Aus dem
Deutschen Krebsforschungszentrum in Heidelberg
Abteilung Biostatistik
(Abteilungsleiter/in: Prof. Dr. Annette Kopp-Schneider)

Strategies for cancer clinical trials with multiple biomarkers.

Inauguraldissertation
zur Erlangung des Doctor scientiarum humanarum (Dr.sc.hum)
an der
Medizinischen Fakultät Heidelberg
der
Ruprecht-Karls-Universität

vorgelegt von

Christina Habermehl geb. Beisel

aus
Darmstadt
2017

Dekan: Prof. Dr. Wolfgang Herzog
Doktormutter: Prof. Dr. Annette Kopp-Schneider

In memory of my father.

Contents

List of Tables

List of Figures

List of Abbreviations

AMLSG	German-Austrian Acute Myeloid Leukemia Study Group
CC	complete case
cf.	confer
DKFZ	Deutsches Krebsforschungszentrum (German Cancer Research Center)
EGFR	epidermal growth factor receptor
Exp	experimental treatment
FLT3	fibromyalgia syndrome-like tyrosine kinase 3
HR	hazard ratio
i.i.d.	independent identically distributed
ITD	internal tandem duplication
JAV	just-another-variable
LR	logistic regression
MAR	missing at random
MCAR	missing completely at random
MNAR	missing not at random
MSE	mean squared error
NPM1	Nucleophosmin-1
pdf	probability density function
PH	proportional hazards
PI	passive imputation
PMM	predictive mean matching
PLR	polytomous logistic regression
RCT	randomized controlled trial
RMSE	root mean squared error
Std	standard-of-care
vs.	versus

Chapter 1

Introduction

Molecular characterization of tumor tissue by omics technologies has been extensively used to unravel the underlying biological mechanisms of cancer and to identify novel predictive biological markers ('biomarkers'), expecting that biomarker-guided treatment decisions for individual patients improve the effectiveness of these treatments. This development imposes the need for new design and analysis concepts of clinical trials that test multiple experimental medical interventions according to the molecular phenotype of a patient in parallel.

1.1 Incorporation of biomarkers into clinical trial design

Traditionally, clinical trials are subdivided into phases I, II, and III. After each phase a decision is made whether the experimental treatment performed well enough to move on to the next phase or, after phase III, if it can be approved for clinical use. However, during the past decade the requirements for clinical trials have started to expand, inducing a development away from the traditional trial setups such as a simple randomized design (Fig. 1.1). New design concepts are being developed that aim to keep up with the rapidly increasing number of potential biomarker-guided medical interventions that need to be evaluated. This poses several challenges and difficulties for the design of clinical trials. The traditional study designs require long time lines and usually only evaluate one experimental treatment at a time. To evaluate the increasing amount of potential medical interventions more efficiently, various modified trial designs have been proposed that also address issues that can arise when incorporating biomarkers into clinical trials.

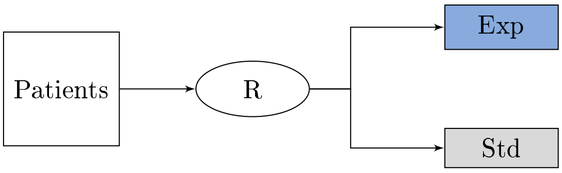


Figure 1.1: Standard randomized design

1.1.1 What is a 'biomarker'?

A biomarker is, as defined by the National Institutes of Health Biomarkers Definitions Working Group, "a characteristic that is objectively measured and evaluated as an indicator of normal biological processes, pathogenic processes, or pharmacologic responses to a therapeutic intervention" (Downing 2001). A biomarker can be something as simple as a patient's blood pressure or it can be more complex, like a certain mutation in a cancer cell, or other characteristics which need elaborate laboratory testing (Strimbu and Tavel 2010). A distinction is made between prognostic and predictive biomarkers. Both are baseline characteristics, but while a prognostic biomarker categorizes patients by the degree of the outcome of interest, a predictive biomarker categorizes patients by the degree of response to a certain treatment or therapy (Gosho et al. 2012).

Prognostic biomarkers are associated with the outcome of a disease regardless of the therapy which was used. While they make it possible to group patients by likely outcome after treatment with standard therapy, they cannot be used to guide the choice of treatment for a particular patient (Gosho et al. 2012). The validation of prognostic biomarkers is rather straightforward and can usually be done retrospectively (Mandrekar and Sargent 2009a). **Predictive biomarkers**, on the other hand, should be validated through a prospective study. They are associated with the response to a certain treatment and hence, ideally, offer the possibility to prospectively identify patients who are likely to benefit from a certain treatment (Mandrekar and Sargent 2009b). The goal of validating predictive biomarkers is to ultimately be able to select an optimal therapy from several options (Mandrekar and Sargent 2009a).

Mandrekar and Sargent (2009b) state that if time and money are sparse, a retrospective validation of a predictive biomarker can be considered. However, it is crucial to use data from a randomized controlled trial (RCT) to ensure comparability of biomarker-positive and biomarker-negative patients. Data from a non-randomized trial (e.g. cohort or single-arm studies) are not suitable for this purpose, since causal effects of the biomarker of interest on the treatment effect cannot be isolated from other potentially confounding factors (Mandrekar and Sargent 2009b). The focus in this thesis will be on prospective validation of predictive biomarkers.

1.1.2 Recent design approaches

Various design approaches have been proposed for predictive biomarker validation and oftentimes, there is more than one term for a specific study design. For consistency purposes, the following sections will use the terminology and definitions used by Mandrekar and Sargent (2009b).

One possibility to incorporate biomarkers into a clinical trial design and to investigate a biomarker-guided therapeutic intervention are so-called **enrichment designs**. In this type of design patients are screened for their biomarker profile as they enter the study. If they test positive for the biomarker of interest they are included in the study, otherwise they are excluded (cf. Figure 1.2). After this screening and selection step, the study proceeds with the biomarker-positive patients only, who are then randomized between an experimental therapy and standard of care. Hence, enrichment designs are essentially simple randomized designs within a subpopulation with an added screening step in the beginning of the study to identify patients belonging to the subpopulation of interest.

The primary hypothesis of enrichment designs typically tests the clinical benefit of the investigated treatment in the biomarker-positive subpopulation only. Hence, this design should only be used if there is reliable evidence that only a subgroup of patients will benefit from the experimental treatment. Otherwise, a benefit for the remainder of the patient population may go undetected. An enrichment design is appropriate and ethical if an experimental therapy only has a moderate benefit for the entire population but a high toxicity or if inclusion of biomarker-negative patients is ethically impossible based on findings of previous studies. Additionally, the assay reproducibility and accuracy should be well established, i.e. the determination of a patient's biomarker status should be reliable and reproducible (Mandrekar and Sargent 2009b).

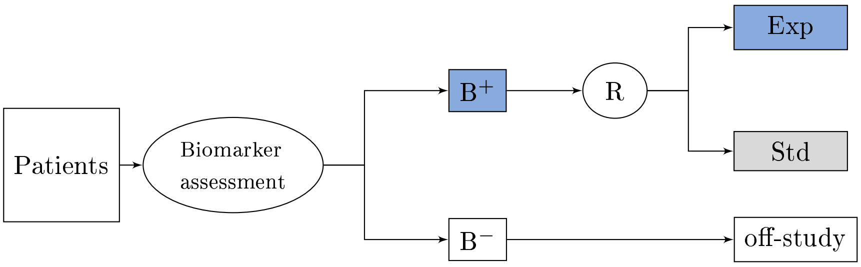


Figure 1.2: Enrichment design

Besides enrichment designs, there are also **all-comers designs**, where all patients meeting the (non biomarker related) eligibility criteria are included in the study, independent of their biomarker status. There are different types of all-comers designs, including the biomarker-based strategy design, biomarker-by-treatment interaction design, hybrid design, and sequential testing strategy design (Mandrekar and Sargent 2009b).

Just as the name **biomarker-based strategy design** suggests, this type of design aims to compare the strategy of a biomarker-guided treatment against randomizing between experimental therapy and standard of care, independent of biomarker status (cf. Figure 1.3). For this purpose, patients are randomized between biomarker-guided and non-biomarker-guided strategy in the beginning of the study. Here, biomarker-guided treatment strategy means that biomarker-positive patients are treated with the experimental therapy and biomarker-negative patients with standard of care. This is based on the assumption that only biomarker-positive patients benefit from the experimental treatment under investigation. Patients in the non-biomarker-guided group are randomized between experimental therapy and standard of care, without assessing their biomarker status.

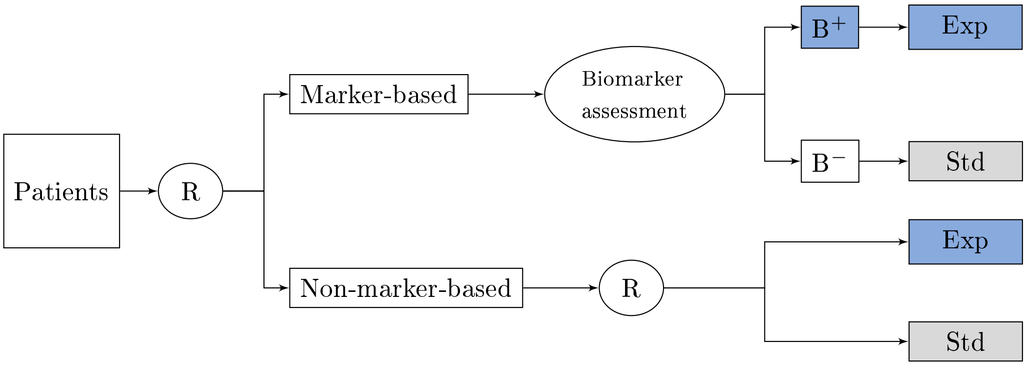


Figure 1.3: Biomarker-based strategy design

The biomarker-based strategy design primarily tests the difference in treatment outcome between the two treatment strategies. There are some discussions regarding the drawbacks of this type of design. In both arms there are biomarker-positive and biomarker-negative patients receiving one of the two on-study treatments. This causes an overlap of patient groups receiving the same treatment within the biomarker-guided and non-biomarker-guided arms. Thus, this type of design is usually less efficient than other randomized designs (Mandrekar and Sargent 2009b). Additionally, the treatment effect of the therapy and the prognostic effect of the biomarker cannot be distinguished (Gosho et al. 2012).

An alternative design, which does not have this overlap issue, is the **biomarker-by-treatment interaction design**, also called **biomarker-stratified design**. This design compares the benefit of the experimental therapy in the biomarker-positive population against the benefit in the biomarker-negative population. Just as for an enrichment design, the patients are screened for their biomarker status upon entering the study and are assigned to a biomarker-positive or a biomarker-negative group accordingly. Within these groups, patients are randomized between experimental therapy and standard of care.

This type of design can be used if there is not enough evidence that the investigated therapy only benefits the biomarker-positive population, as it investigates the experimental therapy within the entire population, stratified by biomarker status (Mandrekar and Sargent 2010). The primary hypothesis of the biomarker-by-treatment interaction design usually either tests the interaction between biomarker and treatment, or alternatively, it tests the treatment benefit separately in each biomarker-group (An et al. 2012).

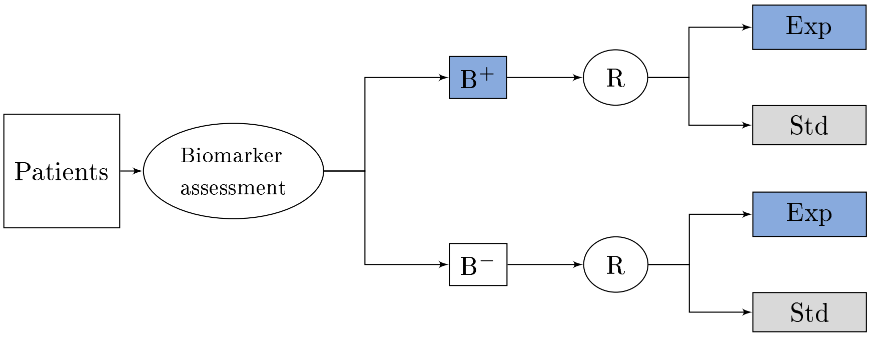


Figure 1.4: Biomarker by treatment interaction design

If it is not possible to treat biomarker-negative patients with the experimental therapy due to ethical constraints, but one still wishes to collect data for these patients, one could use a **hybrid design**, which is a mixture of the enrichment design and the biomarker-by-treatment interaction design. This hybrid design randomizes biomarker-positive patients between treatments, but rather than excluding the biomarker-negative patients, they are kept in the study and treated with standard of care (cf. Figure 1.5).

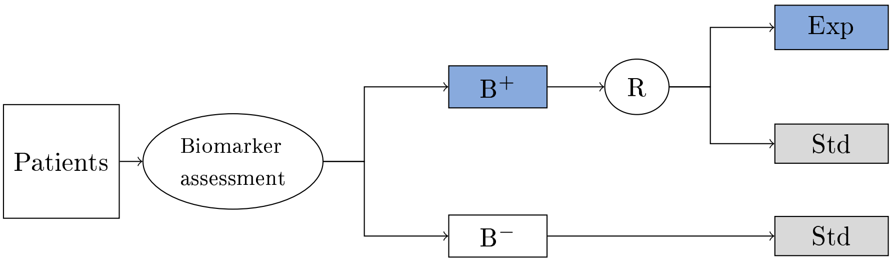


Figure 1.5: Hybrid design

Mandrekar and Sargent (2009b) describe this design as similar to an enrichment design, but providing additional value by including and collecting specimens and follow-up from all patients. This allows using the collected data for retrospective testing for other prognostic biomarkers later on. However, like the enrichment design, the primary hypothesis of the hybrid design only tests the benefit of the investigated treatment in the biomarker-positive population, and hence it is only powered to detect differences in outcomes in this subpopulation (Mandrekar and Sargent 2009b).

In their paper, Mandrekar and Sargent (2009b) discuss another type of design to investigate the treatment effect within the entire population as well as the biomarker-positive subpopulation: the **sequential testing strategy design**. Its general design is similar to a biomarker-by-treatment interaction design, but it allows testing in the overall population as well as in the subpopulation and it is usually based on one of two different testing strategy options. If the experimental treatment is expected to be broadly effective, it is tested in the entire population first and afterwards in the (prospectively defined) biomarker-positive subpopulation. If there is strong prior evidence that the effect of the experimental treatment is much stronger in the biomarker-positive population, the treatment is tested in the subpopulation first (given that the investigated biomarker has a sufficient prevalence). If the analysis within the subpopulation yields significant results, the entire population is tested as well (Mandrekar and Sargent 2009b). These strategies for testing in the overall as well as the subpopulation can help to avoid that a subpopulation that benefits from a new treatment may go unidentified. This could happen when a new treatment is tested in the overall population but only the biomarker-positive patients benefit from the treatment, or if the treatment is only tested in the biomarker-positive subpopulation but it would also benefit the biomarker-negative patients. For analysis strategies where two (or more) hypotheses are tested, the issue of multiple testing can be addressed by utilizing so-called closed testing procedures (Mandrekar and Sargent 2009b, Millen and Dmitrienko 2012).

1.1.3 Clinical trial designs with multiple biomarkers

The next step after incorporating a single biomarker into a clinical trial is to incorporate multiple biomarkers. The motivation behind trials testing multiple biomarkers simultaneously is saving resources compared to running separate trials for each of the biomarkers. Additionally, multiple biomarker trials offer treatment options to a larger percentage of patients who undergo a biomarker-screening, which makes the trial more attractive to potential participants.

Renfro and Sargent (2016) describe three different approaches to multiple biomarker trials, which they refer to as master protocols: trials that view biomarkers as a refinement of a certain tumor type (umbrella trials) and trials that understand biomarkers as a replacement of the tumor type and which therefore recruit patients independent of histology (basket trials). Master protocols that fit neither of these descriptions are referred to as platform trials. The detailed definitions of these three trial concepts can slightly differ, sometimes leading to a certain trial being classified differently, depending on the author. For consistency purposes, the following sections will use the definitions of Renfro and Sargent (2016).

In their paper, Renfro and Sargent (2016) define **basket trials** as trials which are based on the hypothesis that certain biomarkers predict the response to a corresponding treatment better than the tumor type. Therefore, patient eligibility is independent of histology and the 'baskets' patients are assigned to are defined solely by biomarkers (Figure 1.6). Nevertheless, basket trials are typically not entirely independent of tumor type - they can, for example, be restricted to solid tumor types.

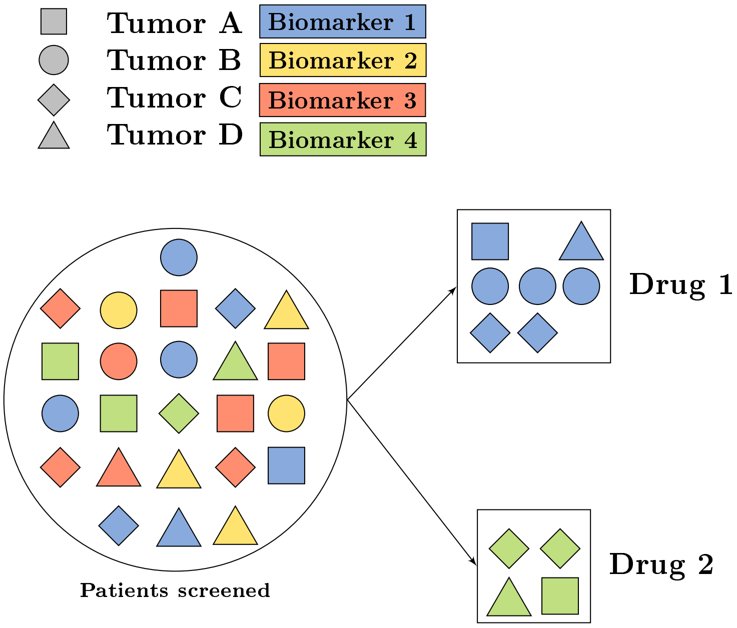


Figure 1.6: Study scheme for a basket trial [1].

[1] Adapted from https://www.bhdsyndrome.org/forum/bhd-research-blog/genetic-sequencing-approaches-to-cancer-clinical-trials

Due to different standard treatments across tumor types, basket trials usually do not include randomization to a standard arm. Examples for basket trials are NCI MATCH (Mullard 2015) and SIGNATURE (Kang et al. 2015), which are both still ongoing (Renfro and Sargent 2016).

Basket trials can be described as "an efficient way of screening experimental therapeutics across multiple patient populations in early-phase drug development" (Mandrekar et al. 2015). Hence, one advantage discussed by Renfro and Sargent (2016) is that these trials offer biomarker-guided treatment for a great variety of tumor types, often even for rare tumor types for which a standalone (randomized) clinical trial would not be possible.

While these advantages sound compelling, Mandrekar et al. (2015) criticize the uncertainty that prevails regarding the statistical planning and analysis of basket trials. Their main point of criticism is the lack of justification of sample sizes, but they also demand taking more measures to take into account the inter-patient and inter-tumor heterogeneity. Additionally, they call for more awareness regarding multiple testing issues. However, above all there is the major limitation as pointed out by Renfro and Sargent (2016): The underlying assumption that biomarkers can predict response to a targeted therapy independent of the tumor type is still just a hypothesis; not a proven concept.

Umbrella trials on the other hand do not rely on this hypothesis, as enrollment is generally restricted to one tumor type (Figure 1.7). Patients are centrally screened and assigned to one of several biomarker-defined subtrials, which can be randomized or single-arm. Examples for umbrella trials include FOCUS4 (Kaplan et al. 2013), ALCHEMIST (Gerber et al. 2015), and LUNG-MAP (Ferrarotto et al. 2015). Again, all of these studies are still ongoing (Renfro and Sargent 2016).

An advantage over basket trials that Renfro and Sargent (2016) point out is that the restriction to a specific tumor type makes umbrella trials less susceptible to inter-tumor heterogeneity. Additionally, inference regarding the considered tumor type can be drawn more easily and, given there is randomization between the experimental and standard treatments, prognostic and predictive effects of the biomarkers can be investigated.

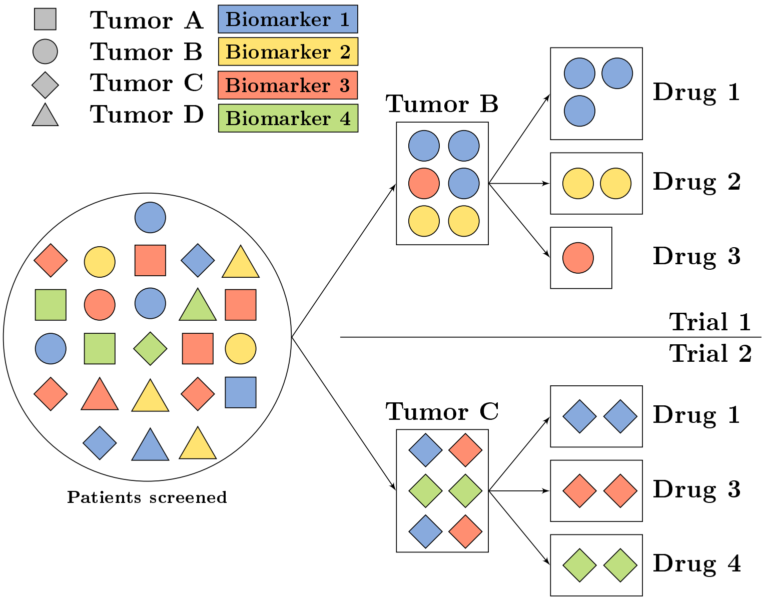


Figure 1.7: Study scheme for an umbrella trial [2].

However, the restriction to a single tumor type can also be a disadvantage: if a rare tumor is subdivided into even smaller biomarker-groups, there could be issues with sufficient accrual and overall progress of the trial (Renfro and Sargent 2016).

Master protocols that do not fit either the basket or the umbrella type trials are referred to as **platform trials** by Renfro and Sargent (2016). According to their definition, platform trials typically comprise a randomized study design with common control arm and many experimental treatment arms that are dynamically added to the study and which can be closed again based on futility or efficacy. Examples for platform trials are SHIVA (Le Tourneau et al. 2015), NCI-MPACT (Do et al. 2015), the BATTLE trials (Liu and Lee 2015), I-SPY2 (Park et al. 2016), and CUSTOM (Lopez-Chavez et al. 2015).

Unlike the aforementioned basket and umbrella trials, some of the listed platform trials have already been completed. One of these trials is the SHIVA trial, whose design was similar to a basket trial, i.e. patient accrual across different tumor types, but it additionally included a randomized comparison between targeted therapy and physician's

[2] Adapted from https://www.bhdsyndrome.org/forum/bhd-research-blog/genetic-sequencing-approaches-to-cancer-clinical-trials

choice. The results of the SHIVA trial were published in 2015 by Le Tourneau et al. (2015). They reported that they were not able to detect a significant difference in progression free survival of the targeted treatment strategy versus treatment with physician's choice. In their paper, they identify four key issues of their trial. Three are of a more logistical nature: usage only of targeted agents marketed in France, treatment with mostly single agents instead of combination therapy, and finally, not being able to revise the assigned therapy (to react to developments/mutations within a patients' tumor). Beyond these three issues, they criticize that the treatment algorithm used was 'unidimensional', meaning that it did not take into account the potential interaction of coexisting biomarkers with the assigned treatment (Le Tourneau et al. 2015).

Another completed platform trial is the CUSTOM trial, again with recruitment across tumor types, where patients were assigned to one of several experimental therapies according to their basket (Lopez-Chavez et al. 2015). Patients that could not be assigned to one of the baskets were treated with standard of care and followed up until death. Each basket was treated as an independent phase II trial, with at least 40% response rate as primary endpoint. Lopez-Chavez et al. (2015) reported that one of the investigated drugs achieved its primary endpoint, one other did not. For all other drugs, completion of accrual was deemed unfeasible. According to the authors, the main weaknesses of the CUSTOM study were the low patient numbers due to the low prevalence of the chosen biomarkers and the lack of an adaptive design. In particular, they would have liked to be able to react to the latest developments by adding new biomarker-arms and/or new treatments to the ongoing study (Lopez-Chavez et al. 2015).

All types of master protocols discussed in this section are quite challenging to plan and execute, since they require a close collaboration between "multiple industry, academic, regulatory, and community oncology stakeholders, often including participation by multiple pharmaceutical companies providing drugs to the same trial" (Renfro and Sargent 2016). The results of the screening need to be sufficiently reliable and the screening process should be financially feasible. Still, the overall number of patients that can be screened is usually limited, simultaneously limiting the number of patients in the individual biomarker-groups. Therefore, it is advantageous to be able to offer experimental treatment to as many patients as possible. If available, offering an experimental therapy to the biomarker-negative patients may help with accrual (Renfro and Sargent 2016).

In summary, attention should be paid to expected sample sizes in the biomarker-positive groups and inclusion of biomarker-negative patients should be considered when appropriate. Another issue with master protocols, as discussed by Le Tourneau et al. (2015) regarding the SHIVA trial, is that patients may qualify for more than one of the biomarker-groups. Therefore, for these cases there should be clear rules regarding the allocation to the biomarker-groups, possibly taking into consideration known interactions between biomarkers and treatments. Finally, due to the rapid developments in the area of biomarkers and targeted therapies, master protocols should be flexible enough to react accordingly, as Lopez-Chavez et al. (2015) concluded from their CUSTOM study.

Aims of this PhD Thesis

The goal of this thesis is to develop and examine the design and analysis of an umbrella-type cancer clinical trial with a time-to-event outcome as primary endpoint. This trial should comprise multiple biomarker-defined subgroups, each testing a distinct experimental therapy versus standard-of care. The design should be flexible enough to accommodate new biomarker-based subgroups as new information (from internal or external data as well as from expert knowledge) emerges. Of primary interest is a proof of efficacy of the biomarker guided strategy versus standard of care, and evaluation of specific multi-arm subgroups, which may be defined by prognostic risk or biological functioning. The focus will be on three issues which arise in multiple biomarker trials.

The first issue considered is **low prevalence** of the biomarkers. As more and more biomarkers are discovered, patient populations are further and further subdivided into biomarker-defined subpopulations. Hence, the biomarkers of interest may have a prevalence that is too low to analyze the data for these subpopulations individually. For this situation, it is aimed to investigate the evaluation of the biomarker-guided treatment strategy rather than evaluating each group separately.

As a consequence of the low prevalence of the biomarkers, a large number of **biomarker-negative patients** should be expected at the screening stage, which is the second issue considered in this thesis. It is aimed to investigate whether inclusion of these patients in the trial and the analysis provides additional benefit, such as improvement of power or reduction of bias.

The third issue considered is the constant discovery of new biomarkers and corresponding biomarker-guided experimental therapies. It is aimed to be able to react to these continuous developments by investigating options to **add new biomarkers** and corresponding therapies to an ongoing study.

