rutherford et al., 2015). If the hazard function hardly changes with a higher number of internal knots, this is a sign of overfitting and a simple model should be preferred (Rutherford et al., 2015). In very large datasets ($n = 30,000$) the AIC was found to often select overly complex models, so in these cases the more stringent BIC should also be considered (Rutherford et al., 2015). With regards to the plausibility of the extrapolated part of the curve, it is recommended that, if available, external evidence (Rutherford et al., 2020) as well as clinical and biological plausibility (Latimer and Adler, 2022) are considered. Note that the censoring function can be modeled by a Royston-Parmar model, but it may also be modeled by a simpler model if the extra flexibility is not deemed necessary here. Friede et al. (2019), for example, suggested that while a Weibull model might be used to model the time-to-event process, a simpler exponential model might be used for the time-to-censoring process.

To carry out the BSSR for an event-driven trial based on PH splines models we propose to use the BSSR algorithm developed by Friede et al. (2019), which has been introduced in Section 2.3.2.

4.2 Hybrid spline-based blinded sample size reestimation

While the main focus of this research project were fully parametric spline models, we also considered a hybrid approach based on spline models. As pointed out in Section 3.2, Hybrid methods combine the non-parametric Kaplan-Meier estimate with a parametric model for the tail of the survival curve. Such an approach is appealing, because it makes use of the fact that an unbiased, non-parametric estimator of the survival function exists for the available follow up. Assume that the interim analysis is performed at some time point $t_{interim} < R$. This may for example be the time point at which half of the necessary number of events have been observed. The last time point at which an event has been observed when performing the interim analysis is denoted by t_*, where $t_* \leq t_{interim}$. Here we suggest to use the Kaplan-Meier estimator until the time point of the last event, t_*, and to extrapolate beyond that based on the PH spline model. So, if $\tilde{S}_{KM}(t)$ denotes the Kaplan-Meier estimate and $\tilde{S}(t)$ denotes the PH spline estimate of the survival function then our hybrid estimate, $\hat{S}(t)$ is defined as

$$\hat{S}(t) = \begin{cases} \tilde{S}_{KM}(t), & t \leq t_* \\ \tilde{S}(t), & t > t_*, \end{cases} \tag{19}$$

This approach is similar to that of Moeschberger and Klein (1985) and Ying et al. (2004), only that they extrapolated based on an exponential or Weibull model, rather than a spline model. The piecewise nature of spline models also makes this method akin to that of Fang and Su (2011), who proposed using a piecewise exponential model to extrapolate survival times beyond the Kaplan-Meier estimate. Indeed, since the extrapolated section of the PH spline model follows a local Weibull distribution it is comparable to using a piecewise Weibull model for extrapolation.

Note that the Kaplan-Meier estimate of the survival function and the spline model estimate are not necessarily joint at time point t_*. The $S(t)$ estimate of the PH spline model at time t_* may be higher or lower than than the Kaplan-Meier estimate. In their exponential hybrid approach, Moeschberger and Klein (1985) addressed this issue by carrying out a restricted maximum likelihood estimation that tied the exponential model estimate to the Kaplan-Meier estimate of $S(t)$ at time t_*. We decided to use the regular maximum likelihood estimate of the PH spline model, but implemented a simple measure to maintain monotonicity of the estimated survival function. When the extrapolated PH spline model estimate of $S(t)$ is higher than the Kaplan-Meier estimate at the last observed event, we simply set those extrapolated values equal to the last observed Kaplan-Meier estimate, $\tilde{S}_{KM}(t_*)$. Once the extrapolated PH spline estimates become smaller than $\tilde{S}_{KM}(t_*)$, the spline based estimates are used. An example of this will be given in Subsection 5.2.5.

The censoring distribution, $G(t)$ can be modeled analogously using the spline hybrid method. Then, for the hybrid splines method the integral from Equation (18) to compute the expected number of events becomes

$$\int\limits_0^{L-l} \hat{G}(t)d\hat{S}(t) = \begin{cases} \int\limits_0^{L-l} \hat{G}(t)d\tilde{S}_{KM}(t), & L-l \leq t_* \\ \int\limits_0^{t_*} \hat{G}(t)d\tilde{S}_{KM}(t) + \int\limits_{t_*}^{L-l} \hat{G}(t)d\tilde{S}(t) & L-l > t_* \end{cases} \tag{20}$$

That is, for patients whose time left in study $(L-l)$ is smaller than t_* the probability of experiencing the event can be calculated using only the non-parametric Kaplan-Meier estimate of the survival function, $\tilde{S}_{KM}(t)$. This might happen when the interim analysis is performed shortly before planned recruitment closure and when the planned follow-up period is not too long. For patients whose time left in the study is longer than that, extrapolation based on the PH spline model is necessary. Rewriting the above expression with an indicator function, the expected number of events in the hybrid spline framework becomes

$$\hat{E}(D) = -\sum_{l=1}^{R} r_l \int_0^{min(t_*, L-l)} \hat{G}(t)d\tilde{S}_{KM}(t) - \sum_{l=1}^{R} r_l I(t_* < L-l) \int_{t_*}^{L-l} \hat{G}(t)d\tilde{S}(t). \tag{21}$$

In our implementation the integrals were again computed using numerical integration using the same procedure as has been presented in Section 4.1. Using Formula (21) the BSSR is then carried out using the same algorithm as before. The hybrid spline model was briefly investigated as part of the simulation study in Section 5. Some observations regarding the performance of the hybrid spline BSSR approach will be presented in Section 5.2.5. However, our analysis will mainly focus on the performance of the fully parametric spline BSSR approach.

5 Simulation study

We carried out a simulation study to investigate the operating characteristics of our proposed flexible BSSR method in comparison with BSSR based on an exponential and Weibull model as well as a fixed sample size design. The simulation set up is based on the simulation study of Friede et al. (2019), who designed it to mirror an event-driven trial for secondary progressive multiple sclerosis (SPMS). This section is structured according to the clinical scenario evaluation (CSE) framework (Benda et al., 2010; Friede et al., 2010). The three components of this framework are: assumptions, design options, and metrics. The assumptions include for example recruitment pattern, hazard ratio and trial length. The design options describe the methodological designs that are investigated. Finally, the metrics are the operating characteristics that are used to compare the different designs.

5.1 Set up of the simulation study

Assumptions The setup of our event-driven trial simulation followed the simulation study of Friede et al. (2019), who based their simulation design on a clinical trial for secondary progressive multiple sclerosis (SPMS). We assumed that a total of 1530 patients would be recruited over an accrual period of 20 months. The treatment allocation was 2:1 for the treatment group relative to the control group. Recruitment was assumed to start slowly in the first month (9 patients) and to increase linearly every month up until month 10 (90 patients per month). Then, we assumed stable recruitment for months 11 to 15 (102 patients per month) and for months 16 to 20 (105 patients per month). If additional recruitment months were added as part of the BSSR algorithm, we assumed that 102 additional patients would be recruited each month. The event rates of the control group were chosen such that the probability of experiencing an event by 24 months was between 20% and 30%. This resulted in 11 different simulation scenarios (20%, 21%, ... 30%), which are illustrated in Figure 1. Proportional hazards were assumed and the hazard ratio for the treatment was 0.7. The probability of experiencing censoring by 24 months was fixed to be 20% for both treatment groups. Friede et al. (2019) considered independent exponentially distributed event and censoring times in their original simulation study. We extended this by also considering Weibull and Gompertz distributed event times. This resulted in 3 different simulation scenarios with regards to the data generating mechanism, which are summarised in Table 1.

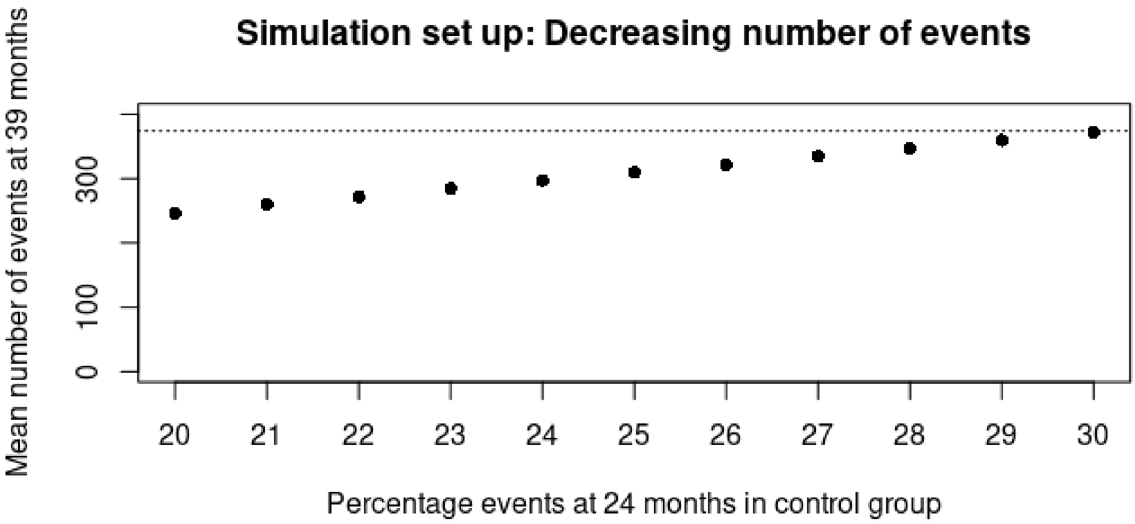

Figure 1: Mean number of events for the 11 different simulation scenarios depending the probability of experiencing an event (1,000 replications per scenario based on exponential survival times). As the 24 month event probability becomes lower than the 30% assumed at the planning stage, the number of events decreases below the 374 necessary events (dotted horizontal line).

The Weibull model and Gompertz model both belong to the class of proportional hazards models and include the exponential model as a special case (Collett, 2015, Ch. 5). The survival function of the Weibull model can be parameterized as $S(t) = e^{-\lambda t^{\gamma}}$ and the corresponding hazard function is $h(t) = \lambda\gamma t^{\gamma-1}$, where λ is known as the scale parameter and γ is known as the shape parameter. When $\gamma = 1$ the model reduces to an exponential model (Collett, 2015, Ch. 5). If $\gamma > 1$ the hazard increases monotonically and if $\gamma < 1$ it decreases monotonically . The survival function of the Gompertz model can be parameterized $S(t) = exp\{\frac{\lambda}{\theta}(1 - e^{\theta t})\}$ and the corresponding hazard function is $h(t) = \lambda e^{\theta t}$, where θ is known as the shape parameter (Collett, 2015, Ch. 5). When $\theta = 0$ the model reduces to an exponential model. If $\theta > 0$ the hazard increases monotonically and if $\theta < 0$ it decreases monotonically (Collett, 2015, Ch. 5).

In the Weibull and Gompertz simulation scenarios we obtained the event percentages of 20%, 21%, ..., 30% by varying the shape parameter of the respective distribution. The reason for varying the shape parameters was that we wanted to generate data that gradually moved away from the exponential scenario with constant hazards. This allowed us to investigate how an exponential model based BSSR performs in such a misspecification scenario, where the assumption of constant hazards is violated. Details on the specific parameter values and shapes of the simulated data in each of the 3 scenarios are provided in the respective results section.

Table 1: Overview of the simulation scenarios based on the exponential, Weibull and Gompertz distribution. In all BSSR simulations the planning assumptions of the event rate in the control group were 30% events at 24 months. The simulated data then violated this assumptions with event percentages gradually decreasing 30%, 29%, ..., 20%. If the percentage was lower than 30%, additional recruitment would be necessary to finish the trial on time.

Scenario	S(t)	G(t)	Simulated % events	Survival hazards
Exponential	Exponential	Exponential	20%-30%	Constant
Weibull	Weibull	Exponential	20%-30%	Increasing or Decreasing
Gompertz	Gompertz	Exponential	20%-30%	Increasing or Decreasing

As for the treatment effect, we simulated data under the assumed alternative scenario of $\theta = 0.7$ and the null hypothesis scenario of $\theta = 1$. Unlike Friede et al. (2019) we did not vary the hazard ratio for the treatment, since both Hade et al. (2010) and Friede et al. (2019) found that BSSR war robust to misspecification of the hazard ratio. Based on the Schoenfeldt formula we can compute that 374 events are necessary to obtain a power of 90% for a two-sided test at 5% significance level and assuming a hazard ratio of 0.7 (Friede et al., 2019). When the planning assumptions hold, 374 events are expected to occur within 39 months of follow-up, so this was set as the trial length. When the event probability at 24 month in the control group starts to be <30%, however, the trial is expected to take longer than 39 months and a design modification is necessary.

Design options A blinded review was carried out at Month 18, shortly before the planned end of the accrual period at Month 20. At that point, 1320 of the planned 1530 patients were already recruited into the study. We compared the following designs:

1. Fixed sample size design: No BSSR is carried out. The study is carried out using the sample size of $n = 1530$ calculated based on the planning assumption.

2. Exponential BSSR: Time-to-event and time-to-censoring are modeled based on independent exponential models for the BSSR at Month 18.

3. Weibull BSSR: Time-to-event is modeled based on a Weibull model and time-to-censoring is modeled based on an exponential model for the BSSR at Month 18.

4. Splines BSSR: Time-to-event and time-to-censoring are modeled based on independent Royston-Parmar PH spline models for the BSSR at Month 18.

5. Hybrid splines BSSR: Time-to-event and time-to-censoring are modeled based on independent hybrid Royston-Parmar PH spline models for the BSSR at Month 18 (results considered separately in Section 5.2.5).

Exponential BSSR and Weibull BSSR correspond to the methods presented by Friede et al. (2019). The splines BSSR methods correspond to the proposed methods in Section 4 of the current thesis. For all BSSR methods the maximum number of additional recruitment months was set to 6 months, which corresponds to a maximum of 612 additional patients. Additional recruitment was considered in discrete steps of 1 month with 102 patients being recruited each month.

With regards to spline BSSR, we had to determine the flexibility of the models in terms of the number of internal knots. Since the model fitting was carried out automatically in the simulations, a sensitivity analysis and careful inspection of the models as has been recommended in practice by Rutherford et al. (2015) was not possible. Instead, we had two options. First, we could carry out the model selection mechanically using information criteria (e.g. select the model with the lowest AIC), as it has been done by Gray et al. (2021) and Kearns et al. (2021). Second, we could specify a fixed number of internal knots beforehand, which would mean that the model has a certain level of default flexibility. We considered both options and report on our results regarding model selection in Subsection 5.2.1 of the results section.

Metrics As operating characteristics of the BSSR we considered the rejection probability (based on a log-rank test), the mean trial duration and the mean number of additional patients. The rejection probability corresponds to either the type I error rate or the power of the study, depending on whether the data were generated under the null or alternative hypothesis. In addition to these BSSR metrics, which stem from Friede et al. (2019), we also considered the mean relative bias in the estimated number of events for each of the BSSR methods. This is relevant for the current study, because the data generating mechanism may now be different from the BSSR model (e.g. exponential BSSR applied to Weibull data), so it is important to understand to what extent the number of events was on average over- or underestimated by a given model. The relative bias in each simulation run was calculated as

$$\frac{\widehat{E(D)} - E(D)}{E(D)}. \tag{22}$$

$E(D)$ is the true expected number of events. It was calculated with Equation (10) based on the true parameter values used for the data generation. $\widehat{E(D)}$ is the estimate of the expected number of events based on the parameter estimates of a given model (e.g. estimated exponential event rate in the exponential BSSR method). For the exponential and Weibull BSSR methods, it was also calculated based on Equation (10) and the event rates were split as proposed by Friede et al. (2019). For the spline BSSR methods, the pooled estimate of the survival function was used directly, as proposed in Equation (18). Finally, in addition to the mean estimates, we also carried out exploratory investigations of the distributions of above mentioned quantities in order to gain a more detailed understanding of the performance of our BSSR methods.

Software All simulations were carried out in R 4.0.2. (R Core Team, 2017). Survival data were generated using the 'simsurv' package (Brilleman et al., 2021). Exponential and Weibull models were fit using the 'survival' package (Therneau, 2020). Spline models were fit using the 'flexsurv' package (Jackson, 2016). The simulations were parallelised using the 'parallel' package (R Core Team, 2017).

5.2 Simulation results

5.2.1 Spline model selection

For the survival function we considered spline models with 1-3 internal knots, so that we could investigate the performance of flexible spline models in the Gompertz misspecification scenarios. The censoring functions was always based on an exponential model, so here we only considered moderately flexible spline models with one knot. Therefore, the following discussion on model selection and numbers of knots is referring to the modeling of the survival function.

We first wanted to investigate how well model selection based on the AIC performs in our simulation scenarios. First, we considered spline model selection in the case of exponentially distributed data. Initially, we considered only moderately flexible spline models with up to one internal knot. That is, the AIC selected between a 0-knot or 1-knot model. The PH spline model is a generalization of the Weibull model and in this simple exponential scenario a 0-knot (Weibull) model would be the most parsimonious model. For exponential data we found that, indeed, in most cases a 0-knot model was selected. However, in about 15% of cases a 1-knot spline model was selected, which can be due to overfitting of local deviations of the data.

Next we considered AIC based model selection for simulated Weibull data. Here, we also considered more flexible spline models with up to 3 internal knots. We found that while a 0-knot model was still the most frequently selected model (75.5%), 1-knot (11.3%), 2-knot (7.2%) and 3-knot (6%) models were also occasionally selected. Again, this was due to the spline models overfitting local deviations, which resulted in a marked increase in the variability of the spline-based estimates.

In contrast to the exponential and the Weibull scenario, in the Gompertz scenario a more complex spline model would be desirable, since a standard Weibull model cannot capture the strong increase or decrease in hazards observed in Gompertz distributed data. However, even in the scenario with the strongest decrease in hazards (smallest Gompertz shape parameter) the AIC selected a 1-knot spline model over a 0-knot spline model only in 28.3% of simulations. This is only a small difference to the 15% of 1-knot models that were already selected in the constant hazards (exponential) scenario. The model selection was even worse in the increasing hazards scenarios, where for the strongest increase in hazard rates (largest Gompertz shape parameter) only in 18.6% of cases 1-knot model was selected over a 0-knot model. The model selection did not improve substantially when more complex spline models with up to 3 internal knots were available. In those cases, still in about 60-70% of cases a 0-knot model was selected. Clearly, the Gompertz data did not sufficiently deviate from the Weibull model (which also has monotonous hazards) for the AIC to select a spline model over a Weibull (0-knot) model. This is problematic, however, since spline models tended to perform better than Weibull models in terms of capturing the strong increase or decrease in hazards observed in the Gompertz scenarios (see Subsection 5.2.4).

We carried out a sensitivity analysis to see if the model selection based on the AIC improves if the interim analysis is carried out at a later point in time, when more events have been observed. We extended the recruitment period by 5 months and also performed the interim analysis 5 months later, at month 23. This resulted in a larger number of interim events, which was hypothesized to improve model selection. Indeed, some improvements were found. In the most extreme decreasing hazards scenario the 1-knot spline model was now selected in 39.8% of cases, compared to previous 28.3%. In the most extreme increasing hazards scenario, the AIC selected the 1-knot spline model in 30.6% of cases, compared to the previous 18.3% of cases. This indicates that the violations of the Weibull assumptions are more likely to be recognized by the AIC as we observe more interim events, which results in a more frequent selection of the 1-knot spline model. When we allowed for the selection of more complex spline models (2-3 internal knots), the results were similar. In the majority of cases a 0-knot spline model was still chosen by the AIC. If a spline model was chosen, it was mostly the 1-knot spline model and the 2- and 3-knot spline models were selected with decreasing frequency. Overall, even though the AIC based model selection somewhat improved with a later interim analysis the results the results are still unsatisfactory. In about 60-70% of cases the AIC still selected a 0-knot model.

Based on these preliminary findings, we concluded that the AIC alone cannot be relied on to the select the most suitable model in the Gompertz misspecification scenario (where a more complex model would be desirable). While overfitting can be a problem in the exponential and Weibull scenarios, the spline model has little use if in the majority of cases a 0-knot Weibull model is selected, even when a more complex model would be more suitable. Therefore, we considered the alternative approach of selecting a fixed number of knots by default without performing any model selection. This means that a certain level of default flexibility of the spline model is specified, which allows to capture distributions that deviate from a Weibull model. However, a more flexible model is also more prone to overfitting and so the variance of the model estimates should be considered carefully. In the results that follow, we fit spline models with the number of knots fixed to either 1, 2 or 3 knots and compared their performance to that of exponential and Weibull models.

5.2.2 Exponential data

For illustration, the modeling of the survival function for one simulation iteration is shown in Figure 2. As we can see, the various BSSR models differ slightly in terms of their extrapolated survival. However, all BSSR models correctly predicted that fewer events than anticipated in the planning assumptions (black dotted lines) would occur in this example. The extrapolated survival function was used in the BSSR algorithm of the respective design to decide whether additional recruitment would be necessary.

