Adjustment for exploratory cut-off selection in randomized clinical trials with survival endpoint

Heiko Götte1 | Marietta Kirchner2 | Meinhard Kieser2

1 Merck Healthcare KGaA, Darmstadt, Germany
2 Institute of Medical Biometry and Informatics, University of Heidelberg, Heidelberg, Germany

Abstract

Defining the target population based on predictive biomarkers plays an important role during clinical development. After establishing a relationship between a biomarker candidate and response to treatment in exploratory phases, a subsequent confirmatory trial ideally involves only subjects with high potential of benefiting from the new compound. In order to identify those subjects in case of a continuous biomarker, a cut-off is needed. Usually, a cut-off is chosen that resulted in a subgroup with a large observed treatment effect in an exploratory trial. However, such a data-driven selection may lead to overoptimistic expectations for the subsequent confirmatory trial. Treatment effect estimates, probability of success, and posterior probabilities are useful measures for deciding whether or not to conduct a confirmatory trial enrolling the biomarker-defined population. These measures need to be adjusted for selection bias. We extend previously introduced Approximate Bayesian Computation techniques for adjustment of subgroup selection bias to a time-to-event setting with cut-off selection. Challenges in this setting are that treatment effects become time-dependent and that subsets are defined by the biomarker distribution. Simulation studies show that the proposed method provides adjusted statistical measures which are superior to naïve Maximum Likelihood estimators as well as simple shrinkage estimators.

KEYWORDS
approximate Bayesian computation, cut-off selection, mixed exponential distribution, probability of success, time to event

1 | INTRODUCTION

Predictive biomarkers play an important role in the development of new therapies. In early phases of development, there is often an idea about a biomarker candidate based on the mode of action. Preclinical experiments might have identified a relationship between the biomarker candidate and response to the treatment. Thus, it is assumed that subjects with certain biomarker values have a greater benefit from receiving the new therapy. Ideally, such a biomarker is already included in the design of an early phase clinical trial. However, this requires that the biomarker is identified before start of the trial and the biomarker values can be determined at subject inclusion into the trial. We consider here a situation where the biomarker data are only available for data
analysis. Statistical analyses can only be planned up to a certain degree when using a novel biomarker and/or a novel assay in combination with a new compound in a certain indication. We focus on a continuous biomarker. After identifying a relationship between the biomarker and the treatment effect in a randomized early phase trial, a cut-off point defining the biomarker subset for further development is needed. In the given situation, there is no predefined value to be applied as cut-off point and a data-driven selection is applied.

In general, defining cut-off points is a three-step process: (a) the potential cut-off points have to be determined; (b) the cut-off points have to be evaluated in a training set and one or more cut-off points are selected in a data-driven approach; (c) validation has to be done in an independent validation set. The approach proposed in this manuscript fits between steps (b) and (c): the question is addressed whether it is worth to invest in a validation study, usually a confirmatory trial in the selected biomarker subpopulation. This issue is related to the question about the reliability of exploratory subgroup analyses prior to investing in a subsequent confirmatory trial. Previous work relies rather on nonquantitative criteria (Oxman & Guyatt, 1992; Sun, Briel, Walter, & Guyatt, 2010; Sun, Ioannidis, Agoritsas, Alba, & Guyatt, 2014). Quantitative approaches have looked at what can be expected by chance (Schou & Marschner, 2015) or at probabilities that all reported positive analyses are truly positive ones for a trial reporting at least one positive subgroup effect (Burke, Sussman, Kent, & Hayward, 2015). In order to answer the question about whether or not conducting a confirmatory trial in the selected biomarker subpopulation from a statistical point of view, we consider the following measures: posterior probability of success (PoS) for a future trial, posterior probabilities about the true effect, and treatment effect estimates. Selection-adjusted versions of these measures based on an Approximate Bayesian Computation (ABC) approach were introduced in Götte, Kirchner, Sailer, and Kieser (2017b). With overlapping subsets and data-driven selection procedures, a classical Bayesian approach using Monte Carlo simulations is difficult to implement as the analytical determination of likelihood and/or prior distribution is quite challenging. However, simulation of the subset selection procedure based on a continuous biomarker is rather simple. This is exactly the situation where the ABC approach can be very helpful. It provides an approximation of the posterior distribution by accepting simulation scenarios which are in line with the observed data. Thus, the ABC posterior is a subset of the realizations of the prior distribution of the parameter of interest and reflects all the information about the true parameter given the data. An excellent overview of the ABC approach is given in Sunnäker et al. (2013). In Götte et al. (2017b), several binary biomarker subgroup factors and a binary endpoint were considered. In the present manuscript, this work is extended to cut-off point selection and time-to-event endpoints. Compared to the setting of binary biomarker subgroup factors and a binary endpoint, new challenges are that observed as well as true treatment effects become time-dependent and that overlapping subsets are defined by the biomarker distribution. In Götte et al. (2017b), the computationally intensive ABC approach was compared with the strongly biased Maximum Likelihood (ML) approach. In this manuscript, we consider as a third option a shrinkage estimator which is computationally simple, and which has the intention to account for multiple independent subsets.

The rest of the manuscript is organized as follows. In Section 2, a motivating example illustrates the problem setting. In Section 3, after evaluation of alternative approaches, ML, shrinkage, and the ABC approaches have crystallized as most interesting for application in this manuscript. Posterior PoS for a future trial, posterior probabilities about the true treatment effect, and treatment effect estimates based on these three approaches will be introduced. In Section 4, a simulation study is conducted and application for the motivating example is presented. The manuscript closes with a discussion in Section 5.

2 | MOTIVATING EXAMPLE

We consider a simulated data set which is inspired by a real data problem. In a randomized phase II trial (experimental and control group), overall survival data were observed in 130 subjects with 82 events. Using a standard Cox model, a hazard ratio (HR) of 0.791 ($HR < 1$ favoring the experimental group) resulted. HRs below 0.75 would be considered as relevant. Using a noninformative prior, the posterior probability that the true HR is less than 0.75 is 41%. A future confirmatory phase III trial with 373 events would have a probability of success of 55% (further details about the calculation of these measures will be provided in the next section). Based on these results, further development of the experimental drug in this population might not be regarded as promising. However, data from a potential predictive biomarker were observed as well. There is a biological rationale for an association between the biomarker and the response to treatment. The biomarker values range between 0 and 300. The observed biomarker distribution is shown in the Online Supporting Information Section A.6. We evaluate the functional relationship between the biomarker and the treatment effect (Figure 1) using subpopulation treatment effect pattern plot (STEPP) and fractional polynomials (Sauerbrei, Royston, & Zapien, 2007); further details about these approaches are given in Section A.4).
Both the STEPP and the fractional polynomial approach show a trend toward increasing treatment effects with increasing biomarker values. Bootstrap samples suggest high uncertainty in relationship between biomarker and treatment effect. The STEPP approach suggests rather strong treatment effects ($HR < 0.6$) for values of around 240 and above. Fractional polynomials show more moderate treatment effects; the red curve based on the original data lies above 0.6 throughout. Both approaches show strong effects in the opposite direction for biomarker values below 100. Due to the observed trend, it seems more promising to continue development in a biomarker-restricted population, that is, a population with sufficient potential for marketing authorization (high posterior probability that true effect is at least minimally relevant ($HR < 0.75$) and high PoS for following confirmatory trial). This would require identifying a biomarker cut-off.

With fractional polynomials as well as STEPP, it is difficult to identify a cut-off directly from the corresponding plots. STEPP shows deviations from the overall treatment effect (center group) when excluding the lowest and highest biomarker values. Thus, effects might be diluted if the true cut-off is rather around the median/mean of the biomarker values. In contrast, the fractional polynomials approach is based on a model with best fit to the whole data set. The best fit is influenced by the biomarker distribution. If the true biomarker subgroup contains only 30% of the data, the best model might rather reflect the best fit to the data of the complement of the subgroups while the treatment effects in the biomarker subgroup might not be adequately estimated. For both approaches, only the functional relationship is investigated and treatment effects within biomarker subgroups cannot directly be read from the plots.

Therefore, we will identify the biomarker population by a different data-driven procedure. It is well known that (most) data-driven selection procedures lead to overpromising results. Therefore, the primary goal in this manuscript is to find selection-adjusted versions of the treatment effect estimate, posterior probability, and probability of success.