For this purpose an umbrella-type study design is proposed, where enrollment is restricted to a single tumor type. Upon entering the study, each patients' biomarker status is assessed with regard to the biomarkers included in the study (cf. Figure 1.8). According to these results, the patients are assigned to the respective biomarker-defined groups or to a biomarker-negative group, if they cannot be matched with any

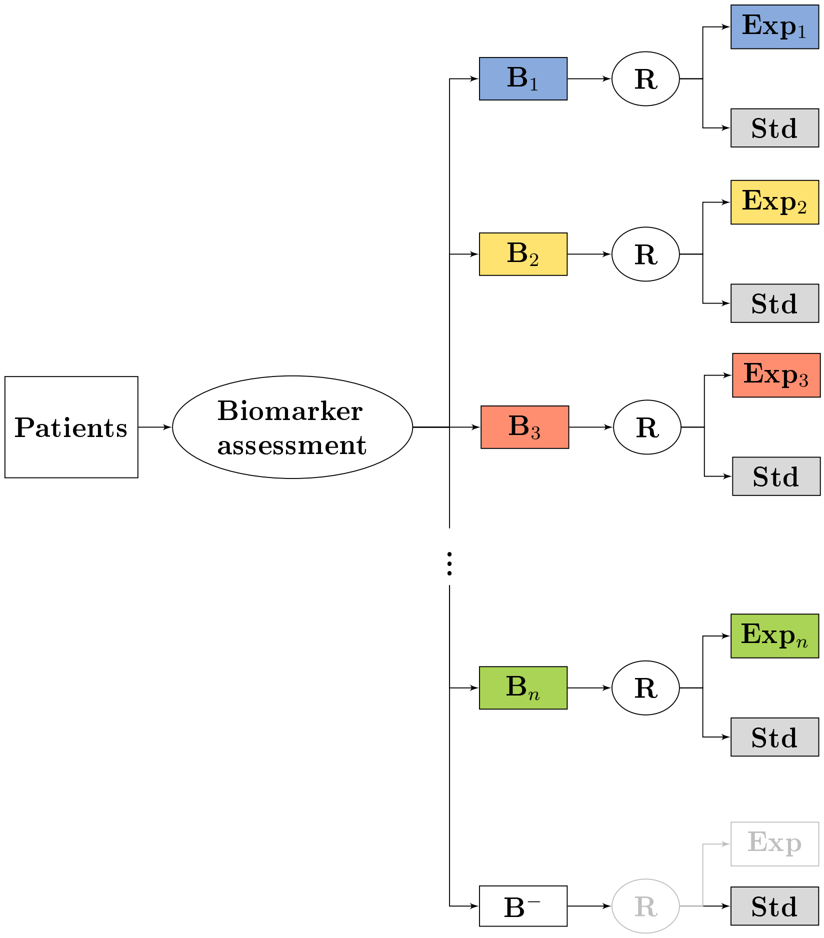


Figure 1.8: Proposed biomarker-guided study design

of the biomarker-groups. If a patient is matched with more than one of the biomarker-defined groups, guidelines should be available on how to proceed, possibly by assigning different priorities to the biomarkers, e.g. based on their prevalence or expected treatment outcomes. Within the biomarker-defined subgroups, patients are randomized between an experimental, i.e. biomarker-guided, therapy and standard of care. The experimental therapies can differ for each group. Due to the restriction to one tumor type, standard of care is assumed to be the same across the biomarker-groups. The biomarker-negative group can either simply be assigned to standard of care or may also be randomized between an experimental treatment and standard of care. During the course of the study, biomarker-arms may be added as new information and biomarkers become available.

Chapter 2

Fundamental Methods

2.1 Survival analysis

This Section is largely part of a paper that has already been published. The relevant passages have been taken verbatim from Beisel et al. (2017). Section 2.1.1.4 has already been published in Habermehl et al. (2017).

For data arising from oncological clinical trials, the endpoint of interest is commonly a survival- or time-to-event endpoint, such as overall survival or progression free survival. Hence, survival analysis is an often used tool. Time-to-event means that a time until a certain event happens is observed. A key difference to other types of data is the so-called 'censoring'. Censoring occurs when the event of interest has not occurred by the time the follow-up ends or if a patient leaves the study prematurely. A patient who was censored at time t is known to not having had the event until time t, but it is not known if or when the event occurred after time t. The survival function, i.e. the probability that a patient survives longer than time t, is commonly denoted by $S(t) = P(T > t)$. The hazard function or hazard rate

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

is the instantaneous rate of experiencing the event of interest at time t given that the patient has survived until time t (e.g. see Kleinbaum and Klein 1996).

2.1.1 Cox proportional hazards model

In 1972, Cox suggested a proportional hazards (PH) model which models the hazard function as

$$\lambda(t) = \lambda_0(t)\exp(\beta x), \tag{2.1}$$

where λ_0 is the baseline hazard, β is the treatment effect, and x the treatment indicator (Cox 1972). As the name suggests, this model relies on the proportional hazards assumption, i.e. the assumption that the hazard ratio is constant over time. The Cox PH model is a common choice to model survival data to estimate the treatment effect of an experimental therapy compared to standard of care.

2.1.1.1 Stratified Cox proportional hazards model

To take into consideration that the biomarkers in the study might be prognostic, the baseline hazards of the biomarker-groups should be allowed to differ. Hence, a stratified Cox PH model (Kalbfleisch and Prentice 1980) may be a more suitable choice, where each biomarker-defined stratum is allowed to have an individual baseline hazard. In the stratified Cox PH model with a single covariate the stratum-specific hazard function is modeled as

$$\lambda_i(t) = \lambda_{0_i}(t)\exp(\beta x), \tag{2.2}$$

where $i = 1, 2, ..., s$ is the stratum indicator and $\lambda_{0_i}(t)$ is the baseline hazard for stratum i.

For the simulation studies in Section 3, the Cox PH model and the stratified Cox PH model were applied using the `R` function `coxph{survival}`.

2.1.1.2 Two-step procedure

Mehrotra et al. (2012) suggested an alternative to the stratified Cox PH model for the case of unequal hazard ratios. They developed a two-step approach which first estimates the treatment effects within the strata separately and then combines them via a weighted average to an estimate for the overall treatment effect:

1. Estimate treatment effect $\widehat{\beta}_i$ for each stratum i individually, using an unstratified Cox PH model, separately fitted to each stratum.

2. Combine the treatment effect estimates for the strata to an overall treatment effect by a weighted mean
$$\widehat{\beta} = \sum_{i=1}^{s} \omega_i \widehat{\beta}_i,$$
with weights ω_i $(\sum_{i=1}^{s} \omega_i = 1)$.

These weights can, for example, be defined by the proportion of patients in the strata, i.e. $\omega_i = g_i$, where g_i is the proportion of total sample size in stratum i. Under the assumption of homogeneous treatment effects, using these weights for the strata gives similar results to the stratified Cox PH model. Allowing to adjust the weighting of the strata makes this procedure more flexible and can reduce the bias of the estimator in the case of heterogeneous treatment effects (cf. Mehrotra et al. 2012). Hence, the usage of this two-step approach is a possibility to use the Cox PH model in situations when the assumption of homogeneous treatment effects across strata, which is made by the stratified Cox PH model, is violated.

Note that the overall treatment effect β can be considered as the average benefit of a random patient sampled from a mixture distribution with weights ω_i.

To calculate a Wald test statistic (the Wald test will be discussed in more detail in Section 2.1.2.3), one also needs a variance estimate:

$$\hat{V}(\hat{\beta}) = \sum_{i=1}^{s} \omega_i^2 \hat{V}(\hat{\beta}_i),$$

where $\hat{V}(\hat{\beta}_i)$ is the variance estimate for $\hat{\beta}_i$ from the Cox PH model fitted to stratum i.

2.1.1.3 Frailty model

Another alternative to the stratified Cox PH model is the shared frailty model, a type of random effects model (e.g. see Duchateau and Janssen 2007). Instead of considering different baseline hazards for the strata, this model extends the Cox PH model by including an unobservable random variable W which acts multiplicatively on the common baseline hazard across all strata, λ_0. For each stratum i the random variable W achieves an outcome w_i, i.e. this model assumes homogeneity within the strata but heterogeneity across strata. The stratum-specific hazard function is modeled as

$$\lambda_i(t) = w_i \lambda_0(t) \exp(\beta x). \tag{2.3}$$

The most common distribution choices for the frailty variable W are the gamma distribution and the lognormal distribution. The probability density function for the lognormal distribution with $E(Y) = \exp(\gamma/2)$ and $Var(Y) = \exp(2\gamma) - \exp(\gamma)$ is

$$f(y) = \frac{1}{y\sqrt{2\pi\gamma}} \exp\left(-\frac{(\log y)^2}{2\gamma}\right), \tag{2.4}$$

with $\gamma > 0.$

The probability density function for the gamma distribution with $E(Y) = 1$ and $Var(Y) = \theta$ is

$$f(y) = \frac{y^{\frac{1}{\theta}-1} e^{-\frac{y}{\theta}}}{\Gamma(\frac{1}{\theta})\theta^{\frac{1}{\theta}}}. \tag{2.5}$$

The shape and scale parameters of this gamma distribution are $\frac{1}{\theta}$ and θ, respectively.

The popularity of the Gamma distribution is mainly due to mathematical and computational convenience, as it is easy to derive closed-form expressions of the survival and hazard functions. The lognormal frailty model is more computationally intensive. It requires the solution of numerical integrals, because there exists no closed-form expression of the marginal likelihood. With increasing computer power, the lognormal distribution has become a popular alternative due to its close connection to random effects and mixed effects models (Wienke 2010).

For the simulation studies in Section 3, the shared frailty model was applied using the `R` function `coxme{coxme}`.

2.1.1.4 Firth correction

Firth (1993) addressed the issue of bias of maximum likelihood estimates caused by small sample sizes and rare events. To reduce this bias, he suggested using a penalized likelihood based on a modified score function. Heinze and Schemper (2001) formulated the modified score function, which is oftentimes referred to as Firth correction or Firth penalty, for Cox regression:

$$U(\beta)^* = U(\beta) + 0.5\,trace\left[I(\beta)^{-1}\left\{\partial I(\beta)/\partial\beta\right\}\right], \tag{2.6}$$

where $U(\cdot)$ is the score function and $I(\cdot)$ is the Fisher information matrix. This modified score function is related to the penalized log likelihood function $\log L(\beta)^* = \log L(\beta) + 0.5 \log|I(\beta)|$.

For the simulation studies in Section 3, the Cox PH model with Firth correction was applied using the `R` function `coxphf{coxphf}`.

2.1.2 Significance tests

There are several tests available to test the significance of the estimated treatment effect, i.e. to test the null hypothesis H$_0$: $\beta = \beta_0$ against the alternative hypothesis H$_1$: $\beta \neq \beta_0$. The `coxph` procedure from the `R`-package `survival`, which can be used for all of the analysis methods discussed above, gives 3 different test statistics and corresponding p-values: the score test, the Wald test, and the likelihood ratio test. In the following, the focus will be on the former two.

2.1.2.1 Asymptotic stratified log-rank test

The stratified log-rank test statistic is given by

$$Z = \frac{\sum\limits_{i=1}^{s} O_i - E_i}{\sqrt{\sum\limits_{i=1}^{s} V(O_i - E_i)}} \sim N(0,1), \tag{2.7}$$

where $i = 1, ..., s$ are the strata, and

$$O_i - E_i = \sum_j \sum_k (o_{ijk} - e_{ijk}), \tag{2.8}$$

where j and k are the treatment groups and failure times, respectively, and o_{ijk} and e_{ijk} are the observed and expected events at time k in treatment group j and stratum i.

For the simulation studies in Section 3, the asymptotic log-rank test was applied using the R function `survdiff{survival}`.

2.1.2.2 Score test

If no other covariates are included in the model, and the single covariate in the model is categorical, the score test is identical to the log-rank test (Therneau and Grambsch 2000). Rao's score test (Rao 1948) tests the hypothesis $H_0 : \beta = \beta_0$ and its test statistic is given by

$$Z = \frac{U(\beta_0)}{\sqrt{I(\beta_0)}} \sim N(0, 1), \tag{2.9}$$

where $U(\beta_0)$ is the score function, i.e. the derivative of the log-likelihood function with respect to β at β_0, and $I(\beta_0)$ is the Fisher information.

The stratified version of the log-rank test is used to test the null hypothesis that there is no difference between two populations with respect to the probability of a specific event, controlling for a stratification variable. The test statistic of the stratified log-rank test approximately follows a standard normal distribution (cf. Kleinbaum and Klein 1996). However, just as for the unstratified log-rank test, this assumption may not apply for small sample situations, resulting in loss of power. For small samples, the exact log-rank test by Mehta et al. (1992) is a suitable alternative, because it is based on permutation. The downside of this test is that due to the permutation, the exact log-rank test becomes computationally expensive quite fast as the sample size increases. Alternatively, the approximate version of the exact log-rank test can be used, which approximates the exact log-rank test via Monte Carlo resampling (Strasser and Weber 1999).

For the simulation studies in Section 3, the exact/approximate log-rank test was applied using the R function `logrank_test{coin}`.

2.1.2.3 Wald test

The Wald test and the score test are asymptotically equivalent, however the Wald test has been referred to as less reliable in finite samples by several authors, such as Therneau and Grambsch (2000) and Agresti (2007). The test statistic of the Wald test is given by

$$Z = \frac{\hat{\beta} - \beta_0}{\sqrt{Var(\hat{\beta})}}. \tag{2.10}$$

In contrast to the score test, the Wald test statistic not only depends on β_0 (cf. Equation 2.9), but also on the estimate $\hat{\beta}$. Therefore, it can be used when comparing the performance of Mehrotra's two-step procedure (cf. Section 2.1.1.2) to the regular stratified Cox PH model. Note that the score test would give the same test statistic for both cases, since it only depends on the treatment effect under the null, β_0.

2.1.3 Sample size calculation for survival trials

2.1.3.1 Schoenfeld's formula

A well-known formula for sample size calculation for the Cox PH model is the Schoenfeld formula (Schoenfeld 1983). In his paper, Schoenfeld showed that the sample size n needed to compare two survival distributions is given by:

$$n = \frac{(z_{1-\alpha/2} + z_{1-\beta})^2}{r(1-r)(\log HR)^2\, q}, \tag{2.11}$$

where α and $1-\beta$ are significance level and required power of the test, respectively, $z_{1-\alpha/2}$ and $z_{1-\beta}$ are the $1-\alpha/2$ and $1-\beta$ percentiles of the standard normal distribution, respectively, r is the proportion of patients randomized to the standard treatment arm, HR is the minimal detectable hazard ratio of the experimental and standard treatments, and q is the expected probability to experience the event of interest. An advantage of Schoenfeld's formula is that, beyond the proportional hazards assumption, it does not rely on a specific survival distribution.

Note that Schoenfeld's formula neither allows stratification nor stratum specific hazard ratios.

2.1.3.2 Palta and Amini's formula

A few years later, Palta and Amini (1985) extended Schoenfeld's formula to allow for stratification. They generalized Equation 2.11 for $m > 1$ strata:

$$n = \frac{(z_{1-\alpha} + z_{1-\beta})^2}{\mu^2}, \tag{2.12}$$

where

$$\mu = \frac{\sum_{i=1}^{s} g_i \int_0^\infty \log HR_i(t)\, r_i(1-r_i)v_i(t)dt}{\sqrt{\sum_{i=1}^{s} g_i \int_0^\infty v_i(t)r_i(1-r_i)dt}}, \tag{2.13}$$

where g_i is the proportion of total sample size in stratum i, r_i is the proportion of patients in stratum i randomized to standard of care, and the function $v_i(t)$ is the density function of an observed event. In their paper, $v_i(t)$ is given for exponential survival with hazard rate λ_{j_i}, for treatment j, and stratum i, accrual and follow-up periods of length a and f, respectively, and uniform patient entry:

$$v_i(t) = \begin{cases} r_i\lambda_{0_i}\exp(-\lambda_{0_i}t) + (1-r_i)\lambda_{1_i}\exp(-\lambda_{1_i}t) & t \leq f \\ r_i\frac{a+f-t}{a}\lambda_{0_i}\exp(-\lambda_{0_i}t) + (1-r_i)\frac{a+f-t}{a}\lambda_{1_i}\exp(-\lambda_{1_i}t) & f < t \leq a+f \\ 0 & t > a+f, \end{cases} \tag{2.14}$$

keeping in mind that for administrative censoring C the probability $\mathsf{P}(C > t) = \frac{a+f-t}{a}$ for $f < t \leq a + f$. This integrates to

$$\begin{aligned} V_i = \int_0^{a+f} v_i(t)dt = r_i &\left[1 - \frac{1}{a\lambda_{0_i}}\Big(\exp\left[-\lambda_{0_i} f\right] - \exp\left[-\lambda_{0_i}(a+f)\right]\Big)\right] \\ &+ (1-r_i)\left[1 - \frac{1}{a\lambda_{1_i}}\Big(\exp\left[-\lambda_{1_i} f\right] - \exp\left[-\lambda_{1_i}(a+f)\right]\Big)\right]. \end{aligned} \tag{2.15}$$

Note that Equation 2.15 is the probability for a patient in stratum i to experience the event of interest during follow-up: Consider a study with uniform accrual, an accrual period $[0, a]$, a follow-up time f, and no loss to follow-up. Then we have a uniform distribution for administrative censoring, C, with density $\frac{1}{a}\mathbb{1}_{[f,a+f]}(t)$. Assuming exponentially distributed survival times T, the probability q for a patient to experience an event of interest throughout the study is given by

$$\begin{aligned} q = \mathsf{P}(T \leq C) &= \int_0^{\infty} P(T \leq t)\frac{1}{a}\mathbb{1}_{[f,a+f]}(t)dt \\ &= \int_f^{a+f} \left[1 - \exp(-\lambda t)\right]\frac{1}{a}dt \\ &= 1 - \frac{1}{\lambda a}\Big(\exp\left[-\lambda f\right] - \exp\left[-\lambda(a+f)\right]\Big). \end{aligned} \tag{2.16}$$

Coming back to the formula by Palta and Amini (1985), they give a simplified version of their formula (Equation 2.13) assuming proportional hazards over time:

$$\mu = \log HR \sqrt{\sum_{i=1}^{s} g_i r_i (1-r_i) V_i}, \tag{2.17}$$

where HR is the hazard ratio, and V_i is the integral of $v_i(t)$ over the study length, i.e. the probability of not being censored in stratum i (see Equation 2.15).

2.1.3.3 Lachin's formula

Schoenfeld's - and consequently Palta and Amini's - sample size formula is derived using the score statistic, testing whether the hazard ratio is different from 1. Another sample size formula, which in contrast is based on the Wald test, was introduced by Lachin in 1981. It tests the difference in hazard rates between the standard and experimental treatment arms for unstratified, exponentially distributed survival time data:

$$n = \left[\frac{z_{1-\alpha/2}\sqrt{\Phi(\overline{\lambda})\left(\frac{1}{r}+\frac{1}{1-r}\right)} + z_{1-\beta}\sqrt{\Phi(\lambda_1)\frac{1}{1-r}+\Phi(\lambda_0)\frac{1}{r}}}{\lambda_1 - \lambda_0} \right]^2, \qquad (2.18)$$

where r is the proportion of patients randomized to the standard treatment arm, λ_j is the hazard rate of treatment arm j (for the unstratified case), $\overline{\lambda} = (\lambda_1 + \lambda_0)/2$, and, assuming exponential survival, uniform patient entry, and (only) administrative censoring

$$\Phi(\lambda) = \frac{\lambda^2}{q} = \lambda^2 \left[1 - \frac{1}{\lambda a}\Big(\exp\left[-\lambda f\right] - \exp\left[-\lambda(a+f)\right] \Big) \right]^{-1}, \qquad (2.19)$$

where a and f are the accrual and follow-up time, respectively. For details on the probability for a patient to experience an event of interest throughout the study, q, see Equation 2.16.

2.1.3.4 Lachin and Foulkes' formula

Later on, Lachin extended his sample size formula together with Foulkes, to allow for nonuniform patient entry, loss to follow-up, noncompliance and stratification (Lachin and Foulkes 1986). Note that Lachin's formula and the extension by Lachin and Foulkes is based on the assumption of exponential survival, which makes it less flexible than the other two formulas.

Their extension of the formula for stratified trials with two strata uses a pooled estimator as test statistic, which is calculated as a weighted average over the strata of the within-stratum differences in hazard rates:

$$\overline{\widehat{\lambda}_0 - \widehat{\lambda}_1} = \nu_1(\widehat{\lambda}_{1_1} - \widehat{\lambda}_{0_1}) + \nu_2(\widehat{\lambda}_{1_2} - \widehat{\lambda}_{0_2}) \qquad (2.20)$$

where $\widehat{\lambda}_{j_i}$ is the estimated hazard rate for treatment j in stratum i and ν_i are weights that are inversely proportional to the variances of the within-stratum hazard rate differences:

$$\nu_i = \frac{1}{\sigma_{0,i}^2}\left(\frac{1}{\sigma_{0,1}^2} + \frac{1}{\sigma_{0,2}^2}\right)^{-1} \tag{2.21}$$

where

$$\sigma_{0,i}^2 = \frac{\psi}{N_i}, \tag{2.22}$$

where N_i is the total sample size of stratum i, and

$$\psi = \Phi(\overline{\lambda}_i)\left(\frac{1}{r_i} + \frac{1}{1-r_i}\right), \tag{2.23}$$

with $\overline{\lambda_i} = (\lambda_{1_i} + \lambda_{0_i})/2$ and Φ as defined in Equation 2.19.

With the pooled estimator, the sample size can be calculated as:

$$n = \left[\frac{z_\alpha\sqrt{\Omega^{-1}} + z_\beta\sqrt{\Omega^{-2}\sum_{i=1}^{2} g_i\left(\Phi(\lambda_{1_i})\frac{1}{1-r} + \Phi(\lambda_{0_i})\frac{1}{r}\right)\left(\Phi(\overline{\lambda}_i)\left(\frac{1}{r} + \frac{1}{1-r}\right)\right)^{-2}}}{\overline{\lambda_1 - \lambda_0}}\right]^2, \tag{2.24}$$

where λ_{j_i} is the hazard rate for treatment $j \in \{0,1\}$ in stratum i, $\overline{\lambda_i} = (\lambda_{1_i} + \lambda_{0_i})/2$, and

$$\Omega = \frac{g_1}{\Phi(\overline{\lambda}_1)\left(\frac{1}{r} + \frac{1}{1-r}\right)} + \frac{g_2}{\Phi(\overline{\lambda}_2)\left(\frac{1}{r} + \frac{1}{1-r}\right)}.$$

2.2 Data generation for simulation studies

2.2.1 Generating survival time data

For the generation of survival times, the baseline hazards, hazard ratios, proportion of patients in each stratum, and the randomization probability are predetermined. Survival is assumed to be exponential and the randomization probability between treatments is set to 0.5 for all strata. Patients are assigned to biomarker-group i by drawing from a multinomial distribution according to the prespecified proportions of the strata, and are then randomized equally to the treatment arms. For every patient k, three times were generated:

The survival time for patient k in stratum i obtaining treatment x was generated according to Bender et al. (2005):

$$t_k = -\frac{\log(U_k)}{h_0 \exp(\beta_i\, x)}, \tag{2.25}$$

where $x \in \{0, 1\}$ is the treatment indicator, $U_k \sim \text{Unif}(0, 1)$ and $h_0 = \lambda_0$ for the Cox PH model, $h_0 = \lambda_{0_i}$ for the stratified Cox PH model, and $h_0 = w_i \lambda_0$ for the shared frailty model.

Times to administrative censoring, $t_{\text{ad},k}$, were generated by drawing from a uniform distribution $\text{Unif}(f, a + f)$ with accrual period a and follow-up time f.

A time for random censoring $t_{\text{cens},k}$ for each patient k was generated from an exponential distribution with hazard λ_{cens}, calculated by

$$\lambda_{\text{cens}} = \frac{p_{\text{cens}}\, \widetilde{\lambda}}{1 - p_{\text{cens}}}, \tag{2.26}$$

where $\widetilde{\lambda} = \sum_{i=1}^{s} g_i(\lambda_{1_i} + \lambda_{0_i})/2$ is an "average" hazard rate, where λ_{j_i} is the hazard rate for treatment j in stratum i, and p_{cens} is the expected proportion of random censoring among patients.

Realized "overall survival" (os) is then derived as the minimum of these three generated times, i.e. $os_k = \min(t_k, t_{\text{ad},k}, t_{\text{cens},k})$. If $t_k = os_k$, patient k is assigned status 1 (dead), and 0 (alive) otherwise. All simulation studies were carried out in R.

2.2.2 Generating missing at random data

To generate missing data that is missing at random (MAR), first a complete data set is generated as described above. Then, the probability of missingness for variable y, i.e. $\Pr(y = \text{missing}|x)$, can be modeled, e.g. using a logistical model. If the missingness is modeled to depend on a single other observed variable, the probability is given by

$$\Pr(y_k = 1|x_k) = \frac{\exp(c + \beta x_k)}{1 + \exp(c + \beta x_k)}, \tag{2.27}$$

where y is the variable for which missing data shall be generated, and x is the variable upon which the missingness depends. Now a binary variable can be generated which indicates whether y is missing by drawing from a binomial distribution with probability p_k. Whenever this variable is 1, the corresponding data point of y is deleted (Van Buuren 2012).

If a certain proportion of missing data, p_{miss}, is targeted, c and β need to be determined that satisfy $f(c, \beta) = p_{\text{miss}}$, where $f(c, \beta)$ is the mean of the probabilities:

$$f(c, \beta) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{1 + \exp(c + \beta x_i)}. \tag{2.28}$$

This can, for example, be done by fixing β at some value, and then solving

$$f(c, \beta) - p_{\text{miss}} = 0 \tag{2.29}$$

for c. Then, using c and the fixed value for β, the probabilities for missingness can be obtained from Equation 2.27.

Chapter 3

Results

3.1 Development of a concept for a trial with multiple biomarkers

Some sections of this chapter are part of papers that have already been published. For Section 3.1 and Section 3.3, the relevant passages have been taken verbatim from Habermehl et al. (2017), and for Section 3.2 the relevant passages have been taken verbatim from Beisel et al. (2017).

The optimal choice of the trial design is an important step within the process of developing a trial. However, there is not a single answer which design is best - it depends on various factors. One type of multiple-biomarker trials, which were introduced in Section 1.1.3, are the so-called umbrella trials, where enrollment is generally restricted to one tumor type. The biomarker-defined subtrials of an umbrella trial are usually analyzed individually, treating each subtrial as an independent trial, due to the heterogeneity caused by the different biomarkers and different experimental therapies targeting these biomarkers. Biomarkers can cause heterogeneity in the study population if they are either prognostic, predictive, or both.

An example for an umbrella trial is the FOCUS4 trial (Kaplan et al. 2013), which is aimed at patients with colorectal cancer. One of the biomarkers investigated in the trial is the BRAF-mutation, which is indicative of a poor prognosis in this tumor type. The excessive signaling caused by this mutation can be targeted by an inhibitor of the BRAF-protein. However, in colorectal cancer, signaling through the epidermal growth factor receptor (EGFR) plays a role. Therefore, the treatment investigated in FOCUS4 is a combination of BRAF-inhibitor and EGFR-inhibitor, with and without a MEK-inhibitor. For more details see Kaplan et al. (2013).