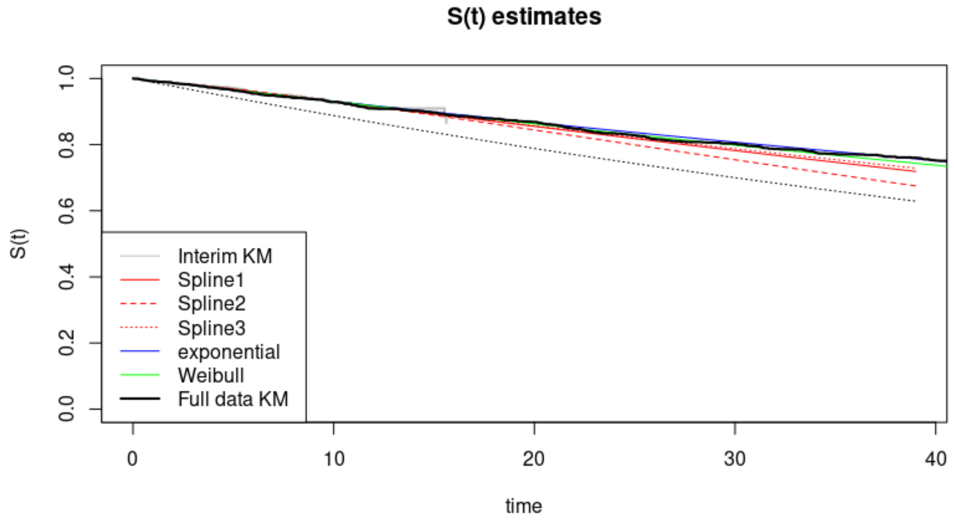

Figure 2: Example of the modeling of the survival function in one simulation run. In this example, data were generated from an exponential distribution with 20% events at 24 months. The fixed design used the planning assumptions of exponential survival with 30% events at 24 months (black dotted line). The different parametric models were fit based on the data available at the interim analysis at Month 18, illustrated by the grey Kaplan-Meier curve. Spline models were fit with 1, 2 and 3 knots (denoted by Spline1, Spline2 and Spline3, respectively).

The different simulation scenarios based on the exponential model are depicted in Figure 3. The exponential event rate in the simulation scenarios ranged from 0.0093 (20% events at 24 months) to 0.0149 (30% events at 24 months).

The mean relative bias in the estimated number of events at trial end based on the different BSSR models is depicted in Figure 4. Across the exponential scenarios with varying event probabilities at 24 months, no bias is apparent for the true exponential model and for the 1-knot spline model. The Weibull model seems to have a very small overestimation, but this is presumably due to sampling error. The spline models with 2 and 3 knots have a small but consistent overestimation of up to 2.5%.

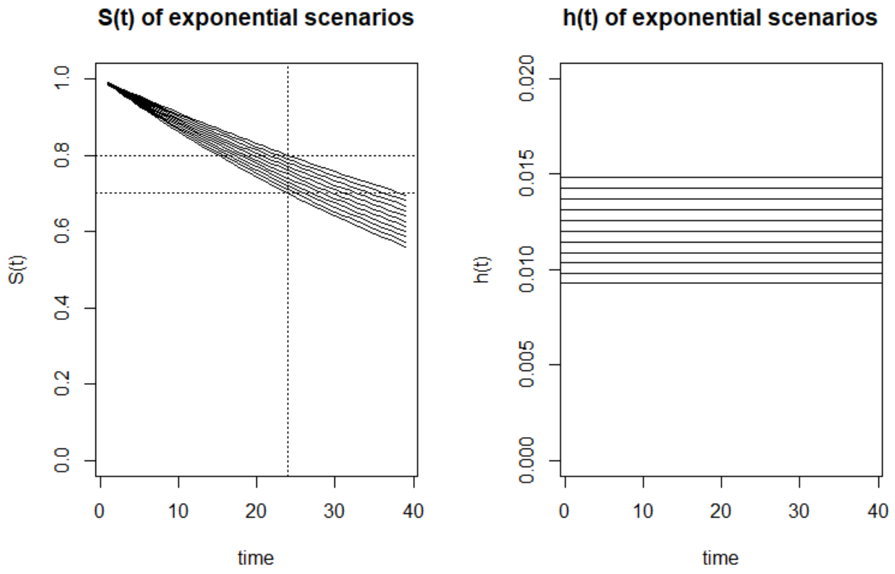

Figure 3: Survival and hazard functions of the simulated exponential scenarios. The vertical dotted line indicates the reference month 24, at which the probability of experiencing the event is between 20% and 30% (horizontal dotted lines).

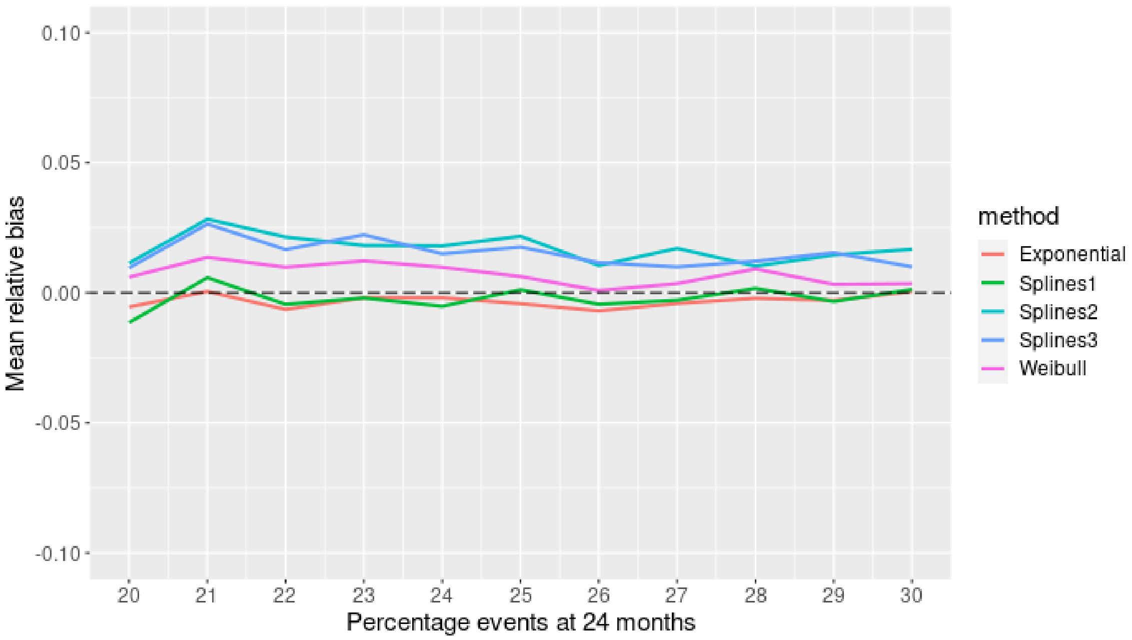

Figure 4: Mean relative bias in the estimated number of events at trial end in the simulated exponential scenarios based on the BSSR models (1,000 replications per scenario). Splines models were fit with 1, 2 and 3 knots (denoted by Splines1, Splines2 and Splines3, respectively). Note that no sample size reestimation was carried out in the fixed design. The expected number of events in the simulation scenarios ranged from 246.8 (20% events) to 372.3 (30% events).

To better understand the performance of the different methods we can consider boxplots of the iteration-wise relative bias for the 20% events scenario in Figure 5. As we can see, the variance of the estimates across the 1,000 replications clearly increases as we fit more complex models. Moreover, overestimation of the number of events seems to be as we fit increasingly complex models. For the 2- and 3-knot models there are very large overestimations of more than 100% in the most extreme cases. This occurs due to complex spline models overfitting observed downwards trends in the interim data, which are a mere artifact of sampling error. The extreme outliers are only observed for overestimation, but not for underestimation. This is presumably a feature of our simulation set up. Since only 20% to 30% of patients typically experience an event in the first 24 months, extrapolation at month 18 has more potential to underestimate survival than to overestimate it.

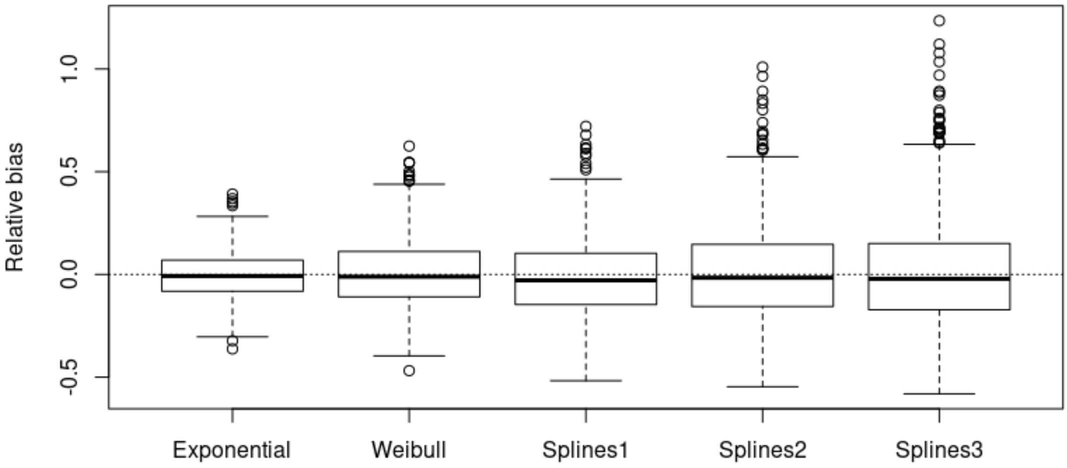

Figure 5: Boxplots of the relative bias in the estimated number of events at trial end in the simulated exponential scenario with 20% events at 24 months (1,000 replications).

After considering the bias in the expected number of events it is important to inspect how this modeling performance translates into design modifications and trial lengths via the BSSR algorithm. Figure 6 shows how many additional patients were recruited on average and Figure 7 shows how long the trial took to finish on average based on the different designs. The mean number of additional patients is relatively similar between the models. However, as the percentage of events at 24 months decreases, the more

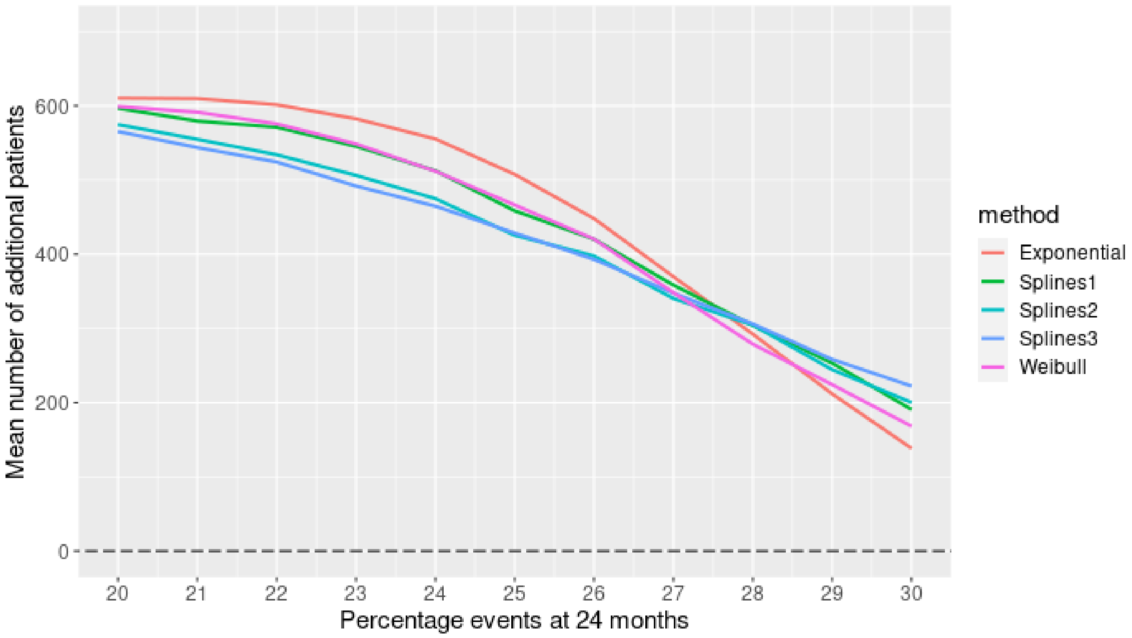

Figure 6: Mean number of additional patients in the exponential scenarios added by the different BSSR models (1,000 replications per scenario). The maximum number of patients that could be added was 612. Note that no patients were added in the fixed design.

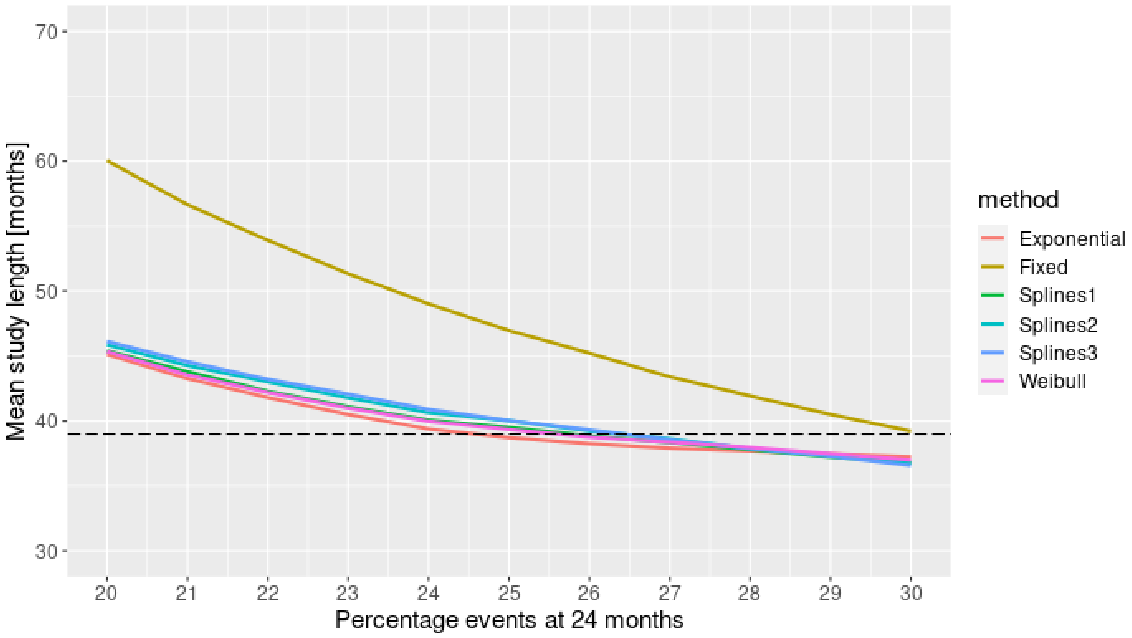

Figure 7: Mean study lengths in the exponential scenarios based on the BSSR models and the fixed design (1,000 replications per scenario). The trial finished once 374 events were observed and the goal was to finish in 39 months (black dotted line).

complex models tended to recruit fewer patients on average. Since the more complex models tended to have more occasions of overestimating the number of events (see Figure 5 for example), the respective BSSR algorithm added fewer patients on these occasions.

With regards to the trial length, this resulted in marginally increased mean trial durations for the more complex designs (see Figure 7). Overall, however, the mean study lengths are very similar between all the BSSR models. In the most extreme misspecification scenario with 20% events at 24 months, all BSSR methods took on average around 45-46 months. Note that all designs took longer than the goal of 39 months, including the exponential BSSR. This is because the number of additional patients was limited to 612. As can be seen in Figure 6, the exponential BSSR added 612 patients essentially every time in the more extreme misspecification scenarios. The small differences between the BSSR models seem particularly negligible in comparison with the vastly increased average trial duration of the fixed design. With 20% events at 24 months the fixed design took on average about 60 months to finish. Therefore, all BSSR methods finished far sooner than the fixed sample size design when the planning assumptions were wrong.

While we do not see large differences in trial lengths between the BSSR models in terms of the average duration, it is worthwhile to compare the distribution of the study lengths. Figure 8 shows boxplots of the trials lengths of the different design options in the exponential scenario with 20% events at 24 months. The Weibull and 1-knot spline model have a small number of outliers with an increased trial length, but this effect becomes more pronounced in the complex 2- and 3-knot spline models. In this scenario, the number of simulated trials that took longer than 50 months was 9 (exponential), 23 (Weibull), 27 (Splines1), 61 (Splines2) and 80 (Splines3). In the fixed design, all simulated trials in this scenario took more than 50 months to finish. The difference between the BSSR methods is related to the outliers in the estimated number of events, which were shown in Figure 5. Since with the more complex spline models a large overestimation happened more frequently, these trials more frequently added too few (or even no) patients. This, in turn, resulted these markedly increased trial lengths observed occasionally. So, while the average trial lengths are similar, the risk of overfitting and consequently increased trial lengths is higher for the more complex spline models. The 1-knot spline model, however, seemed to perform similarly to the Weibull model.

Finally, we also considered the type I error rate when simulating data under the null hypothesis (see Table 2). For all 6 designs the type I error rates are close to the

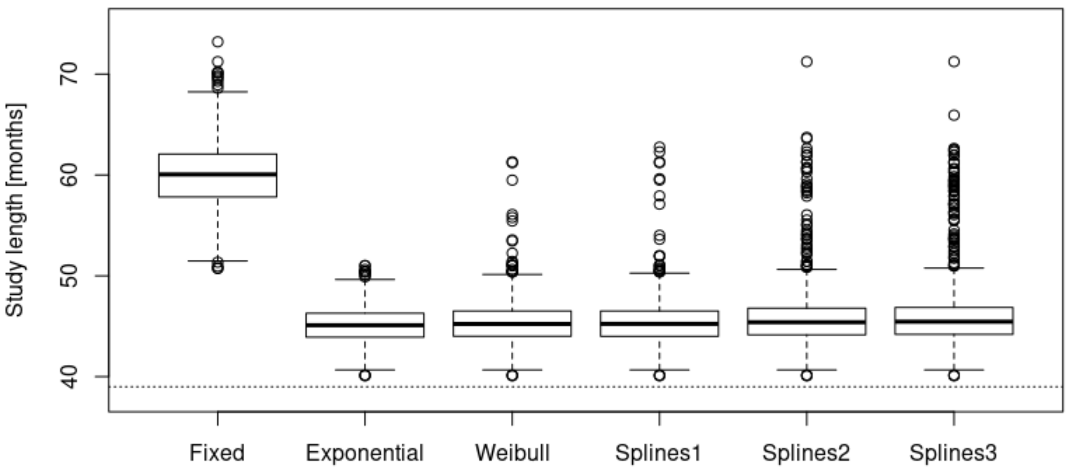

Figure 8: Boxplots of the study lengths of the different designs in the simulated exponential scenario with 20% events at 24 months (1,000 replications).

Table 2: Simulated type I error rates summarized across the 11 exponential simulation scenarios with event probabilities from 20% to 30% (10,000 simulations per scenario). The nominal significance level is 0.05 (two-sided). SD is the standard deviation of the rejection probabilities across the 11 scenarios.

Method	Mean	SD	Minimum	Maximum
Fixed Design	0.0499	0.0028	0.0452	0.0537
Exponential BSSR	0.0505	0.0028	0.0455	0.0536
Weibull BSSR	0.0504	0.0026	0.0451	0.0535
Splines BSSR with 1 knot	0.0501	0.0024	0.0463	0.0540
Splines BSSR with 2 knots	0.0500	0.0023	0.0459	0.0536
Splines BSSR with 3 knots	0.0502	0.0023	0.0471	0.0530

nominal significance level of 0.05 and deviations seem to be within the simulation error (10,000 replications per scenario). The mean rejection probability was very close to 0.005, spanning a range from 0.0499 to 0.0505. The deviations from this occurred in both direction, ranging from 0.0459 to 0.0540. The standard deviations of the rejection probability across the 11 scenarios ranged from 0.0023 to 0.0028. These results look similar to those of Friede et al. (2019), only that in our case the deviations are slightly larger. In their original simulation (also based on exponential data) the standard deviation was between 0.0004 and 0.0008 (Friede et al., 2019). This difference is likely due to the larger number of replications that was used in their simulation study. To investigate the type I error rate they used 100,000 replications per scenario as opposed to the 10,000 replications used in our simulation. To get a more detailed understanding of the influence of the number of replications on the variance of our estimates, the Monte Carlo error could be considered (Koehler et al., 2009). The Monte Carlo error is the standard deviation of the Monte Carlo estimator (in this case the rejection probability) and it can be estimated for example based on a jackknife or bootstrap procedure (Koehler et al., 2009). However, an extensive analysis of this simulation error was beyond the scope of the current thesis. Overall, there we found no indication of an increased type I error rate for the BSSR methods in the simulation scenarios that were considered here.

To summarize the results of the exponential simulation, we found that in general all BSSR methods performed similarly well and were mostly unbiased. However, the 2- and 3-knot spline models had a markedly increased variance and more outliers due to overfitting, which negatively affected the BSSR performance in some simulation runs.

5.2.3 Weibull data

The different simulation scenarios based on the Weibull distribution are depicted in Figure 9 and Figure 10. Note that the hazard of a Weibull distribution can be either monotonically decreasing or increasing. We were interested in both scenarios and how the misspecification of constant hazards would affect the performance of the exponential BSSR. In the decreasing hazards scenario (Figure 9) an exponential distribution with 30% events at 24 months is chosen as a reference point (this coincides with planning assumptions). Then, the Weibull shape parameter is gradually decreased until the percentage of events at 24 months is 20%. That is, in the shape parameter starts at 1 (30% events) and decreases down to 0.852 (20% events). It is expected that here the exponential BSSR will overestimate the number of events in the scenarios with a smaller shape parameter. This might lead to fewer patients being recruited, which could result in increased trial lengths.