### METHODS

#### 3.1 General setting

Let us consider a clinical development program which is composed of one exploratory phase II trial and one subsequent confirmatory phase III trial. In both trials, the experimental treatment $E$ is compared to the control $C$ with respect to a time-to-event endpoint, and 1:1 randomization is used. In phase II, besides the primary analysis based on the full population, the primary endpoint is evaluated in subgroups defined by cut-off points of a continuous biomarker. Let $y_{obs,i} = (t_i, cens_i, treat_i, x_i)$ be the data observed for subject $i$ in the phase II trial, $i = 1, \ldots, n$, where $t_i$ is the observed survival time, $cens_i = 1$ if $t_i$ is event time,
treat_i is the assigned treatment group (treat_i = 1 for treatment E, treat_i = 0 for C), and x_i is the value of a continuous biomarker for subject i.

We assume the observed survival time is t = \min(t^*, d). Event times t^* are based on realizations from a Weibull model. The hazard of subject i is

\[ h(t | \text{treat}_i, x_i) = \lambda e^{(a_1 \text{treat}_i + a_2 X_i + a_{12} \mathbb{I}_{x_i > c_x \omega}) \nu t^{\nu - 1}} \]

with shape parameter \( \nu \), scale parameter \( \lambda \), treatment effect \( a_1 \), biomarker effect \( a_2 \) and interaction effect \( a_{12} \). Let \( \mathbb{I}_A \) be the indicator function with support A and \( c_{x \omega} \) is the “true” cut-off point. Independent censoring time \( d \) follows an exponential distribution (Weibull distribution with \( \nu = 1 \)). Note that \( t = \min(t^*, d) \) is also Weibull distributed.

The treatment effect is captured by the negative log HR and is denoted by \( \theta \). Our goal is to investigate the common treatment effect of the analysis population. The individual hazard is constant, increasing or decreasing (no indication for u-shape hazards). HRs are constant over time in the Weibull model. It holds, \( \theta = a_1 \) for subjects with \( x_i \leq c_{x \omega} \) and \( \theta = a_1 + a_{12} \) for subjects with \( x_i > c_{x \omega} \). However, if the analysis population contains subjects from below and above \( c_{x \omega} \) the corresponding HRs or population hazards might not be proportional anymore. As \( c_{x \omega} \) is unknown some of the populations that are evaluated during the cut-off selection procedure might lead to time dependent HRs \( \theta_k = \theta_\nu(t) \) (see Section 3.6.2 for more details). An overview of the notation used in this manuscript is provided in Table 1.

### 3.2 Cut-off selection

Different approaches exist for cut-off selection. Here, a short overview is given. The first approach considers only the biomarker distribution. If the distribution is, for example, bimodal, a natural and often biologically reasonable choice for the cut-off would be one that separates both modes/subdistributions (see, for example, Budczies et al., 2012). The approach has the advantage that no overestimation of treatment effects would be expected. However, as no outcome data are included, this cut-off represents not necessarily the best split in terms of treatment effects. The same applies for cut-offs defined based on prognostic effects. Some approaches are described in Budczies et al. (2012). We will not consider one of these approaches. However, when choosing the potential cut-off points, we take the biomarker distribution into consideration (see below).

Approaches which take the observed treatment effects into account either focus on finding a subset with enhanced treatment effect (e.g., Foster, Taylor, & Ruberg, 2011; here, the approach would reduce to a regression tree with one continuous covariate) or on finding the strongest interaction effect (e.g., He, 2014). He (2014) also proposed using a change-point model based on the functional relationship between biomarker and treatment effect to identify a cut-off.

Assuming there is only one cut-off, all of these approaches lead most probably to very similar results (the subset with the largest treatment effect corresponds often to the maximum interaction effect and the best fitting change-point model). As our primary focus is on finding appropriate measures for decision making (taking data-driven approaches into account), we consider only the approach for cut-off selection which seems to be related to our problem statement: continue development with a primary focus on finding appropriate measures for decision making (taking data-driven approaches into account), we consider only the approach for cut-off selection which seems to be related to our problem statement: continue development with a population with sufficient potential for marketing authorization (high posterior probability that true effect is at least minimum relevant and high PoS for confirmatory trial).

The proposed procedure is as follows: Subgroup factors will be defined by applying single cut-off values to a transformation of \( x_i \) dividing the full population into two subsets, below and above the cut-off. The observed values \( x \) will be transformed to \( z \) by calculating percentiles. Advantage of the percentile approach is that it focuses rather on the number of selected subjects than on differences in \( X \). If the distribution of \( X \) strongly differs from a uniform distribution, large differences in \( X \) might lead to only small differences in the number of selected subjects or vice versa. We consider cut-off points \( c_x \) between 0.3 and 0.7 in 0.1-steps so that there are \( K = 5 \) cut-off points. The corresponding cut-off points \( c_{xx} \) on the \( x \)-scale are defined based on \( \hat{G}(c_{xx}) = c_x \), where \( \hat{G} \) is the empirical distribution function. As cut-off points below 30% and above 70% would lead to either very small subsets (with \( n = 130 \) (82 events) as in the motivating example, there are only around 19 subjects (12 events) per treatment group in these subsets) or to subsets very close to the full population, we will exclude these subsets. Repeating this binary splitting with \( K \) potential cut-off points, there are maximally \( 2K \) overlapping subsets. For each subset, a treatment effect estimate will be determined based on a Cox regression model with treatment as single covariate (see Section 3.5.1). The trial population in phase III is defined based on the selected subset in phase II, the subset with the maximum observed treatment effect.

In order to reflect common practice, we assume that a subgroup population is only selected if the observed effect in this subpopulation is considerably high (\( HR < 0.8 \)) while the complement shows an irrelevant effect (\( HR > 0.85 \)). The true effect should be below 0.75 to be considered as relevant. However, to allow for some random variation and to avoid overstrictness, \( HR < 0.8 \) is selected for the observed effect in the subpopulation.
### TABLE 1 Overview of notation

| General setting, simulation of survival times (compare Sections 3.1 and 4) |
|---------------------------------------------------------------|
| $t_i$ | Observed survival time for subject $i$ |
| $cens_i$ | Censoring factor for subject $i$: 0 = no censoring, 1 = censored observation |
| $d$ | Censoring time |
| $t^*$ | Event time |
| $x_i$ | Value of a continuous biomarker for subject $i$ |
| $\text{treat}_i$ | Treatment group: 1 = treatment E, 0 = control C |
| $y_{obs,i} = (t_i, cens_i, \text{treat}_i, x_i)$ | Observed data for subject $i$ |
| $(\lambda, \nu)$ | Scale and shape parameter of the Weibull model |
| $(a_1, a_2, a_{12})$ | Treatment, biomarker, interaction effect (survival model) |
| $c_x$ | “True” cut-off point |
| $K$ | Number of cut-off points |
| $N_{\text{sim}}$ | Number of simulated “observed” data sets |
| $G(x)$ | Biomarker distribution |
| $t_h$ | Time point for determination of the true effects in the selected subset |

### Quantities of interest and adjusted measures (compare Sections 3.3 and 3.5)

| $\theta_k$ | True treatment effect in subset $k$ |
| $\hat{\theta}_k^{\text{ML}}$ | Maximum Likelihood (ML) estimator in subset $k$ |
| $\hat{\theta}_i$ and $\hat{\theta}_i K_{\text{ML}}$ | ML estimator in subset $i$ and its complement with $i$ or $i + K$ being the index of the selected subset |
| $\hat{\theta}_{\text{select}}$ | True treatment effect in the selected subset |
| $\hat{\theta}_{\text{select}}^{\text{ML}}$ | Future observation of the treatment effect in phase III for the selected subset |
| $\hat{\theta}_{\text{select}}^{\text{ML}}$ | ML estimator in the selected subset |
| $\hat{\theta}_{\text{select}}^{\text{ABC}}$ | ABC-adjusted estimator in the selected subset |
| $\theta_{\text{rel}}$ | Minimum relevant effect |
| $d_k$ | Number of events in subset $k$ |
| $d_{\text{select,III}}$ | Expected events in phase III |
| $q$ | Success criterion applied in phase III, for example, $q = \Phi^{-1}(1 - \alpha) \cdot se(\hat{\theta}_{\text{select}})$ |
| $se(\hat{\theta}_{\text{select}})$ | Standard error of $\hat{\theta}_{\text{select}}$ |