Since colorectal cancer is one of the most common cancer types, recruitment of patients should not be an issue. However, completion of accrual may become unfeasible when this type of design is used for a less prevalent disease. In this case, small sample sizes within the subtrials have to be expected, as well as many biomarker-negative patients at the initial screening stage, i.e. patients which test negative for all biomarkers considered in the trial. The small sample sizes may make it unfeasible to treat the subtrials as independent and analyze them individually. Moreover, the small sample sizes can lead to biased treatment effect estimates. This imposes the need to investigate alternative approaches for the analysis of such a trial, and possibly for the study design itself.

The following sections will discuss several options for multiple-biomarker trials and give examples for which situations the design option would be an appropriate choice. Furthermore, the issue of whether or not biomarker-negative patients should be included in the study will be discussed.

3.1.1 Biomarker-negative patients

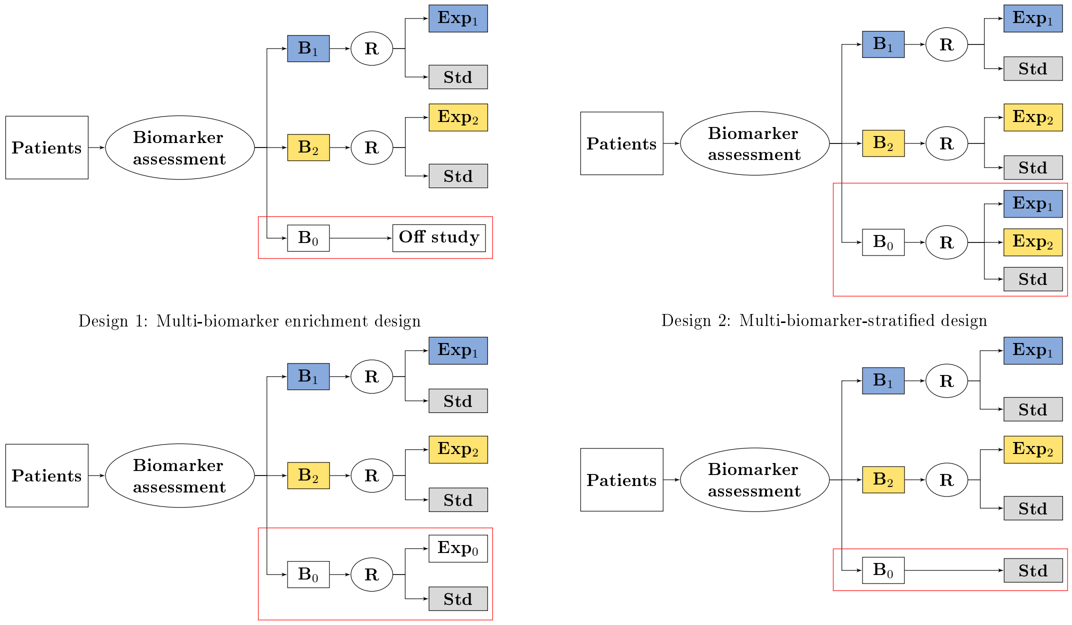
Before considering specific design features, a decision should be made whether to use an enrichment-type or an all-comers design, i.e. it should be decided if patients that cannot be matched with one of the relevant biomarkers are excluded from the trial or if they are included in a separate trial arm. Excluding biomarker-negative patients may seem like the most cost-effective option at first. But it should be taken into consideration that the initial step of determining a patient's biomarker profile already entails costs. Excluding these patients after this step means spending resources on valuable information that is not going to be used afterwards. Especially for trials with lower prevalence biomarkers, where it is expected to encounter numerous biomarker-negative patients throughout the accrual period, it might be worth considering to at least include these patients for follow-up purposes, if possible. This way, a valuable database of patients for which the biomarker-profile has already been assessed can be obtained. This database can later be utilized for retrospective analyses and identification of new potential biomarkers.

3.1.2 Design choices

A popular strategy for designing a multiple-biomarker trial, such as an umbrella trial, is to exclude biomarker-negative patients at the screening stage. This basically leads to a multiple-biomarker enrichment design (Figure 3.1, Design 1). For less prevalent biomarkers, this means disregarding many biomarker-negative patients for whom information about their biomarker profile was already gathered during the screening process. Alternatively, a study design could be used that includes the biomarker-negative patients in the study. There are several possibilities to do so.

A possibility that offers the most information about the biomarkers investigated in the study is to randomize the biomarker-negative patients between all the experimental therapies targeting the biomarkers in the study and standard of care (Figure 3.1, Design 2). This way, these experimental therapies can be investigated in the biomarker-positive and the biomarker-negative population. Additionally, this design allows drawing conclusions about prognostic, as well as predictive properties of the biomarkers. Essentially, this design is a biomarker-stratified design (also called biomarker-by-treatment interaction design) with more than one biomarker. While this design can supply useful information about the properties biomarkers, it simultaneously limits the number of biomarkers that can be investigated in the study, since each added biomarker adds another treatment arm for the biomarker-negative patients. Furthermore, treating biomarker-negative patients with an experimental therapy which targets a biomarker that they do not have can be an issue. Based on prior evidence it needs to be decided for each of these treatments whether this treatment strategy is ethically tenable.

Another option to include biomarker-negative patients in a multiple-biomarker trial is a stratified randomize-all design, where biomarker-negative patients are randomized between a (different) experimental therapy and standard of care (Figure 3.1, Design



Design 1: Multi-biomarker enrichment design

Design 2: Multi-biomarker-stratified design

Design 3: Stratified randomize-all design

Design 4: Multi-biomarker hybrid design

Figure 3.1: Schematic study designs for a multiple-biomarker trial, simplified to two biomarkers, with different options for biomarker-negative patients.

3). This way, the resulting study has multiple strata with two treatment arms each. While the analysis of such a design is fairly straight forward, it is important to take into consideration the heterogeneity within the study population. On the one hand, this can be caused by prognostic and predictive biomarkers, and on the other hand by the different treatments with, most likely, different treatment effects. For smaller sample sizes, which is likely to be the case for lower prevalence biomarkers, the analysis concept for the resulting data could be an evaluation of the overall biomarker-guided treatment strategy, rather than performing separate analyses for each biomarker. This will be discussed in more detail in Section 3.2. This evaluation of an overall treatment strategy should be performed in conjunction with subsequent subgroup analyses to avoid false conclusions about individual biomarkers. The subgroup analyses could either be of exploratory nature or preplanned, utilizing multiple comparison strategies. One option would be a serial chain procedure (Millen and Dmitrienko 2012) where the primary hypothesis is first tested at a predetermined significance level, α. If the null hypothesis is rejected, the unspent α from the main hypothesis can be reallocated to the remaining secondary hypotheses, since chain procedures are a class of closed testing procedures. The α can be split equally or weighted between the remaining hypotheses. Multiple comparison procedures will be discussed in more detail in Section 3.5. In practice, the usability of this stratified randomize-all design depends on availability of a suitable experimental therapy for the biomarker-negative patients.

The remaining alternative to excluding biomarker-negative patients at the screening stage is a 'multi-biomarker hybrid design' where these patients are kept in the study and treated with standard of care (Figure 3.1, Design 4). This allows gathering follow-up data for the biomarker-negative patients as well as potentially using this data for retrospective identification of new biomarkers. Additionally, the prognostic properties of the biomarkers can be investigated by comparing the standard of care arms. If the biomarkers in the study are assumed to be non-prognostic, the data for the biomarker-negative patients can be used in the analysis by pooling the three standard of care arms. For prognostic biomarkers the inclusion of the biomarker-negative patients in the analysis becomes more complicated. This will be discussed in more detail in Section 3.3. A practical drawback of this multi-biomarker hybrid design is that biomarker-negative patients potentially have a larger risk of dropping out of the study if other studies become available that offer an experimental therapy.

3.2 Stratified randomize-all design

The categorization of cancer types into subtypes leads to a stratification of the study population into multiple subtrials investigating different experimental therapies. In umbrella trials, these subtrials are usually analyzed individually, treating each subtrial as an independent trial. While this is feasible for common cancer types, it may be difficult to recruit enough patients to each subtrial to obtain statistically meaningful results within a reasonable time frame.

An alternative approach could be a proof of efficacy of the overall treatment strategy (e.g. a biomarker-guided treatment strategy) as primary hypothesis before looking at the subtrials individually. However, this approach entails several difficulties. The subpopulations cannot simply be assumed to be homogeneous across all subtrials, especially if various disease subtypes and different treatments are investigated in the trial.

Obviously, it cannot be assumed that different treatments will have similar treatment effects. Some of the factors defining the subtypes, such as targeted mutations, might be prognostic or predictive and have an impact on the outcome or the treatment effect, respectively. For this situation, the assumption that the treatment effect is the same across all strata does not seem appropriate. However, the Cox PH model, its stratified version, and most sample size formulas rely on this assumption and a violation may result in deviation from the desired level of power.

In the following sections, the performance of different methods for sample size calculation and data analysis under heterogeneous treatment effects will be investigated. With regard to sample size calculation, the commonly used sample size formula by Schoenfeld (1983) is compared to a formula by Lachin and Foulkes (1986), and an extension of Schoenfeld's formula by Palta and Amini (1985). Possibilities for statistical modeling of heterogeneity are stratification by factors, the assumption of a probability distribution of the inter-patient or inter-strata variation, or the inclusion of covariates in the regression model. With a focus on the former two options, the widely used (stratified) Cox PH model (Kalbfleisch and Prentice 1980), a two-step analysis approach by Mehrotra et al. (2012), and the lognormal shared frailty model (Duchateau and Janssen 2007) will be considered as potential methods for data analysis which attempt to adjust for inter-strata heterogeneity.

3.2.1 Data modeling and analysis

The study design considered in the following is a stratified design with s strata. The strata are denoted by B_i, $i \in \{1, 2, ..., s\}$, and are defined by biomarkers which are targeted by one of the stratum-specific experimental treatments investigated in the study (see Figure 3.2). Upon entering the study, the patients' biomarker-status is determined and they are assigned to the strata accordingly. Patients matched with neither of the biomarkers are assigned to the biomarker-negative stratum. For the case that a patient is matched with more than one biomarker, there should be predefined priorities for the biomarkers, such that the patient can be distinctly allocated to one of the biomarker-defined strata. These priorities could, for example, be defined by biomarker prevalence or expected treatment outcomes.

The proportion of patients in stratum B$_i$ is denoted by $g_i \in [0, 1]$. Within each stratum i, patients are randomized between the stratum-specific experimental therapy (Exp$_i$) and standard of care (Std), with probabilities $1 - r_i$ and r_i, respectively. Patient entry is assumed to be uniform throughout the accrual time a. Patients are then monitored for the event of interest. Patients that are still in the study after follow-up time f are subject to administrative censoring. It is assumed that additional random censoring can occur. The hazards of death at time t for patients in stratum i receiving treatment j are denoted by $\lambda_{j_i}(t)$, where $j = 1$ for experimental treatment or $j = 0$ for standard of care.

As mentioned in Section 2.1.1, a common choice for modeling data and estimating a corresponding treatment effect is the Cox PH model, or, for a stratified study population, the stratified Cox PH model. An alternative to the stratified Cox PH model is the shared frailty model, which was introduced in Section 2.1.1.3. An advantage of this model is that it is able to treat heterogeneity between strata without requiring specific assumptions about stratum specific prognostic effects.

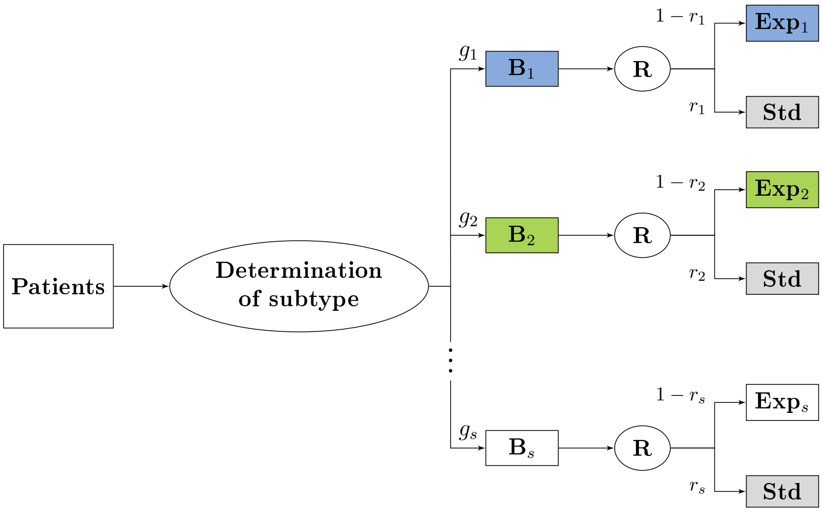


Figure 3.2: Schematic study design.

Mehrotra et al. (2012) suggested another alternative to the stratified Cox PH model for the case of unequal hazard ratios. They developed a two-step approach which first estimates the treatment effects within the strata separately and then combines them via a weighted average to an estimate for the overall treatment effect (see Section 2.1.1.2). Allowing to adjust the weighting of the strata makes this procedure more flexible and can reduce the bias of the estimator in the case of heterogeneous treatment effects (cf. Mehrotra et al. 2012). Hence, the usage of this two-step approach is a possibility to use the Cox PH model in situations when the assumption of homogeneous treatment effects across strata, which is made by the stratified Cox PH model, is violated. Note that the overall treatment effect β can be considered as the average benefit of a random patient sampled from a mixture distribution with weights ω_i.

Two commonly used tests for the treatment effect in Cox regression are the score test and the Wald test. Since Mehrotra et al. (2012) use the Wald test in their two-step approach, both tests are considered in the subsequent simulation study.

More details on the methods discussed in this section can be found in Section 2.1.1.

3.2.2 Sample size calculation

An advantage of the well-known Schoenfeld formula, which was introduced in Section 2.1.3.1, is that, beyond the proportional hazards assumption, it does not rely on a specific survival distribution. But Schoenfeld's formula neither allows stratification nor stratum specific hazard ratios. To be able to examine the performance of Schoenfeld's formula in a scenario with heterogeneous treatment effects, an "average" hazard ratio is calculated that can be used in the formula:

$$\overline{HR} = \frac{-\log\left(\sum_{i=1}^{s}(1-r_i)g_i \exp\left[-\lambda_{0_i}\,\left(\frac{a}{2}+f\right)\,HR_i\right]\right)}{\overline{\lambda}_0\left(\frac{a}{2}+f\right)}, \tag{3.1}$$

where $i \in \{1, ..., s\}$ are the strata, λ_{0_i} is the baseline hazard for stratum i, HR_i is the hazard ratio for stratum i, g_i is the proportion of patients in stratum i, and r_i is the randomization probability to the standard treatment in stratum i. For $\overline{\lambda}_0$ see Equation 3.4.

The first step in the derivation of Equation 3.1 is finding the survival function of the patient population across all strata. An "average" survival function $\widetilde{S}(t|x)$ can then be found by integrating over the strata given the treatment option:

$$\widetilde{S}(t|x) = \sum_{i=1}^{s} {r_i}^{1-x}(1-r_i)^x \, g_i \; \exp\big[-\lambda_{0_i} \, t \exp(\log HR_i \cdot x)\big], \tag{3.2}$$

where $x \in \{0,1\}$ is the treatment indicator.

Next, the observed survival function is considered, which contains an average baseline hazard $\overline{\lambda}_0$ and an average hazard ratio $\overline{HR}$:

$$\overline{S}(t|x) = \exp\big[-\overline{\lambda}_0 \, t \exp(\log \overline{HR} \cdot x)\big]. \tag{3.3}$$

Now, $\overline{HR}$ can be found by equating the two survival functions (Equations 3.2 and 3.3), i.e. $\widetilde{S}(t|x) \overset{!}{=} \overline{S}(t|x)$, at $t = \frac{a}{2} + f$, where a and f are accrual and follow-up time, respectively, i.e. t is the average time a patient is under observation.

In a first step, it is necessary to solve Equation 3.3 for $\overline{\lambda}_0$ (for $x = 0$), which yields

$$\overline{\lambda}_0 = \frac{-\log\left(\sum_{i=1}^{s} r_i g_i \exp\left[-\lambda_{0_i}\left(\dfrac{a}{2}+f\right)\right]\right)}{\frac{a}{2}+f}. \tag{3.4}$$

Equation 3.4 can then be used to solve Equation 3.3 for $\overline{HR}$ (for $x = 1$), which yields Equation 3.1.

Palta and Amini (1985) extended Schoenfeld's formula to allow for stratification, testing the null hypothesis $\log(HR_i) = 0 \; \forall i$, where HR_i is the stratum-specific hazard ratio (see Equation 2.17). The formula by Palta and Amini does require an assumption about the distribution of survival times, but it is not restricted to the exponential distribution. To consider other survival distributions, one simply needs to adjust V_i accordingly (see Equations 2.14 and 2.15).

In their paper, Palta and Amini give a simplified version of their formula by assuming equal hazard ratios across strata. However, using their general formula (see Equation 2.17), and assuming unequal but constant hazard ratios over time, one can obtain a sample size formula that allows for unequal hazard ratios across strata:

$$\mu = \frac{\sum_{i=1}^{s} g_i \log HR_i \, r_i(1-r_i)V_i}{\sqrt{\sum_{i=1}^{s} g_i r_i (1-r_i) V_i}}, \tag{3.5}$$

where HR_i is the hazard ratio for stratum i, and V_i is the integral of $v_i(t)$ over the study length, i.e. the probability of not being censored in stratum i (see Equation 2.15).

The extension of Lachin's formula by Lachin and Foulkes (1986), allowing for stratification (Equation 2.24), which was given for two strata, was generalized for a case with s strata:

$$n = \left[\frac{z_\alpha\sqrt{\Omega^{-1}} + z_\beta\sqrt{\Omega^{-2}\sum_{i=1}^{s} g_i\left(\Phi(\lambda_{1_i})\frac{1}{1-r} + \Phi(\lambda_{0_i})\frac{1}{r}\right)\left(\Phi(\overline{\lambda}_i)\left(\frac{1}{r}+\frac{1}{1-r}\right)\right)^{-2}}}{\overline{\lambda_1 - \lambda_0}}\right]^2,$$

where λ_{j_i} is the hazard rate for treatment $j \in \{0,1\}$ in stratum i, $\overline{\lambda_i} = (\lambda_{1_i} + \lambda_{0_i})/2$,

$$\Omega = \left(\sum_{i=1}^{s} \frac{g_i}{\Phi(\overline{\lambda}_i)\left(\frac{1}{r} + \frac{1}{1-r}\right)}\right),$$

and

$$\overline{\lambda_1 - \lambda_0} = \sum_{i=1}^{s} \nu_i (\lambda_{1_i} - \lambda_{0_i}).$$

More details on the sample size formulas discussed in this section can be found in Section 2.1.3.

Unfortunately, there is currently no closed sample size formula available for a shared frailty model of the kind discussed in Section 2.1.1.3. Therefore, the required sample size was calculated empirically as follows: The first step for an empirical sample size calculation is to choose the required power of the test (ρ_r), and to pick a starting value for the sample size, e.g. from a sample size formula which is expected to provide a reasonable initial estimate. Then, a sufficient number of data sets is simulated, e.g. $10,000$ data sets, under a specific alternative hypothesis, according to the planned study design. Subsequently, these data sets are analyzed by the chosen data analysis method, e.g. the lognormal shared frailty model. The actual power (ρ_a) is then obtained from the percentage of rejected null hypotheses. If the actual power is within the required accuracy range of the required power, i.e. if $\rho_a = \rho_r \pm 0.01$, the calculation is completed and no further iteration steps are needed. If the actual power is outside this range, the sample size needs to be adjusted. This adjustment should be a predetermined rule, e.g. $n_{\mathsf{new}} = n_{\mathsf{old}}(1 - (\rho_{\mathsf{a}} - \rho_{\mathsf{r}}))$. The sample size is iteratively adjusted until the actual power reaches the required power with the desired accuracy, which was chosen here as 0.01.

3.2.3 Simulation study: Heterogeneous treatment effects

To compare the performance of the sample size formulas and the analysis methods for the case of heterogeneous treatment effects a simulation study with three strata is carried out. The following sections explain the study design, parameter setup and data generation before presenting and discussing the results.

3.2.3.1 Study design

A scenario is considered with two biomarkers, each targeted by a corresponding experimental therapy, which is to be included in the study. Furthermore, patients that cannot be matched with either biomarker should also be included in the study to test another, more broadly aimed experimental therapy. This leads to a study design with three strata, denoted by B$_i$, $i \in \{0, 1, 2\}$, where B$_1$ and B$_2$ are each comprised of patients matched with biomarker 1 or 2, respectively. All other patients are allocated to B$_0$, the biomarker-negative patients. Patients in strata B$_1$, B$_2$, and B$_0$ are randomized between the corresponding stratum specific experimental therapy (Exp$_1$, Exp$_2$, and Exp$_0$, respectively) and standard of care (Std).

3.2.3.2 Data generation

The data were generated as described in Section 2.2.1. For the baseline hazards, initially a common baseline hazard was chosen (0.05) and then the baseline hazards for B$_1$ and B$_2$ were multiplied by factors 0.8 and 1.2, respectively, to simulate stratification. For the remaining simulation parameters see Table 3.1. Note that a rough correction for loss to follow-up due to random censoring was made for all sample size formulas by dividing the calculated sample size by $1 - p_{\text{cens}}$, where p_{cens} is the expected proportion lost to follow-up. For this simulation, p_{cens} was set to 0.05.

Table 3.1: Parameters for the simulation study using a design with three biomarker-groups, denoted by B_i, $i \in \{0, 1, 2\}$.

Fixed simulation parameters	
Accrual time (months), a	24
Follow-up time (months), f	36
Proportion random censoring, p_{cens}	0.05
Treatment allocation ratio	$1:1$
Hazard ratio B$_1$, $\exp(\beta_1)$	$0.8, 0.7, 0.6, 0.5, 0.4, 0.3$
Hazard ratio B$_2$, $\exp(\beta_2)$	$0.8, 0.7, 0.6, 0.5, 0.4$
Number of simulations	10,000
Parameters for biomarker-groups $(\mathbf{B_0}, \mathbf{B_1}, \mathbf{B_2})$	
Proportion of patients, g_i	(0.5, 0.25, 0.25)
Baseline hazards, λ_{0_i}	(0.05, 0.04, 0.6)

If one wishes to simulate a stratum specific random effect in the patient data, as is assumed by the lognormal shared frailty model, the stratification factors for the baseline hazard, mentioned in the beginning of this section, can be replaced by numbers drawn from a lognormal distribution for each stratum. The choice of mean and variance of the lognormal distribution determines the intensity of the random effect, e.g. using a lognormal distribution with mean 0 and variance 0.15 on the log scale results in a relatively minor random effect. The data used in the next section is simulated without a random effect present. Results for random effects data can be found in the sensitivity analysis in Section 3.2.3.4. The simulation study was carried out in R (Version 3.2.2).

3.2.3.3 Results of the simulation study

In this section, the formulas for sample size calculation by Schoenfeld, Palta and Amini, and Lachin and Foulkes are compared. Afterwards, the different analysis methods are evaluated with respect to the power to detect a significant overall treatment effect, given specific hazard ratio scenarios.

For each parameter constellation of the simulation study, the required sample size for a power of 0.8 is determined with Schoenfeld's, Palta and Amini's and Lachin and Foulkes' sample size formula. For each calculated sample size, 10,000 data sets are simulated and then analyzed using the stratified Cox PH model, Mehrotra's two-step approach, and the shared frailty model. The empirical power for each method is assessed as percentage of rejected null hypotheses. The sample size formulas as well as the analysis methods, will be compared with respect to their compliance to the power used at the planning stage.

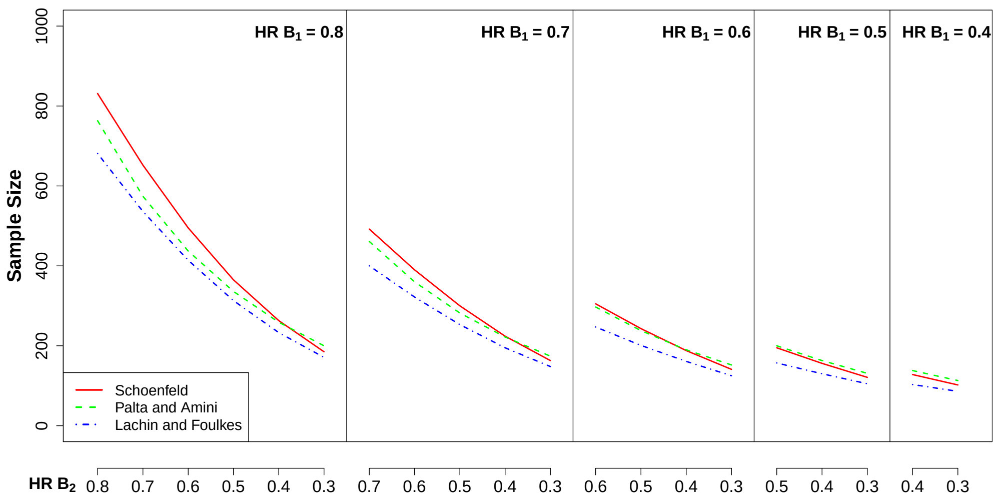


Figure 3.3: Comparison of sample size formulas for different hazard ratio scenarios: Sample sizes calculated from formulas by Schoenfeld, Palta and Amini, and Lachin and Foulkes.

A comparison of the sample sizes calculated with the formulas by Schoenfeld, Palta and Amini, and Lachin and Foulkes for different hazard ratio scenarios can be seen in Figure 3.3. The formula by Lachin and Foulkes yields the smallest sample size in all cases. The sample size calculated by Palta and Amini's formula mostly lies between the ones calculated by the other two formulas, except for the scenarios with the most extreme differences in hazard ratios. For those scenarios, the sample size by Schoenfeld is slightly smaller. The numerical results can be found in the Appendix.

The results of the empirical sample size calculation (see Section 3.2.2) are not shown in Figure 3.3, because Palta and Amini's formula provided an adequate sample size for the lognormal shared frailty model to reach the required power (see Table 3.2). The empirically calculated sample size only differed in one case, but not by much (3 patients less). Hence, it was decided that the (computationally expensive) empirical calculation of the sample size for the shared frailty model is not necessary. Note that this conclusion may not be valid for other scenarios, e.g. for more diverse baseline hazards.

Table 3.2: Numerical results for the empirical sample size calculation for the lognormal shared frailty model.