In the increasing hazards scenario (Figure 10) an exponential distribution with 20% events at 24 months is chosen as a reference point (note that this is not the exponential distribution from the planning assumptions, since there 30% events is assumed). Then,

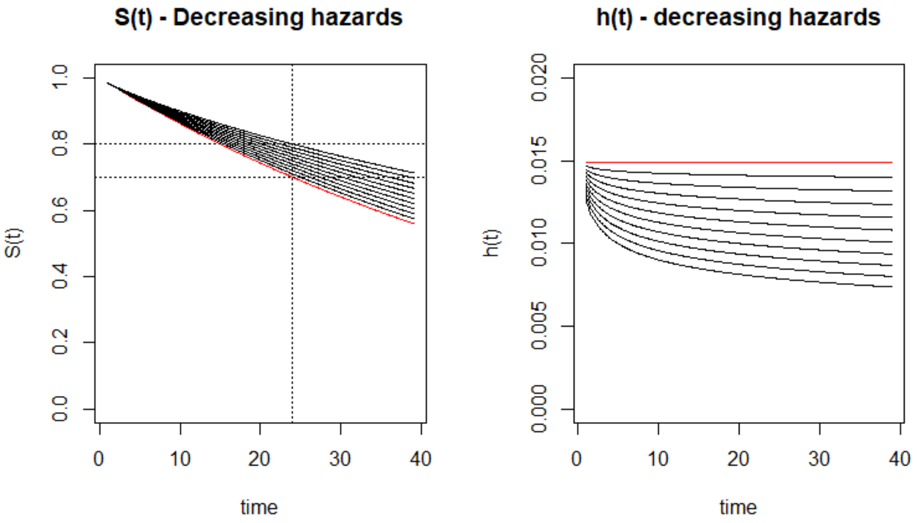

Figure 9: Survival and hazard functions of the simulated Weibull scenarios with decreasing hazards. The red line indicates the exponential distribution with 30% events at 24 months, which is chosen as reference point (Weibull shape parameter equal to 1). From, there the Weibull shape parameter is gradually decreased until there are only 20% events at 24 months. The resulting decreasing hazards are shown in the right panel.

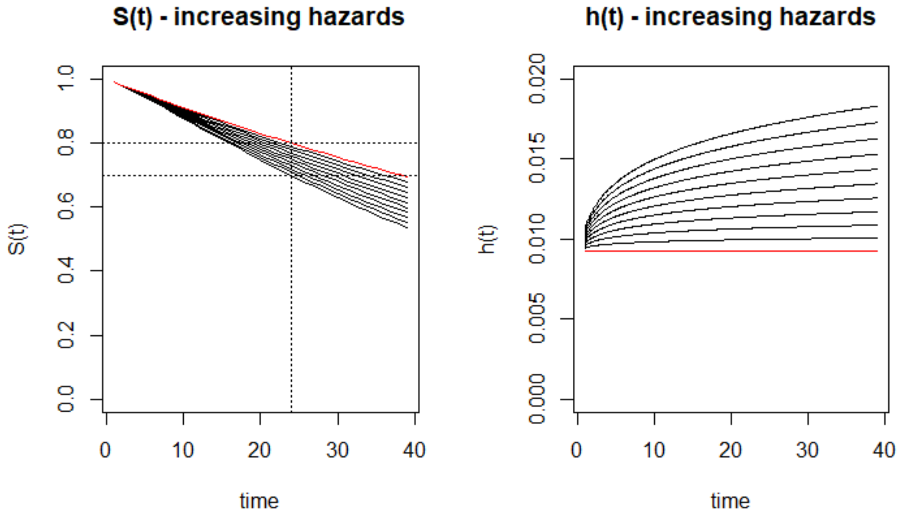

Figure 10: Survival and hazard functions of the simulated Weibull scenarios with increasing hazards. The red line indicates the exponential distribution with 20% events at 24 months, which is chosen as reference point (Weibull shape parameter equal to 1). From there, the Weibull shape parameter is gradually increased until there are 30% events at 24 months. The resulting increasing hazards are shown in the right panel.

the Weibull shape parameter is gradually increased until the percentage of events at 24 months is 30%. That is, in the shape parameter starts at 1 (20% events) and increases to 1.148 (30% events). It is expected that here the exponential BSSR will underestimate the number of events in the scenarios with a larger shape parameter. This might lead to too many patients being added, which could result in unnecessary additional costs for the trial sponsors.

Weibull decreasing hazards The mean relative bias in the estimated number of events at trial end in the decreasing hazards scenarios is shown in Figure 11. The Weibull model and all of the spline models performed similarly to previously in the exponential simulation. Again, the Weibull and 1-knot spline model both seem unbiased, while there is some very minor overestimation in the more complex spline models. As anticipated, however, the exponential model becomes increasingly biased, as the simulated data gradually depart from the constant hazards assumption (at 30% events) toward decreasing hazards (29%, 28%, ..., 20%). In the scenario with the most strongly decreasing hazards (20% events) the exponential model on average overestimated the number of events at trial end by 15.6%. The distributions of the relative bias of the estimates for this scenario are shown as boxplots in Figure 12. Again, we see that for the 2- and 3-knot spline models there is an increase in outliers that overestimated the number of events at trial end. Moreover, we can see how the exponential BSSR systematically overestimated the number of events in this misspecification scenario.

The resulting trial characteristics can be found in Figure 13 (mean additional patients) and Figure 14 (mean trial length). As before, the Weibull and 1-knot model are nearly indistinguishable in terms of their mean number of additional patients and their mean study length. Again, the 2- and 3-knot spline models tended to on average add slightly fewer patients as the percentage of events decreases. This is again due to the occasional rather extreme outliers in terms of overestimation of the expected number of events, which was caused by overfitting. Due to these outliers, even in the 20% events scenario, there were a number of iterations in which the 2- and 3- knot spline models did not add any patients.

The exponential model added on average fewer patients than all other models in most of the simulation scenarios, though the difference towards the complex spline models diminishes towards to the lowest event percentages. The exponential model typically

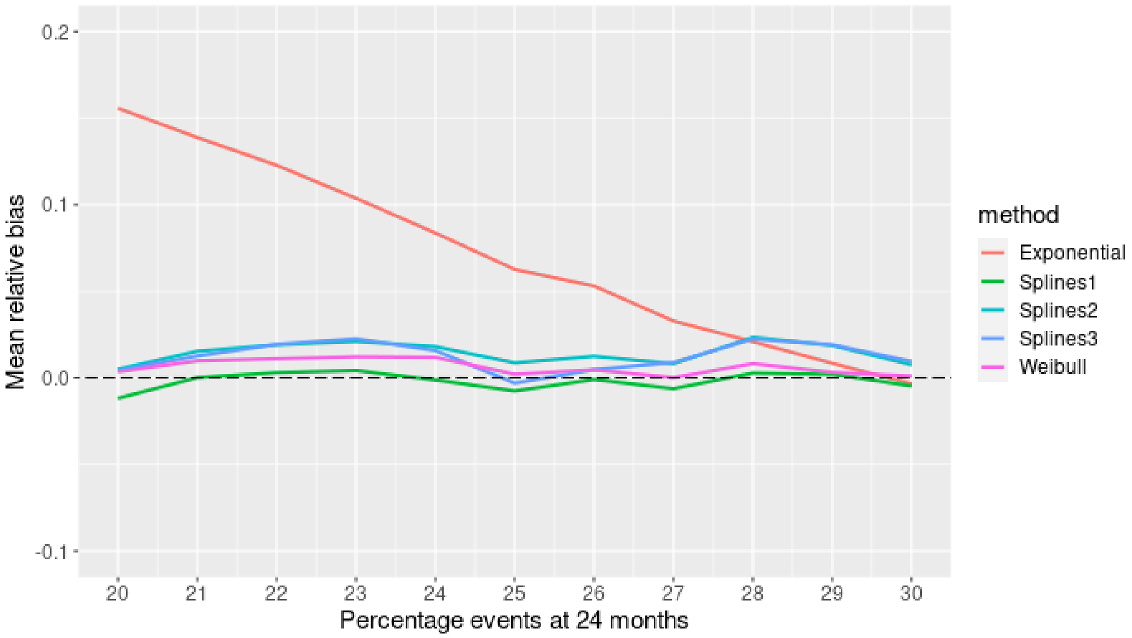

Figure 11: Mean relative bias in the estimated number of events at trial end in the simulated decreasing hazards Weibull scenarios based on the BSSR models (1,000 replications per scenario). Splines models were fit with 1, 2 and 3 knots (denoted by Splines1, Splines2 and Splines3, respectively). Note that no sample size reestimation was carried out in the fixed design. The expected number of events in the simulation scenarios ranged from 245.1 (20% events) to 372.3 (30% events).

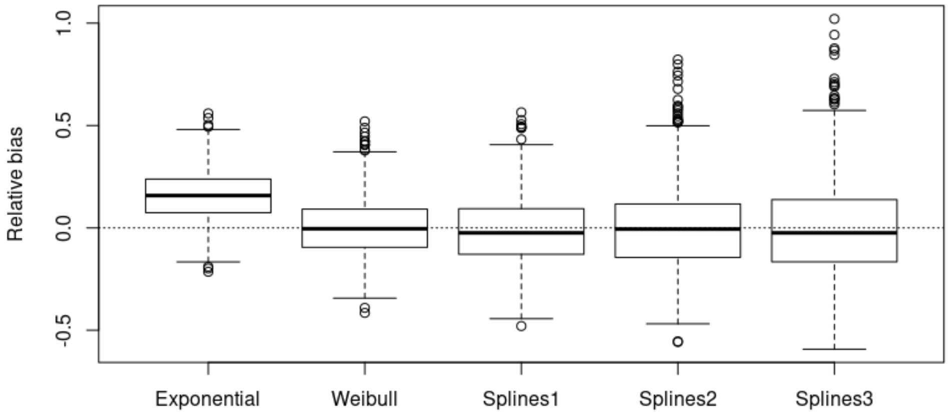

Figure 12: Boxplots of the relative bias in the estimated number of events at trial end in the simulated decreasing hazards Weibull scenario with 20% events at 24 months (1,000 replications). The Weibull shape parameter in the simulation scenario was 0.852.

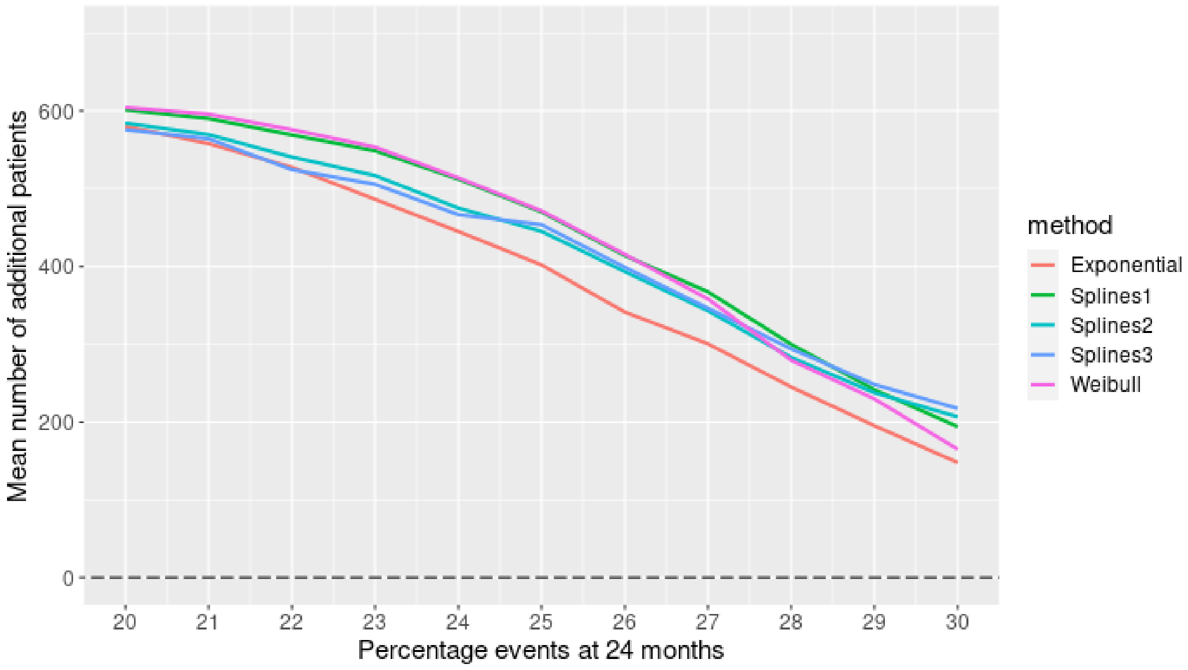

Figure 13: Mean number of additional patients in the Weibull scenarios with decreasing hazards added by the different BSSR models (1,000 replications per scenario). The maximum number of patients that could be added was 612. Note that no patients were added in the fixed design.

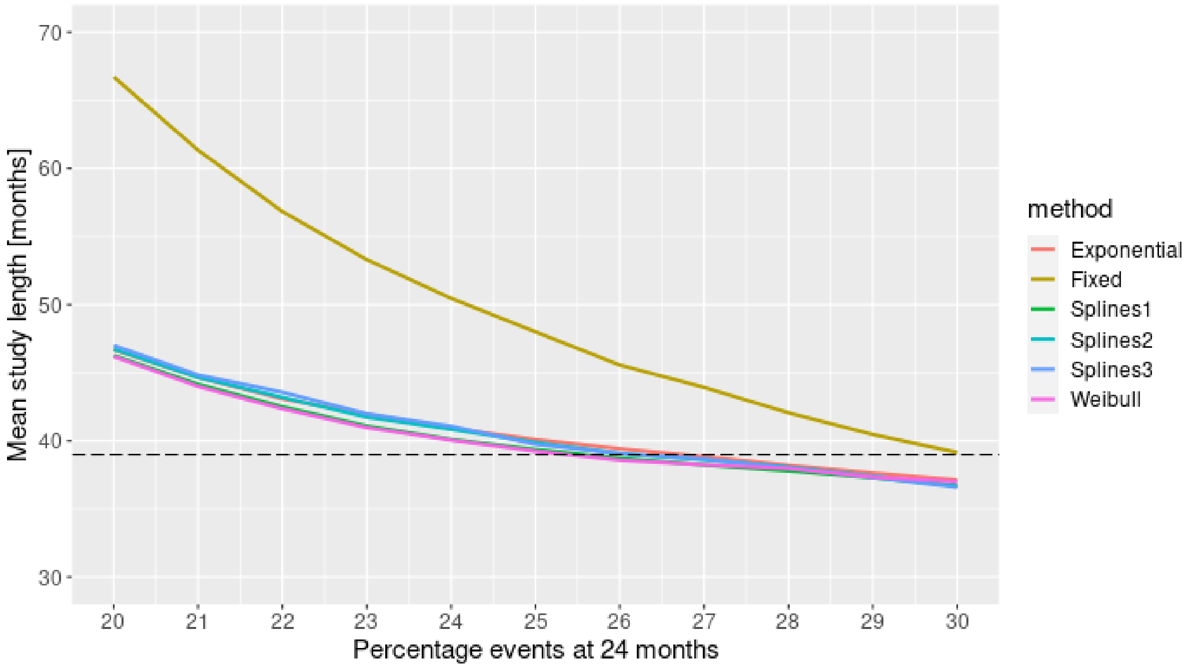

Figure 14: Mean study lengths in the Weibull scenarios with decreasing hazards based on the BSSR models and the fixed design (1,000 replications per scenario). The trial finished once 374 events were observed and the goal was to finish in 39 months (black dotted line).

added fewer patients due to the overestimation that was observed in Figure 11. In terms of average trial length, this resulted in a similar performance to the complex spline models, which tended to take minimally longer than Weibull and 1-knot spline models. However, this difference is again very small, typically about 1 month or less. That is, the biased estimates of the exponential model did not lead to markedly worsened trial outcomes, as the exponential BSSR still added sufficiently many patients.

Figure 15 shows the boxplots of the trial lengths of the 20% events scenario and we can see that all BSSR methods appear similar. Again, some more outliers are observed for the 2- and 3-knot spline models. The number of trials that took longer than 50 months to complete in this scenario are 90 (exponential), 51 (Weibull), 55 (Spline1), 91 (Spline2) and 107 (Spline3). Again, all trials in the fixed design took more than 50 months in this scenario. On average the fixed design took on 66.7 months to finish in this scenario. For practical purposes, all BSSR methods performed very similar in terms of reducing the trial duration. However, there was a slightly larger risk of prolonged trials for the complex spline models due to overfitting and occasionally for the exponential model due to biased estimates.

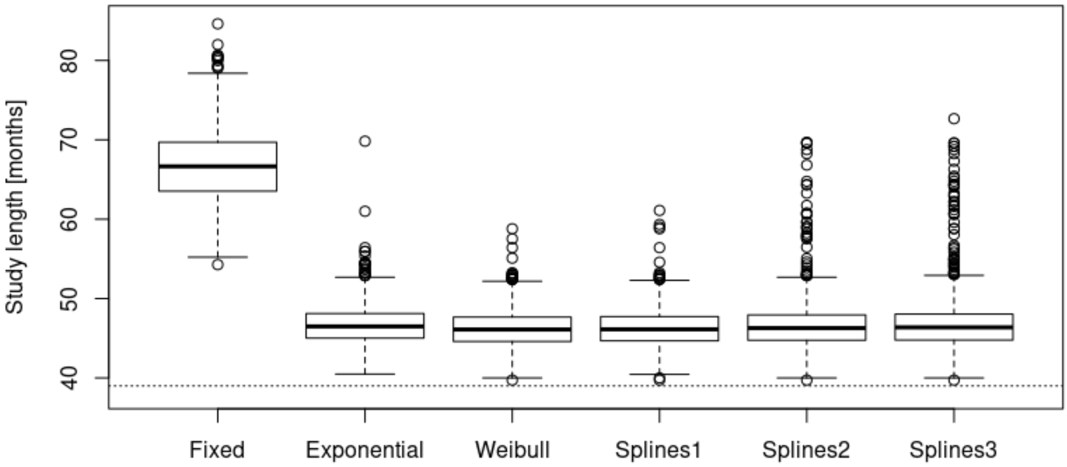

Figure 15: Boxplots of the study lengths of the different designs in the simulated decreasing hazards Weibull scenario with 20% events at 24 months (1,000 replications). The Weibull shape parameter in the simulation scenario was 0.852. The trial finished once 374 events were observed and the goal was to finish in 39 months (black dotted line).

With regards to the type I error rate, we again found no indication of an increased rejection probability for the BSSR designs when data were simulated under the null hypothesis (10,000 replications per scenario). The table of the simulated type I error rates for the Weibull scenarios can be found in Appendix A.

Weibull increasing hazards The mean relative bias in the estimated number of events at trial end in the increasing hazards scenarios is shown in Figure 16. For the Weibull model and the spline models the results are essentially the same as have been reported in the previous section on the decreasing hazards scenarios. As anticipated, however, the exponential model becomes increasingly biased as we move away from the constant hazards assumption (at 20% events) towards increasing hazards (21%, 22%, ..., 30%). In the scenario with the most strongly increasing hazards (30% events) the exponential model underestimated the number of events on average by 13.7%. The distributions of the relative bias of the estimates for this scenario are shown as boxplots in Figure 17. The results look similar to the boxplots in the decreasing hazards scenario except that the distribution of the bias of the exponential method now shows a negative, rather than positive bias.

Since we observed an underestimation here the effect on the BSSR and the trial length should be the opposite of what was found in the decreasing hazards scenario. The mean number of additional patients is depicted in Figure 18 and the mean trial length in Figure 19. Figure 18 shows that the exponential BSSR on average added many more patients than all the other BSSR methods. This difference becomes more pronounced the more the distribution of the simulated data deviated from an exponential model. In the most extreme scenario (30% events, Weibull shape parameter of 1.148) the exponential model added on average 428.9 Patients, whereas all other BSSR models added less than 250 patients on average. This is highly problematic for the exponential model, because in the 30% events scenario no additional patients are needed. The addition of some patients is expected, since even unbiased models will occasionally estimate a lower number of events at trial end due to sampling error. However, because the exponential model was biased in its estimation of the expected number of events it systematically added too many patients. In practice, this would results in a large increase of costs for the study sponsor, which are not justified.

With regards to the trial length (Figure 19), we can see that as a consequence of this

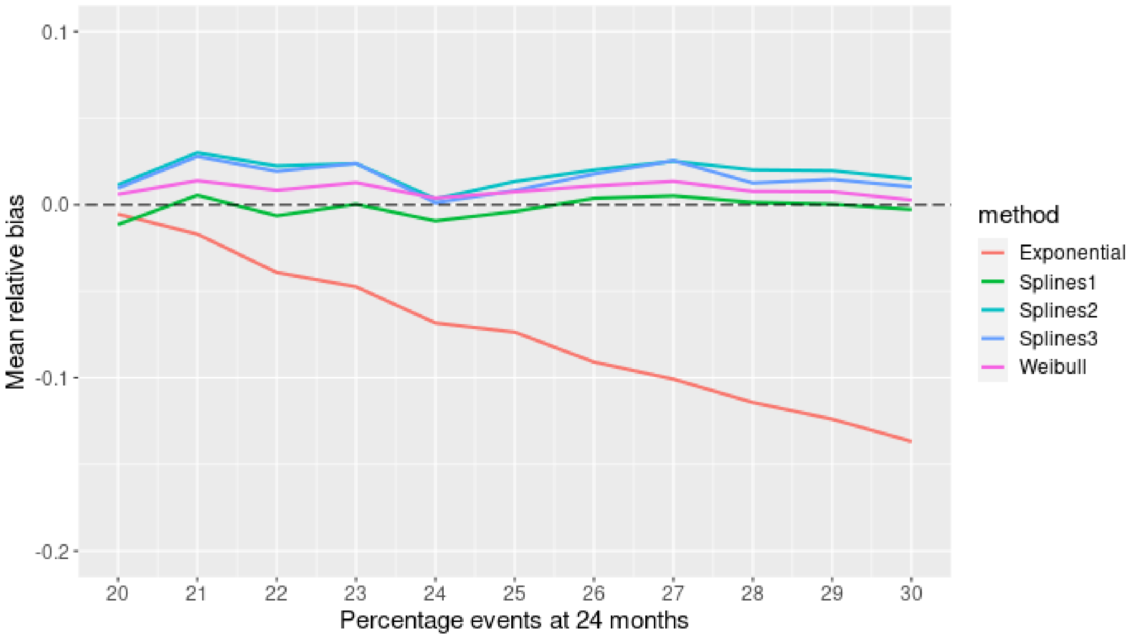

Figure 16: Mean relative bias in the estimated number of events at trial end in the simulated increasing hazards Weibull scenarios based on the BSSR models (1,000 replications per scenario). Splines models were fit with 1, 2 and 3 knots (denoted by Splines1, Splines2 and Splines3, respectively). Note that no sample size reestimation was carried out in the fixed design. The expected number of events in the simulation scenarios ranged from 246.8 (20% events) to 374.9 (30% events).