### Specific notation for the shrinkage estimate (compare Section 3.5.2)

| $\tau^2$ | Variance assumed for $\hat{\theta}_i$ and $\hat{\theta}_i K_{\text{ML}}$ |
| $B$ | Weight in the formula for the shrinkage estimator |
| $\hat{\theta}_{\text{weighted}}$ | Observed treatment effect in the full population |
| $V_{\text{shrink}}$ | Variance of posterior distribution |

### ABC-specific notation (compare Section 3.5.3)

| $\pi_{\text{ABC}}(\theta_{\text{select}} | S_{\text{obs}})$ | ABC-based approximation of the posterior distribution for $\theta_{\text{select}}$ given the observed data $S_{\text{obs}}$ |
| $N_{\text{sim,ABC}}$ | Number of ABC simulations |
| $S_{\text{obs}} = (S_{\text{full}}, S_{\text{select}})$ | Summary statistic for the observed data (full = full population, select = selected subset) |
| $s_j = (s_{\text{full}}, s_{\text{select}})$ | Summary statistic for simulation run $j$ |
| $\varepsilon = (\varepsilon_\lambda, \varepsilon_{a_1}, \varepsilon_{\nu})$ | Tolerance for the three parameters scale, treatment effect, shape |
| $w_j$ | Acceptance factor: 1 = accept simulation run, 0 = reject |

### 3.3 | Quantities of interest

Let $\theta_k (k = 1, \ldots, 2K)$ be the true treatment effect in subset $k$. If one of the subsets is selected, then the true effect in the selected subset is $\theta_{\text{select}} = \sum_{k=1}^{2K} \theta_k \mathbb{I}_{\text{subset } k \text{ is selected}}$. As the subset selection is data-driven, $\theta_{\text{select}}$ depends on the observed data. Thus, if a study is conducted with different subjects, a different subset might be selected and $\theta_{\text{select}}$ differs. If we extend this view to studies at a different time, with different study centers, there is also variation or uncertainty around the true treatment effects.
Consequently, we move to a Bayesian perspective and assume that the $2K$-dimensional vector $(\theta_1, \ldots, \theta_{2K})$ follows a joint normal distribution which depends on the observed biomarker distribution $f(x)$ as well as the distribution for $(\alpha, a_1, a_2, a_{12}, \nu)$. Note that these distributions could also reflect a portfolio view, that is, a different treatment for a different biomarker at a different time with different study centers.

As a treatment effect measure, we are interested in two quantities: the conditional expectation for a specific subset $i$, $\theta_{\text{select}} = \theta_i | \text{subset } i \text{ is selected}$, as well as the unconditional expectation $\theta_{\text{select}} = \sum_{k=1}^{2K} \theta_k P(\text{subset } k \text{ is selected})$. In the following, if there is no distinction made, the statements hold for both quantities.

Three quantities based on treatment effect in the selected subset can be relevant for decision making after phase II: the observed treatment effect estimate (or median of the posterior distribution) in the selected subset, $\hat{\theta}_{\text{select}} = \sum_{k=1}^{2K} \hat{\theta}_k P(\text{subset } k \text{ is selected})$, the posterior probability $P(\theta_{\text{select}} > \theta_{\text{rel}} | \text{data})$, and the PoS in phase III $\int_{-\infty}^{\infty} P(\hat{\theta}_{\text{select}} > q | \theta_{\text{select}}) f(\theta_{\text{select}} | \text{data}) d\theta_{\text{select}}$. Here, $\theta_{\text{rel}}$ denotes the minimum relevant effect, $\hat{\theta}_{\text{select}}$ is the future observation of the treatment effect in phase III for the selected subset, and $q$ defines the success criterion applied in phase III. If success is defined as rejecting $H_0$ then $q = \Phi^{-1}(1-\alpha) \cdot \text{se}(\hat{\theta}_{\text{select}})$, where $\Phi^{-1}(1-\alpha)$ is the $(1-\alpha)$-quantile of the standard normal distribution and $\text{se}(\hat{\theta}_{\text{select}})$ is the standard error of $\hat{\theta}_{\text{select}}$.

Note that if selection is based on $(\hat{\theta}_1, \ldots, \hat{\theta}_{2K})$, the mean for a specific selected subset $i$ is $E(\hat{\theta}_{\text{select}} | \text{subset } i \text{ is selected}) = \frac{1}{P(i \text{ is selected} | \theta)} \int_{-\infty}^{\infty} \hat{\theta} \cdot P(i \text{ is selected} | \theta, \hat{\theta}) f(\hat{\theta} | \theta) d\hat{\theta}$, with $\theta$ being the $2K$-dimensional vector $(\theta_1, \ldots, \theta_{2K})$. The mean for any selected subset is $\sum_{k=1}^{2K} E(\hat{\theta}_{\text{select}} | \text{subset } i \text{ is selected}) P(i \text{ is selected} | \theta)$.

In Götte, Kirchner, and Sailer (2017a), more details about the mean for the PoS of the selected subgroup are presented, and the selection bias and how to calculate the selection probabilities for na"ive treatment effect estimates (see Section 3.5.1) are defined. Thus, an adjusted estimate could be derived by subtracting the calculated bias. However, even under the assumption of an approximate normal distribution of the effect estimates determination of the mean and variance of this multivariate normal distribution and furthermore multidimensional integrals are rather difficult to calculate (at some point, a certain kind of simulation is needed).

Therefore, in the following, we provide a short review of potential alternative adjustment methods.

### 3.4 Overview of potential adjustment methods

Due to the data-driven procedure, unadjusted results will be overoptimistic. Therefore, we evaluate different adjustment methods. The overview is not expected to be comprehensive but provides a guidance for the choice of the adjustment methods further evaluated in this manuscript.

A correction for overoptimism of a selection procedure is often performed by approaches for multiplicity adjustment. This is a related but different topic. It focuses on the number of subgroups and/or independent decision bounds (e.g., there is a significance level applied to each subgroup effect). The approaches usually provide adjusted estimates for all subgroups. In the phase II setting, the selection of one or more subgroups for further development would be based on adjusted estimates. Thus, bias of comparing subgroup effects with each other (e.g., selecting the maximum effect) is ignored. Approaches in the Bayesian setting were, for example, proposed by Berger, Wang, and Shen (2014) and Bornkamp, Ohlssen, Magnusson, and Schmidli (2016). These approaches are computationally rather intense. A simpler Bayesian approach for independent (nonoverlapping) subgroups is based on hierarchical models. Computation is quite simple, and effects are shrunk toward the overall mean. However, no adjustment for a specific selection procedure is possible.

Resampling was also proposed. The basic idea is that the distribution for the treatment effect in selected subset cannot be derived in a straightforward way. However, bootstrap approaches, for example, can provide an approximation to this distribution, identify the magnitude of selection bias, and subtract it from the unadjusted estimate (e.g., Foster et al., 2011; Rosenkranz, 2014; Simon & Simon, 2017). The selection procedure is taken into account and it is assumed that all information about the underlying true distributions is reflected in the data. Thus, prior information cannot be included but is also not required. Rosenkranz (2014) discusses that bias could not be completely eliminated with bootstrap approaches (if fixed true effects are assumed). The observed mean squared errors sometimes favor the unadjusted estimates and sometimes the bootstrap-adjusted estimates. Related approaches were proposed by Jiang, Freidlin, and Simon (2007) and He (2014). However, they only consider the permutation distribution of the maximum test statistic. They do not provide adjusted treatment effect estimates.

None of the presented methods combine classical Bayesian analysis with the option to account for the applied selection procedure. ABC adjustment as presented in Götte et al. (2017b) fulfill these criteria. In the scenarios they considered, bias was eliminated if the underlying true treatment effect distribution was also used for their ABC simulation. Mean squared error of ABC estimate was consistently lower than for unadjusted ML estimates. Therefore, we extend their approach to time-to-event setting with cut-off selection. From the presented methods, we select the shrinkage approach as it is computationally simple and allows Bayesian analyses. The basic idea behind the resampling approaches is similar to the ABC approach. However, resampling
does not allow to account for prior beliefs in homogeneous or heterogeneous populations and the level of bias reduction seems to vary. Following the argumentation of Dane, Spencer, Rosenkranz, Lipkovich, and Parke (2019), one can assume that the quality of adjustment depends on how often the most influential data points are included in the bootstrap samples. Thus, the level of adjustment rather reflects whether the observed effects are based on a few or on many observations. In addition, the level of computational complexity is similar to the ABC approach.