Input	**Power**	**Sample size**			**Iterations**
HR	Shared	Empirical	Palta	Lachin	Iteration
B_0, B_1, B_2	Frailty	estimate	Amini	Foulkes	steps
0.8 0.8 0.8	0.805	763	763	627	1
0.8 0.8 0.7	0.798	574	574	465	1
0.8 0.8 0.6	0.796	437	437	357	1
0.8 0.8 0.5	0.800	336	336	280	1
0.8 0.8 0.4	0.803	259	259	224	1
0.8 0.8 0.3	0.802	197	200	182	2
0.8 0.7 0.7	0.800	461	461	378	1
0.8 0.7 0.6	0.797	360	360	295	1
0.8 0.7 0.5	0.792	282	282	235	1
0.8 0.7 0.4	0.796	222	222	191	1
0.8 0.7 0.3	0.801	174	174	157	1
0.8 0.6 0.6	0.807	297	297	246	1
0.8 0.6 0.5	0.798	238	238	199	1
0.8 0.6 0.4	0.799	190	190	163	1
0.8 0.6 0.3	0.806	152	152	135	1
0.8 0.5 0.5	0.805	200	200	168	1
0.8 0.5 0.4	0.801	163	163	140	1
0.8 0.5 0.3	0.803	131	131	117	1
0.8 0.4 0.4	0.797	138	138	120	1
0.8 0.4 0.3	0.799	113	113	101	1

Subsequently, the methods for sample size calculation were compared with respect to compliance to the desired power level for each of the data analysis methods. Figure 3.4 shows the power for all three sample size calculation methods using the exact log-rank test as reference analysis method which does not depend on asymptotic assumptions. Equivalent comparisons were also made for the stratified Cox PH model, the two-step approach, and the lognormal shared frailty model. Using these analysis methods yields similar results (see Figures 3.5-3.7).

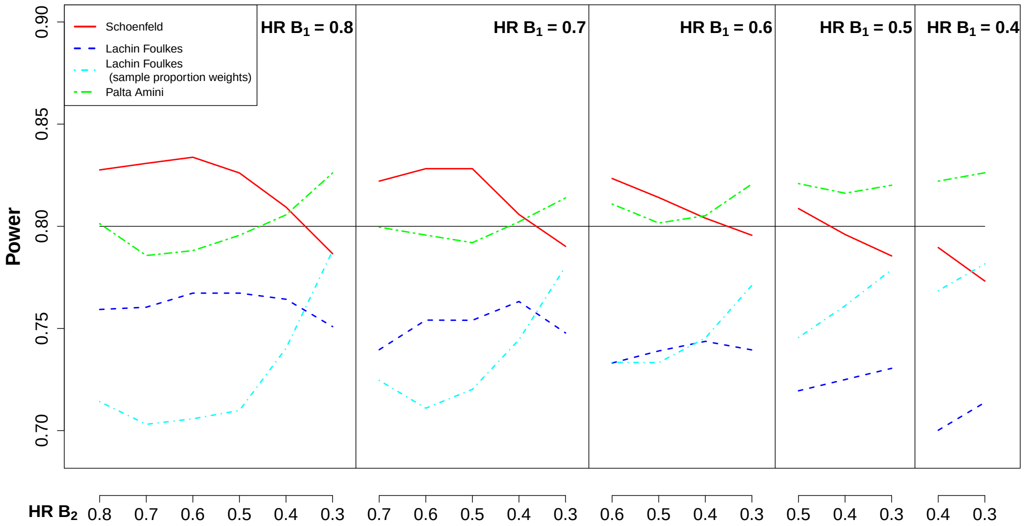


Figure 3.4: Power of the exact log-rank test to detect a true treatment effect for the different sample size formulas and different hazard ratios.

The sample size formula by Schoenfeld appears to overestimate the required sample size when the hazard ratios are similar and then begins a downward trend as the hazard ratios become more heterogeneous (see Figure 3.4). One has to keep in mind though, that this formula does not take the stratification into consideration and an averaged hazard ratio has to be used in the formula (see Equation 3.1). Therefore, this behavior was not surprising. Note that even though hazard ratios in the first scenario are homogeneous ($HR_1 = HR_2 = HR_0 = 0.8$), the baseline hazards are still heterogeneous due to stratification, which causes the non-compliance to the desired power level.

For the case of a stratified population, the sample size formula by Lachin and Foulkes uses a pooled estimator of the within-stratum differences in hazard rates. For this pooled estimator, they define optimal weights for each within-stratum difference (see Equations 2.20 - 2.23). However, the formula for these weights requires knowledge of the stratum-specific sample sizes. To be able to calculate these weights, in a first step the sample size was calculated empirically (analogous to the empirical sample size calculation for the shared frailty model in Section 3.2.3.2) before using the formula given by Lachin and Foulkes in the second step. This made the usage of this formula computationally more expensive than the other two. Additionally, one might question why one would use the formula at all if the required sample size was already calculated empirically. The resulting sample size from Lachin and Foulkes' method is too small to reach the desired power of 0.8. The power curve has a downward tendency as the hazard ratios of B_1 and B_2 get smaller, i.e. Lachin and Foulkes' formula does not seem to handle heterogeneous treatment effects well.

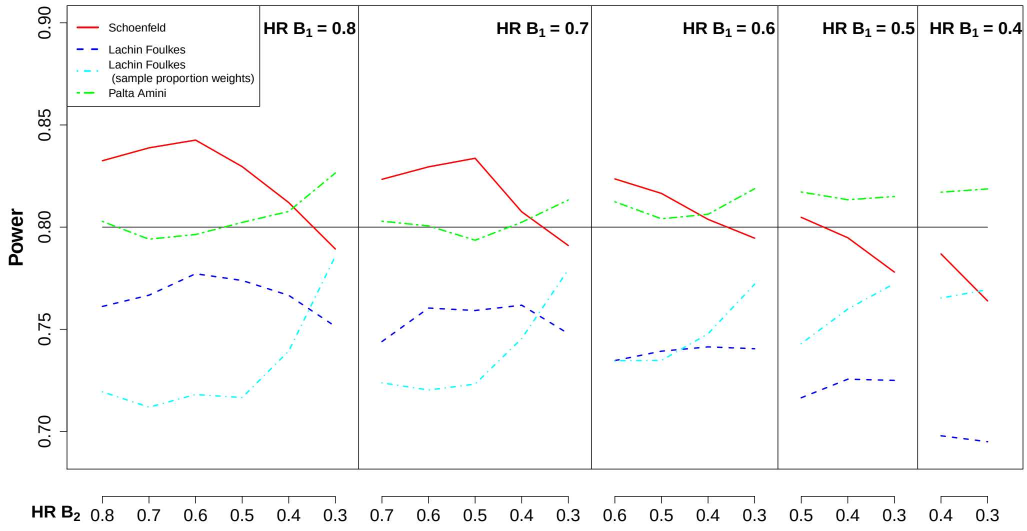


Figure 3.5: Power of the two-step approach to detect a true treatment effect for different hazard ratios.

Two alternate attempts, both avoiding empirical calculation, were made to reduce computation time and potentially improve the resulting power. The first attempt to replace the empirical calculation in the first step was to use the original formula by Lachin (1981) that does not account for stratification (see Equation 2.18). For the hazard rates, weighted means of the stratum specific hazard rates were used, using the expected sample proportion per stratum as weights. As before, the resulting sample size was then used to calculate the weights for the pooled estimator of the within stratum differences. This did not change the sample size calculated with Lachin and Foulkes' formula in the second step, even though the sample sizes used for the calculation of the pooled estimator of the within stratum differences were quite different: The former (empirical) sample size from step 1 was always larger than the final sample size in step 2, while the latter (from Equation 2.18) was always smaller. Hence, the formula for the weights seems to be rather robust regarding the stratum specific sample size. While this did not improve the sample size with respect to power, this shows that empirical calculation is not necessary, because it suffices to use a very rough estimate of the stratum specific sample size.

Another simulation was run where the weights used by Lachin and Foulkes were replaced with sample size proportions, i.e. the pooled estimator of the overall difference in hazard rates is a weighted average of the stratum specific hazard differences, weighted by the expected sample proportions of the strata. For similar and moderately heterogeneous treatment effects this sample size yields a lower power than the previous one. But, surprisingly, the resulting power increases as the treatment effects diverge, improving the power by up to 0.08 compared to using the original weights.

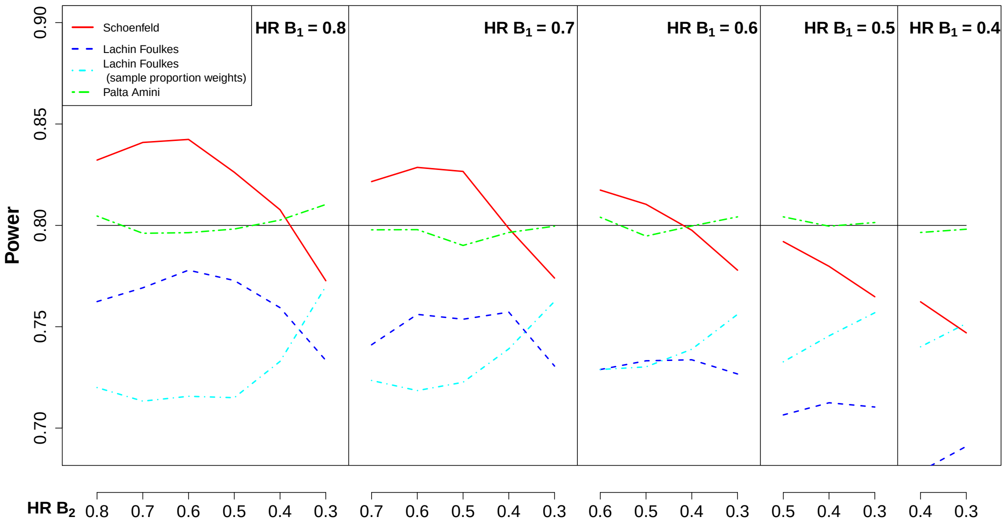


Figure 3.6: Power of the lognormal shared frailty model to detect a true treatment effect for the different sample size formulas and different hazard ratios.

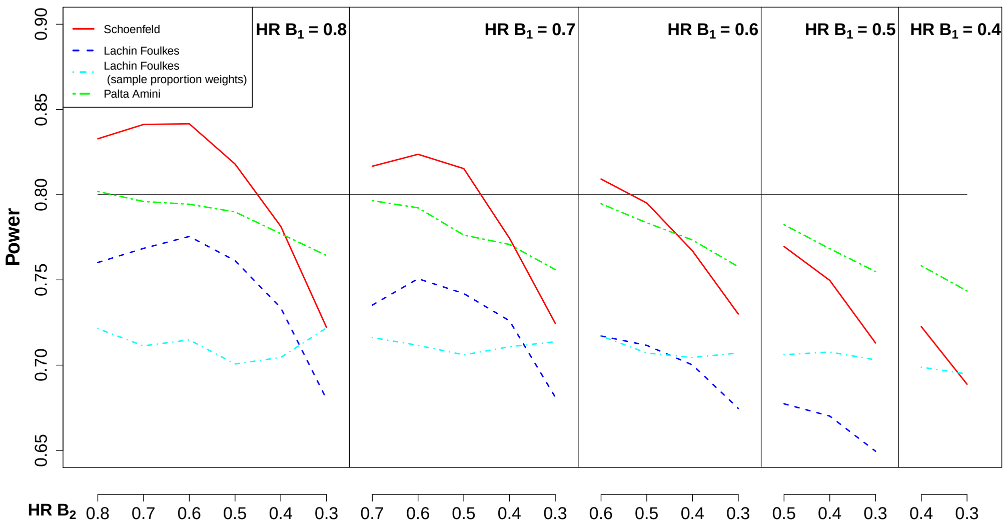


Figure 3.7: Power of the stratified Cox PH model to detect a true treatment effect for the different sample size formulas and different hazard ratios.

Nevertheless, it is still more than 0.02 short of reaching the desired power of 0.8. Hence, Lachin and Foulkes is an unreliable method for sample size calculation in the case of heterogeneous treatment effects.

The formula by Palta and Amini, allowing for unequal hazard ratios, yields the most reliable sample size: The actual power matches the expected power the closest. For similar hazard ratios it neither over- nor underestimates the sample size. As the hazard ratios become more heterogeneous, the power curve takes on a slight upwards trend (for the exact log-rank test). This is a considerable improvement over the other sample size formulas, which exceed or undercut the desired power level by up to 0.1. In conclusion, Palta and Amini's sample size formula performs best for most scenarios and was used to compare the different analysis methods in the following sections. If one is only interested in a rough number for the required sample size, the slightly easier to compute method by Schoenfeld is also acceptable in most cases.

Prior to the comparison of the analysis methods, it was verified that all methods control the type I error rate. For a sample size of 10,000 subjects and 10,000 simulations, the type I error rates ranged between 0.486 and 0.494.

The power to detect a significant treatment effect when using the sample size calculated from Palta and Amini's formula is compared for the stratified Cox PH model, the shared frailty model, and the two-step approach by Mehrotra et al. (see Figure 3.8). Additional to these three methods, the stratified asymptotic log-rank test and the exact log-rank test are included as well, to reveal potential failure of asymptotics. The asymptotic properties are also investigated as part of the sensitivity analysis in Section 3.2.3.4. Note that the approximate version of the exact log-rank test was used (for details see Section 2.1.2.2). The different analysis methods perform similarly for large sample sizes and minor heterogeneity of hazard ratios. As the sample size gets smaller and the hazard ratios become more heterogeneous, the power curves of the methods increasingly diverge. As expected, the curves differ the most for the most extreme scenario, with a minimum of 0.748 (stratified Cox) and a maximum of 0.829 (approximate log-rank). For the complete numerical results see Table 3.3.

The stratified Cox regression and the asymptotic log-rank test perform the worst. In the most extreme scenarios, the power is about 0.1 below the desired level of 0.8. Note that since there are no other covariates included in the model, the score-statistic of the stratified Cox regression would yield the same curve as the asymptotic stratified

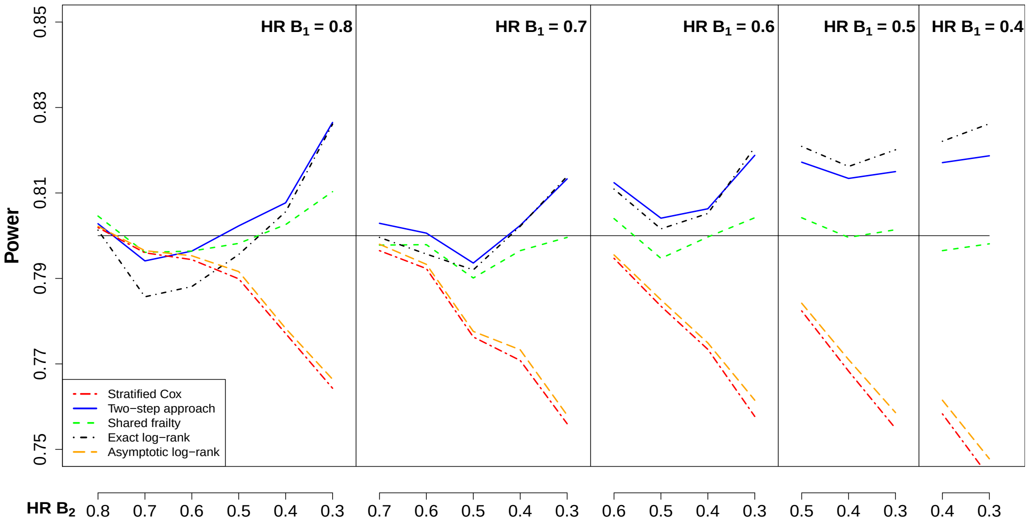


Figure 3.8: Power of the stratified Cox PH model, Mehrotra's two-step approach, the lognormal shared frailty model, and stratified exact and asymptotic log-rank test to detect a true treatment effect when using Palta and Amini's sample size formula under different hazard ratios.

log-rank test. The difference between the two curves is caused by usage of the Wald-statistic for the stratified Cox regression. Since Mehrotra et al. (2012) use the Wald test in their two-step approach, the Wald test was also used for the stratified Cox PH model for a fair comparison.

The shared frailty model, the two-step approach, and the approximate log-rank test perform similarly for small to moderate differences in hazard ratios. For the most extreme scenarios, the shared frailty model yields a slightly lower power than the other two, but does not drop considerably below the desired power.

Mehrotra et al. (2012) suggested two different weighting options for the second step of their two-step approach: "sample size weights", which use the sample proportions of the strata as weights, and "minimum risk weights", which are intended to minimize the mean squared error when estimating β. Both weighting options were tested, but

Table 3.3: Numerical results for the power comparison of the stratified Cox PH model, Mehrotra's two-step approach, the lognormal shared frailty model, and stratified exact and asymptotic log-rank test to detect a true treatment effect when using Palta and Amini's sample size formula under different hazard ratios.

Input		**Power**					
HR	Sample	Strat. Cox	Two-step	Frailty	Frailty	Exact	Asympt.
B_0, B_1, B_2	Size	Wald	Wald	lognorm	Gamma	log-rank	log-rank
0.8 0.8 0.8	763	0.802	0.803	0.805	0.804	0.802	0.802
0.8 0.8 0.7	574	0.796	0.794	0.796	0.795	0.785	0.796
0.8 0.8 0.6	437	0.794	0.796	0.796	0.795	0.788	0.795
0.8 0.8 0.5	336	0.790	0.802	0.798	0.797	0.795	0.792
0.8 0.8 0.4	259	0.777	0.808	0.803	0.805	0.806	0.778
0.8 0.8 0.3	200	0.764	0.827	0.810	0.816	0.825	0.766
0.8 0.7 0.7	461	0.796	0.803	0.798	0.797	0.799	0.798
0.8 0.7 0.6	360	0.792	0.801	0.798	0.797	0.795	0.793
0.8 0.7 0.5	282	0.776	0.794	0.790	0.790	0.791	0.778
0.8 0.7 0.4	222	0.771	0.802	0.796	0.799	0.802	0.773
0.8 0.7 0.3	174	0.756	0.813	0.800	0.805	0.814	0.758
0.8 0.6 0.6	297	0.795	0.812	0.804	0.805	0.812	0.795
0.8 0.6 0.5	238	0.783	0.804	0.795	0.795	0.803	0.785
0.8 0.6 0.4	190	0.773	0.806	0.800	0.802	0.805	0.775
0.8 0.6 0.3	152	0.758	0.819	0.804	0.809	0.821	0.761
0.8 0.5 0.5	200	0.782	0.817	0.804	0.805	0.820	0.784
0.8 0.5 0.4	163	0.768	0.813	0.800	0.802	0.816	0.771
0.8 0.5 0.3	131	0.755	0.815	0.801	0.806	0.820	0.759
0.8 0.4 0.4	138	0.758	0.817	0.796	0.803	0.824	0.761
0.8 0.4 0.3	113	0.744	0.819	0.798	0.804	0.827	0.748

the resulting weights did not differ by much. Therefore, the simpler weights, i.e. the sample proportions, were used for the results presented here.

Overall, the shared frailty model, the two-step analysis, and the approximate log-rank test do not suffer loss of power for any of the scenarios and are hence the preferable choice over the asymptotic log-rank test and the stratified Cox PH model when dealing with heterogeneous treatment effects.

A plot where the roles of the hazard ratios of B_1 and B_2 are interchanged shows similar results and is shown in the sensitivity analysis in Section 3.2.3.4.

The results for the sample size formulas other than the formula by Palta and Amini are shown in Figures 3.9 - 3.11. While Schoenfeld's sample size formula performs well for all analysis methods for small to moderate differences in hazard ratios, the power declines for all methods as the difference between the hazard ratios increases. For the most extreme hazard ratio scenarios, all methods yield a power below 0.8.

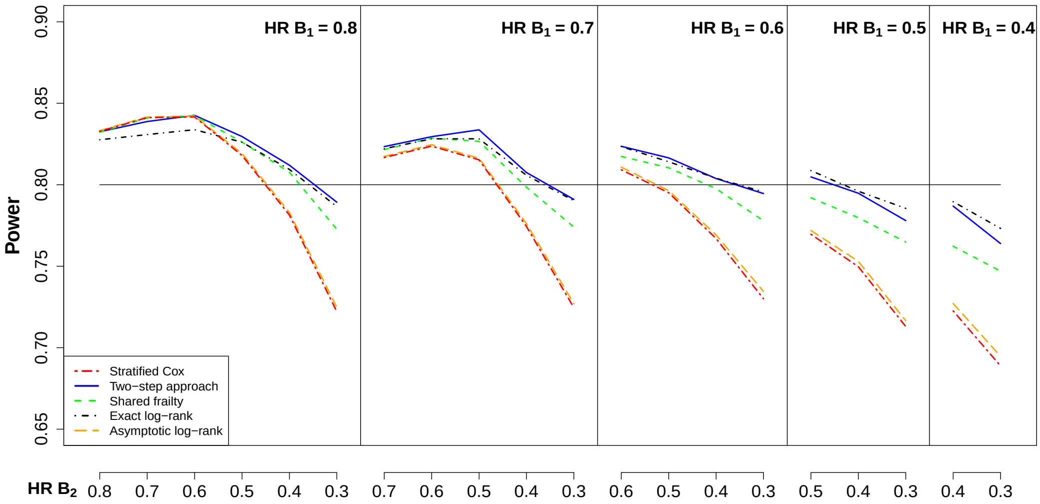


Figure 3.9: Power of the stratified Cox PH model, Mehrotra's two-step approach, the lognormal shared frailty model, and stratified exact and asymptotic log-rank test to detect a true treatment effect when using Schoenfeld's sample size formula under different hazard ratios.

With Lachin and Foulkes' sample size formula, all analysis methods have less than 0.8 power to detect a true treatment effect for all hazard ratio scenarios and for both weighting options for the sample size formulas (Figures 3.10 and 3.11).

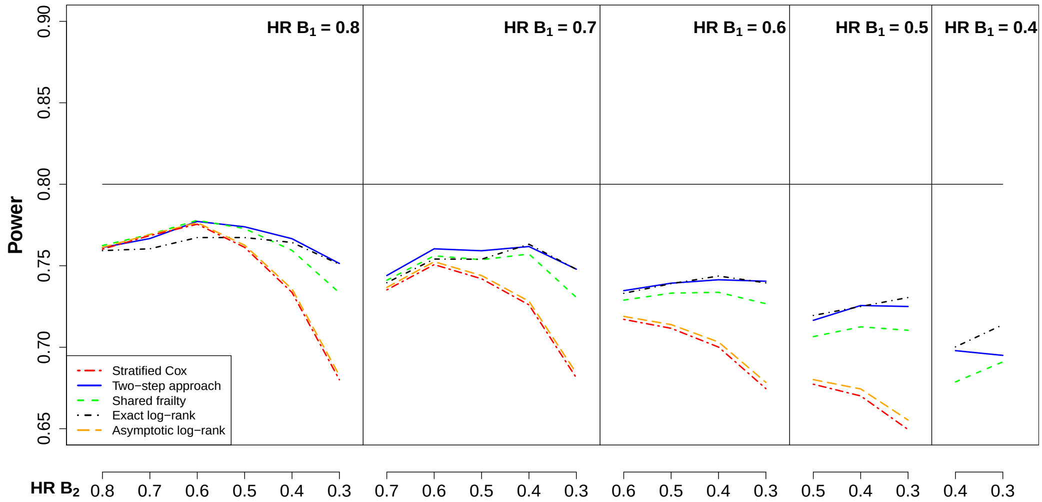


Figure 3.10: Power of the stratified Cox PH model, Mehrotra's two-step approach, the lognormal shared frailty model, and stratified exact and asymptotic log-rank test to detect a true treatment effect when using Lachin and Foulkes' sample size formula under different hazard ratios.

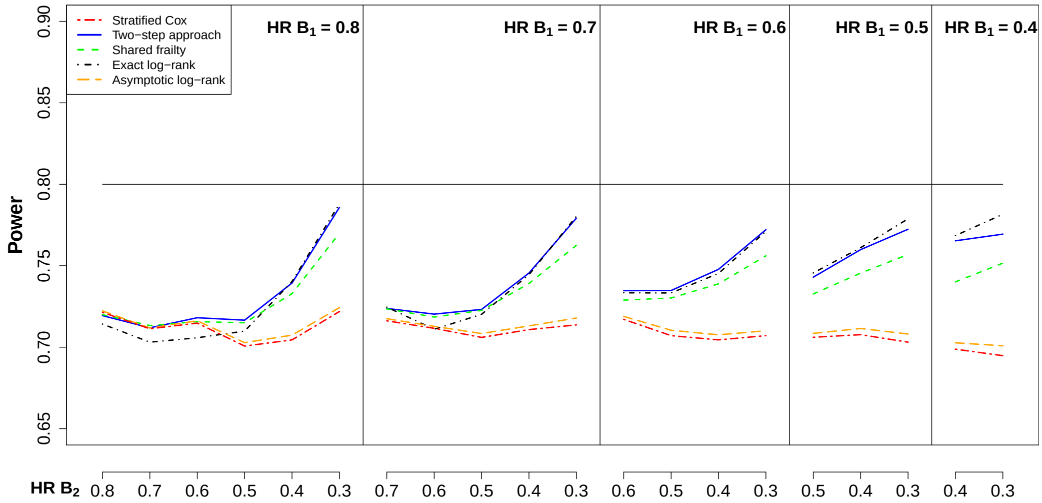


Figure 3.11: Power of the stratified Cox PH model, Mehrotra's two-step approach, the lognormal shared frailty model, and stratified exact and asymptotic log-rank test to detect a true treatment effect when using Lachin and Foulkes' sample size formula (with sample proportions as weights) under different hazard ratios.

3.2.3.4 Sensitivity analysis

Sensitivity analyses were performed to investigate the robustness of the results under different parameter settings and violations of assumptions.

Comparing the asymptotic and the exact log-rank test for the scenarios with the largest differences in hazard ratios, and taking into consideration the calculated sample sizes, an obvious question is whether the reason for the poorer performance of the log-rank test and the stratified Cox PH model is that the assumptions regarding asymptotic properties are not met. For the most extreme scenarios, the calculated sample size is below 300, which, with prevalences of 0.5, 0.25, and 0.25 for B_0, B_1, and B_2 respectively, results in stratum sizes of less than 75 patients for B_1 and B_2. Hence, asymptotic assumptions are problematic in these cases. Another small simulation study was carried out with equal hazard ratios across strata but smaller sample sizes. The results in Table 3.4 show that there is indeed some loss of power, but for an overall sample size of 115 patients (which is very close to the 113 patients in the most extreme case considered), the loss of power is minor. E.g., the stratified Cox PH model has a power of 0.77 as opposed to a power of 0.744 with heterogeneous hazard ratios (see Table 3.3). For control of the type-I-error rate, the exact log-rank test should be considered instead if the strata sizes are expected to be small, i.e. in the double digits, and there is too little data available for reliable approximations. The shared frailty

Table 3.4: Check of asymptotic properties: Numerical results for the power comparison of the stratified Cox PH model, Mehrotra's two-step approach, the lognormal shared frailty model, and stratified exact and asymptotic log-rank test to detect a true treatment effect when using Palta and Amini's sample size formula under homogeneity of hazard ratios but under small sample sizes.

Input		**Power**				
HR B_0, B_1, B_2	Sample Size	Strat. Cox Wald	Two-step Wald	Frailty Lognorm	Exact log-rank	Asympt. log-rank
0.70 0.70 0.70	306	0.801	0.806	0.809	0.804	0.802
0.65 0.65 0.65	213	0.792	0.796	0.800	0.794	0.794
0.60 0.60 0.60	154	0.787	0.793	0.797	0.791	0.790
0.55 0.55 0.55	115	0.779	0.788	0.796	0.789	0.784
0.50 0.50 0.50	87	0.770	0.777	0.794	0.786	0.776
0.45 0.45 0.45	67	0.756	0.745	0.784	0.772	0.762
0.40 0.40 0.40	53	0.759	0.683	0.789	0.778	0.769

model also appears to be an appropriate choice and additionally offers the possibility to include covariates in the model, which is an advantage over the exact log-rank test.