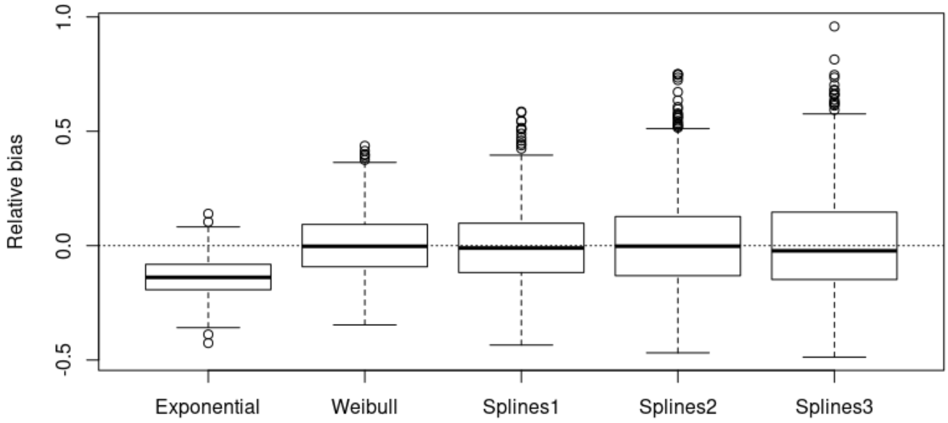

Figure 17: Boxplots of the relative bias in the estimated number of events at trial end in the simulated increasing hazards Weibull scenario with 30% events at 24 months (1,000 replications). The Weibull shape parameter in the simulation scenario was 1.148.

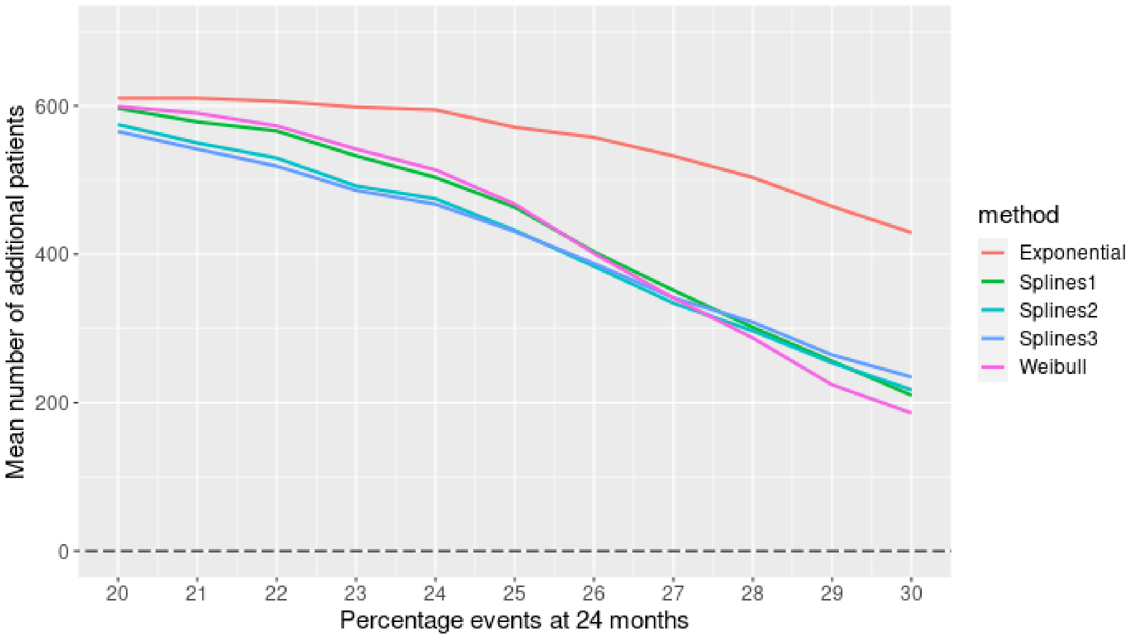

Figure 18: Mean number of additional patients in the Weibull scenarios with increasing hazards added by the different BSSR models (1,000 replications per scenario). The maximum number of patients that could be added was 612. Note that no patients were added in the fixed design.

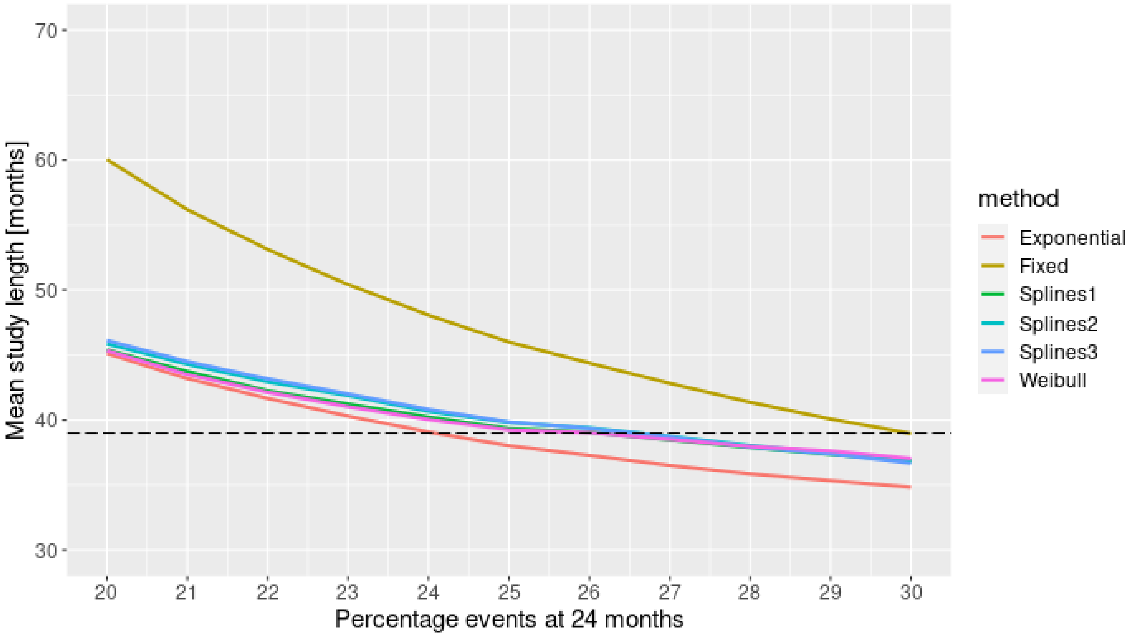

Figure 19: Mean study lengths in the Weibull scenarios with increasing hazards based on the BSSR models and the fixed design (1,000 replications per scenario). The trial finished once 374 events were observed. The trial finished once 374 events were observed and the goal was to finish in 39 months (black dotted line).

excess recruitment the exponential BSSR method finished on average after 34.8 months in the 30% events scenario. That means it finished on average 5 months early, as opposed to the roughly 2 months that the other, unbiased methods finished early. This illustrates that this extreme recruitment of the additional patients was not necessary based on the underlying data-generating mechanism.

To present a more detailed overview of how the patterns of additional recruitment differed between the BSSR methods, a barchart of the distribution of recruitment numbers for the 30% event scenario is shown in Figure 20. Note that additional recruitment occurred in terms of discrete steps of one additional month of recruitment with 102 patients each. Therefore, additional recruitment proceeded in steps of 102, 204, ..., 612, depending on how much lower the estimated number of events based on the interim data was compared to the necessary number of events (see BSSR algorithm in Section 2.3). In the barchart we can see that all of the relatively unbiased methods primarily added no additional patients. Due to occasional outliers caused by overfitting 2- and 3 knot spline models display a moderate increase in iterations where 612 patients were added. However, they still added 0 patients in most cases. In contrast to this, the exponential BSSR added 612 patients the most frequently due its systematic underestimation of the expected number of events.

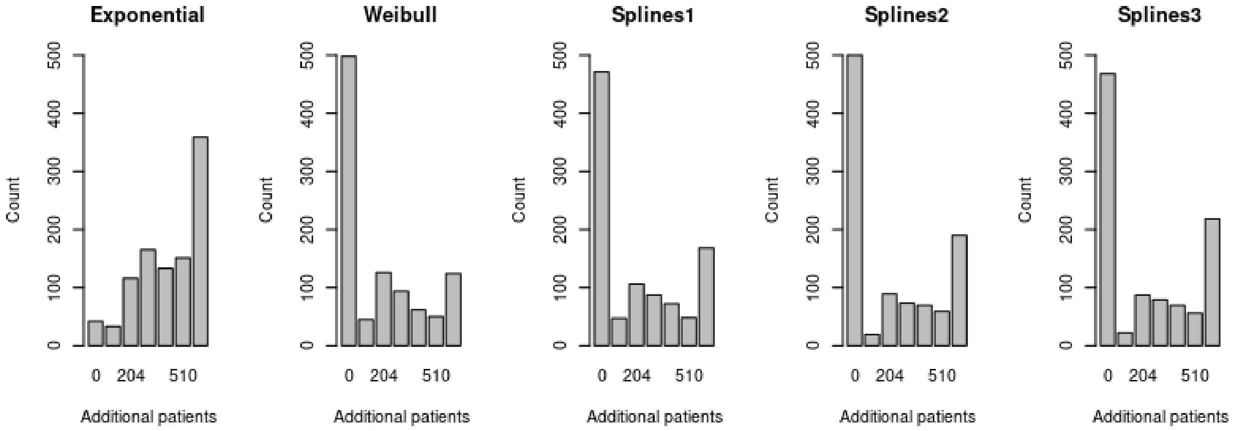

Figure 20: This barchart counts how frequently 0, 102, 204, ..., 612 patients were added by the different designs in the increasing hazards Weibull scenario with 30% events at 24 months (1,000 replications). This corresponded to 0, 1, 2, ..., 6 months of additional recruitment, respectively. The Weibull shape parameter in the simulation scenario was 1.148.

To summarize the results of the Weibull simulations, we found that in the decreasing hazards scenarios the exponential BSSR tended to overestimate the number of events. However, in our simulation this did not dramatically affect the BSSR performance and the average trial duration, since it was still comparable to that of the other, unbiased BSSR methods. In the increasing hazards scenario the exponential BSSR tended to underestimate the number of events. In this case, it had a large effect on the BSSR performance, since a larger number of additional patients was recruited when this was not justified. This shows that, when the assumptions of constant hazards is violated, using an exponential model for BSSR can lead to unnecessary costs for the trial sponsors. Complex 2- and 3-knot spline models were generally unbiased, but their variance was increased and overfitting of local deviations led to a large bias in individual simulation runs. In these individual occasions, this could lead to additional costs (when too many patients are added) or an increased trial duration (when too few patients are added). In all Weibull simulations the 1-knot spline model seemed almost indistinguishable from the true Weibull model in its BSSR performance. This is promising, since the added flexibility of this model did not seem to have any relevant drawbacks in the BSSR scenarios considered here.

5.2.4 Gompertz data

Like the Weibull distribution, the Gompertz distribution can either be monotonically increasing or decreasing. Again, we considered both situations and simulated the data analogously to the simulation of the Weibull data described in Section 5.2.3. The decreasing hazards Gompertz scenarios are depicted in Figure 21. Here an exponential distribution with 30% events at 24 months is chosen as a reference point and the Gompertz shape parameter is then gradually decreased until the percentage of events is 20%. That is, the shape parameter starts at 0 (30% events) and decreases down to -0.427 (20% events). Note that the hazard decreases further in the Gompertz scenarios compared to the Weibull scenarios. We expected that in this setting both the exponential and the Weibull BSSR would overestimate the number of events. The spline BSSR was anticipated to be more robust than the Weibull BSSR.

The increasing hazards Gompertz scenarios are depicted in Figure 22. In this setting an exponential distribution with 20% events at 24 months is chosen as a reference point and the Gompertz shape parameter is then gradually increased until the percentage of

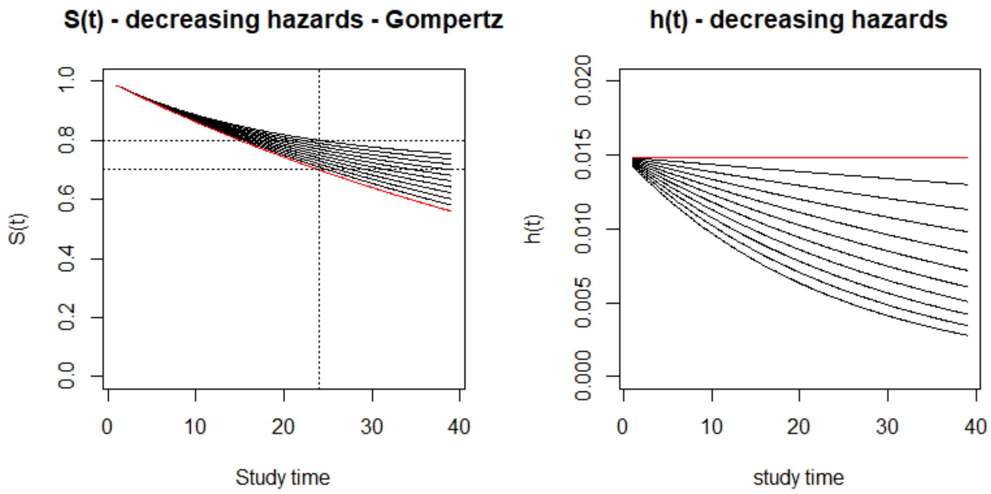

Figure 21: Survival and hazard functions of the simulated Gompertz scenarios with decreasing hazards. The red line indicates the exponential distribution with 30% events at 24 months, which is chosen as reference point (Gompertz shape parameter equal to 0). From there the Gompertz shape parameter is gradually decreased until there is only 20% events at 24 months. The resulting decreasing hazards are shown in the right panel.

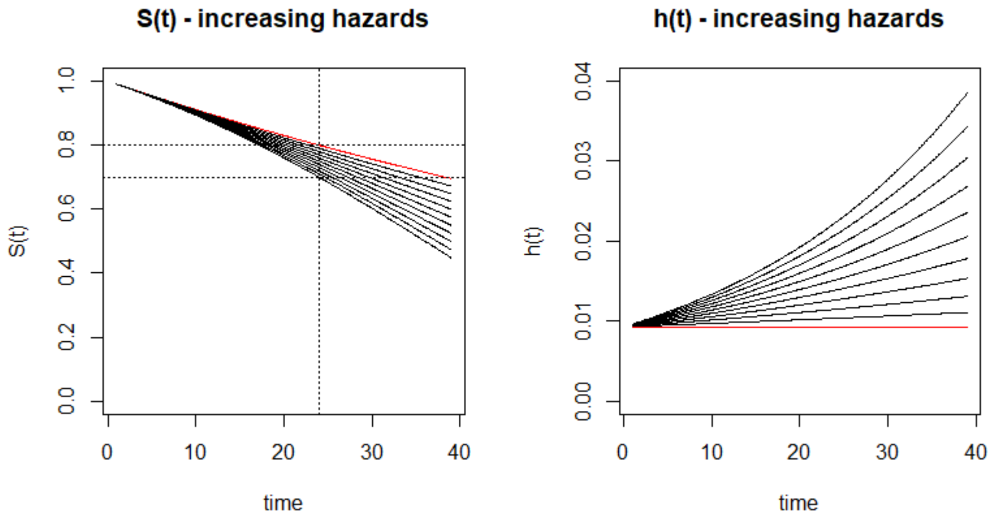

Figure 22: Survival and hazard functions of the simulated Gompertz scenarios with increasing hazards. The red line indicates the exponential distribution with 20% events at 24 months, which is chosen as reference point (Gompertz shape parameter equal to 1). From there the Gompertz shape parameter is gradually increased until there is 30% events at 24 months. The resulting increasing hazards are shown in the right panel.

events is 30%. That is, the shape parameter starts at 0 (30% events) and increases up to 0.0364 (30% events). Similarly to the decreasing hazards scenario, the Gompertz hazards here increase further than those in the corresponding Weibull scenarios. It was expected that both the exponential and Weibull BSSR would underestimate the number of events in this setting. Again, we expected the spline BSSR to be more robust than the other methods in this misspecification scenario.

Gompertz decreasing hazards Due to the strong decrease in hazards in the most extreme Gompertz scenarios (see Figure 21) it could take the simulated trials a very long time to finish. In fact, the strongly decreasing hazards sometimes led to numerical problems in our simulations. Simulating data from a Gompertz distribution with decreasing hazards is not trivial. Some authors (see for example Collett (2015, Ch. 5)) restrict the Gompertz distribution to only have positive shape parameters, which results in increasing hazards. With a negative shape parameter the hazard may decrease too quickly and the survival probability is not ensured to decrease to 0 as time increases to infinity (Jackson, 2016). We addressed this issue by implementing a maximum trial duration of 200 months for the decreasing hazards Gompertz scenarios. At 200 months, patients who had not yet experienced the event were censored. Moreover, if the necessary number of 374 event had not occurred by the time of the maximum trial duration (200 months) the final analysis was carried out regardless. This was relevant primarily for the fixed design in the Gompertz scenarios with the lowest event rates. Since trials could now also finish before the necessary number of events had been observed, we also examined the power of the different design options. For this purpose, we increased the number of replications to 10,000 for the decreasing hazards Gompertz scenarios, so that more accurate rejection probabilities could be obtained. This maximum trial duration was only relevant in the decreasing hazards Gompertz scenarios, since only here such a time constraint was necessary. In the other simulation scenarios the necessary number of events always occurred before 200 months.

The mean relative bias in the estimated number of events at trial end in the decreasing hazards Gompertz scenarios is shown in Figure 23. The results differ from our previous simulation results in the exponential and Weibull settings, because now all BSSR methods exhibit at least some bias once the distribution moves away from the constant hazards assumption (30% events). As the shape parameter of the Gompertz distribution

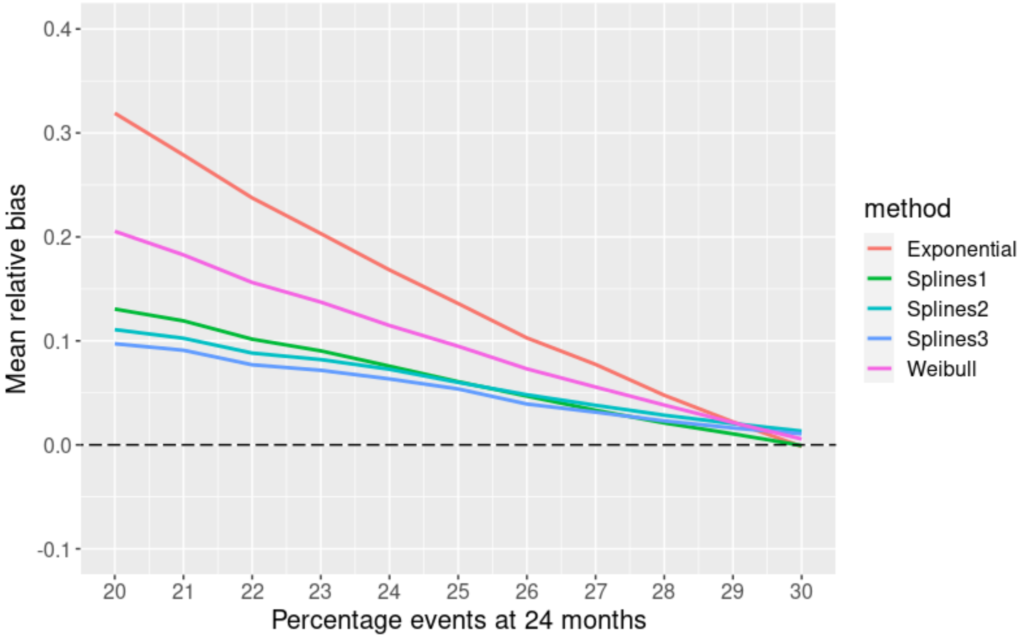

Figure 23: Mean relative bias in the estimated number of events at trial end in the simulated decreasing hazards Gompertz scenarios based on the BSSR models (10,000 replications per scenario). Splines models were fit with 1, 2 and 3 knots (denoted by Splines1, Splines2 and Splines3, respectively). Note that no sample size reestimation was carried out in the fixed design. The expected number of events in the simulation scenarios ranged from 239.1 (20% events) to 372.3 (30% events).

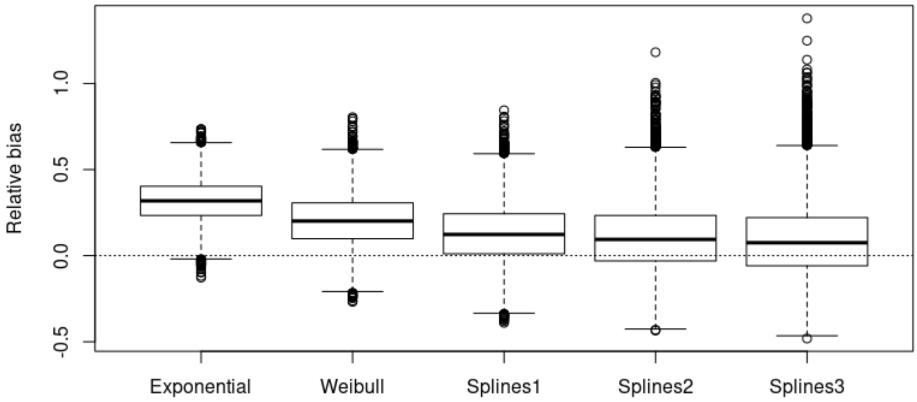

Figure 24: Boxplots of the relative bias in the estimated number of events at trial end in the simulated decreasing hazards Gompertz scenario with 20% events at 24 months (10,000 replications). The Gompertz shape parameter in the simulation scenario was -0.427.

decreases, all BSSR methods start to overestimate the number of events. This effect is the most pronounced for the exponential model, which overestimates the number of events by 31.9% in the most extreme simulation setting (20% events, Gompertz shape parameter of -0.427). The Weibull model overestimated the number of events by a somewhat smaller amount of 20.5 % in the same scenario. The spline models had the lowest bias and overestimated events by between 9.1% and 11.9% in that scenario. The distributions of the relative bias of the estimates for the 20% events scenario are shown as boxplots in Figure 24. Here we can see that the distribution of estimates is generally less biased for the spline models as indicated by the median being closer to 0. However, we also again see the trend of increased variance and outliers for the more flexible 2- and 3-knot spline models.