A naïve, unadjusted approach will be used for comparison. In the following section, we introduce the three different approaches evaluated in this manuscript: naïve (partial likelihood), shrinkage, and ABC.

3.5 Measures evaluated in this manuscript

We assume there is no explicit prior information about treatment effects on overall survival available. This is often the case as previous phase I trials are usually single arms trials with very limited sample size, short follow-up time, and inclusion of subjects from various indications. Especially, the association with the biomarker was not investigated in phase I. If it is a “first-in-class” drug, there is also no competitor data available.

Therefore, for the ML approach noninformative priors are used. For shrinkage, we use an empirical Bayes approach and exploit the observed data to define prior distributions.

For the ABC approach, more parameters are considered, and consequently more prior distributions are required. Here, available data about the control group could be utilized. Noninformative priors are rather difficult to use when simulating and analyzing individual survival data. Simulated data become “too extreme” and unrealistic and the parameter estimates from fitted models might not exist. As a compromise, uniform priors are used. They consider all values within an interval as equally likely, however, unrealistic outcomes can be avoided by appropriate choice of the intervals. Further details will be provided in Section 5.

3.5.1 Naïve measures

Naïve statistical measures ignoring the selection procedure are based on the ML approach. Cox regression models with treatment as single independent variable are fitted for each subset. These models are the standard for the primary analysis as well as for subgroup analyses in clinical trials with time-to-event endpoints and are therefore used here. The resulting ML estimator as single independent variable are fitted for each subset. These models are the standard for the primary analysis as well as for


developed in this manuscript: naïve (partial likelihood), shrinkage, and ABC.

3.5.2 Measures based on shrinkage estimates

A simple shrinkage estimator \( \hat{\theta}_{\text{select, shrink}} \) is derived from a hierarchical model. In order to determine \( \hat{\theta}_{\text{select, shrink}} \), we consider the ML estimators from Section 3.5.1: the selected subset, \( \hat{\theta}_{\text{select, ML}} \), and its complement denoted by \( \hat{\theta}_{\text{select, ML}^c} \) with \( i \) or \( i + K \) being the index of the selected subset. We assume the expectations for both estimates follow the same distribution: \( \hat{\theta}_k \sim N(\theta, \tau^2), k \in \{i, i + K\} \). We use an empirical Bayes approach and estimate \( \theta \) and \( \tau^2 \) based on the current data. An estimator for \( \theta \) is the weighted sum of \( \hat{\theta}_{\text{ML}} \) and \( \hat{\theta}_{\text{select, ML}} \) (see below). For estimating \( \tau^2 \), we consider methods from meta-analysis. The ML-estimator of the selected subset, \( \hat{\theta}_{\text{select, ML}} \), and of its complement as well as the corresponding asymptotic variances are used, that is, \( \hat{\theta}_{\text{ML}}, \hat{\theta}_{\text{select, ML}}, \text{var}(\hat{\theta}_{\text{ML}}), \text{var}(\hat{\theta}_{\text{select, ML}}) \). We use the “SJ”-estimator (“model error variance estimator”) (Sidik & Jonkman, 2005, 2007) obtained from the function \texttt{rma} from the R package \texttt{metafor} for calculating \( \tau^2 \). This estimator has the advantage that it does not tend to underestimate the true between-heterogeneity (Veroniki et al., 2015). Thus, if there are actually
differential effects among subgroups, $\hat{\theta}_{\text{select shrink}}$ would have less tendency for overcorrection. The shrinkage estimator for the selected subset is a weighted average of the observed treatment effect in the selected subset and in the full population:

$$\hat{\theta}_{\text{select shrink}} = \hat{\theta}_{\text{select ML}} \cdot (1 - B) + \hat{\theta}_{\text{weighted}} \cdot B$$

with $B = \frac{\text{var}(\hat{\theta}_{\text{ML}})}{\text{var}(\hat{\theta}_{\text{ML}}) + \tilde{\tau}^2}$ or $\frac{\text{var}(\hat{\theta}_{\text{KML}})}{\text{var}(\hat{\theta}_{\text{KML}}) + \tilde{\tau}^2}$, respectively, and $\hat{\theta}_{\text{weighted}} = \frac{(\text{var}(\hat{\theta}_{\text{KML}}) + \tilde{\tau}^2) \hat{\theta}_{\text{ML}} + (\text{var}(\hat{\theta}_{\text{KML}}) + \tilde{\tau}^2) \hat{\theta}_{\text{KML}}}{\text{var}(\hat{\theta}_{\text{ML}}) + \tilde{\tau}^2 + \text{var}(\hat{\theta}_{\text{KML}}) + \tilde{\tau}^2}$.

The posterior distribution is a normal distribution with mean $\hat{\theta}_{\text{select shrink}}$ and variance $V_{\text{shrink}} = B \times \left( \tilde{\tau}^2 + B \times \left( \frac{1}{\text{var}(\hat{\theta}_{\text{ML}}) + \tilde{\tau}^2} + \frac{1}{\text{var}(\hat{\theta}_{\text{KML}}) + \tilde{\tau}^2} \right)^{-1} \right)$ (Neuenschwander, Wandel, Roychoudhury, & Bailey, 2015). Thus, posterior probability and PoS are calculated based on the same formula as for the ML estimator, just replacing $\hat{\theta}_{\text{select ML}}$ with $\hat{\theta}_{\text{select shrink}}$ and $\text{var}(\hat{\theta}_{\text{ML}})$ with $V_{\text{shrink}}$.

Note that this shrinkage estimator ignores the additional subsets that could potentially be selected as well as the used selection procedure. If a different selection procedure identifies the same subset, then $\hat{\theta}_{\text{select shrink}}$ would still be the same. In contrast, the ABC-adjusted estimator described in the next subsection takes all overlapping subgroups as well as the used selection procedure into account.

Nevertheless, we include this shrinkage estimator to evaluate whether the computational effort associated with the ABC estimate is actually needed or whether a simpler estimator would provide similar performance in the setting considered here.

### 3.5.3 | Measures based on ABC-adjusted estimates

We extend the ABC-adjusted measures previously introduced in Götte et al. (2017b) to a time-to-event setting. The ABC-adjusted measures are based on an approximation of the posterior distribution for $\theta_{\text{Select}}$ given the observed data which are denoted as $p_{\text{ABC}}(\theta_{\text{Select}} | \mathbf{S}_{\text{obs}})$, where the observed data are summarized in a statistic $\mathbf{S}_{\text{obs}}$.

Below is a high-level overview of an algorithm to derive ABC-adjusted measures. Further details will be provided in Section 4. In addition, an algorithm for practical implementation is provided in Section A.4. Let $c_{\text{sh}}$ be the “true” cut-off point, and $\epsilon$ be a tolerance parameter for accepted simulations. $N_{\text{sim ABC}}$ simulations will be performed including the following steps:

1. Sample from prior distributions for parameters in the Weibull model $(\lambda_j, a_{1j}, a_{2j}, a_{12j}, \nu_j)$ for $j = 1, \ldots, N_{\text{sim ABC}}$.
2. Generate data: $y_{\text{sim},ij} = (t_{ij}, cens_{ij}, treat_{ij}, x_{ij})$ with the observed time $t_{ij} = \min(t_{ij}^*, \text{censoring time})$ and $cens_{ij} = 1$ if $t_{ij}$ is event time. The theoretical event time $t_{ij}^*$ is based on the Weibull hazard model.

$$h(t^{*}|\text{treat}_{ij}, x_i) = \lambda e^{(a_{1j}\text{treat}_{ij} + a_{2j}x_i)}(\nu_i t_i^{\nu_i - 1})$$