In Figure 3.12, the role of the hazard ratios for B_1 and B_2 was interchanged, i.e. within each of the plot windows, the hazard ratio for B_2 is fixed and the hazard ratio for B_1 varies, rather than the other way around (as in Figure 3.8). The results shown in the plot are similar to Figure 3.8. The minor differences that can be seen could be caused by the different baseline hazards of the two strata.

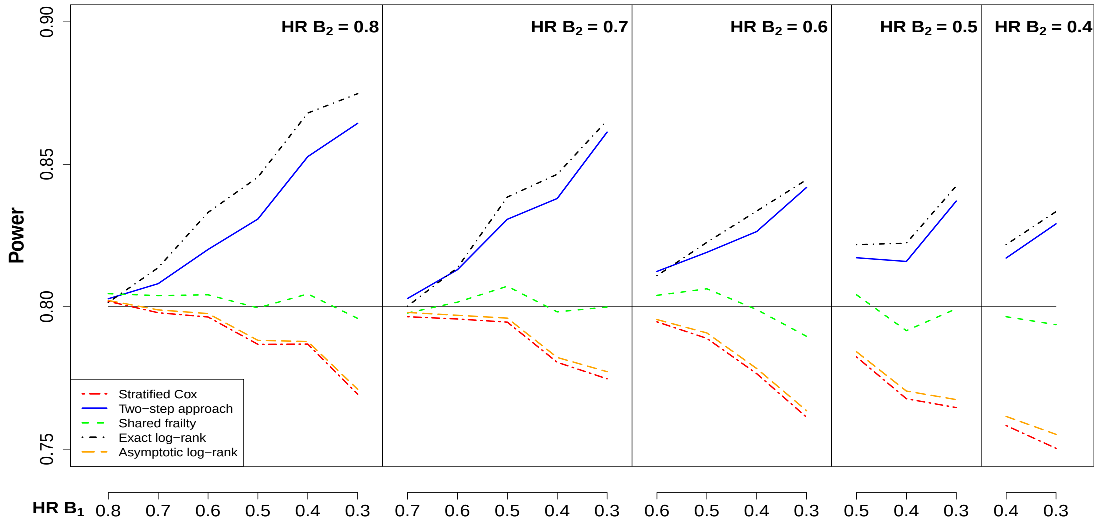


Figure 3.12: Role of hazard ratios of B_1 and B_2 exchanged: Power of the stratified Cox PH model, Mehrotra's two-step approach, the lognormal shared frailty model, and stratified exact and asymptotic log-rank test to detect a true treatment effect when using Schoenfeld's sample size formula under different hazard ratios.

It was also investigated whether larger differences in baseline hazards, i.e. stronger stratification, have an impact on the results. Stratification factors 1, 0.5, 1.5 were used, leading to baseline hazards 0.05, 0.025, 0.075 for B_0, B_1, and B_2, respectively. Figure 3.13 shows that there is some impact, but especially for the more extreme hazard ratio scenarios the two-step procedure and the approximate log-rank test still have a power close to 0.8.

As discussed in section 3.2.3.2, it is possible to simulate a stratum specific random effect in the data, as is assumed by the shared frailty model. The results for a minor, a moderate, and a stronger random effect (with lognorm(0, 0.15), lognorm(0, 0.3), and lognorm(0, 0.5), respectively) can be seen in Figures 3.14, 3.15, and 3.16, respectively. Note the different scaling of the y-axes compared to the previous sections.

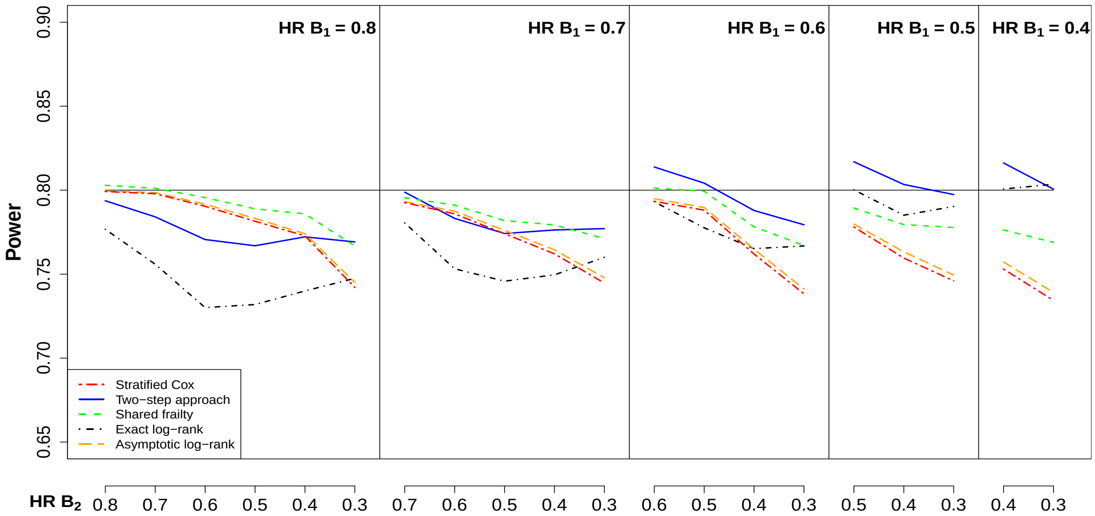


Figure 3.13: Power of the stratified Cox PH model, Mehrotra's two-step approach, the lognormal shared frailty model, and stratified exact and asymptotic log-rank test to detect a true treatment effect when using Palta and Amini's sample size formula under different scenarios for stronger stratification (stratification factors 1, 0.5, 1.5 leading to baseline hazards 0.05, 0.025, 0.075 for B_0, B_1 and B_2, respectively).

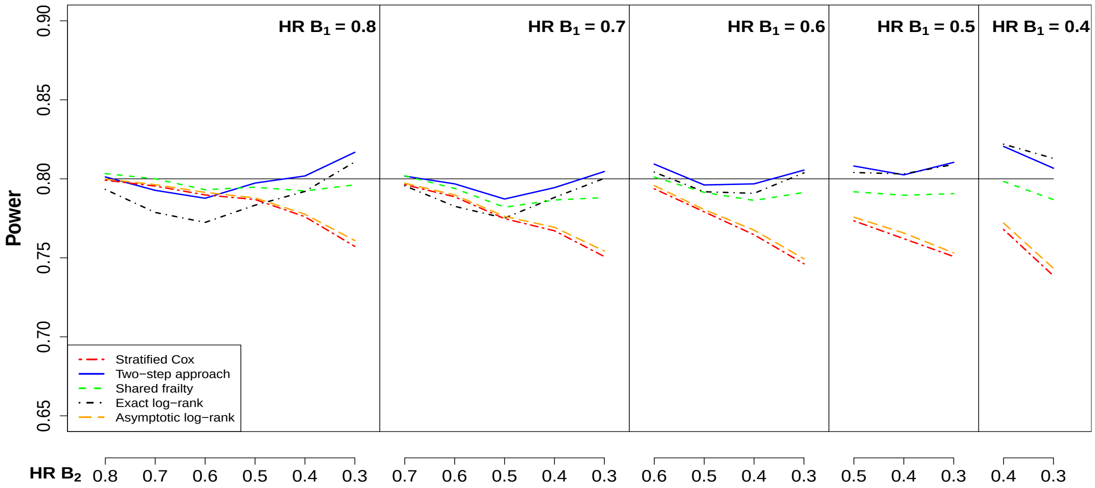


Figure 3.14: Power of the stratified Cox PH model, Mehrotra's two-step approach, the lognormal shared frailty model, and stratified exact and asymptotic log-rank test to detect a true treatment effect when using Palta and Amini's sample size formula under different hazard ratio scenarios and data with minor random effect (lognorm(0, 0.15)).

While the minor random effect does not have much of an impact compared to Figure 3.8, the analysis methods do suffer some power loss in the presence of a moderate random effect. The largest power loss happens for the strong random effect. It can also be observed that the shared frailty model loses less power than the other methods, while the exact log-rank test does not appear to handle the random effect very well.

Finally, it was investigated how the analysis methods perform under a misspecified censoring distribution, under dependent censoring and under misspecification of the survival distribution.

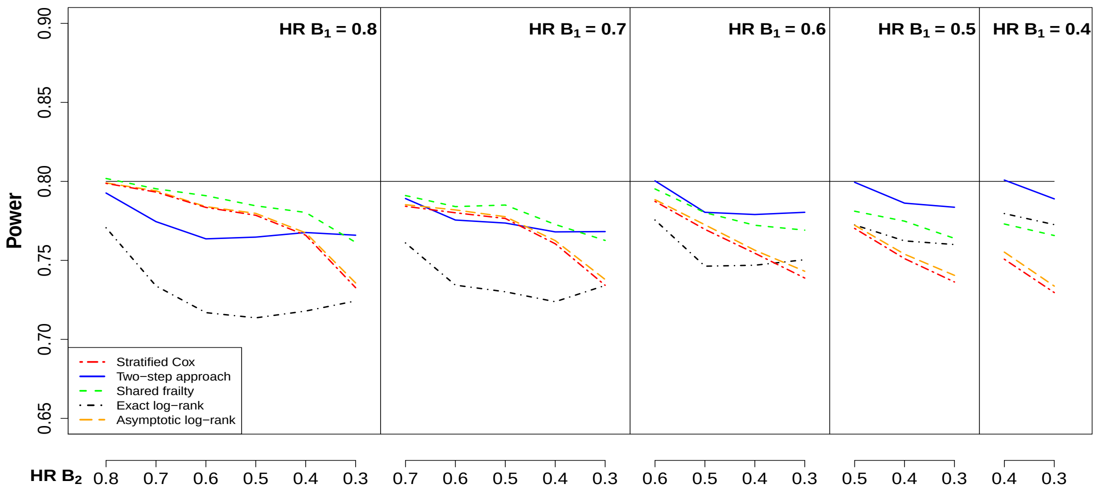


Figure 3.15: Power of the stratified Cox PH model, Mehrotra's two-step approach, the lognormal shared frailty model, and stratified exact and asymptotic log-rank test to detect a true treatment effect when using Palta and Amini's sample size formula under different hazard ratio scenarios and data with moderate random effect (lognorm(0, 0.3)).

Figure 3.17 shows the results for Weibull distributed censoring when exponential censoring is assumed and Figure 3.18 shows the results for dependent censoring when independent censoring assumed. In both cases, there is only a minor impact on the results compared to Figure 3.8. Figure 3.19, on the other hand, shows that the models are sensitive to a misspecified survival distribution. The data were simulated with Weibull distributed survival times with a shape parameter of 0.8 while the methods assume exponential survival.

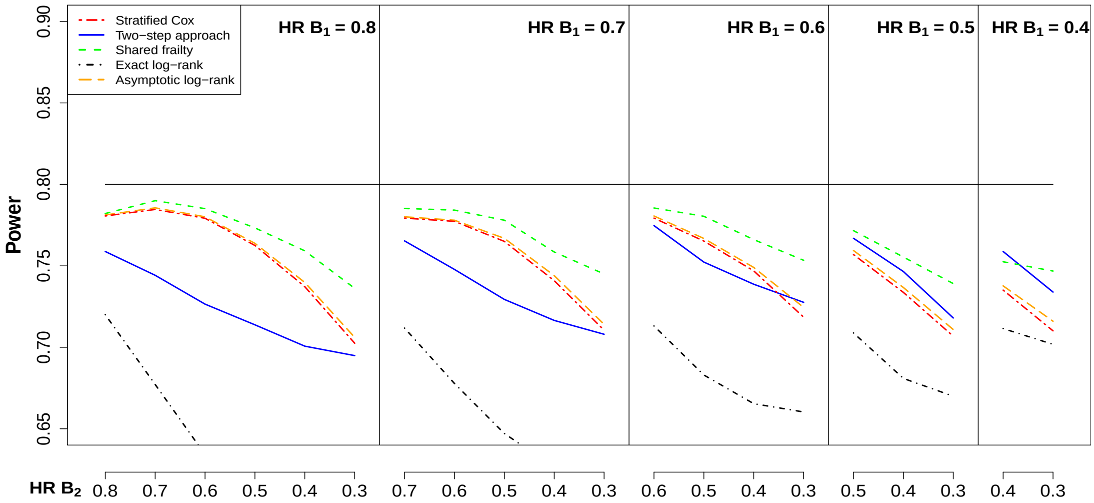


Figure 3.16: Power of the stratified Cox PH model, Mehrotra's two-step approach, the lognormal shared frailty model, and stratified exact and asymptotic log-rank test to detect a true treatment effect when using Palta and Amini's sample size formula under different hazard ratio scenarios and data with strong random effect (lognorm(0, 0.5)).

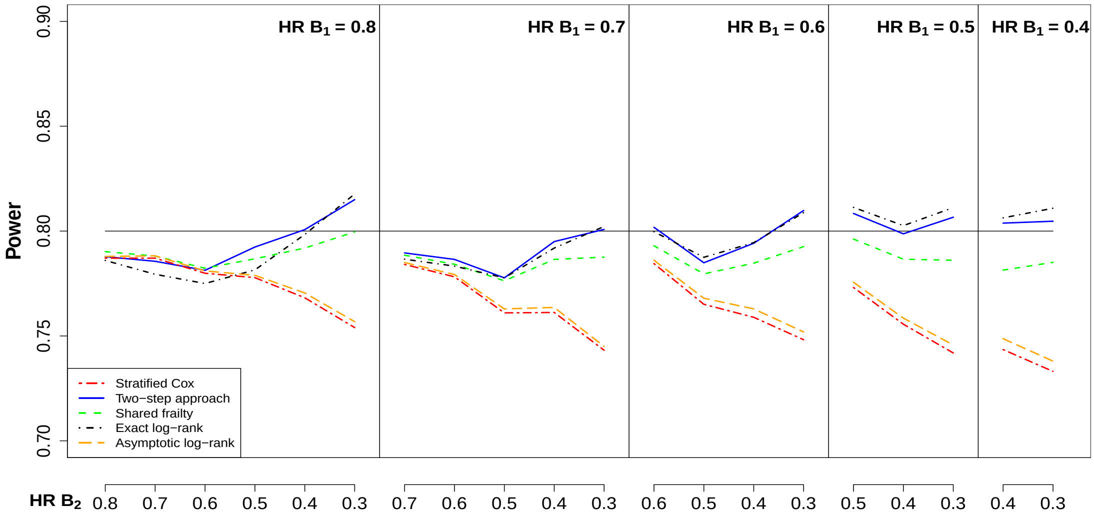


Figure 3.17: Sensitivity to misspecified censoring distribution: Weibull distributed censoring with shape 0.8. Power of the stratified Cox PH model, Mehrotra's two-step approach, the lognormal shared frailty model, and stratified exact and asymptotic log-rank test to detect a true treatment effect when using Lachin and Foulkes' sample size formula.

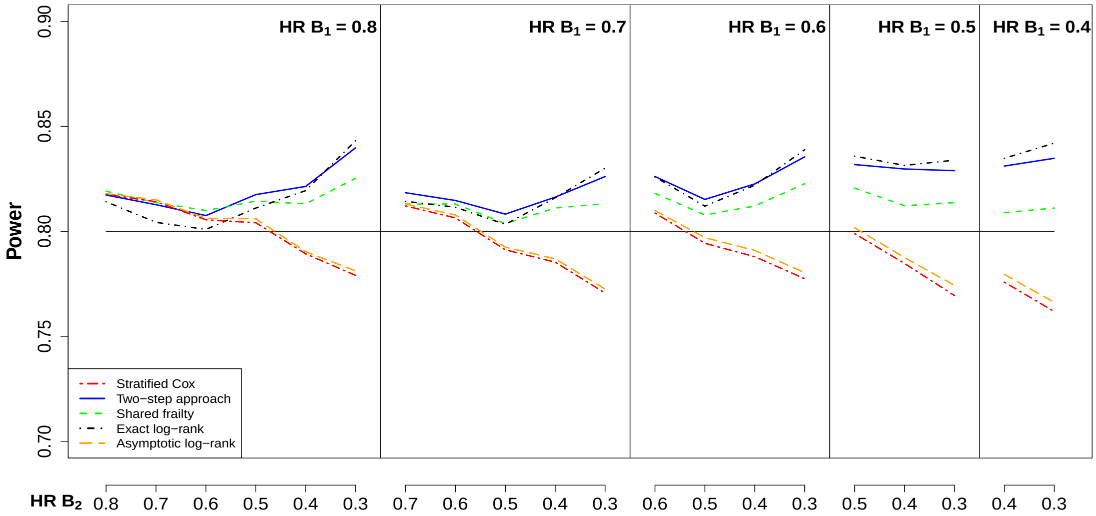


Figure 3.18: Sensitivity to non-independent censoring. Power of the stratified Cox PH model, Mehrotra's two-step approach, the lognormal shared frailty model, and stratified exact and asymptotic log-rank test to detect a true treatment effect when using Lachin and Foulkes' sample size formula.

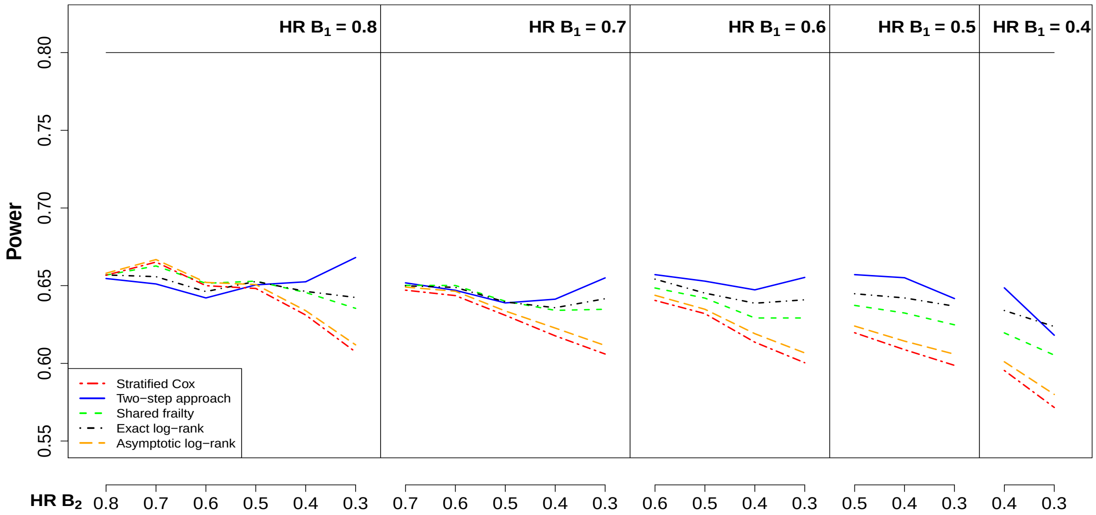


Figure 3.19: Sensitivity to misspecified survival distribution: Weibull distributed survival with shape 0.8. Power of the stratified Cox PH model, Mehrotra's two-step approach, the lognormal shared frailty model, and stratified exact and asymptotic log-rank test to detect a true treatment effect when using Lachin and Foulkes' sample size formula.

3.2.4 Data example

The approach of investigating several biomarker-defined groups with corresponding stratum-specific experimental treatments within a single clinical trial has only just emerged in recent years. Therefore, most trials of this kind are either still in the planning or recruitment stage, making it difficult to obtain data for a study design as it is considered here. For illustration purposes, data were taken from three studies carried out by the German-Austrian Acute Myeloid Leukemia Study Group (AMLSG) and subsets of the data were combined to be used as example data set. For stratum B_1, patients from the 06-04 study (Tassara et al. 2014) with mutated Nucleophosmin-1 (NPM1) were chosen (n=40). Stratum B_2 is comprised of patients from the 07-04 study (Schlenk et al. 2016) with internal tandem duplication mutations of FMS-like tyrosine kinase 3 (FLT3-ITD) (n=139). The data from the study HD98B (Schlenk et al. 2004) were used for stratum B_0, excluding patients with mutated NPM1 and FLT3-ITD (n=144). This resulted in a data set with 323 patients. The survival curves for the three strata can be seen in Figure 3.20.

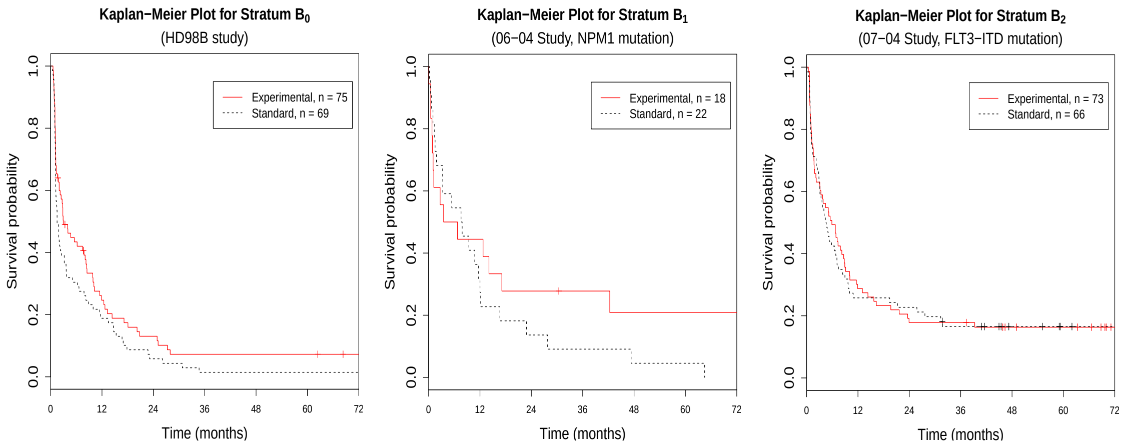


Figure 3.20: Kaplan-Meier plots for event-free survival for the data from the German-Austrian Acute Myeloid Leukemia Study Group (AMLSG). The x-axes were cut at 6 years.

Just like the simulated data, the data set was then analyzed using the asymptotic and exact log-rank test, the stratified Cox PH model, the two-step approach, and the shared frailty model. The resulting hazard ratio estimates for event-free survival were

0.81 for the stratified Cox PH model, 0.82 for the two-step approach, and 0.83 for the lognormal shared frailty model (Table 3.5). The hazard ratios for the individual strata were also estimated with 0.74, 0.68, and 0.96 for B_0, B_1, and B_2, respectively. Due to the small sample sizes, especially in B_1, one might consider using an unconventional α of 0.1. Then, rejection of the null hypothesis in this example depends on the analysis method used. While the null hypothesis would be rejected using the stratified Cox PH model, the two-step approach, and the asymptotic log-rank test, it cannot be rejected using the lognormal shared frailty model and the exact log-rank test. Note that the performance of the shared frailty model could be influenced by more complex underlying baseline hazards than the constant ones in the simulation study.

The performed overall analysis can be understood as assessing the benefit of using targeted therapies in the overall patient population. Such an overall analysis should be performed in conjunction with subsequent subgroup analyses to avoid drawing false conclusions, as it could have been the case for B_2 in this example. But especially in small sample situations, an evaluation of the overall targeted treatment strategy can be a useful tool to guide further analysis and procedure.

Table 3.5: Hazard ratios with corresponding p-values for the data from the German-Austrian Acute Myeloid Leukemia Study Group (AMLSG) estimated with the stratified Cox PH model, the two-step approach, and the shared frailty model. Additionally, p-values are given for the stratified exact and asymptotic log-rank test.

Analysis	Method	HR	90% C.I.	p-value
Overall	Stratified Cox	0.81	(0.67, 0.99)	0.08
	Lognormal shared frailty	0.83	(0.68, 1.01)	0.11
	Two-step approach	0.82	(0.67, 0.99)	0.09
	Exact log-rank			0.13
	Asymptotic log-rank			0.08
Individual	Cox B_0	0.74	(0.55, 0.98)	0.08
	Cox B_1	0.68	(0.38, 1.20)	0.26
	Cox B_2	0.96	(0.71, 1.31)	0.83

From the results of this study, it can be concluded that the assumption of homogeneous treatment effects is not appropriate in this context. Hence, the heterogeneity of treatment effects supports using the sample size formula by Palta and Amini instead of the formula by Lachin and Foulkes, since the latter does not facilitate the option to use stratum-specific treatment effects within the formula. Furthermore, the two-step analysis seems like a reasonable first step in a multi-stage analysis. Looking at stratum 2 with a hazard ratio of 0.96, one could also consider using an adaptive design for future studies to be able to stop strata that perform poorly early in the study.

3.3 Multiple biomarker hybrid design

When a design with multiple biomarkers, such as the multiple biomarker hybrid design, is used for a less prevalent disease, it may be difficult to recruit enough patients to each subtrial to obtain statistically meaningful results within a reasonable time frame. In this case, small sample sizes within the subtrials have to be expected, as well as many biomarker-negative patients at the initial screening stage, i.e. patients which test negative for all relevant biomarkers. The small sample sizes may make it unfeasible to analyze the subtrials individually. Moreover, the small sample sizes can lead to biased treatment effect estimates. This imposes the need to investigate alternative approaches for the analysis of such a trial, and possibly for the study design itself. Measures should be taken to reduce the potential bias of the treatment effect estimates. Additionally, with an expected large group of biomarker-negative patients, it seems reasonable to explore options to include them in such a trial and potential benefits to the trial through their inclusion, such as collection of additional data, improving power, or reducing bias.

For the following sections, Design 4 from Figure 3.1 will be considered. The biomarker-groups are denoted by B_i, $i = \{0, 1, 2\}$, where B_0 stands for biomarker-negative. Upon entering the study, the patients' biomarker-profile is determined and they are assigned to the biomarker-groups accordingly. The proportion of patients in biomarker-group i is denoted by g_i. With a total sample size of n patients, this results in a total number of $n_i = g_i n$ patients in group i. Within B_1 and B_2, patients are randomized between the biomarker-specific experimental therapy (Exp_1 and Exp_2, respectively) and standard of care (Std), with probabilities $1 - r_i$ and r_i, respectively. Patients in B_0 are treated with standard of care only. The baseline hazard for patients in biomarker-group i is denoted by λ_{0_i}. The hazards of death at time t for biomarker-positive patients in biomarker-group i receiving treatment j are denoted by $\lambda_{j_i}(t)$, where $j = 1$ for experimental treatment or $j = 0$ for standard of care.

3.3.1 Data modeling and analysis

A prognostic biomarker can be modeled by allowing different baseline hazards for the biomarker-groups (stratified Cox PH model). If a biomarker is predictive on the other hand, it causes the treatment effects for the biomarker-groups to be different, which cannot be modeled using the stratified Cox PH model, since it assumes homogeneous treatment effects. The inclusion of a single treatment arm for biomarker-negative patients in the regression model for the multi-biomarker hybrid design (Figure 3.1, Design 4) adds the difficulty that data analysis methods for a stratified analysis, such as the stratified Cox PH model, the shared frailty model, or the two-step approach by Mehrotra et al. (2012), as used in Section 3.2 for the stratified randomize-all design, are not applicable in this situation.