With regards to the BSSR characteristics, Figure 26 shows how many patients were added on average by the different models. As could be expected (due to the overestimation of the number of events at trial end), the exponential BSSR on average added the fewest patients compared to the other methods. The Weibull BSSR added somewhat more patients than the exponential model across all event percentages. The spline models added the most patients across all scenarios. In the 20% events scenario (Gompertz shape parameter of -0.427) the exponential model added 472, the Weibull model 548 and the spline models between 563 and 574 patients on average. Figure 27 shows the the distribution of the additional recruitment in the 20% events scenario. It shows that the spline models added the maximum of 612 patients more frequently and the number decreases for the Weibull and exponential model, where a smaller number of patients was recruited more often. Moreover, we can see that for the 2- and 3-knot spline models there is a small increase in the number of iterations where no patients were added. This is due to the more extreme overestimations of the number of events caused by occasional overfitting.

As for the mean trial length (Figure 25), the fixed design took an extremely long time when the Gompertz shape parameter became small. In the most extreme Gompertz scenario (20% events, Gompertz shape parameter of -0.427) the fixed design always ran until the maximum trial duration of 200 months. That is, the necessary number of 374 events was never observed within a trial duration of 200 months and the final analysis had to be carried out with insufficient events for the fixed design. While all BSSR methods performed better than that, the exponential BSSR still took on average 86.4 months in

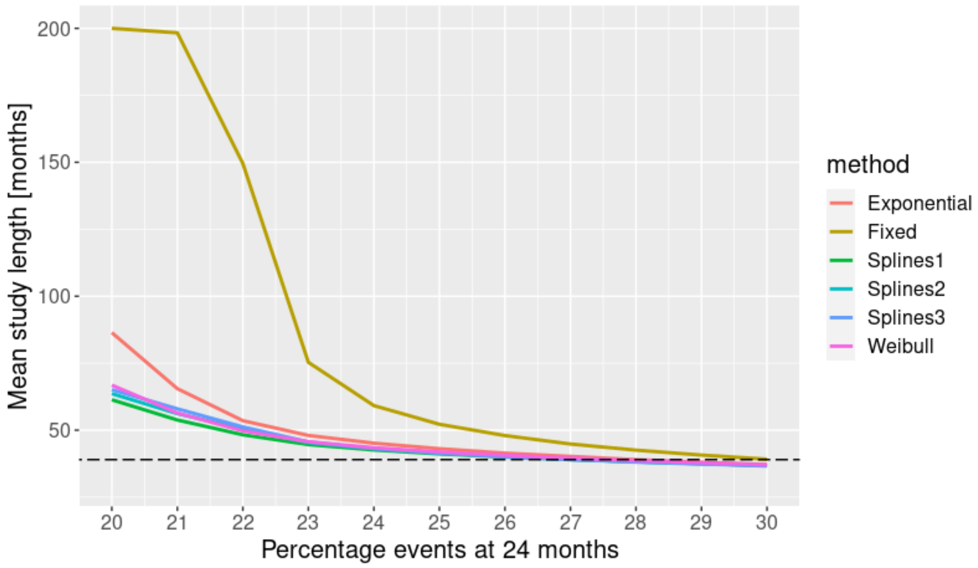

Figure 25: Mean study lengths in the Gompertz scenarios with decreasing hazards based on the BSSR models and the fixed design (10,000 replications per scenario). The trial finished once 374 events were observed or after a maximum trial duration of 200 months (only relevant for the decreasing hazards Gompertz scenarios). The goal was to finish in 39 months (black dotted line).

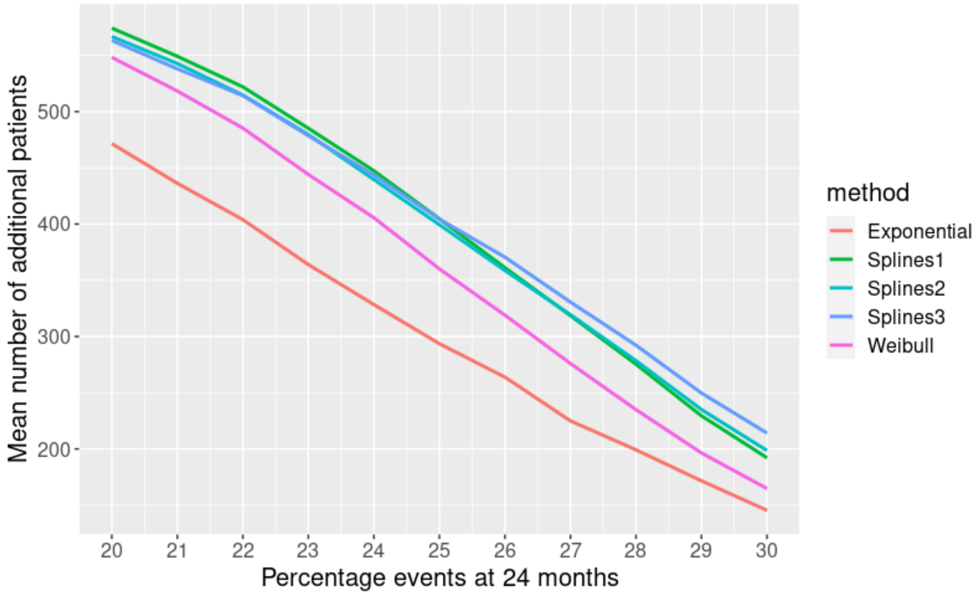

Figure 26: Mean number of additional patients in the Gompertz scenarios with decreasing hazards added by the different BSSR models (10,000 replications per scenario). The maximum number of patients that could be added was 612. Note that no patients were added in the fixed design.

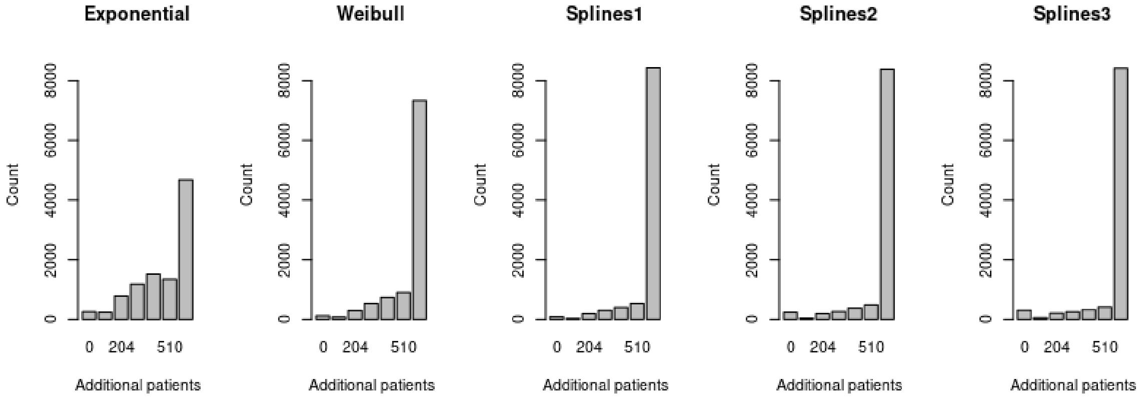

Figure 27: This barchart counts how frequently 0, 102, 204, ..., 612 patients were added by the different designs in the decreasing hazards Gompertz scenario with 20% events at 24 months (10,000 replications). This corresponded to 0, 1, 2, ..., 6 months of additional recruitment, respectively. The Gompertz shape parameter in the simulation scenario was -0.427.

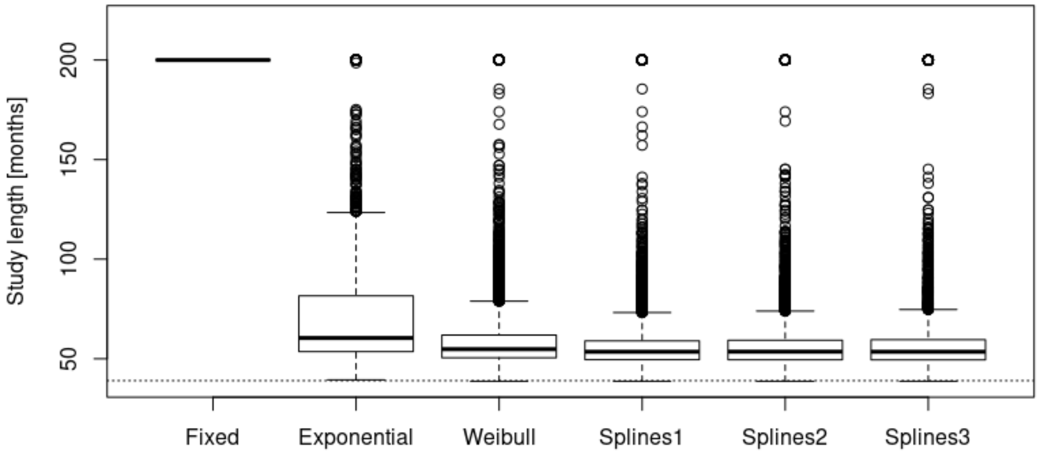

Figure 28: Boxplots of the trial lengths of the different designs in the simulated decreasing hazards Gompertz scenario with 20% events at 24 months (10,000 replications). The trial finished once 374 events were observed or after a maximum trial duration of 200 months (only relevant for the decreasing hazards Gompertz scenarios). The goal was to finish in 39 months (black dotted line). The Gompertz shape parameter in the simulation scenario was -0.427.

the 20% events scenario. In this extreme scenario the other BSSR methods finished far sooner. The Weibull BSSR took on average 66.8 months and the spline models took 61.3 months (Splines1), 63.6 months (Splines2) and 65.1 months (Splines3). Therefore, the

1-knot spline model performed best in terms of average trial duration. Looking at the distribution of the trial lengths for the 20% events scenario in Figure 28, we can see the large discrepancies between the fixed design (which always finished at the maximum trial duration of 200 months) and the BSSR designs. We can also see the worse performance of the exponential model, for which trial durations are severely skewed upwards.

Since the more flexible 2- and 3-knot spline models had more iterations where no patients were added (due to extreme overestimation), there was also a higher proportion of iterations in which the necessary number of events was not observed within 200 months. In the 20% events scenario, this happened this happened in 8.1% and 9.7% of the simulated trials for the 2- and 3-knot spline models, respectively. In contrast, this only happened in 5.8% of the simulated trials for the 1-knot spline model. In the fixed design, 374 events were never observed within 200 months in the 20% events scenario.

Table 3 shows the average number of events available in the different designs at the final analysis for event percentages between 20% and 24%. When the percentage of events was larger than 24%, all simulated trials recorded 374 within 200 months regardless of the design. As we can see, in the 20% events scenario the fixed design carried out the log-rank test based on only 308.4 events on average. In the scenario with 21% events this increased to 338.9, but it was still far below the required 374 events. In contrast, the smallest average number of events of all BSSR designs was 370.2 events for the exponential BSSR in the 20% events scenario. Therefore, all BSSR methods performed considerably better than the fixed design in terms of ensuring sufficiently many events are observed by the end of the trial.

Since the fixed design trials often finished before the necessary number of 374 events was observed, we expected this to reflect in a diminished statistical power for the log-rank test at the final analysis. The simulated power of the fixed design and the different BSSR methods is depicted in Figure 29. It was estimated as the proportion of trials in which the null hypothesis of no treatment effect was rejected (data were generated under the assumed alternative of $\theta = 0.7$). The power of the designs should be at least 90% if the necessary 374 events are observed. As anticipated, the power of the fixed design is lower than 90% in the scenarios with low event rates. In the 20% and 21% events scenarios the power of the fixed design was 88.4% and 86%, respectively. In contrast, the power was between 91% and 92% for the BSSR designs. Interestingly, the power was on average slightly larger than 90% in the simulation scenarios where 374 were always observed. This

Table 3: Average numbers of events available at the final analysis for the different designs in the Gompertz decreasing hazards scenarios with 20% to 24% events. The final analysis was carried out either when 374 events had been observed or at the maximum trial duration of 200 months. When the percentage of events was larger than 24%, 374 events were available in all simulated trials for all design options, so these results are omitted from the table.

Model	**20%**	**21%**	**22%**	**23%**	**24%**
Fixed Design	308.4	338.9	366.8	373.9	374
Exponential model	370.2	372.9	373.9	374	374
Weibull model	372.4	372.5	373.7	374	374
Spline model with 1 knot	373	373.4	373.9	374	374
Spline model with 2 knots	372	372.9	373.8	374	374
Spline model with 3 knots	371.4	372.5	373.7	374	374

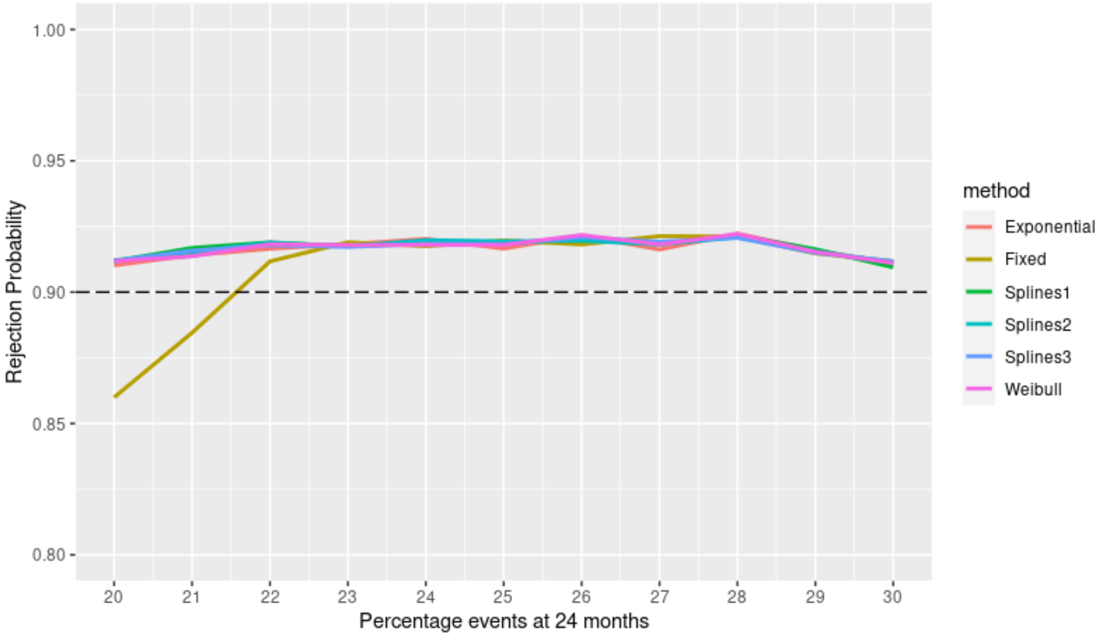

Figure 29: Power of the log-rank test based on the fixed design and the different BSSR models in the Gompertz scenarios with decreasing hazards (10,000 replications per scenario). The simulated power was estimated as the proportion of trials in which the null hypothesis of no treatment effect was correctly rejected. The power of the designs should be at least 90% if the necessary 374 events are observed.

is in line with the findings of Friede et al. (2019), who also found the simulated power in their original simulation to be slightly larger than 90%. In the current simulation, we used the same number of necessary events as Friede et al. (2019), which was based on the design of the motivating study in Multiple Sclerosis (Kappos et al., 2018). Upon closer inspection we noticed, however, that the necessary number for our design based on the Schoenfeldt formula was 372 (and not 374). This could explain why the designs that finished after 374 events were slightly overpowered.

With regards to the type I error rate, we again found no indication of an increased rejection probability for the BSSR designs when data were simulated under the null hypothesis (10,000 replications per scenario). The table of the simulated type I error rates for the Gompertz scenarios can be found in Appendix A.

Gompertz increasing hazards The mean relative bias in the estimated number of events at trial end in the increasing hazards Gompertz scenarios is shown in Figure 30. As before in the decreasing hazards Gompertz scenarios, all BSSR methods now exhibit some bias under this misspecification. The increasing Gompertz shape parameter results in hazards that increase strongly over time (see Figure 22) and all BSSR methods underestimated the number of events in these simulation scenarios. The exponential BSSR again had the largest bias, underestimating the number of events at trial end by 25.6% in the scenario with the largest Gompertz shape parameter (30% events, Gompertz shape paremter of 0.0364). The average bias of the Weibull BSSR was considerably lower at 13.7%. The 1-knot spline model had a slightly lower average bias of 10.9% and the 2- and 3-knot spline models had the lowest average bias of 6.5% and 6.3%, respectively. The distributions of the relative bias of the estimates for this scenario are shown as boxplots in Figure 31. Again, the results look similar to the decreasing hazards Gompertz scenario in Figure 24, only that the bias is now negative.

The different degrees of underestimation are reflected in the average number of patients added by the different BSSR methods, which are plotted in Figure 32. The exponential BSSR almost always added about 600 patients on average regardless of the percentage of events at 24 months. The average number of additional patients for the remaining models decreases somewhat as the percentage of events in the data increases. As to be expected based on the bias, the 2- and 3-knot spline models added the fewest patients on average, followed by the 1-knot spline model and the Weibull model. In

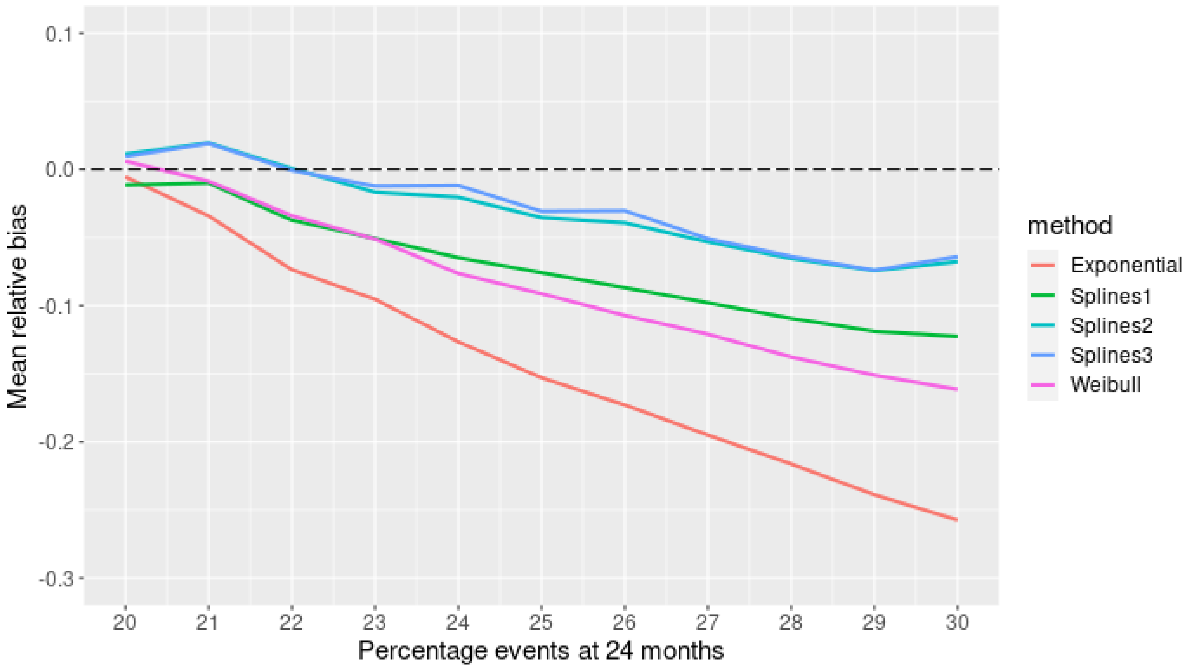

Figure 30: Mean relative bias in the estimated number of events at trial end in the simulated increasing hazards Gompertz scenarios based on the BSSR models (1,000 replications per scenario). Splines models were fit with 1, 2 and 3 knots (denoted by Splines1, Splines2 and Splines3, respectively). Note that no sample size reestimation was carried out in the fixed design. The expected number of events in the simulation scenarios ranged from 246.8 (20% events) to 389.9 (30% events).

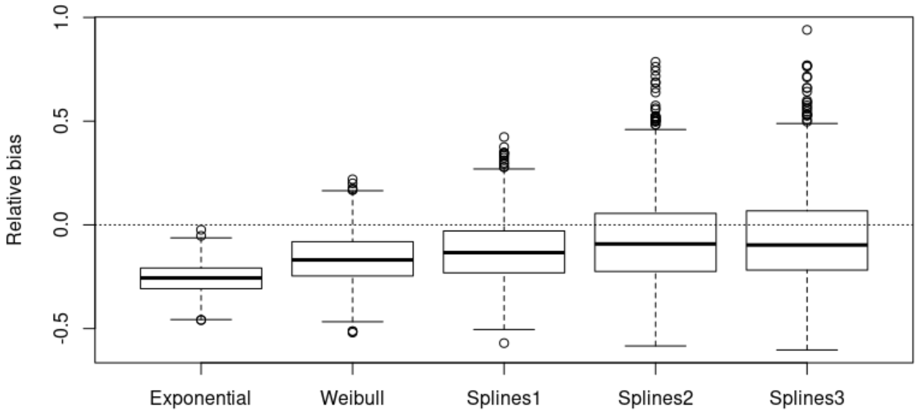

Figure 31: Boxplots of the relative bias in the estimated number of events at trial end in the simulated increasing hazards Gompertz scenario with 30% events at 24 months (1,000 replications). The Gompertz shape parameter in the simulation scenario was 0.0364.