3. Determine the selected subset based on $y_{\text{sim},ij}$ according to the predefined selection rule and calculate summary statistics $s_j = (s_{\text{full},j}, s_{\text{select},j})$. Use the same selection rule to determine $\mathbf{S}_{\text{obs.select}}$ and $\mathbf{S}_{\text{obs.temperature}}$ based on the observed data.
4. If $\mathbf{S}_{\text{obs.select}} \in [s_{\text{select},j} - \epsilon, s_{\text{select},j} + \epsilon]$ and $\mathbf{S}_{\text{obs.temperature}} \in [s_{\text{full},j} - \epsilon, s_{\text{full},j} + \epsilon]$, then $w_j = 1$ (accept); otherwise $w_j = 0$ (reject).
5. Determine adjusted measures:
   a. $\hat{\theta}_{\text{select ABC}}$ is the median of the posterior distribution $p_{\text{ABC}}(\theta_{\text{select}} | \mathbf{S}_{\text{obs}})$, that is, $\hat{\theta}_{\text{select ABC}} = \inf_{\theta_{\text{select}}} \left\{ \frac{1}{N_{\text{sim ABC}} \mathbf{w}_j} \sum_{j=1}^{N_{\text{sim ABC}}} w_j 1_{\theta_j \leq m} \geq 0.5 \right\}$.
   b. The adjusted posterior probability is determined by $\inf_{\theta_{\text{select}}} \left\{ \frac{1}{N_{\text{sim ABC}} \mathbf{w}_j} \sum_{j=1}^{N_{\text{sim ABC}}} w_j 1_{\theta_j > \theta_{\text{rel}}} \right\}$.
   c. The adjusted PoS is determined by $\int_{-\infty}^{\infty} P(\hat{\theta}_{\text{select}} > \Phi^{-1}(1 - \alpha) se(\hat{\theta}_{\text{select}})) f_{\text{ABC}}(\theta_{\text{select}} | \mathbf{S}_{\text{obs}}) d\theta_{\text{select}}$, where $f_{\text{ABC}}(\theta_{\text{select}} | \mathbf{S}_{\text{obs}}) d\theta_{\text{select}}$ is determined by $\int_{-\infty}^{\infty} P(\hat{\theta}_{\text{select}} > \Phi^{-1}(1 - \alpha) se(\hat{\theta}_{\text{select}})) w_j$.

### 3.6 | Specific aspects of ABC approach

The steps needed to determine ABC adjusted measures require further explanation (see Section 3.5.3). We will further elaborate on steps 4 and 5.
3.6.1 | Choice of acceptance criteria

In Step 4, summary statistics and distance measures are needed. In the binary setting, the summary statistics \( s_j \) used to select the samples from the simulation data set, that is, \( w_j = 1 \), were the response rates in both treatment arms, for the selected as well as for the full population (Götte et al., 2017b). The analog statistics for the time-to-event setting are the estimated hazard rates, that is, the number of events divided by the sum of observation times. These are the ML estimates for the hazard under an exponential distribution. The exponential distribution assumes constant hazard over time, it is a special case of the Weibull distribution with \( \nu = 1 \).

For the Weibull model with fixed \( \nu \), there is a generalization involving the sufficient statistic \( \sum_i r_i \) with \( j \in \{ C, E \} \). However, assuming \( \nu \) follows a random distribution to our knowledge no such sufficient summary statistic exists. Therefore, we compare the parameter estimates from the Weibull model:

\[
\begin{align*}
   s_{\text{select}} &= (\hat{\lambda}_{\text{select}}, \hat{a}_{1\text{select}}, \hat{\nu}_{\text{select}}) \\
   s_{\text{full}} &= (\hat{\lambda}_{\text{full}}, \hat{a}_{1\text{full}}, \hat{\nu}_{\text{full}}).
\end{align*}
\]

In order to determine the necessary number of ABC simulation runs \( N_{\text{sim\,ABC}} \) as well as appropriate values of \( \epsilon = (\epsilon_\lambda, \epsilon_{a_1}, \epsilon_{\nu}) \), we follow the approach of Götte et al. (2017b) and consider a setting without selection where alternative ways to determine posterior distributions are available. We fit a Weibull regression model with treatment as the single covariate (further details are given in the Online Material A1). We evaluated \( N_{\text{sim\,ABC}} = 100,000 \) as well as 1,000,000, \( \epsilon = (\epsilon_\lambda, \epsilon_{a_1}, \epsilon_{\nu}) \) being either \((\infty, 0.01, \infty), (\infty, 0.005, \infty), (0.05, 0.005, 0.15), (0.025, 0.005, 0.1), (0.01, 0.005, 0.1)\). We found that \( N_{\text{sim\,ABC}} = 100,000 \) as well as 1,000,000 and \((\epsilon_{a_1}, \epsilon_{a_2}, \epsilon_{\nu}) = (0.025, 0.005, 0.1)\) provide posterior distributions which are in line with Weibull models based on an Markov Chain Monte Carlo (MCMC) approach (see Figures A1 in the online material).

Therefore, we will use \( N_{\text{sim\,ABC}} = 1,000,000 \) and \((\epsilon_{\lambda}, \epsilon_{a_1}, \epsilon_{\nu}) = (0.25, 0.005, 0.1)\) as a starting point for the selected subset as well as the full population. As there are additional restrictions in the selection rule (see Section 3.2) and more parameter/prior distributions are involved (see Section 4.2), there will be less simulation samples left for the ABC posterior than in the above simplified setting with treatment as single covariate. Therefore, in Section 4 we will further investigate the choice of \((\epsilon_{\lambda}, \epsilon_{a_1}, \epsilon_{\nu})\) if \( N_{\text{sim\,ABC}} = 1,000,000 \) is fixed for computational reasons.

3.6.2 | Calculation of true effects

In order to get an estimate of the posterior distribution \( \pi_{\text{ABC}}(\theta_{\text{select}}|s_{\text{obs}}) \) in step 5 we need to determine the true negative log(HR) for the selected subset, \( \theta_{\text{select}}(t) \).

In Section 3.2, we define cut-off points \( c_{xx} \) on the x-scale based on \( \hat{G}(c_{xx}) = c_{xx} \), where \( \hat{G} \) is the empirical distribution function. In order to calculate the true hazards, we need to consider the theoretical percentiles: \( Z = G(X) \). For a given “true” cut-off \( c_{xx} \) on the x-scale there is a corresponding “true” cut-off on \( z \), called \( c_{xx}; c_{xx} = G(c_{xx}) \). The model from Section 3.1 becomes

\[
   h(t|\text{treat}_i, z_i) = \lambda e^{(a_1 \cdot \text{treat}_i + a_2 \cdot (z_i > c_{xx}) + a_{12} \cdot \text{treat}_i \cdot (z_i > c_{xx}) \cdot t^{1-1}).
\]

We are interested in \( \theta_{h}(t) = -\log \left( HR(t|\psi_k(z)) \right) = -\log(\text{h(t|1, } \psi_k(z)) + \log(\text{h(t|0, } \psi_k(z))) \right), \) where \( \psi_k(z_i) = \{ z_i > c_k \} \) (for \( k = 1, \ldots, K \)) or \( \psi_k(z_i) = \{ z_i < c_k \} \) (for \( k = K + 1, \ldots, 2K \)), respectively, is a subset of the full population defined based on covariate \( z \).

Note that \( HR \) is constant over time if all subjects in the subset \( \psi_k(z) \) have the same hazard function. Once there is a mixture of subjects with biomarker values below and above the “true” cut-off point \( c_{xx} \), \( HR(t) \) becomes time-dependent. The reason is that the probability that a random subject belongs to one of the populations \( p_1(t) \) or \( p_2(t) \), that is, falls above or below \( c_{xx} \), changes over time. Subjects from a population with greater hazard experience events earlier than the subjects from a population with lower hazard. Therefore, \( p_1(t) \) and \( p_2(t) \) are time-dependent. In Mendenhall (1957), it is shown that the hazard for a random subject from the mixture distribution is a weighted sum of the population hazards, where the weights are \( p_1(t) \) and \( p_2(t) \). The calculation of the hazards involves an integration over \( Z \) which requires a comparison of the observed cut-off \( c_k \) and the “true” cut-off \( c_{xx} \) (see further details about the calculation in Section A.2).