There are two possibilities to include the data from the biomarker-negative patients in the analysis. For non-prognostic biomarkers the three standard of care arms could simply be pooled. For prognostic biomarkers, however, this approach is not appropriate. Since it is quite common for biomarkers to be prognostic, it was investigated to include the biomarker status in the Cox PH model as a factor variable (with dummy variables b_1 and b_2) to account for the prognostic effect, as an alternative to using a stratified Cox PH model.

To evaluate the benefit of this strategy and of including biomarker-negative patients the following approaches were compared:

Approach 1: Separate models for both biomarkers, using data only from B_1 patients *or* B_2 patients, respectively:

$$\lambda_1 = \lambda_{0_1} \exp(\beta_1 x_1) \qquad \text{(sample size: } n_1\text{, all patients in B}_1\text{)}$$
$$\lambda_2 = \lambda_{0_2} \exp(\beta_2 x_2) \qquad \text{(sample size: } n_2\text{, all patients in B}_2\text{)}$$

The parameters β_1 and β_2 represent the treatment effects for Exp_1 and Exp_2, respectively. This is equivalent to using an enrichment design with two biomarkers with separate analyses for the biomarkers.

Approach 2: A model performing a combined analysis, using data from B$_1$ patients *and* B$_2$ patients, but excluding biomarker-negative patients (B$_0$):

$$\lambda = \lambda_{0_1} \exp(\gamma_2 b_2 + \beta_1 x_1 + \beta_2 x_2) \quad \text{(sample size: } n_1 + n_2\text{, all patients in B}_1 \text{ and B}_2\text{)}$$

The parameters β_1 and β_2 represent the treatment effects for Exp$_1$ and Exp$_2$, respectively, and γ_2 is the prognostic effect of B$_2$ (with dummy variable b_2) with B$_1$ as reference. This is equivalent to using an enrichment design with two biomarkers with a combined analysis for the biomarkers.

Approach 3: A model performing a combined analysis, using the entire data set (B$_1$, B$_2$, and B$_0$)

$$\lambda = \lambda_0 \exp(\gamma_1 b_1 + \gamma_2 b_2 + \beta_1 x_1 + \beta_2 x_2) \qquad \text{(sample size: } n\text{, all patients)}$$

The parameters β_1 and β_2 represent the treatment effects for Exp$_1$ and Exp$_2$, respectively, and γ_1 and γ_2 are the prognostic effects of B$_1$ and B$_2$ (with dummy variables b_1 and b_2), respectively, with B$_0$ as reference. This uses all data available from the multi-biomarker hybrid design.

Note that these three approaches use different sample sizes due to the exclusion of biomarker-groups in Approaches 1 and 2.

For lower prevalence biomarkers, n_1 and n_2 have to be expected to be rather small compared to n. While estimates of maximum-likelihood methods, such as Cox regression, are asymptotically unbiased, this is not necessarily the case for finite samples (cf. Cordeiro and McCullagh 1991). Hence, especially for small samples, the estimates of the Cox regression can be biased. Langner et al. (2003) investigated the relationship of bias to sample size for logistic and Cox regression. In their simulation study they found that the bias from maximum likelihood methods depends on sample size, but also on baseline hazard and treatment hazard ratio. They found a strong bias for extreme baseline risks and extreme treatment hazards, and also for small numbers of events in the control group.

The bias of estimators due to small sample sizes and rare events can be reduced by using a penalized likelihood based on a modified score function proposed by Firth (1993), the so-called Firth correction or Firth penalty (see Section 2.1.1.4). The bias and root mean squared error (RMSE) of estimators when using the Firth penalty in situations with a small number of events was investigated by Lin et al. (2013). They reported that for a small number of events per variable, Firth's approach had less absolute value of relative bias and a smaller mean squared error (MSE) compared to the Cox PH model.

3.3.2 Simulation study: Small sample size bias

A simulation study was conducted to compare the performance of the three approaches discussed in Section 3.3.1 with respect to bias, standard deviation and RMSE of the different parameter estimates, and to investigate whether there is a benefit of including biomarker-negative patients in the study. Moreover, it was examined to use the Firth correction to reduce the small sample size bias of the parameter estimates.

3.3.2.1 Study design and data generation

The data were generated as described in Section 2.2.1. The hazard ratios of B$_1$ and B$_2$, $\exp(\beta_1)$ and $\exp(\beta_2)$, were varied between 0.8 and 0.4. Note that, due to the single treatment arm, no hazard ratio can be specified for B$_0$. For B$_0$, the baseline hazard λ_0 was chosen to be 0.05 and then the baseline hazards for B$_1$ and B$_2$, λ_{0_1} and λ_{0_2}, were determined by multiplying 0.05 by factors 0.5 and 2, respectively, i.e., $\gamma_1 = \log(0.5)$ and $\gamma_2 = \log(2)$. This was done to simulate a biomarker indicative of a favorable and a poor prognosis, respectively. Survival and random censoring were assumed to be exponential and the allocation ratio between treatments for B$_1$ and B$_2$ was set to 1:1. Time will be measured in months. Accrual time a was chosen to be 24 months and follow-up time f 36 months. For the remaining simulation parameters see Table 3.6. For this simulation study, the censoring proportion p_{cens} was set to 0.05.

The resulting mean overall censoring proportions over 10,000 simulation runs can be found in Table 3.7.

Table 3.6: Parameters for the simulation study using a design with three biomarker-groups. Note that due to the single treatment arm there is no hazard ratio for B_0.

Fixed simulation parameters	
Accrual time (months), a	24
Follow-up time (months), f	36
Proportion random censoring, p_{cens}	0.05
Treatment allocation ratio	$1:1$
Hazard ratio B$_1$, $\exp(\beta_1)$	$0.8, 0.7, 0.6, 0.5, 0.4$
Hazard ratio B$_2$, $\exp(\beta_2)$	$0.8, 0.7, 0.6, 0.5, 0.4$
$\exp(\gamma_1)$	0.5
$\exp(\gamma_2)$	2
Sample size, n	100, 150, 250, 1,000
Parameters for biomarker-groups $(\mathbf{B_0}, \mathbf{B_1}, \mathbf{B_2})$	
Proportion of patients, g_i	(0.5, 0.25, 0.25)
Baseline hazards, λ_{0_i}	(0.05, 0.025, 0.1)

Different sample sizes were considered in the simulation study: smaller sample sizes with 100 and 150 patients in the study, a moderate sample size with 250 patients, and a large sample size with 1,000 patients.

The three approaches discussed in section 3 were compared with respect to bias, standard deviation and RMSE of the estimates $\hat{\beta}_1, \hat{\beta}_2, \hat{\gamma}_1$ and $\hat{\gamma}_2$ out of $S = 10,000$ simulations. The bias and standard deviation of the estimates were calculated as

$$bias(\hat{\beta}) = \frac{1}{S}\sum_{l=1}^{S} \hat{\beta}_l - \beta \qquad \text{and} \qquad s(\hat{\beta}) = \sqrt{\frac{1}{S-1}\sum_{l=1}^{S}(\hat{\beta}_l - \bar{\hat{\beta}})^2}.$$

The MSE was then calculated as $MSE(\hat{\beta}) = bias(\hat{\beta})^2 + s^2(\hat{\beta})$ and the RMSE as $RMSE(\hat{\beta}) = \sqrt{MSE(\hat{\beta})}$. As a result of the consideration of extreme scenarios, the algorithm did not converge for some simulation runs due to lack of events in one group.

These runs were excluded from the analyses (between 0 and 1% of the runs, depending on the hazard ratios of B_1 and B_2).

To simulate a non-constant baseline hazard for the sensitivity analysis, data were simulated where the hazard function is given by a Weibull distribution with shape parameters 0.4 and 5. A shape parameter of 1 corresponds to the exponential distribution. For better comparability, λ_0 was adjusted such that the same number of events is reached at 60 months for all shape parameters, i.e. $\lambda_0 = 0.05 \cdot 60/60^{\mathsf{shape}}$. Additionally, the simulations were run without censoring to ensure equal numbers of events.

Table 3.7: Mean number of events over 10,000 simulations for the individual treatment arms and overall censoring proportion (both administrative and random) for n = 100, 150, 250, and 1,000.

$n = 100$							
HR of B_1	Events B_1 Exp	Events B_1 Std	cens. prop.	HR of B_2	Events B_2 Exp	Events B_2 Std	cens. prop.
0.8	7.3	8.3	0.17	0.8	11.7	12.0	0.17
0.7	6.7	8.3	0.18	0.7	11.5	12.0	0.17
0.6	6.0	8.3	0.19	0.6	11.2	12.0	0.18
0.5	5.3	8.3	0.19	0.5	10.8	12.0	0.18
0.4	4.5	8.3	0.20	0.4	10.1	12.0	0.19

$n = 150$							
HR of B_1	Events B_1 Exp	Events B_1 Std	cens. prop.	HR of B_2	Events B_2 Exp	Events B_2 Std	cens. prop.
0.8	10.9	12.4	0.17	0.8	17.6	18.0	0.17
0.7	10.0	12.4	0.18	0.7	17.3	18.1	0.17
0.6	9.1	12.4	0.19	0.6	16.9	18.1	0.18
0.5	8.0	12.4	0.19	0.5	16.2	18.1	0.18
0.4	6.7	12.4	0.20	0.4	15.1	18.1	0.19

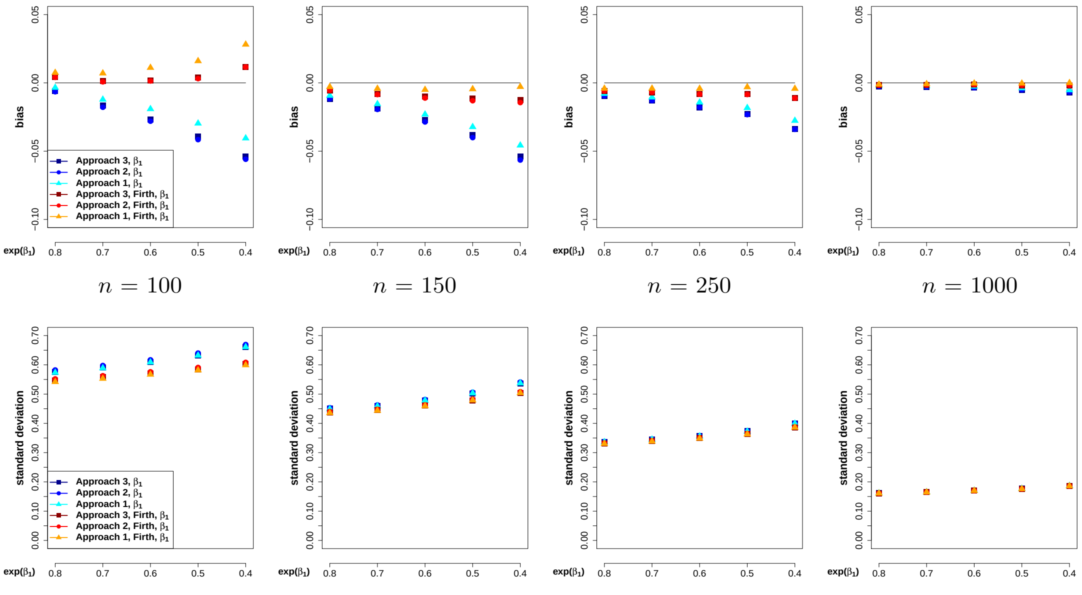
$n = 250$							
HR of B_1	Events B_1 Exp	Events B_1 Std	cens. prop.	HR of B_2	Events B_2 Exp	Events B_2 Std	cens. prop.
0.8	18.1	20.6	0.17	0.8	29.4	30.0	0.17
0.7	16.7	20.6	0.18	0.7	28.9	30.1	0.17
0.6	15.0	20.6	0.19	0.6	28.2	30.1	0.18
0.5	13.2	20.6	0.19	0.5	27.1	30.1	0.18
0.4	11.1	20.6	0.20	0.4	25.3	30.1	0.19

$n = 1,000$							
HR of B_1	Events B_1 Exp	Events B_1 Std	cens. prop.	HR of B_2	Events B_2 Exp	Events B_2 Std	cens. prop.
0.8	72.5	82.3	0.17	0.8	118.1	120.4	0.17
0.7	66.7	82.3	0.18	0.7	116.1	120.4	0.17
0.6	60.3	82.3	0.19	0.6	113.1	120.5	0.18
0.5	52.9	82.4	0.19	0.5	108.6	120.6	0.18
0.4	44.7	82.4	0.20	0.4	101.6	120.7	0.19

3.3.2.2 Comparison of analysis approaches

The three approaches discussed in Section 3.3.1 were compared with respect to bias, standard deviation and RMSE of the different parameter estimates out of 10,000 simulations. The bias correction by Firth (1993) was applied in an additional analysis and the bias, standard deviation, and RMSE were again compared for the three approaches.

For β_1, the treatment effect for the biomarker indicative of a favorable prognosis (B_1), the bias and standard deviation of the estimate $\hat{\beta}_1$ for different sample sizes are shown in Figure 3.21. The Figures showing the RMSE of $\hat{\beta}_1$ (and also of the other estimates) can be found in the Appendix, since there were only minor visible differences between standard deviation and RMSE. For all approaches and sample sizes, it can be observed that bias, standard deviation, and RMSE increase in absolute terms as the hazard ratio for B_1, $\exp(\beta_1)$, gets smaller, i.e. as the treatment effect gets larger. Without Firth correction, Approach 1 yields a slightly smaller bias than Approach 2 and 3, which perform similarly. The differences between the approaches get smaller as the sample size increases. For standard deviation and RMSE all three approaches perform similarly.



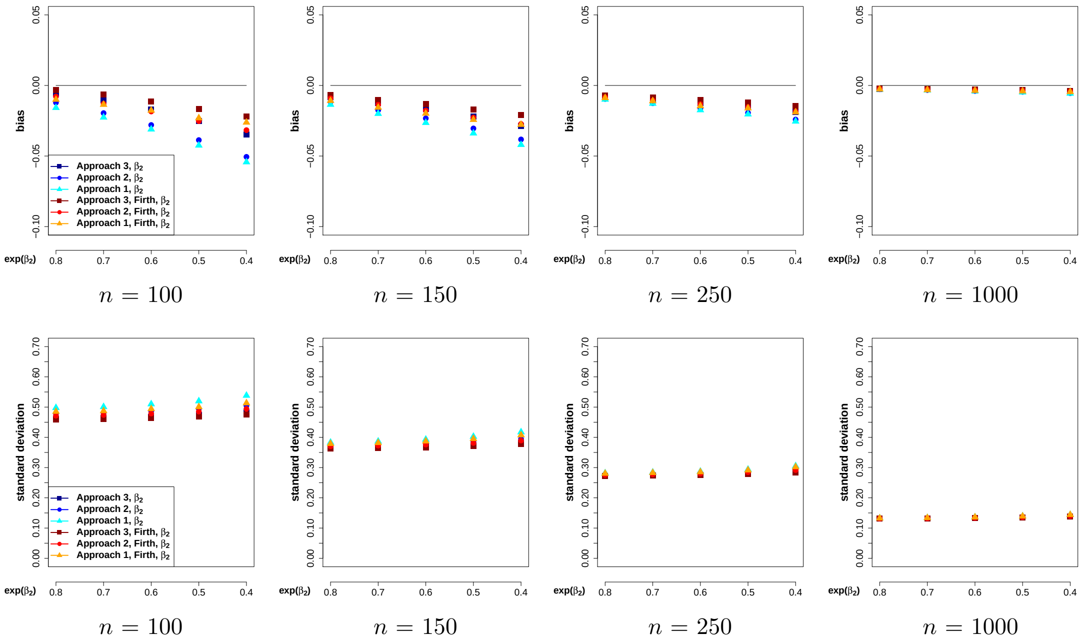
$n = 100$ $n = 150$ $n = 250$ $n = 1000$

$n = 100$ $n = 150$ $n = 250$ $n = 1000$

Figure 3.21: Bias and standard deviation of the estimate of log hazard ratio β_1, the treatment effect estimate for the biomarker indicative of a favorable prognosis (B_1), using a Cox PH model without and with Firth correction for sample sizes 100, 150, 250, and 1,000.

The Firth correction was applied to all three approaches. It is able to reduce the bias for all three approaches but appears to over-correct the bias for Approach 1, and for $n = 100$ also for Approach 2 and 3. The Firth correction offers a slight reduction in standard deviation and RMSE for all three approaches, which all perform similarly.

For β_2, the treatment effect for the biomarker indicative of a poor prognosis (B_2), the bias and standard deviation of the estimate $\hat{\beta}_2$ for different sample sizes are shown in Figure 3.22. For all approaches and sample sizes, it can be observed that the bias, as well as standard deviation and RMSE, increases in absolute terms as the hazard ratio for the treatment effect of biomarker 2, $\exp(\beta_2)$, gets smaller.



$n = 100$ $n = 150$ $n = 250$ $n = 1000$

$n = 100$ $n = 150$ $n = 250$ $n = 1000$

Figure 3.22: Bias and standard deviation of the estimate of log hazard ratio β_2, the treatment effect estimate for the biomarker indicative of a poor prognosis (B_2), using a Cox PH model without and with Firth correction for sample sizes 100, 150, 250, and 1,000.

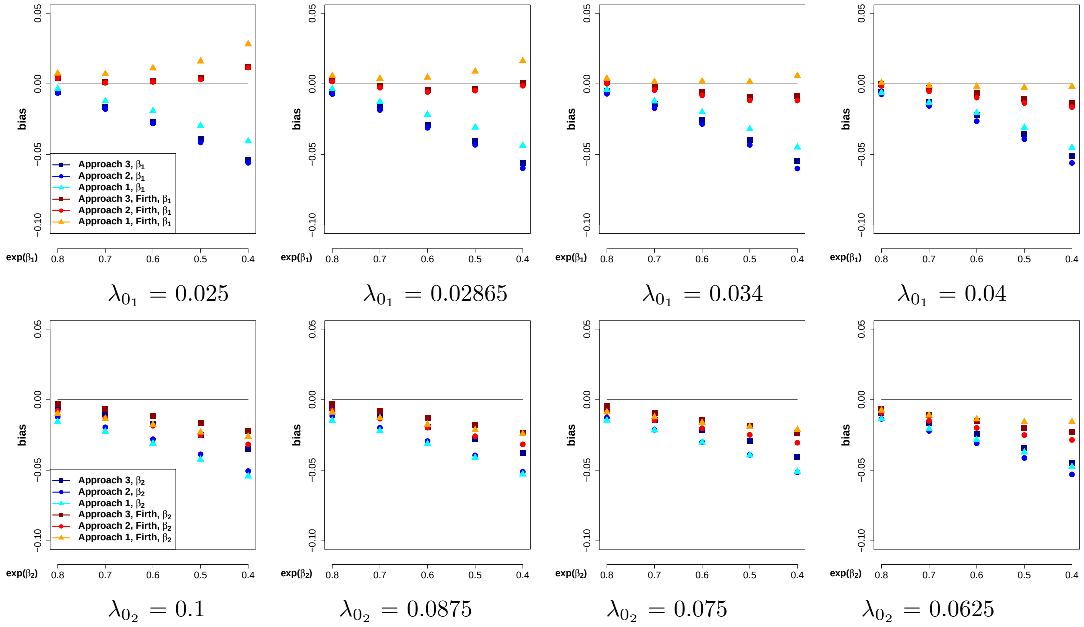
Comparing the three approaches, it can be seen that for the small to moderate sample sizes, the bias of $\hat{\beta}_2$ is approximately similar for Approach 1 and Approach 2, while Approach 3 yields a smaller bias. This difference gets larger as the treatment effect increases. With increasing sample size, the differences in bias of $\hat{\beta}_2$ between the three approaches get smaller. For $n = 1,000$ there is no longer a visible difference. Approach 1 yields the largest standard deviation and RMSE, and Approach 3 the smallest. Again, these differences get smaller with increasing sample size.

The Firth correction was applied to all three approaches. The results are also shown in Figure 3.22. Just as before, for all approaches and sample sizes, the bias, standard deviation, and RMSE increase in absolute terms as the hazard ratio for biomarker 2, $\exp(\beta_2)$, gets smaller.

For $n = 100$ and $n = 150$, a difference in bias between Approaches 1 and 2 can now be observed. For the larger hazard ratios, Approach 2 yields a slightly smaller bias, but as the hazard ratio decreases, the differences get smaller and for the smaller hazard ratios, Approach 1 yields the smaller bias. For standard deviation and RMSE, there is a similar situation for the estimates without Firth correction: Approach 1 yields the largest and Approach 3 the smallest result, but the differences between the approaches are now smaller and again get smaller with increasing sample size. Note that the true value of β_1 does not affect the bias and standard error for $\hat{\beta}_2$ and vice versa.

A comparison of bias, standard deviation, and RMSE of $\hat{\gamma}_1$ and $\hat{\gamma}_2$ for Approach 3 is shown in Figure A.3 in the Appendix. The estimates are almost constant for each sample size and do not change with increasing bias of the treatment effect estimates. Overall, the bias, standard deviation, and RMSE of $\hat{\gamma}_1$ and $\hat{\gamma}_2$ get smaller with increasing sample size. Note that bias and standard deviation of $\hat{\gamma}_1$ and $\hat{\gamma}_2$ are not affected by the values of β_1 and β_2. This allows the conclusion that, while bias and error for $\hat{\gamma}_1$ and $\hat{\gamma}_2$ do depend on the overall sample size, they do not depend on the number of events in the treatment arms.

Figures 3.23 and 3.24 show the behavior of the bias of $\hat{\beta}_1$, $\hat{\beta}_2$, $\hat{\gamma}_1$, and $\hat{\gamma}_2$ for different baseline hazards. For both Figures, the baseline hazards for B$_2$ correspond to multiplying $\lambda_{0_0} = 0.05$ with $\exp(\gamma_2)$, where $\gamma_2 = \log(2)$, $\gamma_2 = \log(7/4)$, $\gamma_2 = \log(3/2)$, and $\gamma_2 = \log(5/4)$, respectively. The baseline hazards for B$_1$ correspond to multiplying λ_{0_0} with $\exp(\gamma_1)$, where $\gamma_1 = \log(\frac{1}{2})$, $\gamma_1 = \log(4/7)$, $\gamma_1 = \log(2/3)$, and $\gamma_1 = \log(4/5)$, respectively.



$\lambda_{0_1} = 0.025$ $\lambda_{0_1} = 0.02865$ $\lambda_{0_1} = 0.034$ $\lambda_{0_1} = 0.04$

$\lambda_{0_2} = 0.1$ $\lambda_{0_2} = 0.0875$ $\lambda_{0_2} = 0.075$ $\lambda_{0_2} = 0.0625$

Figure 3.23: Comparison of bias of the estimates of log hazard ratio β_1 and log hazard ratio β_2 for different baseline hazards for $n = 100$, using a Cox PH model without and with Firth correction for different baseline hazards.

Figure 3.23 shows that there is not much difference for the different baseline hazards looking at the bias of $\hat{\beta}_1$ for the three approaches without Firth correction. For the model with Firth correction, the upwards bias that can be observed in the first plot for $\lambda_{0_1} = 0.025$ gets smaller as the baseline hazard increases, i.e. as λ_{0_1} approaches

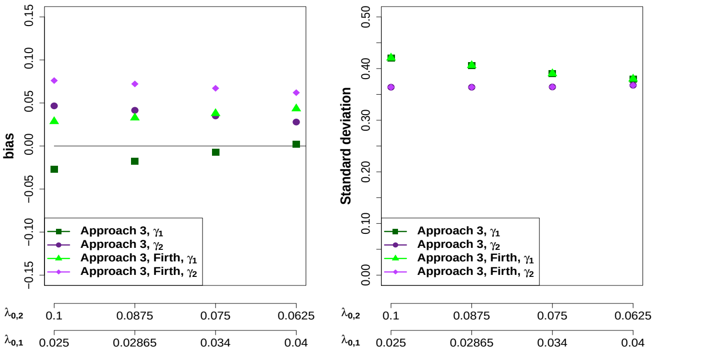


Figure 3.24: Comparison of bias and standard deviation of the estimates of γ_1 and γ_2 for different baseline hazards, using a Cox PH model without and with Firth correction for Approach 3 and for different baseline hazards.

λ_0. For the standard deviation of $\hat{\beta}_1$, there is not much of a difference between the plots for the different baseline hazards, but overall the standard deviation decreases as the baseline hazard increases and it is smaller for the models with Firth correction (see Figure A.4 in the Appendix).

For $\hat{\beta}_2$ there are also only slight differences between the plot for bias and standard deviation of $\hat{\beta}_2$ for the different baseline hazards. But for both models, with and without Firth correction, Approach 1 improves in bias as λ_{0_2} decreases, i.e., approaches λ_0. The plots for the standard deviation of $\hat{\beta}_2$ look similar for all baseline hazards. Figure 3.24 shows that the difference of bias and standard deviation between $\hat{\gamma}_1$ and $\hat{\gamma}_2$ gets smaller as the difference between the baseline hazards gets smaller.

3.3.2.3 Sensitivity analysis

Further simulations were run to investigate the robustness of the estimators against violations of model assumptions and change of parameters. Besides different baseline

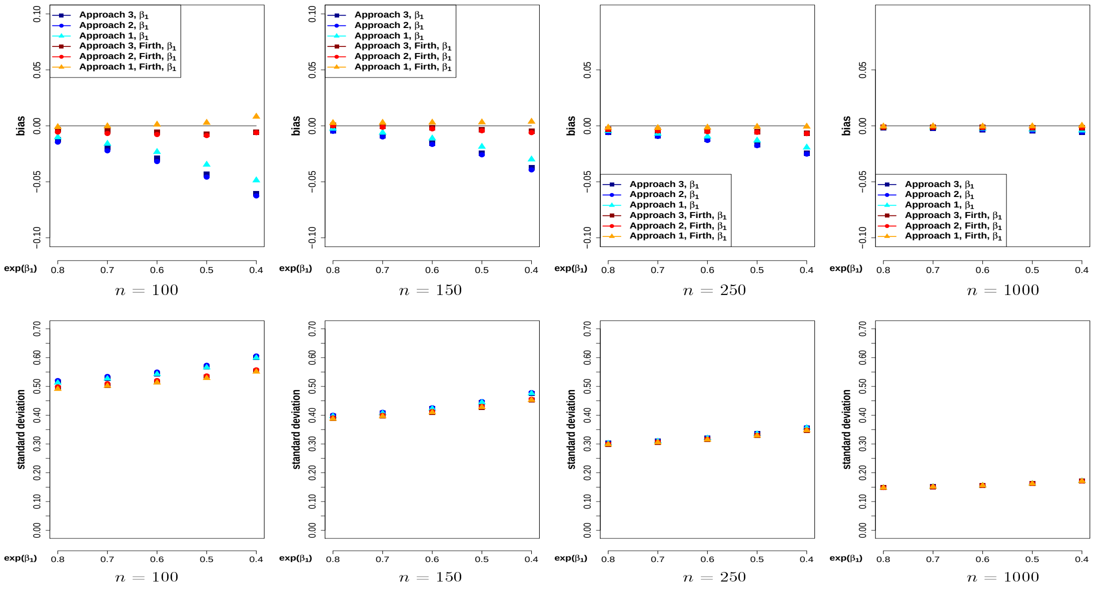


Figure 3.25: Different patient proportions: Bias and standard deviation of the estimate of log hazard ratio β_1, the treatment effect estimate for B_1, using a Cox PH model without and with Firth correction when the patient proportions in the biomarker-groups B_0, B_1, and B_2 are 0.6, 0.3, and 0.1, respectively.

hazards for the biomarkers, the bias and standard deviation were also investigated for different biomarker prevalences. Figures 3.25 and 3.26 show the bias and standard deviation of $\hat{\beta}_1$ and $\hat{\beta}_2$ for patient proportions 0.6, 0.3, and 0.1 in biomarker-groups B_0, B_1, and B_2, respectively.