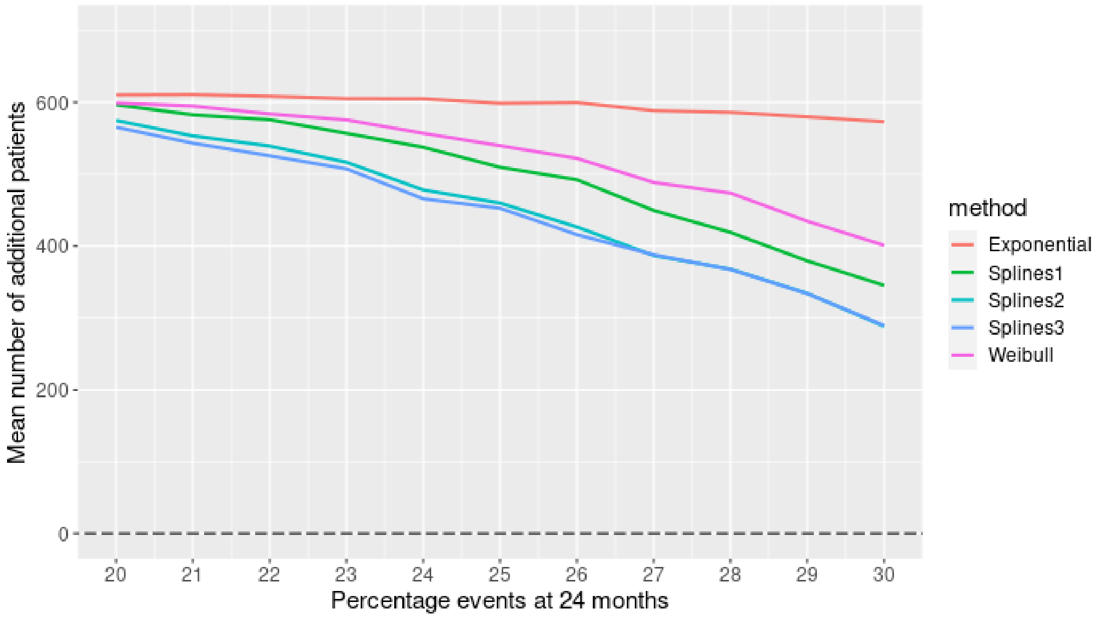

Figure 32: Mean number of additional patients in the Gompertz scenarios with increasing hazards added by the different BSSR models (1,000 replications per scenario). The maximum number of patients that could be added was 612. Note that no patients were added in the fixed design.

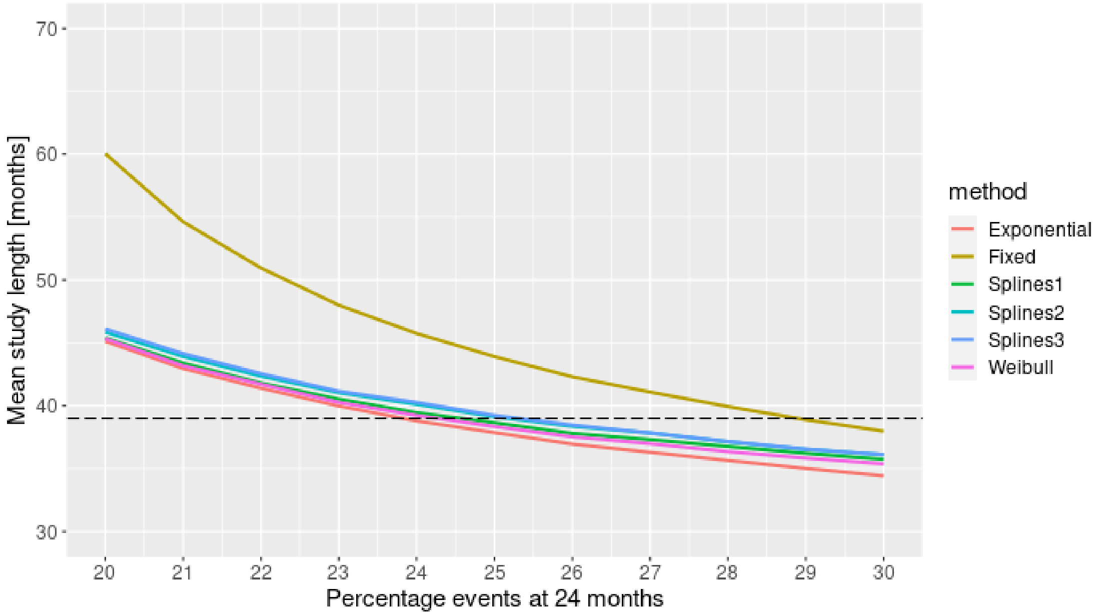

Figure 33: Mean study lengths in the Gompertz scenarios with increasing hazards based on the BSSR models and the fixed design (1,000 replications per scenario). The trial finished once 374 events were observed. The trial finished once 374 events were observed and the goal was to finish in 39 months (black dotted line).

the most extreme increasing hazards Gompertz scenario (30% events, Gompertz shape paremter of 0.0364) the 2- and 3-knot spline models added on average 288 and 299 patients. The 1-knot spline model added on average 345 patients and the Weibull model 401 patients.

As we can see with regards to the average trial length in Figure 33, the trials typically finished earlier than necessary in the scenarios with higher event percentages. That is, the underestimation in the number of expected events at trial end lead to the BSSR methods recruiting too many additional patients. In practice, this would lead to unjustified additional costs for the study sponsors. The patterns of additional recruitment in the 30% events scenario are shown in Figure 34. Here we can see clearly that the exponential model added the maximum of 612 patients in almost all simulation runs, even though no additional recruitment was necessary in this scenario. In the other designs, the pattern of additional recruitment across the iterations is more balanced, but all off them still added 612 patients in a large number of iterations. All in all though, the BSSR methods became more robust as the model complexity increased. Note that in this in-

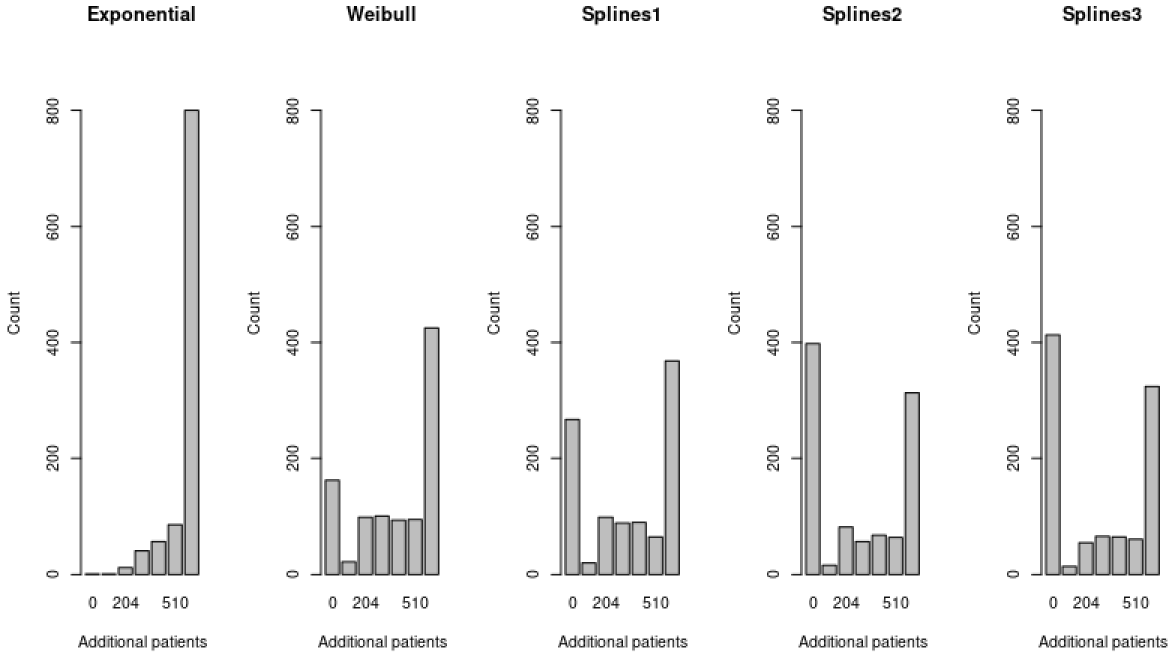

Figure 34: This barchart counts how frequently 0, 102, 204, ..., 612 patients were added by the different designs in the increasing hazards Gompertz scenario with 30% events at 24 months (1,000 replications). This corresponded to 0, 1, 2, ..., 6 months of additional recruitment, respectively. The Gompertz shape parameter in the simulation scenario was 0.0364.

creasing hazards scenario the typical overestimation outliers of the 2- and 3-knot spline models were less problematic, since this only resulted in no additional recruitment here. However, these strongly biased outliers were still present and they would be problematic if the recruitment of additional patients was in fact necessary.

To conclude the results of the Gompertz simulation, we found that the fixed design performed very poorly in the decreasing hazards Gompertz scenarios. The necessary number of 374 events was frequently not observed within 200 months, which resulted in an underpowered trial. Moreover, we found that for both the decreasing and increasing hazards scenarios the exponential model was the worst-performing BSSR method. The Weibull BSSR method had a somewhat larger bias than the splines methods, but performed decently in terms of mean trial length and mean additional patients. The flexible 2- and 3-knot spline models had the lowest bias, but still suffered from occasional overfitting, which led to extreme overestimations of the number of events. The 1-knot spline model had the best overall performance, because it was affected less by overfitting while having only a moderate bias.

5.2.5 Hybrid spline method

After having extensively reviewed the results of the splines BSSR method, we wanted to briefly consider the hybrid extension that was presented in Subsection 3.2. Here we will focus on the exponential simulation scenarios. Figure 35 shows an example simulation run, in which a spline model with one knot and a hybrid spline model with one knot were fit. The initial part of the hybrid method is based on the interim data and is estimated by the Kaplan-Meier estimator. The extrapolated part is then based on the 1-knot spline model and is identical for the spline and hybrid spline method.

To compare the performance of the spline and the hybrid spline method, we carried out simulations based on exponentially distributed data as discussed in Subsection 5.1. The mean relative bias in the estimated number of events at trial end for the spline and hybrid spline BSSR methods are shown in Figure 36. The performance of the different spline methods is overall quite similar. The fully parametric spline model with one internal knot seems to perform best, since the other methods all seem to display some minor overestimation on average. Among these models, the overestimation is the least pronounced for the hybrid spline methods with one internal knot.

We can get a more detailed understanding of the performance of the different spline

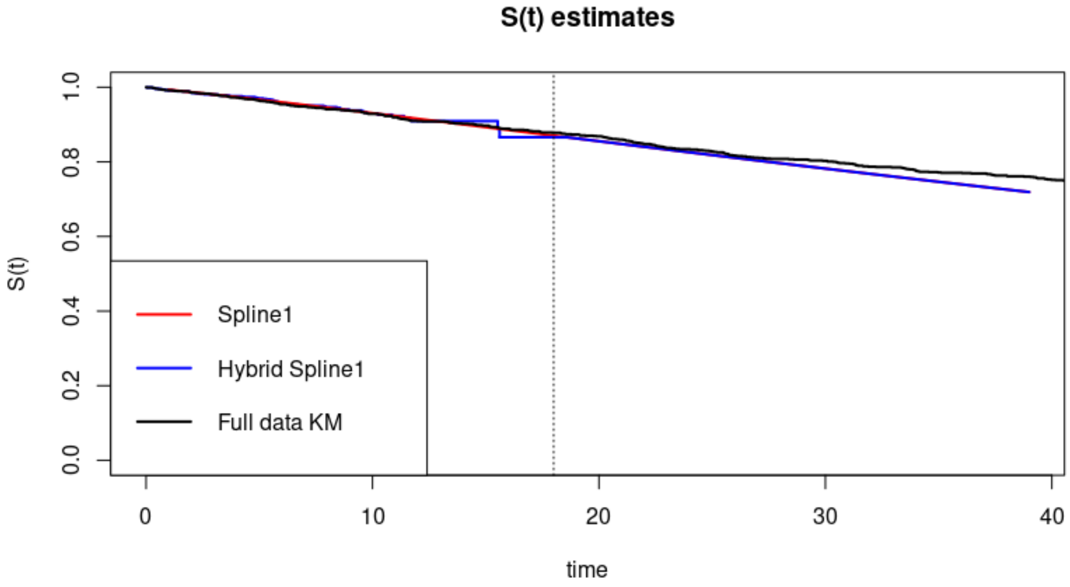

Figure 35: Example of a spline model (spline1) and a hybrid spline model (hybrid spline1) with one internal knot fit to the same dataset. For the interim part of the data the hybrid method uses the Kaplan-Meier estimator and the extrapolation tail is based on the 1-knot spline model. In this example, data were generated from an exponential distribution with 20% events at 24 months. The black dotted line shows the timing of the interim analysis at Month 18.

methods by considering the iteration-wise output. Figure 37 shows boxplots of the relative bias of the spline and hybrid spline BSSR methods in the exponential simulation scenario with 20% events at 24 months. The direct comparison between the respective spline and hybrid spline methods with the same number of knots shows that the hybrid methods tend to have a slightly larger variance well as a larger number of outliers. This difference is the most prominent for spline models with only one internal knot. Here, the fully parametric spline model only has a small number of outliers compared to the equivalent hybrid version with one internal knot. For the fully parametric spline model, there were 10 iterations in which the number of events was overestimated by more than 50%. In the corresponding hybrid method it was already 27 iterations. Moreover, in the hybrid method there were some occasions in which the number of events was overestimated by more than 100%. In contrast, such extreme overestimations did not occur with the fully parametric spline method with one knot.

This increased variance and increased occurrence of outliers is possibly due to the increased variance of the Kaplan-Meier estimator at the end of the available follow-up, when only few patients are at risk (Dudley et al., 2016). Figure 38 shows an example of

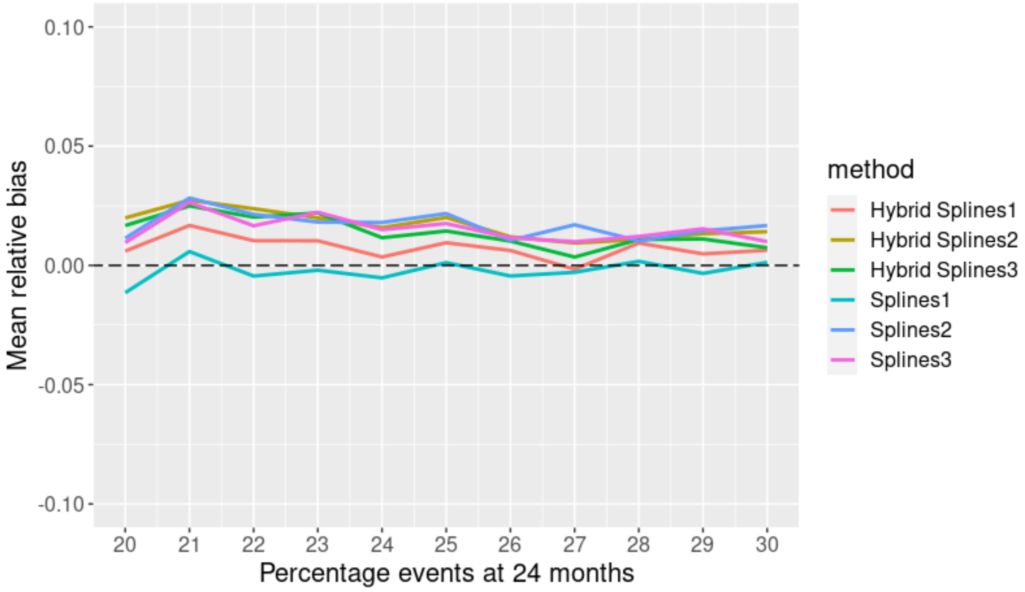

Figure 36: Mean relative bias in the estimated number of events at trial end in the simulated exponential scenarios based on the spline and hybrid spline BSSR models (1,000 replications per scenario). Both spline models (Splines1, Splines2, Splines 3, respectively) and hybrid spline models (Hybrid Splines1, Hybrid Splines2, Hybrid Splines3, respectively) were fit with 1, 2 and 3 knots, respectively. The expected number of events in the simulation scenarios ranged from 246.8 (20% events) to 372.3 (30% events).

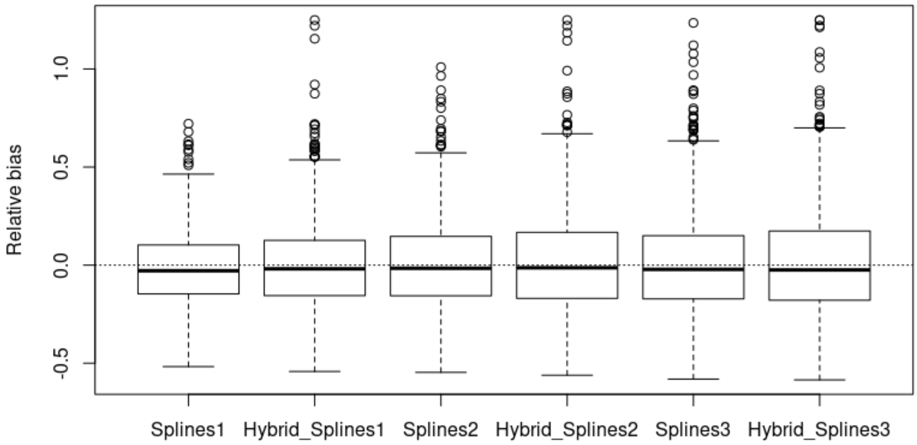

Figure 37: Boxplots of the relative bias of spline and hybrid spline BSSR methods in the estimated number of events at trial end in the simulated exponential scenario with 20% events at 24 months (1,000 replications). The results of the hybrid spline methods are placed next to the corresponding fully parametric spline model with the same number of knots.

such a problematic scenario, where the Kaplan-Meier estimate based on the interim data includes a sharp drop at the end of the initial follow-up. As elaborated in Subsection 3.2, the last available Kaplan-Meier estimate of the survival function may be smaller than following spline estimates for the extrapolated part. As explained above, to maintain monotonicity we set the extrapolated values equal to the last observed Kaplan-Meier estimate as long as the spline estimates were larger than that value. Since the last available Kaplan-Meier estimate was so low here, the extrapolated part remained constant in this example, because the extrapolated spline tail always indicated a larger survival probability. A scenario as shown in this example might have been the reason for the extreme overestimations that were occasionally observed for the hybrid spline model with one internal knot compared to the respective fully parametric spline model. Overall though, the spline and hybrid spline BSSR methods were rather similar, except for the slight increase in variance and outliers for the hybrid method.

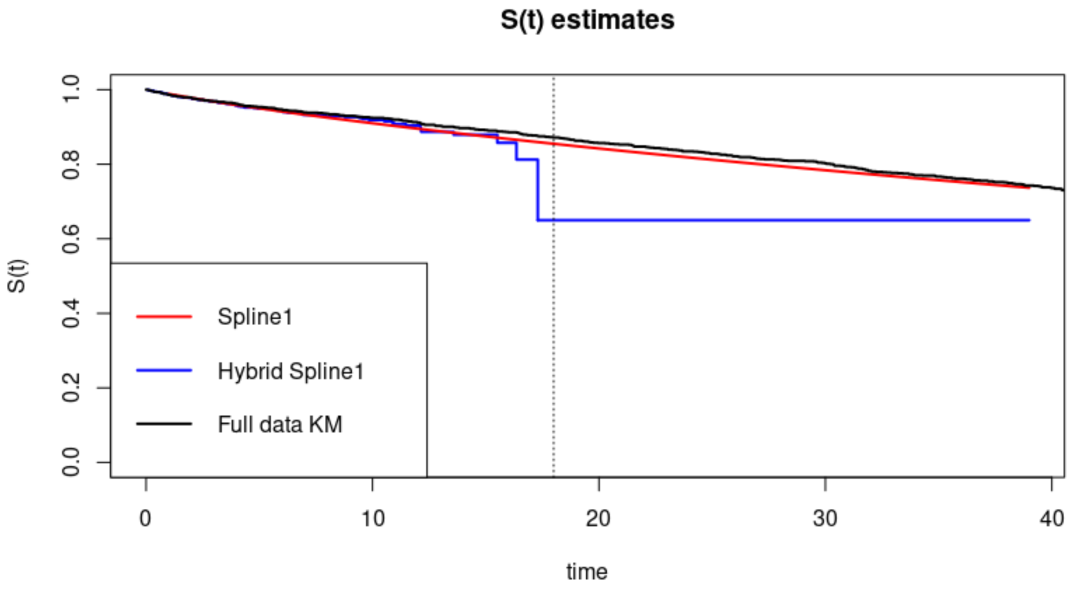

Figure 38: Another example of a spline model (Spline1) and a hybrid spline model (Hybrid Spline1) with one internal knot fit to the same dataset. In this example, the hybrid method has a poor performance due to the sharp drop in the Kaplan-Meier estimate based on the interim data.

6 Case study: Multiple Sclerosis data

The original BSSR method for time-to-event trials developed by Friede et al. (2019) was motivated by a clinical trial on secondary progressive multiple sclerosis (SPMS). To further examine our proposed spline based BSSR method, we decided to also apply it to a relevant real world data. We also considered the data of the SPMS trial, which has been published by Kappos et al. (2018). Note that we also analysed another dataset on relapsing remitting MS (Kappos et al., 2010), which has also been considered by Friede et al. (2019). However, the results were very similar to the ones for the SPMS trial, so we will omit them from the discussion here.