In summary, \( HR(t|\psi_k(z)) \) depends on the time point \( t \), on the “true” cut-off point, on the selected cut-off point as well as on \((\lambda, a_1, a_2, a_{12}, \nu) \). We investigated how all these aspects interact by considering a limited set of parameter constellation (for example, assuming constant individual hazards \( \nu = 1 \)).
As a result, the variation in $HR(t)$ for a fixed cut-off point is rather low (see Figures A2a and A2b). For example, if a predictive biomarker is nonprognostic, then the impact of time-dependency can be neglected. Greatest variation in $HR(t)$ is observed for large effects in both $a_2$ and $a_{12}$. The greater the distance to the “true” cut-off, the greater is the variation in $HR(t)$; maximum variation is given for the full population. If the treatment effect is the greatest in the subset with the worst prognosis, then impact of time-dependency is greatest (compare Figure A2a bottom right).

In the following, we will consider one fixed time point for the ABC-adjustment. This time point is varied in our simulations to investigate the impact of the choice of this parameter on the adjusted measures.

4 | APPLICATION

A simulation study compares the distribution of the different estimators of treatment effect, posterior probability, and PoS for the simulation between estimated and true values. We consider as true treatment effect $\theta_{select}$ (see Section 3.6.2) and as true PoS the power value $P(\theta_{select} > \Phi^{-1}(1 - a) \se(\theta_{select})|\theta_{select})$. Knowing the true treatment effect for the selected population, $\theta_{select}$, means that it is also known whether or not $\theta_{select}$ is greater than $\theta_{rel}$. Therefore, we compare $\frac{\theta_{select}}{\theta_{rel}}$ with the estimated posterior probability. Note that for the comparison of the different estimator, we consider the unconditional expectation from Section 3.3: $\theta_{select} = \sum_{k=1}^{2K} \theta_k P(subset \: k \: is \: selected)$.

For the ABC-estimator, we focus on the choice of the time point for determination of the true effect in the selected subset as well as the impact of the choice of the prior distributions. In addition, we evaluate whether $(\epsilon_{\lambda}, \epsilon_{a_1}, \epsilon_{a_{12}}) = (0.025, 0.005, 0.1)$ is an appropriate choice.

First, we describe the general setup that is common for the ABC simulation and the observed data generation. Then we describe the assumptions which are adapted for the observed data simulation in order to investigate its impact. After describing the selection procedure, we present the results of the simulation study and come back to the motivating example and present the adjusted measures for that data set.

4.1 | Setup for ABC and observed data simulation

We consider a simulation setup which is in line with the motivating example. We assume that 130 subjects in the full population will be randomized in 13 months with a constant recruitment rate. The final analysis, where the full population as well as all subgroups are evaluated, will be performed in each simulation after 82 events occurred in the full population (must be at least 1 month after end of recruitment) or 48 months after first subject randomized, whatever occurs first. Using the formula of Schoenfeld (1983), this number of events would be needed for a log-rank test with 1:1 randomization, one-sided $\alpha = .1$ and power 80% assuming a HR of 0.625 (median time to event 10 months in control, 16 months in experimental treatment group). Based on these assumptions and a 5% drop out rate after 15 months (if no events would occur), the number of events are expected to be reached after 13 months of follow-up.

For the potentially following confirmatory trial in the selected population, we assume 1:1 randomization, one-sided $\alpha = .025$ and power 90% assuming a HR of 0.714 (median time to event 10 months in control, 14 months in experimental). This leads to a required number of 373 events. As minimum relevant effect, $\theta_{rel}$, we use $-\log(0.75)$.

For the comparison of the different estimates based on “observed data,” $N_{sim} = 5.000$ simulations are performed. Treatment group allocation is generated from a Bernoulli distribution with probability $\tau$. The hazard $h(t^*|treat_i, x_i) = \lambda e^{(a_{treat_i} + a_1 x_i + a_{12} x_{12}) t^*}$ is determined based on realizations of the prior distributions for $(\lambda, a_1, a_2, a_{12}, v)$, the distribution of $x_i$ as well as the “true” cut-off point $c_{xoa}$. These parameters are subject to variation and will therefore be described in the following subsection.

With the above hazard function, we generate the theoretical event times $t^*$ based on random samples from an exponential distribution with parameter $-\log(1 - 0.05)/15$. The observed time is the minimum of $t^*$, drop out time, and time between trial entry and clinical cut-off date when 82 events occurred.

4.2 | Differences between ABC simulation and observed data simulation

Several assumptions are needed for the ABC simulation whose validity is unknown in reality. This encompasses the biomarker distribution $G$, the prior distribution for the parameter $(\lambda, a_1, a_2, a_{12}, v)$ as well as the “true” cut-off $c_{xoa}$. In the first step, we use the same assumptions for the generation of the observed data as for the ABC simulation (Scenario 1); this reflects the
ideal situation. Then, we vary the assumptions for the observed data generation (Scenarios 2 and 3). This allows us to check the robustness of the ABC adjustment to inappropriate assumptions. As indicated in Section 3, we will use uniform prior distributions for all parameter except for $\lambda$ in Scenario 1 in order to reflect the uncertainty due to lack of appropriate prior information. Scenario 2 considers prior distribution with greater variances/range than assumed for the ABC simulations while Scenario 3 considers prior distributions with smaller range than assumed for the ABC simulations. Scenario 3 reflects the idea that the ABC prior distributions are wider than the actual distributions. Motivation, further details for the choice of the distributions below as well as plots of the prior distributions will be provided in Section A.3.

In each scenario, we consider different choices for the time point $t$ for determination of the true effects in the selected subset $\theta_{sel}(t)$. In the following, we will denote this time point with $t_h$.

Scenario 1 (identical to ABC simulation):

$c_{x0} = 150$, $\lambda \sim \text{Gamma}(\text{shape} = \frac{\log(0.5)}{10} \cdot 150, \text{scale} = \frac{1}{150})$, $a_1 \sim U(\log(0.6), \log(1.4))$, $a_2 \sim U(\log(0.5), \log(2))$, $a_{12} \sim U(\log(0.5), \log(1.4))$, $\nu \sim U(0.65, 1.35)$. With respect to the distribution function of $X$, $G(x)$, three different Beta distributions are considered leading to different sizes of the true subgroup:

1. $G(x) \sim \text{Beta}(1, 1)$ (true subgroup consists of 50% of overall population),
2. $G(x) \sim \text{Beta}(3, 2)$ (true subgroup consists of 69% of overall population),
3. $G(x) \sim \text{Beta}(2, 3)$ (true subgroup consists of 31% of overall population).

Scenario 2 (different assumptions than for ABC simulation):

$c_{x0} = 110$, $\lambda \sim \text{Gamma}(\text{shape} = \frac{\log(0.5)}{10} \cdot 200, \text{scale} = \frac{1}{200})$, $a_1 \sim N(\log(0.8), \frac{9}{40})$, $a_2 \sim N(\log(0.8), \frac{1}{6})$, $a_{12} \sim N(\log(0.8), 0.38)$, $\nu \sim U(1.05, 1.25)$

1. $G(x) \sim \text{Beta}(3, 2)$ (true subgroup consists of 86% of overall population),
2. $G(x) \sim \text{Beta}(2, 3)$ (true subgroup consists of 53% of overall population),
3. $G(x) \sim \text{Beta}(1, 1)$ (true subgroup consists of 63% of overall population).

Scenario 3 (different assumptions than for ABC simulation):

$c_{x0} = 110$, $\lambda \sim \text{Gamma}(\text{shape} = \frac{\log(0.5)}{10} \cdot 200, \text{scale} = \frac{1}{200})$, $a_1 \sim N(\log(0.8), \frac{4}{350})$, $a_2 \sim N(\log(0.8), \frac{4}{100})$, $a_{12} \sim N(\log(0.8), \frac{1}{80} + \frac{1}{70} + \frac{1}{80} + \frac{1}{70})$, $\nu \sim U(1.05, 1.25)$

1. $G(x) \sim \text{Beta}(3, 2)$ (true subgroup consists of 86% of overall population),
2. $G(x) \sim \text{Beta}(2, 3)$ (true subgroup consists of 53% of overall population),
3. $G(x) \sim \text{Beta}(1, 1)$ (true subgroup consists of 63% of overall population).

In all scenarios, three different time points $t_h = 5, 10, 15$ are applied.