The different prevalences do not seem to have much of an impact on bias and standard deviation of $\hat{\beta}_1$. Both are a bit smaller, which would be expected, given the slightly larger group size. For $\hat{\beta}_2$, the results differ for $n = 100$ and $n = 150$. But given the small numbers of patients in B_2 (10 and 15, respectively), the upwards bias and larger standard deviation that can be seen in those cases is most likely caused by the very small sample size. So it seems that the small prevalence, rather than the differences in prevalence, is the cause for the change in bias and standard deviation.

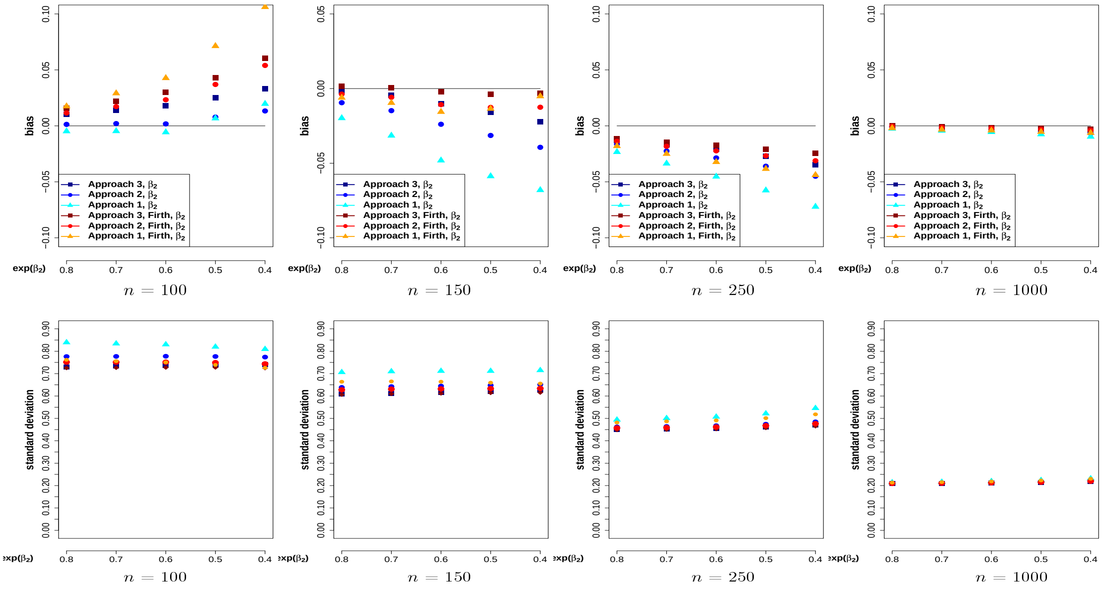


Figure 3.26: *Different patient proportions: Bias and standard deviation of the estimate of log hazard ratio* β_2*, the treatment effect estimate for* B_2*, using a Cox PH model without and with Firth correction when the patient proportions in the biomarker-groups* B_0*,* B_1*, and* B_2 *are 0.6, 0.3, and 0.1, respectively.*

For the non-constant (Weibull distributed) hazard function, shape parameters of the Weibull distribution were chosen to be 0.4, 1 and 5 (cf. Section 3.3.2.1). For the corresponding shape of the resulting survival distribution see Figure 3.27. For better comparability, λ_0 was adjusted such that the same number of events is reached at 60 months for all shape parameters, i.e. $\lambda_0 = 0.05 \cdot 60/60^{\text{shape}}$. Shape parameters smaller than 1 result in survival curves that are steeper at first and then flatter towards the end of the trial, whereas shape parameters greater than 1 result in survival curves which are flatter at first and get steeper towards the end.

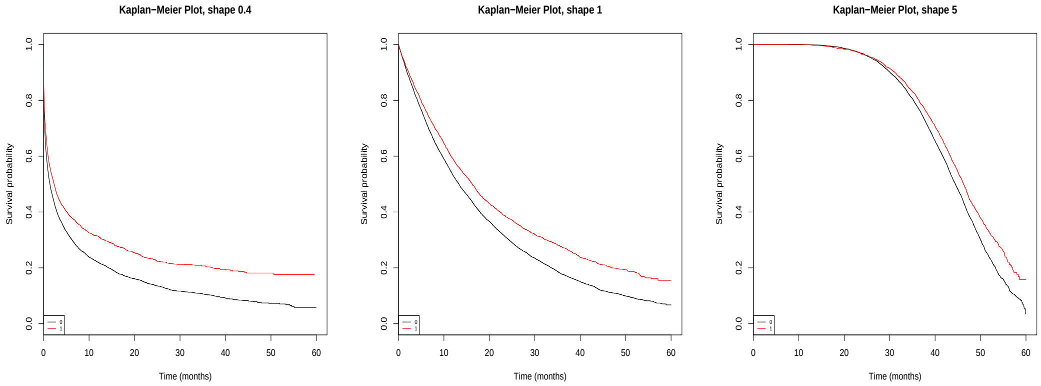
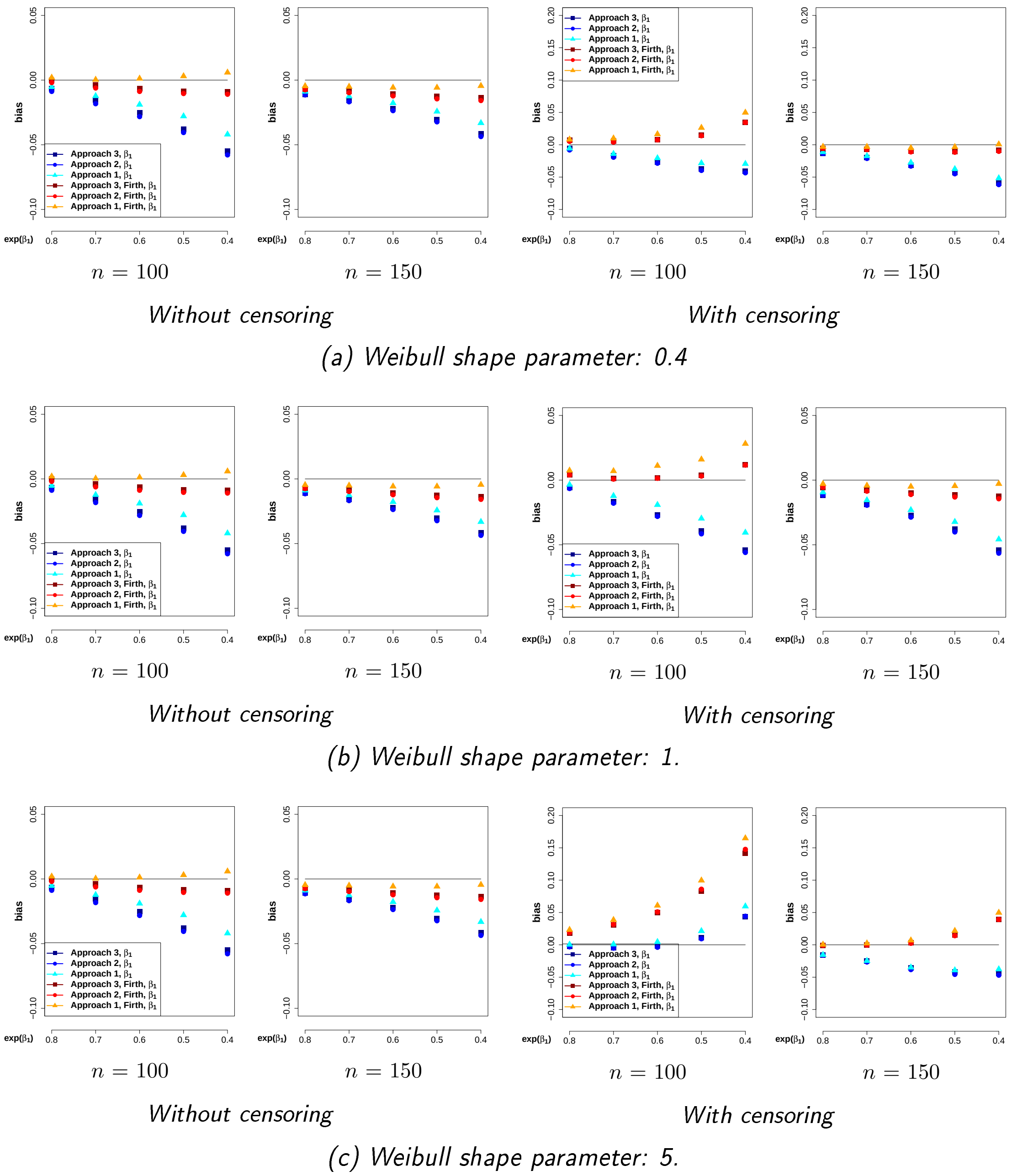


Figure 3.27: Kaplan-Meier plots showing survival distributions for different shape parameters of the Weibull distribution. The red curve shows survival of patients receiving experimental therapy and black stands for standard of care.

For both, $\hat{\beta}_1$ and $\hat{\beta}_2$, not much of a difference can be seen compared to the simulation results in Figures 3.21 and 3.22. The results are shown in Figures 3.28 and 3.29. The slight differences disappear when these simulations are also run without censoring. Hence, it can be concluded that the differences seen between Figures 3.21 and 3.28, and 3.22 and 3.29, respectively, were caused by the different numbers of events, rather than the time-dependent baseline-hazard. With censoring, and therefore with different numbers of events, there are differences in bias and standard deviation.

Additionally, it was verified that all three approaches for β_1 and β_2 without and with Firth correction for the different biomarker prevalences and for the different Weibull shape parameters roughly control the type one error rate. The tables with the numerical results can be found in Tables A.5 and A.6 in the Appendix.



$n = 100$ $n = 150$ $n = 100$ $n = 150$

Without censoring *With censoring*

(a) Weibull shape parameter: 0.4

$n = 100$ $n = 150$ $n = 100$ $n = 150$

Without censoring *With censoring*

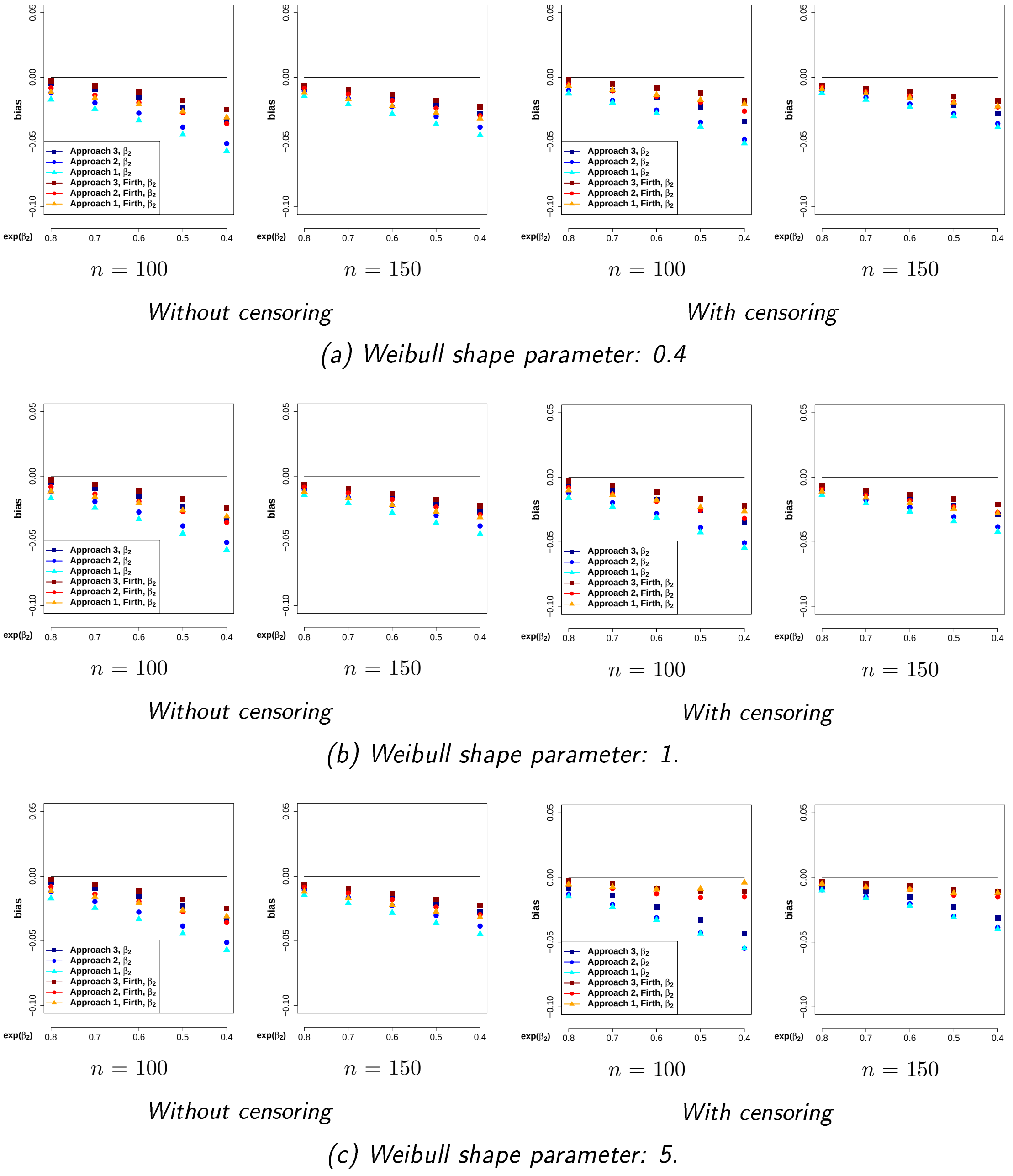
(b) Weibull shape parameter: 1.

$n = 100$ $n = 150$ $n = 100$ $n = 150$

Without censoring *With censoring*

(c) Weibull shape parameter: 5.

Figure 3.28: Bias with and without censoring of the estimate of log hazard ratio β_1 for Weibull distributed hazard function using a Cox PH model without and with Firth correction for sample sizes 100 and 150. Results for sample sizes 250 and 1,000 not shown.



$n = 100$ $n = 150$ $n = 100$ $n = 150$

Without censoring *With censoring*

(a) Weibull shape parameter: 0.4

$n = 100$ $n = 150$ $n = 100$ $n = 150$

Without censoring *With censoring*

(b) Weibull shape parameter: 1.

$n = 100$ $n = 150$ $n = 100$ $n = 150$

Without censoring *With censoring*

(c) Weibull shape parameter: 5.

Figure 3.29: Bias with and without censoring of the estimate of log hazard ratio β_2 for Weibull distributed hazard function using a Cox PH model without and with Firth correction for sample sizes 100 and 150. Results for sample sizes 250 and 1,000 not shown.

3.4 Flexible study designs

The field of biomarker research is rapidly and constantly developing. With a typical trial duration of up to several years, it is desirable to be able to react to the emergence of new potential biomarkers and corresponding experimental therapies during the course of the study, without having to conduct a new, separate clinical trial. Having the flexibility of incorporating a new biomarker and corresponding treatment in an ongoing clinical trial could make their investigation more time- and cost-efficient.

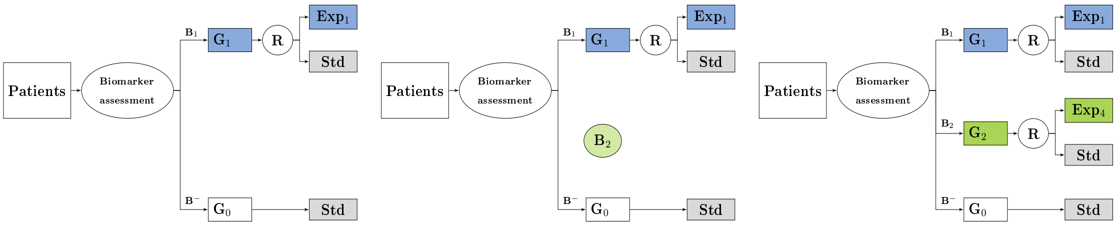
However, adding a new biomarker-group to an ongoing trial needs careful consideration and planning to ensure feasibility and statistical soundness. Note that so far, the biomarker-groups were treated as mutually exclusive. While this is a convenient simplification, it is not necessarily realistic. Often times, patients have more than one biomarker. This will be taken into consideration in the following sections, where several possibilities for the modification of the study design by adding a new biomarker will be presented and statistical considerations and issues will be discussed. To distinguish between biomarkers and biomarker-defined groups, in the following the biomarkers themselves will be denoted as B_i. The biomarker-groups will be denoted as G_i, defined by biomarker B_i. The difference between B_i and G_i is that patients are distinctly allocated to one of the biomarker-groups G_i, but patients within group G_i may also have other biomarkers, additional to B_i.

3.4.1 Inclusion of a new biomarker-group

In the following, a study design with two biomarker-groups is considered, denoted by G_i, $i \in \{0, 1\}$, where G_1 is comprised of patients matched with B_1 (see Figure 3.30a). Upon entering the study, the patients' biomarker-profile is determined and they are assigned to the biomarker-groups accordingly. All patients that cannot be matched with B_1 are allocated to G_0, the biomarker-negative patients. While patients in G_1 are randomized between an experimental therapy, targeting their biomarker, and standard of care, patients in G_0 are treated with standard of care only.

Beyond this initial study design, it is assumed that there is another biomarker, B_2

(see Figure 3.30b). This biomarker is not included at the beginning of the study, but throughout the study, external information becomes available that patients with B$_2$ could potentially benefit from a new experimental treatment Exp$_2$. It is aimed to be able to react quickly to such developments in the field of targeted therapies without going through the lengthy process of planning a new, separate trial for patients with B$_2$. Additionally, there could be concern that the coexistence of the ongoing and the potential new study may impair recruitment rates. Instead, the study shall be planned such that the study protocol allows adding a biomarker-group G$_2$ with experimental therapy Exp$_2$ to the ongoing trial at some point throughout the trial, denoted as t_b, where $0 \leq t_b \leq a$ (see Figure 3.30c). For this situation, the protocol should also include guidelines for an algorithm that prioritizes the biomarkers, such that patients with both biomarkers are distinctly allocated to either G$_1$ or G$_2$. The following sections will discuss potential design options, and practical and statistical considerations and challenges.



(a) Initial study design *(b) New biomarker* *(c) Extended study design*

Figure 3.30: Adding a new biomarker to an ongoing clinical trial

3.4.2 Extension of the study design

One important initial question is whether or not it is possible to retrospectively determine the biomarker status of the patients included in the study up to time t_b with respect to B$_2$. If it is possible to retrospectively determine the biomarker status with respect to B$_2$, it is possible to distinguish between patients with B$_1$ *and* B$_2$ in G$_1$ and patients with B$_1$ *without* B$_2$ in G$_1$, i.e. patients within the already existing biomarker-group can be subdivided into B$_1$+B$_2$ and B$_1$−B$_2$. With this information available, it could be considered excluding the patients with B_1+B_2 from the final analysis. This exclusion could, for example, be based on known interactions between the two biomarkers regarding the treatment, i.e. if B_2 is known to have a positive or negative effect on the response to the given treatment Exp_1. Alternatively, an interaction term could be included in the final model to accommodate for and estimate this effect.

One option to exclude patients from the analysis could be with the goal of not having a change in patient population within the biomarker-groups at time t_b. If B_2 was assigned a higher priority in the allocation algorithm than B_1, this would change the population in group G_1 from $B_1 \pm B_2$ before t_b to $B_1 - B_2$ after t_b, due to patients with B_1+B_2 being assigned to group G_2 instead. Then, in this option, all patients with B_2 already assigned to G_1 would be excluded from the analysis, maintaining consistency in patient population before and after time t_b with respect to B_2.

Additional to the retrospective determination of B_2, it is important to assess the reliability of the external information used. If there is strong evidence, e.g. a large confirmatory study has just revealed that patients with B_1 *and* B_2 do not benefit from treatment Exp_1, exclusion of patients who fit this profile and have already been randomized would be reasonable. Additionally, investigators should consider stopping ongoing treatment for these patients. If, on the other hand, there is only weak evidence, e.g. the information is merely a conjecture from an early phase study, the patients should be kept in the study and treatment should be continued. After the main analysis, an exploratory subgroup analysis should be performed comparing B_1+B_2 and B_1-B_2 patients to see if the collected data support the conjecture.

In the following, options for study design and analysis will be discussed, depending on the reliability of the external evidence and the suspected interaction of the new biomarker with the other biomarker or experimental therapy.

If there is strong evidence that B_2 has a **negative effect on the response** of patients with B_1+B_2 to the experimental therapy Exp_1, it could be considered to exclude all patients with B_1+B_2 already assigned to group G_1 from the analysis and additionally to stop ongoing treatment of these patients (see Fig. 3.31). The patients whose treatment was stopped could possibly be given the new experimental therapy Exp_2 off-protocol. Due to the negative effect of B_2 on response, no new patients with B_1+B_2

should be assigned to group G_1 after t_b. This can be achieved by giving a higher priority to B_2 than to B_1 in the allocation algorithm.

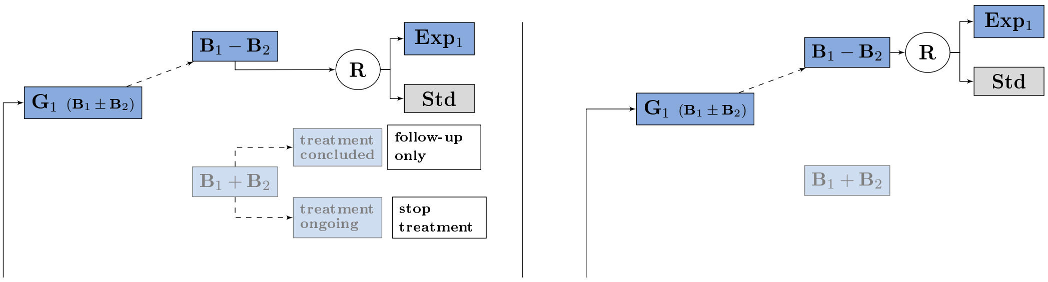


Figure 3.31: Potential approach if there is strong evidence that B_2 has a negative effect on the response of patients with B_1+B_2 to the experimental therapy Exp_1

If the evidence of a negative effect of B_2 on the response is weaker, it is advisable to keep all patients in the analysis and rather aim for strengthening the evidence with the data resulting from the study through an exploratory subgroup analysis, comparing the treatment effect in subpopulation B_1+B_2 against the treatment effect in B_1−B_2 (see Figure 3.32). A challenge here is the change in patient population at t_b, given that B_2 is assigned a higher priority in the allocation algorithm. It should be considered if and how this change can be taken into consideration. If the suspicion is correct that B_2 has a negative effect on the response, there would be two different treatment effects for the experimental therapy, the treatment effect after t_b being larger than before. Additionally, the two subpopulations could also differ with respect to prognostic effects of the biomarkers.

If there is **no evidence that B_2 has an effect on the response** of patients with B_1+B_2 to the experimental therapy Exp_1, there is no necessity of excluding patients from the analysis, since there is no expected difference in treatment effects between patients with B_1+B_2 and patients with B_1−B_2 (see Figure 3.32). However, if B_2 is assigned a higher priority in the allocation algorithm, the population does change at t_b, since patients with B_1+B_2 are no longer assigned to group G_1, but rather to group G_2. In that case, it should be considered accounting for this in the analysis. While the treatment effect is not expected to be different between the two subpopulations, there can still be differences, e.g. due to prognostic factors. After performing the

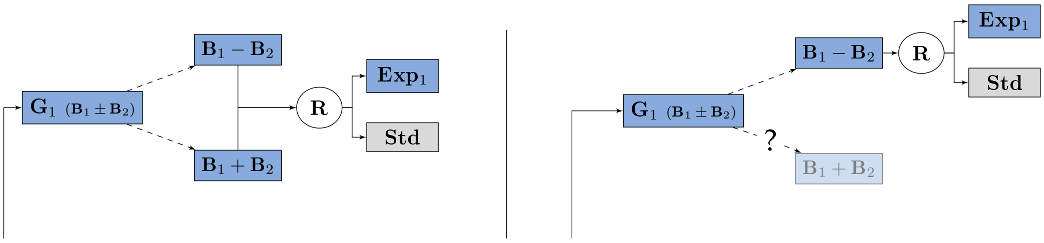


Figure 3.32: Potential approach if there is only weak or no evidence that B_2 has a (negative) effect on the response of patients with B_1+B_2 to the experimental therapy Exp_1

main analysis, exploratory subgroup analyses can be performed, comparing B_1+B_2 and B_1−B_2.

If there is evidence that B_2 has a **positive effect on the response** of patients with B_1+B_2 to the experimental therapy Exp_1, patients should not be excluded. If patients with B_1+B_2 are expected to have a better response to Exp_1 than to Exp_2, it should be considered assigning a higher priority to B_1 than to B_2 in the allocation algorithm. Otherwise, it should be the other way around. Again, a subsequent exploratory subgroup analysis can be performed to compare B_1+B_2 and B_1−B_2 to see if the collected data support the existence of a positive effect of B_2.

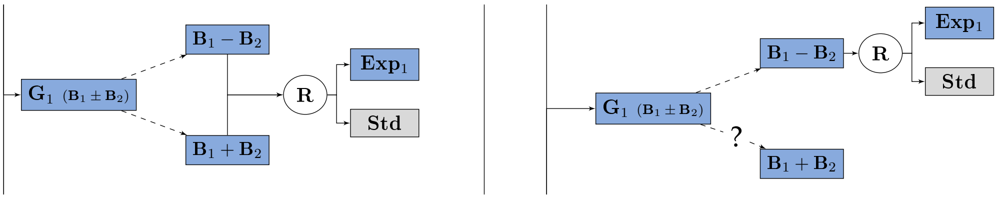


Figure 3.33: Potential approach if there is evidence that B_2 has a positive effect on the response of patients with B_1+B_2 to the experimental therapy Exp_1

When a **new biomarker-group** is added to an ongoing trial, several issues need to be addressed. First of all, the study protocol should state up until which point of the study a new biomarker-group with corresponding treatment can still be added, e.g. until accrual is halfway completed. This already leads to the next issue. Since accrual for this group begins later than for the other groups, it needs to be addressed how this is compensated for. Assigning the highest allocation priority to B_2 could help to somewhat speed up accrual compared to the other groups. But depending on how long after initial begin of accrual G_2 is added, this will probably not be enough. Additionally, the accrual phase could be extended to ensure sufficient accrual to G_2, but this will automatically prolong the overall study duration.

Finally, the group of **biomarker-negative patients** should also be addressed. In this group there is also a change in population, since patients with (only) B_2 are no longer considered 'biomarker-negative' after adding G_2 to the study. This population change should be considered in the analysis.