6.1 Study characteristics and data extraction

Multiple sclerosis (MS) is a chronic neurological condition (Kappos et al., 2018). SPMS is a later stage of the disease, which is associated with a continuous progression of physical disability and neurological deficits (Kappos et al., 2018). The study by Kappos et al. (2018) was a large phase 3, randomised, double-blind and placebo-controlled trial. The design was event-driven and the time-to-event analysis was planned for when a minimum of 374 events had been observed (Kappos et al., 2018). The primary endpoint was 3-month confirmed disability progression (CDP), which was quantified as a relevant increase in the Expanded Disability Status Scale (EDSS) score (Kappos et al., 2018). At the time-to-event analysis 1096 patients had been assigned to the treatment group and 545 patients to the control group (Kappos et al., 2018).

For this case study we considered the data of the control group and wanted to examine how well different parametric models (incl. spline models) capture the observed hazard and survival functions. Since individual patient data (IPD) were not available we had to revert to the available Kaplan-Meier graph. We digitized the graphical data and obtained pseudo-IPD by using Guyot et al.'s (2012) extraction algorithm. The algorithm uses digitized Kaplan-Meier data in conjunction with available information on numbers of events and numbers at risk at given intervals. The pseudo-IPD are then obtained numerically by solving the inverted Kaplan-Meier equations (Guyot et al., 2012). The method has been shown to have a high accuracy in terms of estimating the survival probability based on the reconstructed data (Guyot et al., 2012). We carried out the algorithm based on the R code that has been supplied as an additional file to the publication of Guyot et al. (2012). The digitized data from the Kaplan-Meier plot were obtained by using the *WebPlotDigitizer* app (Rohatgi, 2021). The resulting reconstructed Kaplan-Meier curve is presented in Figure 39.

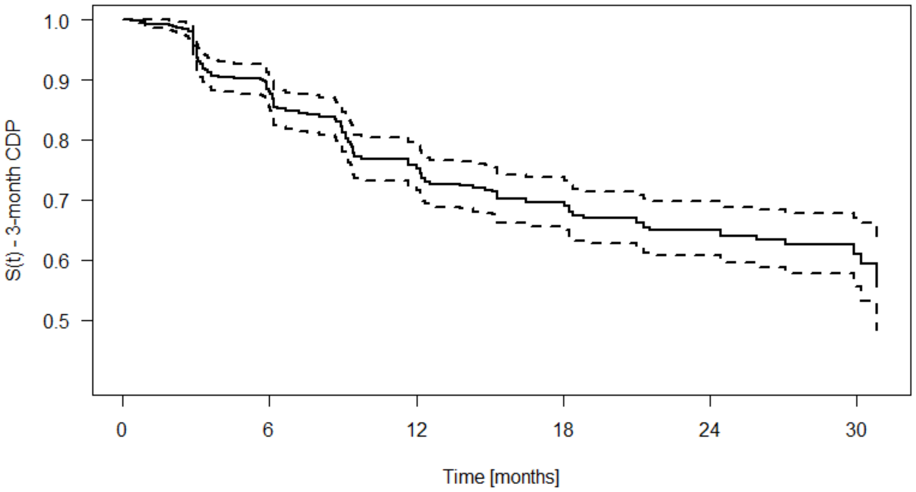

Figure 39: Reconstructed Kaplan-Meier graph for the control group in the SPMS study of Kappos et al. (2018). The data were digitized using the *WebPlotDigitizer* app (Rohatgi, 2021) and pseudo-IPD were obtained by using the extraction algorithm of Guyot et al. (2012).

6.2 Model fits

The fit of various parametric models to the reconstructed Kaplan-Meier graph is shown in Figure 40. We fit standard parametric models (Exponential, Weibull, Gompertz) and Royston-Parmar PH spline models with 1 to 3 internal knots to the pseudo-IPD. Note that a practical application of the BSSR algorithm was not possible, because this would have required knowledge of patients' recruitment and event dates, which were not available to us. Therefore, here we focused on the model fits of the different parametric models and compared their survival extrapolations.

The exponential and Weibull model appear to have a virtually identical fit. This is confirmed by the parameter estimates of the standard parametric models shown in Table 4. The exponential and Weibull event rate parameter is almost the same and the Weibull shape parameter is very close to 1, which implies constant hazards. Both these models, however, do not seem to fit the Kaplan-Meier graph particularly well. The Gompertz model seems to provide a better both visually and also in terms of the AIC and BIC,

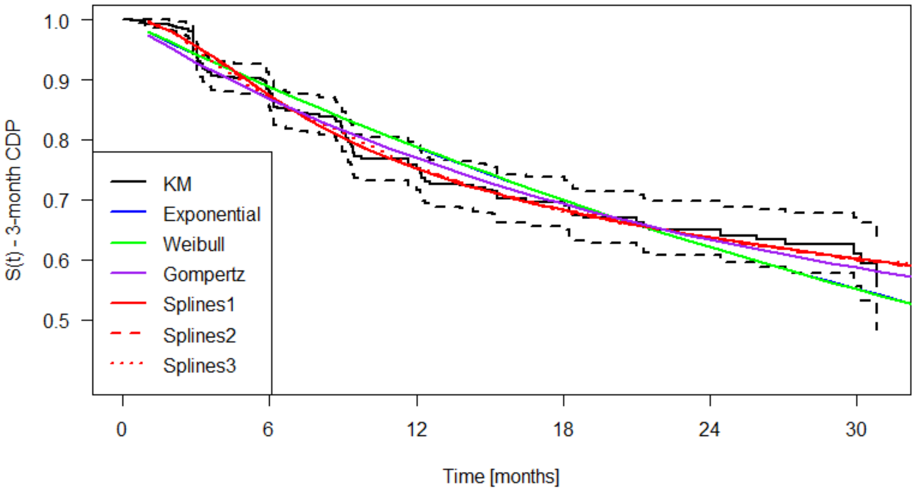

Figure 40: Fit of several parametric models to the reconstructed Kaplan-Meier graph of the SPMS control group data. We fit standard parametric models (Exponential, Weibull, Gompertz) and the Royston-Parmar PH spline model with 1, 2 and 3 knots (Splines1, Splines2, Splines3, respectively).

Table 4: Parameter estimates of the standard parametric models fit to the SPMS control group data.

Model	**Rate parameter**	**Shape parameter**
Exponential model	0.0198	-
Weibull model	0.0199	1.0037
Gompertz model	0.0254	-0.0255

which are presented in Table 5. The Gompertz shape parameter was -0.0255, which implies a decreasing hazard over time.

The resulting hazard functions based on the different models are shown in Figure 41. Here we can see the decrease in hazards over time in the Gompertz model, which was not present in the Weibull model. The closest fit to the observed data was achieved by the spline models. The spline models with different numbers of knots appear very similar. The spline model with one internal knot had the lowest AIC and BIC of all the model considered. That is, this 3 parameter model (see Equation (12)) seemed to have to best fit to the data given its number of parameter used. This seems reasonable, since the added flexibility of the 2- and 3-knot spline models did not result in very different

Table 5: AIC and BIC values for the different parametric models fit to the SPMS control group data. The spline model with one internal knot had the lowest value on both information criteria, indicating the best model fit.

Model	**AIC**	**BIC**
Exponential model	1704.67	1708.97
Weibull model	1706.66	1715.27
Gompertz model	1700.91	1709.52
Spline model with 1 knot	1675.90	1688.80
Spline model with 2 knots	1677.54	1694.75
Spline model with 3 knots	1676.31	1697.82

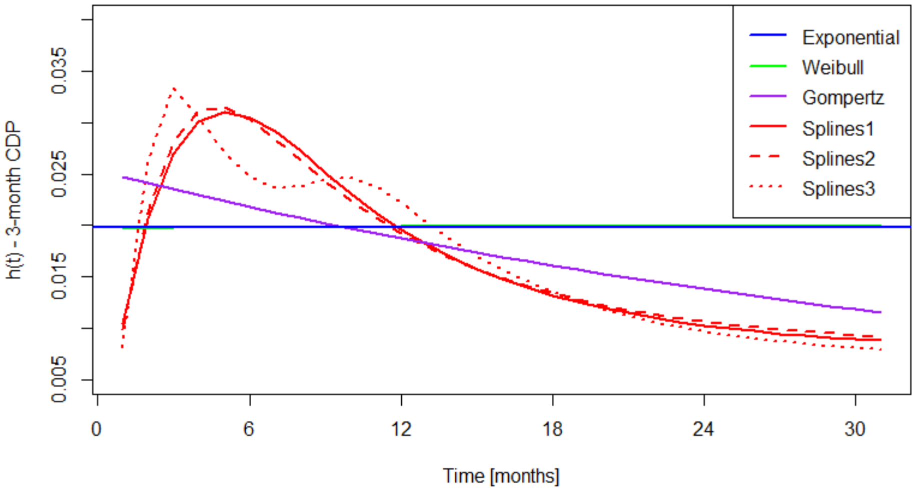

Figure 41: Estimated hazard functions based on the different parametric models fit to the SPMS control group data. We fit standard parametric models (Exponential, Weibull, Gompertz) and the Royston-Parmar PH spline model with 1, 2 and 3 knots (Splines1, Splines2, Splines3, respectively).

estimates of the survival functions in this example.

Interestingly, the spline models exhibit an initial increase in hazards, before they decrease over time like in the Gompertz model. This is possibly due to the nature of the primary endpoint, which was 3-month confirmed disability progression (CDP). That is, there is an inherent lag time in the reporting of the event. Since the event cannot occur within the first 3 months, it is evident that the estimated hazard would initially be low and be followed by an increase. However, this is likely not a feature of the underlying hazard function of the event, but rather a feature of the reporting lag associated with the endpoint.

If the models depicted in Figure 40 would be used for survival extrapolation they would lead to drastically different results. Figure 42 shows the 2-year extrapolated survival curves based on the different models. As we can see, the exponential and Weibull model predict a much larger number of events than the spline models and the Gompertz model. For the exponential and Weibull model the extrapolated survival probability 2 years later was only 33%. In contrast, the Gompertz model predicted 47% survival and the 1-knot spline model predicted 50% survival after 2 years. Thus, if these numbers would be relied on in the context of a BSSR, they would lead to drastically different recruitment responses. If decreasing hazards estimated by the Gompertz and spline models would turn out to be accurate, additional recruitment might be necessary if this was not anticipated at the planning stage. However, if an exponential or Weibull were to be used for the BSSR, potentially necessary additional recruitment would likely not occur. In practice, this could lead to severely increased trial durations. If the novel treatment was effective, this would lead to an increased waiting time for patients in need and a delayed market entry for the drug developers.

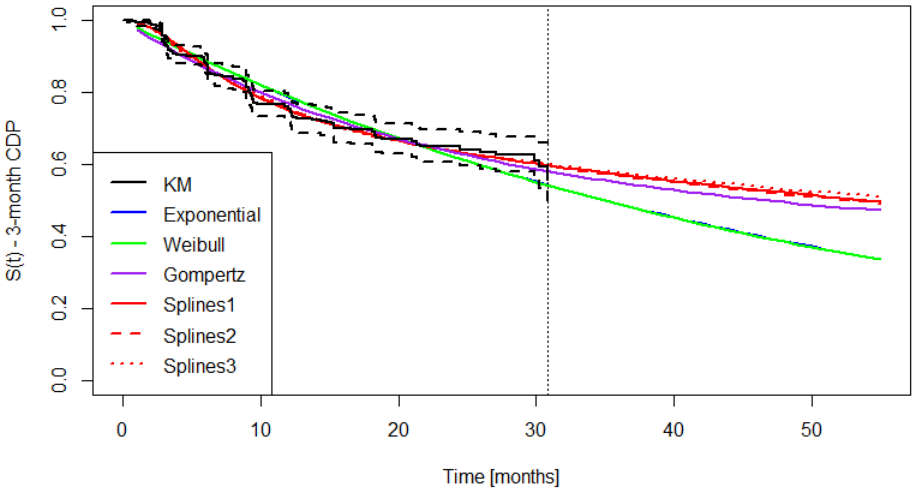

Figure 42: Survival extrapolations based on the different parametric models fit to the SPMS control group data. The vertical black dotted line indicates when the last event was observed in the original dataset.

7 Discussion

The aim of the current thesis was to propose a flexible parametric approach for BSSR in clinical trials with time-to-event outcomes and to compare it to existing parametric approaches. Sample size reestimation method are important tools to ensure the validity of a clinical trial. Here we proposed an extension of Friede et al.'s (2019) parametric BSSR framework by carrying out the extrapolation based on Royston-Parmar PH spline models. We investigated the operating characteristics of our proposed BSSR method in an extensive simulation study as well as in a case study on Multiple Sclerosis. In the simulation study, the simulated data were based on an exponential, Weibull or Gompertz distribution. We compared our proposed flexible spline-based method with BSSR based on standard parametric models (exponential, Weibull) as well as with a fixed design without a BSSR. In the following paragraphs we will discuss our findings in the context of the existing literature, consider limitations and extensions and provide practical recommendations based on our findings. Specifically, we will consider the issue of model selection for spline models, spline based survival extrapolation and implications for BSSR.

7.1 Spline model selection and extrapolation performance

In our results in Subsection 5.2.1 on spline model selection, we found that the AIC selected a 0-knot model in the majority of cases, regardless of whether the data stemmed from a Weibull distribution or a Gompertz distribution. However, spline models (that is, those with >0 knots), were found to be better extrapolators for Gompertz distributed data in our simulation study. This aligns with findings of Rutherford et al. (2020), who report that a good within sample fit (as measured by the AIC, for example) does not guarantee an accurate extrapolation. To illustrate, they show simulated data from a Weibull distribution with unobserved frailty and other-cause mortality increasing with age. In that example, the Weibull and the Royston-Parmar spline model had the highest AIC (compared to a log-normal and generalised gamma model), but in fact had the best extrapolation. This shows that the AIC cannot always be relied on whet it comes to model selection for the purpose of extrapolation.

For complex hazard functions (simulated by means of mixture Weibull distributions), Rutherford et al. (2015) found that the AIC tended to select complex spline models more frequently. However, our Gompertz simulated data seemingly did not deviate sufficiently from a (also monotonic) Weibull distribution for the AIC to select a spline model. Thus, it is possible that the AIC may select spline models more frequently, when the underlying hazards have a more complex shape.

While here the issue was that a suitable, complex model (i.e. spline model) was not chosen, the issue may also be the other way around. Rutherford et al. (2015) found that for moderately large ($n = 3,000$) and large sample sizes ($n = 30,000$) the AIC tended to select unnecessarily complex spline models, when the data actually stemmed from a Weibull distribution. We also found some indications of this in our simulation, where the AIC occasionally (in about 15% of cases) selected a spline model even though the data stemmed from exponential or Weibull distributions. In practice, this could lead to overfitting, which was present especially for the more complex spline model with 2 to 3 knots. Rutherford et al. (2015) reported similar issues, where overly complex spline models had a worse performance in Weibull simulated data due to overfitting.

Overall, in our simulations we found that the moderately flexible 1-knot spline model was similarly robust to overfitting as the Weibull model, but performed slightly better in the misspecification scenarios. Based on these observations, we suggest that when faced with uncertainty, a fixed 1-knot spline model might be a useful compromise. This coincides with the original recommendation of Royston and Parmar (2002), that a 1-knot spline model is often a solid choice. Nevertheless, the AIC may be considered and can provide some indication in favor or against some model choices. If overfitting is a particular concern, the more stringent BIC can be considered.

The results on the relatively robust performance of the spline models correspond to the findings of Kearn et al.'s (2021) simulation results. They found that while the spline models were not always unbiased, they typically had a comparatively small MSE due to their reasonable variance. Naturally, compared to standard parametric models, the variance of flexible spline models is increased. However, especially for 1-knot spline models, we found that there was a good compromise between flexibility and variance. However, we did observe an increase in variability when the spline models became more complex, which occasionally led to extreme outliers.

In our Gompertz simulation we found that the spline BSSR method were slightly more robust in the increasing hazards scenarios, as indicated by a smaller bias. Similarly, in their simulation study Rutherford et al. (2020) also found spline models to be more robust in increasing hazards scenarios compared to decreasing hazards scenarios. Specifically, they simulated Weibull data with unobserved heterogeneity and long survival and found the spline method to be robust to these features in increasing hazards scenarios, but not in decreasing hazards scenarios.

In our case study of SPMS data, we found that the spline model and the Gompertz model seemed to fit well to the observed data. This is in line with the findings of Gray et al. (2021), who also found that spline models fit well to the observed data and also extrapolated well in their real world datasets on various types of cancers. However, they found that in their artificially censored cancer cohort registry data frequently not the PH spline model, but the PO or probit spline models had the best performance. While in the PH spline model, the restricted cubic splines are fit on the log cumulative hazard scale, they are fit on the log odds or the probit scale for the PO and probit spline models, respectively (Royston and Parmar, 2002). Since the restricted cubic splines are linear beyond the last observed uncensored survival time, the different scales on which the spline models are fit imply different extrapolation mechanisms. For the PO spline model, the extrapolated part will follow a local log-logistic distribution and for the probit spline model it will follow a log-normal distribution (Royston and Parmar, 2002). Gray et al. (2021) presented a number of example cohorts with various hazard shapes in greater detail and fit both standard parametric and spline models. They found that in all scenarios some spline model extrapolated well, but it was not always the same spline model. In some instances, the PH spline model outperformed the PO spline model and vice versa. This implies that in practice, fitting the spline models on various scales should be considered. If a Royston-Parmar spline model is fit, the scale should then not only be chosen in terms of optimal within-sample fit, but clinical/ biological plausibility as well as external evidence should be considered for the extrapolated hazard (Latimer and Adler, 2022).

7.2 Implications for blinded sample size reestimation

To our knowledge, the current study was the first to use the Royston-Parmar spline model to extrapolate survival for the purpose of blinded sample size reestimation. This is an important contribution, because the extrapolation performance of spline models has previously only been considered by health economists for the purpose of health technology assessments. In such economic evaluations, the nature of the data at hand and the extrapolation time frame, however, are vastly different compared to an interim analysis setting in a clinical trial. At a blinded sample size review, we will typically have much smaller sample sizes, shorter follow-up times and considerable censoring. Moreover, extrapolation is carried out until the planned end of the trial, whereas a lifetime horizon is frequently considered in economic evaluations (Latimer, 2013). Here, we specifically considered how the Royston-Parmar PH spline model performed as an extrapolator at a blinded review in an event-driven design.

Our simulation study was an extensions of Friede et al.'s (2019) study, who first investigated BSSR for event-driven designs based on exponential, Weibull and piecewise exponential models. In their simulation study, they simulated data based on an exponential distribution and carried out parametric BSSR based on an exponential model. We extended their simulation study in three ways. First, we now had different simulation scenarios, in which the data were simulated from different distributions. In additional to the exponential scenario, we also considered Weibull and Gompertz distributed data. Second, we carried out the parametric BSSR based on different parametric models. We considered the exponential and Weibull BSSR, which have been proposed by Friede et al. (2019), but we now also considered the Royston-Parmar PH spline model with 1 to 3 internal knots. Third, we proposed a modification of Friede's (2019) parametric BSSR approach in that the pooled estimates from the blinded interim data may be used directly for reestimating the sample size (i.e. without splitting up the estimates based on the assumed treatment effect). These extensions of the simulation set up allowed us to address three main questions. First, we were able to investigate how standard parametric BSSR methods perform in a misspecification scenario, where the underlying data follow a different distributions. Second, we could examine to what extent flexible parametric spline models are potentially more robust in such misspecification scenarios. Third, we could investigate whether BSSR can successfully be carried out without splitting the pooled estimates obtained from the blinded interim data.

As for the first question, we found that using a simple exponential model for the BSSR could lead to problems, when the underlying data violated the constant hazards assumption. When the data stemmed from a Weibull distribution, this was less problematic in the decreasing hazards scenarios, because the exponential BSSR mostly still added sufficiently many patients. In the increasing hazards scenarios, however, the exponential BSSR on average added about 200 patients more than the other methods, even when no additional recruitment was necessary. This issue was amplified in the Gompertz scenario, where the exponential BSSR almost always added the maximum of 612 additional patients, even when the trial was in reality expected to finish on time. In the Gompertz simulation the exponential BSSR also struggled a lot in the decreasing hazards scenarios, where it finished on average more than 20 months later than the other BSSR methods.

The Weibull BSSR proved to be much more robust than the exponential method. In the Gompertz misspecification scenarios it was considerably less biased than the exponential method, which resulted in improved trial characteristics. In the decreasing hazards scenarios, a larger number of patients were added, which resulted in markedly reduced trial durations. In the increasing hazards scenarios, fewer unnecessary additional patients were added. However, since the Weibull model was still somewhat biased when the data were generated from a Gompertz distribution, the Weibull BSSR trials characteristics were less favorable here compared to the exponential and Weibull simulation scenarios. In the most extreme decreasing hazards Gompertz scenarios, the trials now took on average about 65 months to finish, compared to about 46 months in the corresponding exponential and Weibull simulation scenarios. In the most extreme increasing hazards Gompertz scenarios, the Weibull BSSR added on average about 400 unnecessary patients, as opposed to about 200 in the corresponding exponential and Weibull simulation scenarios.

As for the second question, the splines BSSR methods were still biased in the Gompertz misspecification scenarios, but less so than the Weibull BSSR method. However, this did not always translate into improved trial characteristics. For more complex 2- and 3-knot spline models overfitting occasional occurred, which sometimes led to extreme outliers in terms of trial duration and additional patient recruitment. The 1-knot spline model, however, did not suffer from this problem as its variance was comparable to that of the Weibull model. Yet, it was less biased than the Weibull BSSR in the Gompertz misspecification scenarios. Consequently, the 1-knot spline BSSR had the best trial characteristics. On average, it finished up to 5 months sooner than the Weibull BSSR method in the strongly decreasing hazards Gompertz scenarios. In the Gompertz scenarios with strongly increasing hazards, the 1-knot spline BSSR added on average about 50 unnecessary patients less than the Weibull. In summary, the 1-knot spline method achieved some improvements in the Gompertz misspecification scenarios compared to the less flexible Weibull BSSR method. Therefore, in practice we recommend that a 1-knot spline model is a useful starting point for a BSSR, when there are a doubts about the validity of a Weibull model.