### 4.3 Results of simulation study

Starting with $(\epsilon_\lambda, \epsilon_{a_1}, \epsilon_v) = (0.025, 0.005, 0.1)$ as suggested in Section 3.6.1 lead to good ABC approximation for Scenario 1.3. However, more than 50% of simulations scenarios lead to two or less observations in the ABC posterior (data not shown here). We further increase $(\epsilon_\lambda, \epsilon_{a_1}, \epsilon_v)$ to $(0.07, 0.006, 0.17)$. The mean difference does not change between estimated and true values on two decimal places. However, the root mean squared error for PoS and posterior probabilities improve while the median number of observations increase to 66 (first and third quartile: 18 and 174). We will therefore use $(\epsilon_\lambda, \epsilon_{a_1}, \epsilon_v) = (0.07, 0.006, 0.17)$ for all other scenarios.

Figure 2 shows boxplots of the difference between estimated and true values for Scenario 1.3. Plots for Scenarios 1.1 and 1.2 are very similar (compare Figures A4 and A7). In Scenario 1, where the distribution to generate the data is identical to the ABC prior distribution, the ABC approach provides very good bias correction as well as reduced variation for the treatment effect estimate compared to the ML and shrinkage approaches (mean and root mean squared error can be found in Table A1).

Thus, using the values chosen for $N_{simABC}$ and $\epsilon$ seem appropriate. When looking at Scenarios 2 and 3, there is a slight overcorrection by the ABC approach (Figures 3 and 4, Figures A5, A6, A8, A9, and Tables A2 and A3). It ranges for the treatment effect between 0.01 for Scenario 2.1 to −0.13 for Scenario 2.2. For Scenario 3, there is always an overcorrection by the ABC approach. This can be explained by the prior distributions. For the ABC simulations, the assumptions for the effects were rather negative compare to Scenario 3 (compare Figure A3). Thus, in terms of mean bias, the ABC approach is sensitive
FIGURE 2  Boxplots for difference between treatment effect estimate, PoS, and posterior probability and corresponding true value for Scenario 1.3. Time points 5, 10, and 15 are considered for ML, Shrinkage, and ABC. Diamonds represent observed mean differences.

FIGURE 3  Boxplots for difference between treatment effect estimate, PoS, and posterior probability and corresponding true value for Scenario 2.3. Time points 5, 10, and 15 are considered for ML, Shrinkage, and ABC. Diamonds represent observed mean differences.

to the choice of the prior distributions. However, for all scenarios the ABC adjusted estimates show lower variation compared to ML and shrinkage approach. In general, sensitivity analyses with different prior distributions can support the interpretation of ABC adjusted estimates.

The shrinkage estimators provide quite consistent results over all simulation scenarios. The bias of the ML estimate for all three measures is cut in half by the shrinkage estimator. Thus, an ad hoc bias corrected estimate could be \( \hat{\theta}_{select_{adhoc}} = \hat{\theta}_{select_{shrink}} - (\hat{\theta}_{select_{ML}} - \hat{\theta}_{select_{shrink}}) \). In terms of HR this means: \( HR_{select_{adhoc}} = \min\left(1, HR_{Full}, \frac{(HR_{select_{shrink}})^2}{HR_{select_{ML}}}\right) \) (assuming that the estimate tends in the opposite direction, that is, \( HR > 1 \) or is less than the effect in the full population is probably not appropriate). However, as we have only considered a very narrow range of simulation scenarios, this estimate could only be used
as a ball park figure in similar settings. The ABC estimator would provide more adequate correction of the selection bias but requires more assumptions and more computation time. Considering all scenarios, the impact of the time point $t_h$ is rather low.

### 4.4 Analyses for motivating example

Subgroup analysis as described in the simulation section was performed for data set from Section 2. Further details and results are presented in the Online Supporting Information. The largest treatment effect is observed for the "above 213 (70% quantile)"-subset: $HR = 0.449$. The effect in the complement "below 213 (70% quantile)"-subset is 1.030. Overall, treatment effects above cut-off increase, that is, $HR$ decrease, with increasing biomarker values.

The approach used so far does not account for the data-driven selection procedure. In order to provide selection-adjusted estimates for the “above 213 (70% quantile)”-subset, ABC posteriors are based on simulation runs where it holds true that the observed $HR$ for the selected subgroup is smaller than 0.80 and larger than 0.85 for the complement. In a further analysis, we are considering in addition only simulation runs where the “greater than 70%-quantile” was selected. Thus, we provide the conditional as well as the unconditional expectations as defined in Section 3.3.

The biomarker distribution is rather right-skewed (see plot A10), we set $X \sim \text{Beta}(3, 2) \cdot 300$ as prior distribution and $c_{x_{0}} = 180$. In the control group below the cut-off, there are 36 events and the sum of observation times is 599. We assume that there is no indication for a strong treatment effect below the cut-off and no indication for a strong prognostic effect. However, given the high level of uncertainty without appropriate prior knowledge we choose the following quantities (similar to scenario 1): ABC prior: $c_{x_{0}} = 180$, $\lambda \sim \text{Gamma}(\text{shape} = 36, \text{scale} = \frac{1}{600}), a_{1} \sim U(\log(0.6), \log(1.4)), a_{2} \sim U(\log(0.5), \log(2)), a_{12} \sim U(\log(0.5), \log(1.4)), \nu \sim U(0.65, 1.35)$.

The parameter from the Weibull model based on the example data are

\[
\begin{align*}
    s_{\text{select}} &= \left( \hat{\lambda}_{\text{select}} = 0.088, \hat{a}_{1,\text{select}} = -0.816, \hat{\nu}_{\text{select}} = 0.984 \right) \\
    s_{\text{full}} &= \left( \hat{\lambda}_{\text{full}} = 0.051, \hat{a}_{1,\text{full}} = -0.210, \hat{\nu}_{\text{full}} = 1.099 \right).
\end{align*}
\]

Note that $\exp(\hat{a}_{1,\text{select}}) = 0.442$ and $\exp(\hat{a}_{1,\text{full}}) = 0.811$.

Table 2 shows the results for shrinkage and ABC analysis (identical for time points 5, 10, 15; success = one-sided $p$-value < .025).

The ABC-adjusted $HR = 0.703$ is in line with the ad hoc estimate based on the shrinkage estimate, $0.571^2/0.449 = 0.726$. If there is no biological rational to expect $HR > 1$ in the “Biomarker < 213”-subset and if there is a lack of alternative treatments, further development of the complement could be reasonable.
## 5 | DISCUSSION

Identification of the target population for confirmatory trials is an essential part of early clinical development programs. We considered a development program where the target population might be defined by one continuous potential predictive biomarker. Appropriate statistical measures are needed to decide whether or not to continue development with a large confirmatory trial in the selected biomarker population.

We showed that ABC-adjusted measures account for the inherent selection procedure and are on average close to their true values in case of appropriate choice of prior distributions. As shown in the simulation study, several assumptions are needed to implement the approach. However, we also showed that simpler shrinkage estimators do not provide an adequate level of bias correction. Shrinkage to the observed mean of the full population is low when the selected subgroup is large and/or the effect in the subgroup and its complement strongly differ.

The higher the number of potential subsets is, the higher is the bias for the subset with the maximum observed effect. Thus, a high number of potential subsets increase the chance of strongly differential effects between selected subgroup and complement. However, when the selected subgroup and its complement show strong differences, the level of shrinkage is rather low. Therefore, the shrinkage approach does not adequately correct the selection bias when several cut-off points are considered.

This effect is even more pronounced for an alternative shrinkage approach in our setting of overlapping subgroups: create the \(K+1\) independent subsets, calculate adjusted shrinkage estimates \(\hat{\theta}_{\text{shrink},m}\) for each independent subset \(m\) with \(m = 1, \ldots, K+1\), and calculate \(\hat{\theta}_{\text{select,shrink}}\) for the selected subgroup based on an inverse variance weighted average of the corresponding \(\hat{\theta}_{\text{shrink},m}\). With 82 events overall, subsets based on the proposed percentile approach can become very small (~8 events expected). This leads to high variation among the subsets (and questionable normal approximation of the estimates). Therefore, the level of shrinkage to the overall mean is very low. Preliminary simulations (not shown here) supported this assumption, therefore, we did not further follow-up on this.

The ABC-adjusted measures adjust for the selection procedure. The selection rule can be refined depending on information from previous investigations. For example, if there is a good biological rational that high biomarker values lead to larger treatment effects, then a population would only be selected if the maximum effect is observed in an “above”-subgroup. Further guidance for application is provided in Section A.4. Overall, in each simulation run different subsets can be selected. In practice, there might be interest in getting adjusted measures specific for the selected subset. In such a case, only simulation runs with the selected subset (for example “above 70% quantile”) would be considered (see also the Discussion in Götte et al., 2017b).