Regarding the third question, we found that our proposed flexible spline BSSR performed well with our proposed BSSR modification, in which the the pooled estimates from the blinded interim data are used directly for the sample size reestimation. We observed no decrease in performance compared to the splitting method of Friede et al. (2019), which was used for the exponential and Weibull BSSR. In our simulation, however, we did not consider both implementations simultaneously for a given BSSR method. A more direct comparison would be possible if, for example, the exponential or Weibull BSSR would be implemented both with and without splitting of the pooled parameter estimates. Such simulations can be a topic of further research. Here we note that our proposed spline BSSR performed well using only the pooled estimates based on the blinded interim data.

An open question based on our simulations is when complex spline models with two or more internal knots should be used for BSSR. In our study we found that their increased flexibility did not translate into more favorable BSSR characteristics. The problem was that their estimates had a markedly increased variance as well as occasional extreme outliers. This can be problematic in practice, if the BSSR recommendations for an individual trial are based on overfitting of local deviations in the interim data. Therefore, it should be carefully considered whether the additional flexibility is necessary. One setting, in which such flexibility might be necessary is in immuno-oncology trials. In such trials complex hazard shapes may occur due to the specific mechanisms of action of novel treatments (Ouwens et al., 2019). When such complex hazard were simulated, Rutherford et al. (2015) found that spline models with 2-4 internal knots were able to best capture survival. Therefore, in application areas where such complex hazards are likely to occur, spline models with a larger number of knots may be a useful tool for BSSR.

When we simulated data under the null hypothesis (10,000 replications per scenario), we found no indication of an inflation of the type I error rate in any of the simulation scenarios that were considered here. The rejection probability was always close to the nominal significance 5%. This is in line with the findings of previous simulation studies for parametric BSSR (Hade et al., 2010; Friede et al., 2019) and non-parametric BSSR (Todd et al., 2012) for time-to-event trials. Similarly to Friede et al. (2019) and Hade et al. (2010) we also confirmed that when the planning assumptions were correct, the BSSR algorithm typically added few or no patients. However, when the different BSSR methods underestimated the number of events (e.g. in the Gompertz increasing hazards scenario), they tended to regularly add more patients than necessary.

We did not consider misspecification scenarios for the assumed hazard ratio of 0.7. However, previous simulations by Friede et al. (2019) and Hade et al. (2010) found that their parametric BSSR procedures were robust to misspecifications of the assumed hazard ratio. Therefore, it is likely that this also holds for our blinded procedure. In fact, our proposed method should be particularly robust since we do not split up the pooled survival estimates in order to reestimate the sample size.

7.3 Extensions and limitations

One extension relating to the above mentioned issue of increased variance would be to compute bootstrap intervals to assess the variability in sample size estimates. Such procedures have been proposed for example by Hade et al. (2010) for parametric BSSR or by Ying et al. (2004) for predicting analysis time in clinical trials with time-to-event outcomes. Bootstrap intervals are obtained by repeatedly sampling with replacement from the sample at hand. The quantity of interest (e.g. the reestimated sample size) is computed in each sample. Then we obtain a distribution of estimates based on the bootstrap samples and we can compute intervals based on the quantiles of this distribution (Held and Sabanés Bové, 2014, Ch. 3).

In the current thesis, we also considered a hybrid spline BSSR approach as an extension of the fully parametric spline BSSR. Hybrid approaches to survival extrapolation combine the non-parametric Kaplan-Meier estimator for the available interim data with an extrapolated tail that is based on a parametric model. Hybrid methods are useful, because an unbiased, non-parametric estimator can be used for a part of the survival function. In our simulation study, we found that the hybrid method generally performed similarly to the fully parametric spline method. This could be expected, since the flexibility of spline models allows them to closely follow the shape of Kaplan-Meier estimate of the observed interim data. However, we also observed that the hybrid method had an increased variance and more outliers compared to the fully parametric splines BSSR. This was possibly due to the increased variability of the Kaplan-Meier at the end of the available follow-up due to the censoring inherent to an interim analysis. One way to address this issue would be not to use the Kaplan-Meier estimate until the last observed event time, but until same earlier time where more patient data are available. Such an approach has been proposed by Gelber et al. (1993), who suggested to use the median follow-up time. Then, more data would be available and the variability in the tail of the Kaplan-Meier estimate would be expected to be smaller.

Another issue that can arise is when the population at hand is heterogeneous. That is, there might be subgroups of patients who respond less or who do not respond at all to a given treatment (Gallacher et al., 2022). In their simulation study, Rutherford et al. (2020) found that spline models can struggle when there is unobserved heterogeneity in the data. An extension that improved the performance of spline models in these scenarios was to include background mortality. Background mortality can be included in a relative survival framework, in which the overall hazard rate is considered as a sum of the background mortality and the disease related mortality (Nelson et al., 2007). An extension of the Royston-Parmar spline model in a relative survival framework has been developed by Nelson et al. (2007). In their simulation study Rutherford et al. (2020) found that (given the availability of appropriate expected mortality rates) including background mortality in a Royston-Parmar spline model improved the extrapolation performance in the simulation scenario with heterogeneous populations.

Finally, there are a number of limitations to the current study. First, there were some numerical problems with the *flexsurvspline()* function from the *flexsurv* package that was used to fit the Royston-Parmar spline models in our simulation. In about 3 out of 1,000 replications the model fitting algorithm did not converge and the results of the resulting iteration could not be used. The resulting error has been reported by some other researchers online and can apparently be avoided by manually changing the default placement of the internal knots. However, implementing such a procedure for our simulation study was beyond the scope of this thesis. Second, in our simulation study we only considered distributions with constant or monotonous hazards. However, one advantage of spline methods is that they can also capture complex hazards with multiple turning points (Latimer and Adler, 2022). The extrapolation performance of the Royston-Parmar models in scenarios with such complex hazards has previously been investigated by Kearns et al. (2021). However, these simulations were done in the context of health economic evaluations and it would useful to see how such complex hazards affect the performance of spline models in a BSSR setting. Third, in our case study we could only compare the model fits of our various candidate models, but we could not test the proposed flexible BSSR algorithm in practice. For this, a complete dataset including recruitment and event times would have been necessary.

Overall, the current thesis provided evidence for the usefulness of flexible spline methods for BSSR in clinical trials with time-to-event data. Royston-Parmar PH splines models were shown to be a useful generalization of the Weibull model, which can make BSSR slightly more robust in the context of misspecification. In the future, this has the potential to help clinical trials finish on time and at less additional costs, which would benefit both the drug developers and patients in need.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE transactions on automatic control*, 19(6):716–723.

Bagiella, E. and Heitjan, D. F. (2001). Predicting analysis times in randomized clinical trials. *Statistics in Medicine*, 20(14):2055–2063.

Bell Gorrod, H., Kearns, B., Stevens, J., Thokala, P., Labeit, A., Latimer, N., Tyas, D., and Sowdani, A. (2019). A review of survival analysis methods used in nice technology appraisals of cancer treatments: consistency, limitations, and areas for improvement. *Medical Decision Making*, 39(8):899–909.

Benda, N., Branson, M., Maurer, W., and Friede, T. (2010). Aspects of modernizing drug development using clinical scenario planning and evaluation. *Drug Information Journal*, 44(3):299–315.

Brilleman, S. L., Wolfe, R., Moreno-Betancur, M., and Crowther, M. J. (2021). Simulating survival data using the simsurv r package. *Journal of Statistical Software*, 97:1–27.

Chen, T.-T. (2016). Predicting analysis times in randomized clinical trials with cancer immunotherapy. *BMC Medical Research Methodology*, 16(1):1–10.

Collett, D. (2015). *Modelling survival data in medical research*. CRC press, third edition.

Cooper, M., Smith, S., Williams, T., and Ibáñez, R. A. (2022). How accurate are the longer-term projections of overall survival for cancer immunotherapy for standard versus more flexible parametric extrapolation methods? *Journal of Medical Economics*, pages 1–41.

Crowther, M. J. and Lambert, P. C. (2014). A general framework for parametric survival analysis. *Statistics in Medicine*, 33(30):5280–5297.

Donovan, M. J., Elliott, M. R., and Heitjan, D. F. (2006). Predicting event times in clinical trials when treatment arm is masked. *Journal of Biopharmaceutical Statistics*, 16(3):343–356.

Draborg, E., Gyrd-Hansen, D., Poulsen, P. B., and Horder, M. (2005). International comparison of the definition and the practical application of health technology assessment. *International Journal of Jechnology Assessment in Health Care*, 21(1):89–95.

Dudley, W. N., Wickham, R., and Coombs, N. (2016). An introduction to survival statistics: Kaplan-meier analysis. *Journal of the Advanced Practitioner in Oncology*, 7(1):91.

Fang, L. and Su, Z. (2011). A hybrid approach to predicting events in clinical trials with time-to-event outcomes. *Contemporary Clinical Trials*, 32(5):755–759.

Friede, T. and Kieser, M. (2001). A comparison of methods for adaptive sample size adjustment. *Statistics in Medicine*, 20(24):3861–3873.

Friede, T., Nicholas, R., Stallard, N., Todd, S. C., Parsons, N., Valdes-Marquez, E., and Chataway, J. (2010). Refinement of the clinical scenario evaluation framework for assessment of competing development strategies with an application to multiple sclerosis. *Drug Information Journal*, 44(6):713–718.

Friede, T., Pohlmann, H., and Schmidli, H. (2019). Blinded sample size reestimation in event-driven clinical trials: Methods and an application in multiple sclerosis. *Pharmaceutical Statistics*, 18(3):351–365.

Gallacher, D., Kimani, P., and Stallard, N. (2022). Biased survival predictions when appraising health technologies in heterogeneous populations. *PharmacoEconomics*, 40(1):109–120.

Gander, W. and Gautschi, W. (2000). Adaptive quadrature — revisited. *BIT Numerical Mathematics*, 40(1):84–101.

Gelber, R. D., Goldhirsch, A., Cole, B. F., and International Breast Cancer Study Group (1993). Parametric extrapolation of survival estimates with applications to quality of life evaluation of treatments. *Controlled Clinical Trials*, 14(6):485–499.

Gray, J., Sullivan, T., Latimer, N. R., Salter, A., Sorich, M. J., Ward, R. L., and Karnon, J. (2021). Extrapolation of survival curves using standard parametric models and flexible parametric spline models: comparisons in large registry cohorts with advanced cancer. *Medical Decision Making*, 41(2):179–193.

Grumberg, V., Roze, S., Chevalier, J., Borrill, J., Gaudin, A.-F., and Branchoux, S. (2022). A review of overall survival extrapolations of immune-checkpoint inhibitors used in health technology assessments by the french health authorities. *International Journal of Technology Assessment in Health Care*, 38(1):e28, 1—9.

Guyot, P., Ades, A., Ouwens, M. J., and Welton, N. J. (2012). Enhanced secondary analysis of survival data: reconstructing the data from published kaplan-meier survival curves. *BMC Medical Research Methodology*, 12(1):1–13.

Hade, E. M., Jarjoura, D., and Wei, L. (2010). Sample size re-estimation in a breast cancer trial. *Clinical Trials*, 7(3):219–226.

Held, L. and Sabanés Bové, D. (2014). *Applied Statistical Inference*. Springer.

Jackson, C. H. (2016). flexsurv: a platform for parametric survival modeling in R. *Journal of Statistical Software*, 70(8):1–33.

Kappos, L., Bar-Or, A., Cree, B. A., Fox, R. J., Giovannoni, G., Gold, R., Vermersch, P., Arnold, D. L., Arnould, S., Scherz, T., et al. (2018). Siponimod versus placebo in secondary progressive multiple sclerosis (expand): a double-blind, randomised, phase 3 study. *The Lancet*, 391(10127):1263–1273.

Kappos, L., Radue, E.-W., O'Connor, P., Polman, C., Hohlfeld, R., Calabresi, P., Selmaj, K., Agoropoulou, C., Leyk, M., Zhang-Auberson, L., et al. (2010). A placebo-controlled trial of oral fingolimod in relapsing multiple sclerosis. *New England Journal of Medicine*, 362(5):387–401.

Kearns, B., Stevenson, M. D., Triantafyllopoulos, K., and Manca, A. (2021). The extrapolation performance of survival models for data with a cure fraction: a simulation study. *Value in Health*, 24(11):1634–1642.

Kieser, M. and Friede, T. (2003). Simple procedures for blinded sample size adjustment that do not affect the type i error rate. *Statistics in Medicine*, 22(23):3571–3581.

Koehler, E., Brown, E., and Haneuse, S. J.-P. (2009). On the assessment of monte carlo error in simulation-based statistical analyses. *The American Statistician*, 63(2):155–162.

Lambert, P. C. and Royston, P. (2009). Further development of flexible parametric models for survival analysis. *The Stata Journal*, 9(2):265–290.

Lan, Y. and Heitjan, D. F. (2018). Adaptive parametric prediction of event times in clinical trials. *Clinical Trials*, 15(2):159–168.

Latimer, N. R. (2013). Survival analysis for economic evaluations alongside clinical trials — extrapolation with patient-level data: inconsistencies, limitations, and a practical guide. *Medical Decision Making*, 33(6):743–754.

Latimer, N. R. and Adler, A. I. (2022). Extrapolation beyond the end of trials to estimate long term survival and cost effectiveness. *BMJ Medicine*, 1(1):1–4.

Liu, X.-R., Pawitan, Y., and Clements, M. (2018). Parametric and penalized generalized survival models. *Statistical Methods in Medical Research*, 27(5):1531–1546.

McClure, L. A., Szychowski, J. M., Benavente, O., and Coffey, C. S. (2012). Sample size re-estimation in an on-going NIH-sponsored clinical trial: the secondary prevention of small subcortical strokes experience. *Contemporary Clinical Trials*, 33(5):1088–1093.

Moeschberger, M. and Klein, J. P. (1985). A comparison of several methods of estimating the survival function when there is extreme right censoring. *Biometrics*, 41(1):253–259.

Nelson, C. P., Lambert, P. C., Squire, I. B., and Jones, D. R. (2007). Flexible parametric models for relative survival, with application in coronary heart disease. *Statistics in Medicine*, 26(30):5486–5498.

Ou, F.-S., Heller, M., and Shi, Q. (2019). Milestone prediction for time-to-event endpoint monitoring in clinical trials. *Pharmaceutical Statistics*, 18(4):433–446.

Ouwens, M. J., Mukhopadhyay, P., Zhang, Y., Huang, M., Latimer, N., and Briggs, A. (2019). Estimating lifetime benefits associated with immuno-oncology therapies: challenges and approaches for overall survival extrapolations. *Pharmacoeconomics*, 37(9):1129–1138.

Peace, K. E. (2009). *Design and analysis of clinical trials with time-to-event endpoints.* CRC Press.

R Core Team (2017). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria.

Rohatgi, A. (2021). Webplotdigitizer: Version 4.5 [computer software]. Retrieved from https://automeris.io/WebPlotDigitizer.

Royston, P., Lambert, P. C., et al. (2011). *Flexible parametric survival analysis using Stata: beyond the Cox model.* Stata Press.

Royston, P. and Parmar, M. K. (2002). Flexible parametric proportional-hazards and proportional-odds models for censored survival data, with application to prognostic modelling and estimation of treatment effects. *Statistics in Medicine*, 21(15):2175–2197.

Rufibach, K. (2021). *eventTrack: Event Prediction for Time-to-Event Endpoints.* R package version 1.0.1, Retrieved from https://CRAN.R-project.org/package=eventTrack.

Rutherford, M. J., Crowther, M. J., and Lambert, P. C. (2015). The use of restricted cubic splines to approximate complex hazard functions in the analysis of time-to-event data: a simulation study. *Journal of Statistical Computation and Simulation*, 85(4):777–793.

Rutherford, M. J., Lambert, P. C., Sweeting, M. J., Pennington, R., Crowther, M. J., Abrams, K. R., and Latimer, N. R. (2020). NICE DSU technical support document 21. Flexible methods for survival analysis. Retrieved from http://www.nicedsu.org.uk.

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

Schumacher, M. and Schulgen-Kristiansen, G. (2008). *Methodik klinischer Studien: Methodische Grundlagen der Planung, Durchführung und Auswertung.* Springer.

Stallard, N., Hampson, L., Benda, N., Brannath, W., Burnett, T., Friede, T., Kimani, P. K., Koenig, F., Krisam, J., Mozgunov, P., et al. (2020). Efficient adaptive designs for clinical trials of interventions for covid-19. *Statistics in Biopharmaceutical Research*, 12(4):483–497.

Therneau, T. M. (2020). *A Package for Survival Analysis in R*. R package version 3.2.7, Retrieved from https://CRAN.R-project.org/package=survival.

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

Todd, S., Valdés-Márquez, E., and West, J. (2012). A practical comparison of blinded methods for sample size reviews in survival data clinical trials. *Pharmaceutical Statistics*, 11(2):141–148.

Whitehead, J. (2001). Predicting the duration of sequential survival studies. *Drug Information Journal*, 35(4):1387–1400.

Whitehead, J., Whitehead, A., Todd, S., Bolland, K., and Sooriyarachchi, M. R. (2001). Mid-trial design reviews for sequential clinical trials. *Statistics in Medicine*, 20(2):165–176.

Ying, G.-S. and Heitjan, D. F. (2008). Weibull prediction of event times in clinical trials. *Pharmaceutical Statistics*, 7(2):107–120.

Ying, G.-S., Heitjan, D. F., and Chen, T.-T. (2004). Nonparametric prediction of event times in randomized clinical trials. *Clinical Trials*, 1(4):352–361.

A Appendix: Additional Simulation results

A.1 Type I error rates for Weibull and Gompertz distributed data

We found no indication of an increased type I error rate in any of the simulation scenarios considered. The results in the tables below are consistent with the results for the type I error rates in the exponential simulation scenarios (Table 2), which were discussed in Subsection 5.2.2.

Table A1: Simulated type I error rates summarized across the 11 decreasing hazards Weibull simulation scenarios with event probabilities from 20% to 30% (10,000 simulations per scenario). The nominal significance level is 0.05 (two-sided). SD is the standard deviation of the rejection probabilities across the 11 scenarios.

Method	Mean	SD	Minimum	Maximum
Fixed Design	0.0507	0.0017	0.00475	0.0528
Exponential BSSR	0.0508	0.0019	0.0469	0.0528
Weibull BSSR	0.0510	0.0015	0.0482	0.0529
Splines BSSR with 1 knot	0.0504	0.0016	0.0478	0.0527
Splines BSSR with 2 knots	0.0508	0.0012	0.0484	0.0525
Splines BSSR with 3 knots	0.0505	0.0018	0.0471	0.0526

Table A2: Simulated type I error rates summarized across the 11 increasing hazards Weibull simulation scenarios with event probabilities from 20% to 30% (10,000 simulations per scenario). The nominal significance level is 0.05 (two-sided). SD is the standard deviation of the rejection probabilities across the 11 scenarios.

Method	Mean	SD	Minimum	Maximum
Fixed Design	0.0497	0.0018	0.0452	0.0517
Exponential BSSR	0.0500	0.0009	0.0485	0.0511
Weibull BSSR	0.0502	0.0011	0.0485	0.0511
Splines BSSR with 1 knot	0.0502	0.0009	0.0491	0.0519
Splines BSSR with 2 knots	0.0500	0.0015	0.0478	0.0528
Splines BSSR with 3 knots	0.497	0.0016	0.0475	0.0518

Table A3: Simulated type I error rates summarized across the 11 decreasing hazards Gompertz simulation scenarios with event probabilities from 20% to 30% (10,000 simulations per scenario). The nominal significance level is 0.05 (two-sided). SD is the standard deviation of the rejection probabilities across the 11 scenarios.

Method	Mean	SD	Minimum	Maximum
Fixed Design	0.0507	0.0024	0.0473	0.0550
Exponential BSSR	0.0506	0.0022	0.0475	0.0549
Weibull BSSR	0.0502	0.0024	0.0465	0.0548
Splines BSSR with 1 knot	0.0497	0.0021	0.0465	0.0543
Splines BSSR with 2 knots	0.0501	0.0022	0.0468	0.0537
Splines BSSR with 3 knots	0.0502	0.0016	0.0473	0.0533

Table A4: Simulated type I error rates summarized across the 11 increasing hazards Gompertz simulation scenarios with event probabilities from 20% to 30% (10,000 simulations per scenario). The nominal significance level is 0.05 (two-sided). SD is the standard deviation of the rejection probabilities across the 11 scenarios.

Method	Mean	SD	Minimum	Maximum
Fixed Design	0.0497	0.0024	0.0452	0.0531
Exponential BSSR	0.0501	0.0021	0.0469	0.0535
Weibull BSSR	0.0498	0.0016	0.0471	0.0528
Splines BSSR with 1 knot	0.0501	0.0022	0.0465	0.0536
Splines BSSR with 2 knots	0.0498	0.0022	0.0471	0.0536
Splines BSSR with 3 knots	0.0498	0.0024	0.0462	0.0547

Eidesstattliche Erklärung

Ich versichere, dass ich die Arbeit selbständig und ohne Benutzung anderer als der angegebenen Hilfsmittel angefertigt habe. Alle Stellen, die wörtlich oder sinngemäß aus Veröffentlichungen oder anderen Quellen entnommen sind, sind als solche kenntlich gemacht. Die schriftliche und elektronische Form der Arbeit stimmen überein.

Göttingen, den 18. Mai 2022

Tim Mori