In this work, we considered a rather simple selection (maximum observed effect with additional restrictions). As described in Section 3.3, there are other approaches to find a biomarker cut-off point and to select a target population with time-to-event endpoints. Extensions of the ABC adjustment to these approaches are straightforward.

We saw that subgroup selection in the time-to-event setting may lead to a mixture population where hazards are not proportional anymore between treatment groups. The aspect of time-dependency in mixture distributions needs to be considered for the ABC approach when the continuous biomarker is assumed to be prognostic as well as predictive.

We assumed the observed data were generated from a Weibull model and used the Weibull model for the ABC simulations. Extensions to other parametric time-to-event models are possible. Like in our approach here, ABC samples can be accepted if observed model parameters are within a certain range of the model parameter from the ABC simulation.
We considered only approaches which use all observations for selection of the cut-off point and the estimation of the resulting subgroup effect, respectively. An alternative approach would be to split the phase II data into training and validation set. However, due to the reduced sample size in the training set the probability of selecting the “correct” subgroup decreases. A treatment effect estimator for the biomarker subset from the validation set would be unbiased but has a large variation (with 82 events in the full population, 5–10 events are included in each treatment arm if 40% of the data are used for the validation set and 50% of subjects are in the selected biomarker subset). Thus, the rMSE (\(\sqrt{\text{bias}^2 + \text{variance}}\)) of the unbiased estimator from the validation set would probably be greater than 0.45 (\(\sqrt{0.2^2 + 4/20} \approx 0.45\)). Using cross-validation, the variation could be reduced. However, the ABC approach would nevertheless be the preferred option: It uses all data, the ABC estimator would only be moderately biased under inappropriate choice of prior and has at the same time small variation. For example, for Scenario 2 the rMSE ranges between 0.25 and 0.34 (see Table A1). Of course, these values are influenced by the distributions used to generate the data (Scenario 3 has even lower rMSE values). However, to bring these number into perspective, a validation set of 50 events would be required to have a corresponding rMSE of \(\sqrt{4/50} \approx 0.28\). This would allow only 32 events for selection. Thus, the ABC approach has a higher probability of selecting the “correct” subgroup while at the same time the resulting treatment effect estimate has higher quality compared to the “splitting into training and validation set”-approach.

ACKNOWLEDGMENTS

The authors would like to thank the reviewers and the Associate Editor for their very helpful and constructive comments, which greatly improved this manuscript, as well as Manuel Cornes, Merck KGaA, for making the R-code more efficient.

CONFLICT OF INTEREST

The authors have declared no conflict of interest.

ORCID

Heiko Götte https://orcid.org/0000-0002-6305-4466
Marietta Kirchner https://orcid.org/0000-0003-0294-2483
Meinhard Kieser https://orcid.org/0000-0003-2402-4333

REFERENCES

Berger, J. O., Wang, X., & Shen, L. (2014). A Bayesian approach to subgroup identification. Journal of Biopharmaceutical Statistics, 24(1), 110–129.
Bornkamp, B., Ohlssen, D., Magnusson, B. P., & Schmidli, H. (2016). Model averaging for treatment effect estimation in subgroups. Pharmaceutical Statistics, 16, 133–142.
Budczies, J., Klauschen, F., Sinn, B. V., Győrffy, B., Schmitt, W. D., & Denkert, C. (2012). Cutoff Finder: A comprehensive and straightforward web application enabling rapid biomarker cutoff optimization. PLoS ONE, 7(12), e51862.
Burke, J. F., Sussman, J. B., Kent, D. M., & Hayward, R. A. (2015). Three simple rules to ensure reasonably credible subgroup analyses. British Medical Journal, 351, h5651.
Dane, A., Spencer, A., Rosenkranz, G., Lipkovich, I., & Parke, T. on behalf of the PSI/EFSPI Working Group on Subgroup Analysis. (2019). Subgroup analysis and interpretation for phase 3 confirmatory trials: White paper of the EFSPI/PSI working group on subgroup analysis. Pharmaceutical Statistics, 18, 126–139.
Foster, J. C., Taylor, J. M. C., & Ruberg, S. J. (2011). Subgroup identification from randomized clinical trial data. Statistics in Medicine, 30, 2867–2880.
Götte, H., Kirchner, M., & Sailer, M. O. (2017a). Probability of success for phase III after exploratory biomarker analysis in phase II. Pharmaceutical Statistics, 16, 178–191.
Götte, H., Kirchner, M., Sailer, M. O., & Kieser, M. (2017b). Simulation-based adjustment after exploratory biomarker subgroup selection in phase II. Statistics in Medicine, 36, 2378–2390.
He, P. (2014). Identifying cut points for biomarker defined subset effects in clinical trials with survival endpoints. Contemporary Clinical Trials, 38, 333–337.
Jiang, W., Freidlin, B., & Simon, R. (2007). Biomarker-adaptive threshold design: A procedure for evaluating treatment with possible biomarker-defined subset effect. Journal of the National Cancer Institute, 99, 1036–1043.
Mendenhall, W. III. (1957). Estimation of parameters of mixed exponentially distributed failure time distributions from censored life test data. Institute of Statistics Mimeos Series. Retrieved from http://www.stat.ncsu.edu/information/library/mimo.archive/ISMS_1957_171.pdf
Neuenschwander, B., Wandel, S., Roychoudhury, S., & Bailey, S. (2015). Robust exchangeability designs for early phase clinical trials with multiple strata. Pharmaceutical Statistics, 15, 123–134.
Oxman, A. D., & Guyatt, G. H. (1992). A consumer’s guide to subgroup analyses. Annals of Internal Medicine, 116, 78–84.
Rosenkranz, G. (2014). Bootstrap corrections of treatment effect estimates following selection. *Computational Statistics and Data Analysis, 69*, 220–227.

Sauerbrei, W., Royston, P., & Zapien, K. (2007). Detecting an interaction between treatment and a continuous covariate: A comparison of two approaches. *Computational Statistics & Data Analysis, 51*, 4054–4063.

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

Schou, I. M., & Marschner, I. C. (2015). Methods for exploring treatment effect heterogeneity in subgroup analysis: An application to global clinical trials. *Pharmaceutical Statistics, 14*, 44–55.

Sidik, K., & Jonkman, J. N. (2005). Simple heterogeneity variance estimation for meta-analysis. *Journal of the Royal Statistical Society, C54*, 367–384.

Sidik, K., & Jonkman, J. N. (2007). A comparison of heterogeneity variance estimators in combining results of studies. *Statistics in Medicine, 26*, 1964–1981.

Simon, R., & Simon, N. (2017). Inference for multimarker adaptive enrichment trials. *Statistics in Medicine, 36*, 4083–4093.

Sun, X., Briel, M., Walter, S. D., & Guyatt, G. H. (2010). Is a subgroup effect believable? Updating criteria to evaluate the credibility of subgroup analyses. *British Medical Journal, 340*, c117.

Sun, X., Ioannidis, J. P. A., Agoritsas, T., Alba, A. C., & Guyatt, G. (2014). How to use a subgroup analysis. *Journal of the American Medical Association, 311*, 405–411.

Sunnäker, M., Busetto, A. G., Numminen, E., Corander, J., Foll, M., & Dessimoz, C. (2013). Approximate Bayesian computation. *PLoS Computational Biology, 9*, e1002803.

Tsiatis, A. A. (1981). A large sample study of Cox’s regression model. *Annals of Statistics, 9*, 93–108.

Veroniki, A. A., Jackson, D., Viechtbauer, W., Bender, R., Bowden, J., Knapp, G., … Salanti, G. (2015). Methods to estimate the between-study variance and its uncertainty in meta-analysis. *Research Synthesis Methods, 7*, 55–79.

SUPPORTING INFORMATION

Additional supporting information including source code to reproduce the results may be found online in the Supporting Information section at the end of the article.

**How to cite this article:** Götte H, Kirchner M, Kieser M. Adjustment for exploratory cut-off selection in randomized clinical trials with survival endpoint. *Biometrical Journal*. 2020;62:627–642. [https://doi.org/10.1002/bimj.201800302](https://doi.org/10.1002/bimj.201800302)