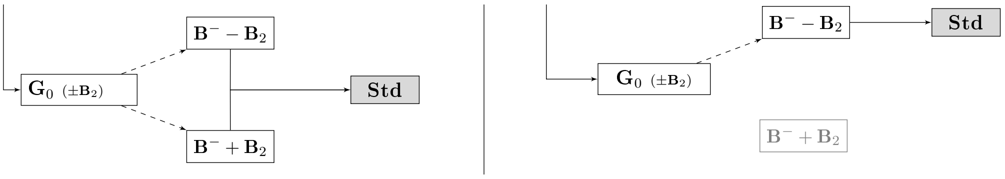


Figure 3.34: Considerations for biomarker-negative patients.

Additional to practical aspects of the design modifications, it is important to look at the options from a statistical point of view and address potential issues and challenges. If it is considered to exclude some of the patients from the main final analysis, it should be discussed whether it is acceptable to do so from a statistical point of view. Additionally, the overall sample size needs to be adjusted since there is a new biomarker-group added which was not included in the initial sample size calculation. Furthermore, it might be necessary to adjust the sample size of the other biomarker-groups if some of the patients in these groups are excluded from the analysis.

There can be many reasons why one might consider to exclude patients from a study or an analysis, but in the following only the exclusion of patients after adding a new biomarker-group will be discussed. Among other situations, Fergusson et al. (2002) consider the case when ineligible patients are mistakenly included in a study who do not meet inclusion criteria. This is not exactly equivalent to the situation considered here, since the patients to be excluded do meet the inclusion criteria for their biomarker-group in the beginning, but that may no longer be true after modification of the study. In their paper, Fergusson et al. (2002) discuss that investigators could avoid bias if patients who were mistakenly included are removed from both treatment arms and the decision to remove the patients is blinded to treatment and outcome, and independent from any events that occurred after randomization. Furthermore, they argue that, if these patients are expected to have a reduced or no response to treatment, their inclusion in the analysis could be a source of random error, reduced power of the study, and a less precise estimate of the treatment effect. This reasoning could also be applied in the situation considered here, especially if external evidence suggests that B_2 has a negative effect on the response to treatment with Exp_1. Additionally, exclusion of patients from the analysis would only depend on external information (e.g. potential interaction with the new biomarker) and information obtained before randomization (assessment of biomarker status).

In a situation where the **new biomarker cannot be determined retrospectively**, it is not possible to exclude patients from the analysis or account for the effect by including an interaction term. In this case it should be considered to perform an analysis that factors in the heterogeneity within the biomarker-groups caused by B_2. If the new biomarker cannot be determined retrospectively for only part of the population (e.g. due to insufficient amounts of stored specimens for these patients), data imputation methods could be considered. This will be discussed in more detail in Section 3.4.5.

If the biomarker status cannot be determined retrospectively for any of the patients, it could be considered to treat the change in patient population as changes in inclusion criteria for the affected biomarker-groups. Lösch and Neuhäuser (2008) discuss the statistical analysis of a trial when an amendment has changed the inclusion criteria and suggest using Fisher's combination test after performing separate statistical tests for the patients recruited before and after the amendment. In their simulation study, they compare the suggested combination test to simple pooling of data with respect to power. Following some simpler scenarios, they consider a case where the treatment effect is different for the two phases and compare the power of the tests for different changes in variance. But they only consider an inflation of variance due to broadening of the inclusion criteria, while in the case considered here, deflation of the variance is more likely, since the inclusion criteria are basically narrowed by no longer assigning patients with B_1+B_2 to group G_1. Hence, the behavior for variance deflation would additionally have to be investigated. There is a subsequent publication by Leuchs and Neuhäuser (2013) who suggest a modified Bauer and Köhne's test after an amendment has changed the inclusion criteria. They compare the performance - with respect to power - of their suggested test to the tests by Liptak, Fisher, Bauer and Köhne, and Edgington. However, they find in their simulation study that their method is not advisable to use in situations with changes in both, treatment effect and variance. In these situations, they find that Fisher's test and Liptak's Z-score approach perform best. Note that both papers compare the tests with respect to power, focusing on hypothesis testing. For estimation of treatment effects, they refer the reader to a publication addressing estimation in flexible two stage designs by Brannath et al. (2006).

In the following, the focus will be on situations where B$_2$ can be fully or partially determined retrospectively.

3.4.3 Approaches for data analysis after adding a new biomarker-group

A study design was considered where the study initially includes one biomarker (B$_1$) that is investigated, with group G$_1$ and the group of biomarker-negative patients, G$_0$. At time t_b, a new, second biomarker, B$_2$, and a corresponding group G$_2$ are added to the study (cf. Figure 3.30). It is assumed that there is evidence that B$_2$ has a negative effect on the response of patients with B$_1$+B$_2$ in G$_1$ to the experimental therapy Exp$_1$. Within G$_1$ and G$_2$, patients are randomized between the biomarker-specific experimental therapy (Exp$_1$ and Exp$_2$, respectively) and standard of care (Std), with probabilities $1-r_i$ and r_i, respectively. Patients in G$_0$ are treated with standard of care only. The baseline hazard for patients in biomarker-group i is denoted by λ_{0_i}. Several new parameters have to be considered:

- **Effect** τ of B$_2$ on Exp$_1$: Factor by which hazard ratio for B$_2$-patients treated with Exp$_1$ differs from hazard ratio for B$_1$-patients treated with Exp$_1$, e.g. if effect$= 1.5$ and HR$= 0.5$ for B$_1$, then HR$_{B_2,Exp_1} = 1.5 * 0.5 = 0.75$.
- **Overlap** l: Proportion of the entire patient population which has both biomarkers (expected number of $n{\cdot}l$ patients), e.g. if both biomarkers have a prevalence of 25% and an overlap of B$_1$ and B$_2$ within the entire population of l =12.5%, 50% of patients within G$_1$ (or G$_2$) have both biomarkers.
- **Time** t_b: Time at which G$_2$ is added to the study.

Additionally, the **priority** for the treatment allocation algorithm needs to be chosen, i.e. which biomarker-group patients are assigned to if they have both biomarkers.

Different models are compared to evaluate whether excluding patients from the analysis or including an interaction term for the effect of B$_2$ on the treatment effect of Exp$_1$ is the better strategy with respect to bias and standard deviation of the treatment effect estimates. Additionally, a combined model, analogous to Approach 3 in Section 3.3, is compared to fitting models for the individual biomarker-groups.

Model 1: Exclude patients with B$_2$
$\lambda(t) = \lambda_0 \exp(\beta_1 x_1)$
Expected sample size: $n_1 - \frac{t_b}{a} n \cdot l$ (G$_1$ excluding patients with B$_2$)

Model 2: Include interaction term for patients with B$_2$
$\lambda(t) = \lambda_0 \exp(\beta_1 x_1 + \beta_{1,2} \mathbb{1}_{\{B_2\}} x_{1,2})$
Expected sample size: n_1 (G$_1$ including patients with B$_2$)

Model 3: Combined model (as discussed in Section 3.3)
$\lambda(t) = \lambda_0 \exp(\gamma_1 b_1 + \gamma_2 b_2 + \beta_1 x_1 + \beta_2 x_2 + \beta_{1,2} \mathbb{1}_{\{B_2\}} x_{1,2})$
Expected sample size: n (all patients)

The parameters β_1 and β_2 are representing the treatment effects for Exp$_1$ and Exp$_2$, respectively and γ_1 and γ_2 are the prognostic effects of B$_1$ and B$_2$, respectively, with B$_0$ as reference. The biomarker status is included as a factor variable (with dummy variables b_1 and b_2). Note that usage of all three models requires that B$_2$ is determinable retrospectively for all patients recruited before time t_b.

3.4.4 Simulation study: Comparison of models

A simulation study was performed to compare bias and standard deviation for all three models to determine if exclusion of patients or including an interaction term is the better option and if using a combined model for all biomarkers can provide additional benefit.

3.4.4.1 Study design and data generation

The data were generated as described in Section 2.2.1 with parameters according to Table 3.8. The time t_b at which G$_2$ is added to the study was chosen to be 12 months, i.e. after half of the accrual time has passed. To take into account the overlap of the biomarkers within the population, an overlap l between the biomarkers was simulated. This overlap was set to be 12.5% out of the entire population. This fixed overlap between the biomarkers can be simulated by drawing from a multinomial distribution. But rather than having three possible outcomes (B$_1$, B$_2$, or B$_0$), there is a fourth possible outcome, which means that a patient has both biomarkers, B$_1$ and B$_2$. If the expected proportions of the biomarkers B$_1$ and B$_2$ in the population are p_1 and p_2, respectively, then the sampling proportions for B$_0$, B$_1$, B$_2$, and B$_1$$\cap$ B$_2$ are $1-(p_1+p_2-l)$, p_1-l, p_2-l, and l, respectively. The resulting proportions g_i in the groups G$_i$ then depend on which biomarker was chosen as priority in the allocation algorithm and at which time t_b G$_2$ was added to the study.

Table 3.8: Parameters for the simulation study using a design with three biomarker-groups, denoted by G$_i$, $i \in \{0,1,2\}$.

Fixed simulation parameters	
Accrual time (months), a	24
Follow-up time (months), f	36
Time t_b (months)	12
Proportion random censoring, p_{cens}	0.05
Treatment allocation ratio	$1:1$
Prevalence p of biomarkers B$_1$ and B$_2$	0.25
Hazard ratio G$_1$, $\exp(\beta_1)$	$0.8, 0.7, 0.6, 0.5, 0.4$
Hazard ratio G$_2$, $\exp(\beta_2)$	0.8
Sample size, n	1,000
Number of simulations	10,000
Overlap l of biomarkers	0.125
Effect τ	1.5
Parameters for biomarker-groups $(\mathbf{G_0}, \mathbf{G_1}, \mathbf{G_2})$	
Baseline hazards, λ_{0_i}	(0.05, 0.025, 0.1)

As mentioned before, it is also taken into consideration that a biomarker could have an effect on the treatment effect of the experimental therapy targeting another biomarker.

For data generation, it was assumed that B_2 has an effect τ on the treatment outcome of patients (that also have B_1) treated with Exp_1, while B_1 has no effect on treatment with Exp_2. This effect was chosen to be $\tau = 1.5$, which means that patients with B_1 and B_2 treated with Exp_1 have a hazard ratio that differs by a factor 1.5 from the hazard ratio for patients with only B_1 treated with Exp_1. A large sample size case with 1,000 patients was considered to avoid small sample size bias (as observed in Section 3.3). Bias and standard deviation of the estimates were calculated as described in Section 3.3.2.1.

3.4.4.2 Comparison of models

The three models discussed in Section 3.4.3 were compared with respect to bias and standard deviation of the estimates $\hat{\beta}_1$, $\hat{\beta}_2$, and $\hat{\beta}_{1,2}$ out of 10,000 simulations. The simulation results in Figure 3.35 show that the three models do not differ by much

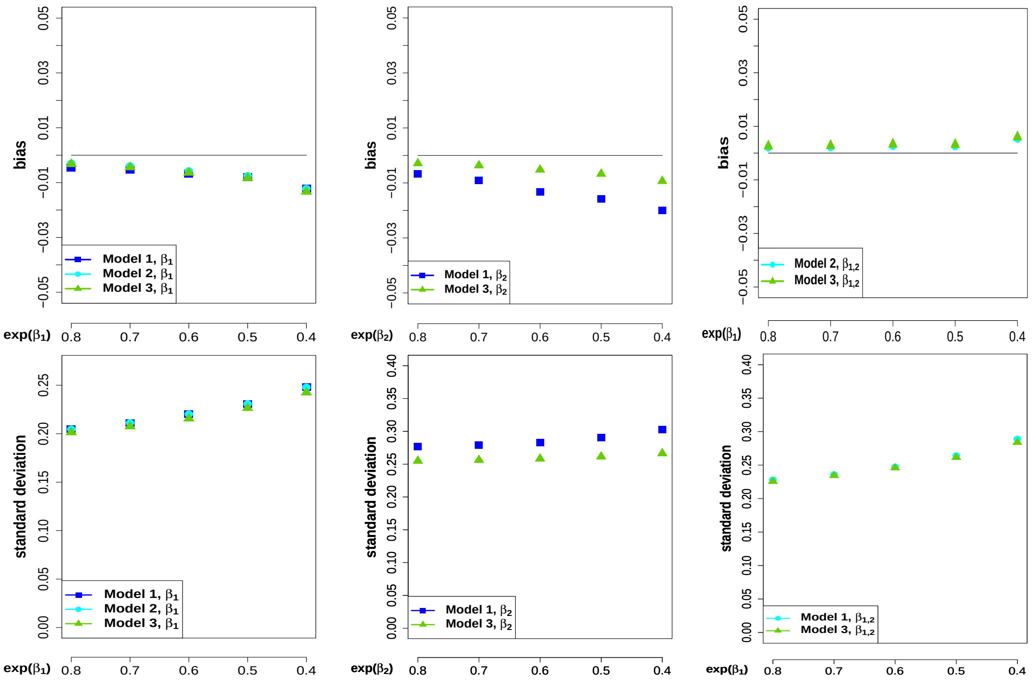


Figure 3.35: Bias and standard deviation of $\hat{\beta}_1$, $\hat{\beta}_2$, and $\hat{\beta}_{1,2}$, using models 1, 2, and 3 with fixed sample size 1,000. Note that for $\hat{\beta}_2$, Model 1 was used with data from G_2.

when it comes to bias of $\hat{\beta}_1$ and $\hat{\beta}_{1,2}$. Model 3, i.e. the combined model, yields a slightly smaller standard deviation for $\hat{\beta}_1$ compared to models 1 and 2. For $\hat{\beta}_{1,2}$, not much of a difference can be seen for the standard deviation. Note that Model 1 cannot be included in the comparison for $\hat{\beta}_{1,2}$. It does not provide an estimate for $\beta_{1,2}$, since it does not include an interaction term. If Model 1 is used for G$_2$, i.e. $\lambda(t) = \lambda_0 \exp(\beta_2 x_2)$, bias and standard deviation of $\hat{\beta}_2$ are larger than for Model 3. Note that Model 2 was not applied and compared for β_2, since there is no biomarker-treatment interaction that can be estimated. Additionally, for this simulation B$_1$ was given priority, i.e. patients with both biomarkers are allocated to G$_1$. Thus, there are no patients with B$_1$ in G$_2$ and Model 1 and Model 2 would yield similar results. Overall, Model 3 appears to be slightly advantageous over the other two models. It yields a smaller bias and standard deviation for $\hat{\beta}_2$ and facilitates estimating β_1, β_2, and $\beta_{1,2}$ simultaneously in one model.

3.4.5 What if B$_2$ is not determinable retrospectively?

In clinical trials missing data is a common issue. There are a multitude of reasons why there may be certain data points missing for a certain patient. When it comes to determining the biomarker status of a patient, common reasons are insufficient amounts of collected specimens, or technical difficulties of the screening procedure. As previously mentioned, usage of all three models discussed int he previous section requires that B$_2$ is retrospectively determinable for all patients recruited before time t_b.

If B$_2$ is not retrospectively determinable for any of the patients, it will not be possible to incorporate this in the model. However, if B$_2$ is only missing for part of the population, data imputation methods can be applied to be able to use data for patients with missing B$_2$ status in the analysis.

In the context of missing data, Rubin (1976) distinguished between missing at random (MAR) and missing completely at random (MCAR). If the data are MCAR, the missingness depends on neither the missing values nor the observed values. In the weaker case, MAR, the missingness does not depend on the missing values but may depend on the observed values. If neither is the case, the data are referred to as missing not at random (MNAR) (Van Buuren 2012). Hereafter, missing data will be assumed to be MCAR or MAR.

3.4.5.1 Data imputation

A common approach to handling missing data is to simply delete all cases with missing data and only include the complete cases (CC) in the analysis. While CC analysis still produces unbiased estimates for data that is MCAR, Van Buuren (2012) argues that if the data is not MCAR, this approach may produce severely biased estimates. In the case of missing B_2 status, it is not unlikely that the data are MAR rather than MCAR, i.e. that the missingness depends on one or more of the observed variables. The missingness could for example depend on the time at which a patient entered the study: The probability of missing B_2 status could be higher the earlier a patient entered the study, e.g. due to decrease of amount or quality of stored specimens over time.

Instead of CC analysis, Van Buuren (2012) recommends using regression imputation, where a regression model is fit using the complete cases to predict the missing values with the resulting equation. One of the advantages of (single) regression imputation over CC analysis pointed out by Van Buuren is that, additional to producing unbiased estimates under MCAR, regression imputation produces unbiased regression weights under MAR, given that the regression model contains the factors influencing the missingness. The variance, however, is underestimated since the estimates do not include an error term and therefore do not provide information about the uncertainty of the imputed values. The extent of underestimation depends on the explained variance and the proportion of missing values in the data (Van Buuren 2012).

Creating several data sets with imputed values including a random component is referred to as multiple imputation and is utilized to account for the uncertainty in the imputed data. There are several different methods to do so. Van Buuren (2012) uses multiple imputation by chained equations, which uses a series of conditional distributions. For each variable, regression models are fitted successively, using the already imputed values for the following regressions. This is done iteratively until the model converges. Each of the resulting complete data sets from the multiple imputation is analyzed separately and afterwards the overall estimate is obtained by averaging the estimates from the individual data sets (Van Buuren 2012). The standard error for this pooled estimate can be obtained from a formula suggested by Rubin (1987):

$$s(\hat{\hat{\beta}}) = \sqrt{\frac{1}{M}\sum_{k=1}^{M} s_k^2 + \left(1 + \frac{1}{M}\right)\left(\frac{1}{M-1}\right)\sum_{k=1}^{M}(\hat{\beta}_k - \hat{\hat{\beta}})^2}, \tag{3.6}$$

where $k = 1, 2, ..., M$ is the k^{th} imputation, $\hat{\beta}_k$ is the estimate from the k^{th} imputed data set, and $\hat{\hat{\beta}}$ is the pooled estimate out of the M imputations.

In the past, there have been some discussions about the number of imputations to use, M. For quite a while, a common recommendation was to use low numbers, such as $M = 5$ or $M = 10$ imputations (Van Buuren 2012). In 2008, Bodner recommended to use approximately the percentage of missing data as number of imputations. This rule of thumb was later also recommended by White et al. (2011). Regarding subsequent analysis with a Cox PH model, White and Royston (2009) recommend using the Nelson-Aalen estimate of the cumulative hazard function as predictor in the imputation model rather than simply using time. Furthermore, White et al. (2011) caution their readers that although multiple imputation gives asymptotically unbiased estimates under MCAR/MAR and a correctly specified model, departures from the MAR assumption and model misspecification may lead to substantial bias.

3.4.5.2 Multiple imputation with interactions

As mentioned in the previous section, the unbiasedness of estimates obtained from multiple imputation depends on the regression model containing all the factors influencing the missingness. White et al. (2011) recommend that the imputation model should include all variables that will later be used in the analysis model, as well as the outcome variable, to avoid bias. Additionally, they state that caution is required when it is intended to include non-linear or interaction terms in the analysis model. Their advice is to include these terms in the imputation model in the correct functional form.

There are several suggestions for imputation with non-linear terms or interaction terms. Seaman et al. (2012) compared different approaches in several simulation studies: passive imputation (PI), predictive mean matching (PMM), and 'just another variable' (JAV). PI only imputes the main effects and then uses these imputed values to calculate the interaction term. PMM on the other hand calculates a predicted value and then draws from a set of actually observed values which are close to the prediction. PMM may be problematic in small sample size cases, because of the limited number of observed values to sample from (White et al. 2011). For JAV the interaction term is treated as 'just another variable' in the imputation model, ignoring its relationship to the main effects. Another option, which is referred to as a 'simple congenial approach' by White et al. (2011), can be used if one of the variables in the interaction term is categorical and completely observed. In that case, it is also possible to split the data into several data sets, one for each level of the categorical variable, and recombine the data after imputation. This method is also referred to as stratify-approach.

Each of these methods has its advantages and disadvantages, which have been investigated and discussed by several authors, such as Von Hippel (2009), White et al. (2011), and Seaman et al. (2012). Von Hippel (2009) applied several methods to a real data example and found that, while PI may yield plausible data but biased estimates, JAV may yield implausible data but unbiased estimates. The biased estimates resulting from the PI approach can be caused by using a linear model for the imputation, which is not suitable for non-linear terms. Hence the estimate for the non-linear term would be biased towards zero (White et al. 2011). Seaman et al. (2012) performed a series of simulation studies and concluded that "JAV is the best of a set of imperfect methods" (Seaman et al. 2012) when it is applied for linear regression with quadratic or interaction terms, but they do not recommend it for logistic regression and caution that JAV may yield biased estimates under MAR. White et al. (2011) point out that the proof of unbiasedness of JAV relies on the MCAR assumption and refer to a simulation study that shows bias for JAV under an extreme MAR mechanism. They also conclude from their simulation studies that it is difficult to recommend one single method, since all methods have their pitfalls. Note that simulation studies in this section use linear or logistic regression as analysis model. They do not investigate the behavior when the analysis model is a Cox PH model.

3.4.6 Simulation study: Missing biomarker status

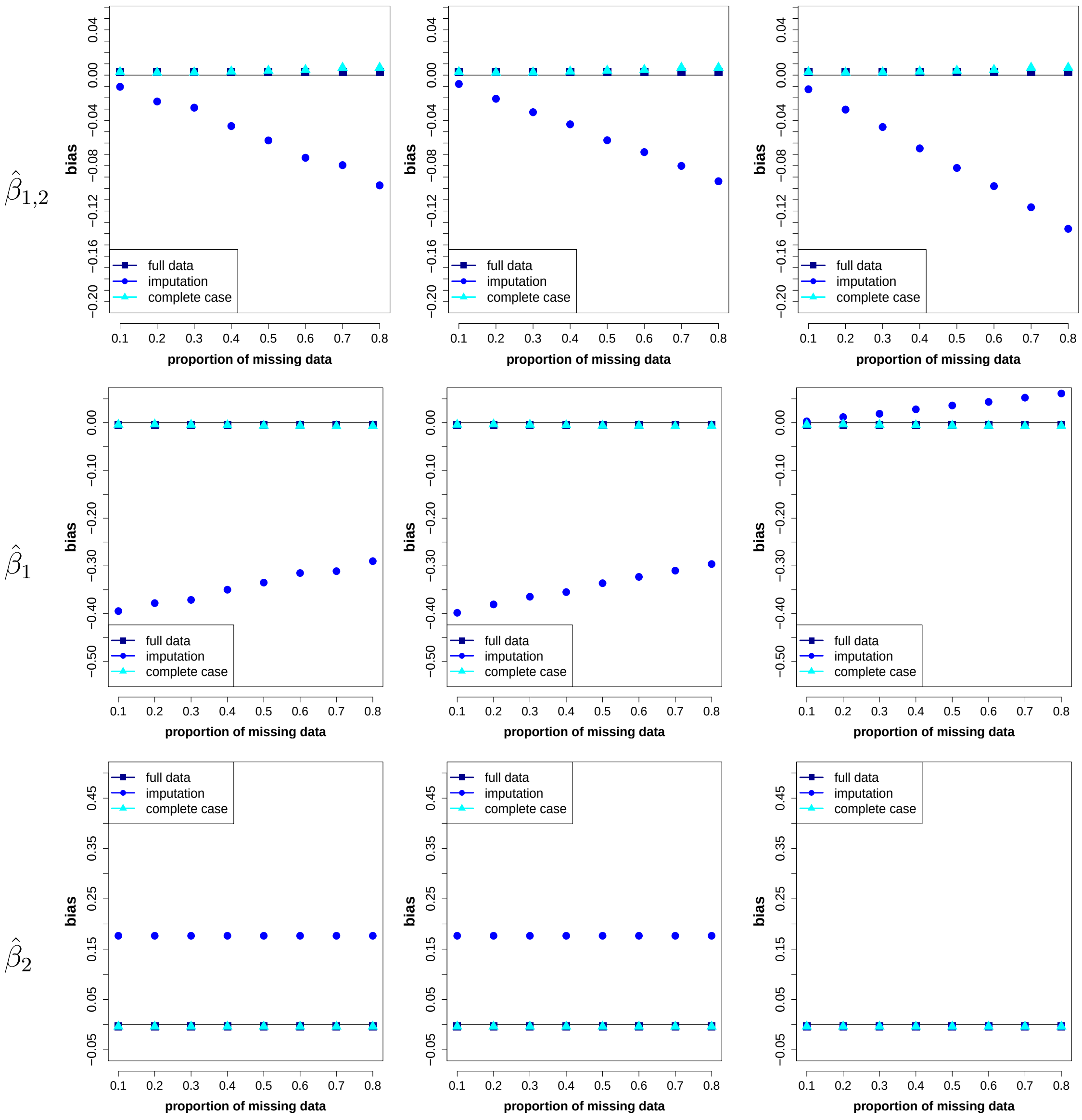
A simulation study was conducted to compare the performance of the different approaches for handling missing data imputation with respect to bias and standard deviation of the different parameter estimates resulting from the combined model (Model 3 from Section 3.4.3). Data was generated according to Section 3.4.4.1, except that the hazard ratios for B_1 and B_2, $\exp(\beta_1)$ and $\exp(\beta_2)$, were fixed at 0.7 to ensure sufficient numbers of events in both groups to minimize bias due to small numbers of events. Missing data was either generated as MCAR or MAR and missing data proportions of $0.1, 0.2, ..., 0.8$ were considered. MCAR data can be generated rather straight forward by using random sampling to delete a given proportion of data points. Generating MAR data is a bit more complex. One or more variables should be determined on which the missingness depends. For this simulation study, the entry time of a patient was chosen. Then MAR data can be generated according to Section 2.2.2. For the data imputation the `R` functions `mice` and `with` from the package `mice` were used. Time as a predictor was replaced by the Nelson-Aalen estimator (Aalen 1978) as suggested by White and Royston (2009).

The bias of the estimates of β_1, β_2, and $\beta_{1,2}$ when using PI, the JAV-approach or the stratify-approach was compared to the bias resulting from CC analysis and from the analysis of the full data set. Three imputation methods to predict the missing values were used and compared for the three imputation approaches: logistic regression (LR), polytomous logistic regression (PLR), and predictive mean matching (PMM). Unless otherwise indicated, missing data are MCAR. The bias of the estimates of γ_1 and γ_2 is not shown in the following figures. Since they are linear terms, there was no notable bias for any of the approaches.

Note that strictly speaking, x_1, x_2 and $x_{1,2}$ in Model 3 (see Section 3.4.3) are all interaction terms. They are created by multiplying biomarker status and treatment variable. If biomarker status and treatment are coded as dummy variables, x_1 is obtained by multiplying dummy variables for B_1 and Exp_1, x_2 is obtained by multiplying B_2 and Exp_2, and finally $x_{1,2}$ is obtained by multiplying B_2 and Exp_1. Usually, all main effects that are part of the interaction are also included in the model. Note that this would not be meaningful in this situation, since there is no general treatment effect that can be estimated. Hence, this may be somewhat of an unconventional model specification, which requires additional care when specifying the imputation and the analysis method.

For the first simulation PI was used, i.e. interactions were not included in the imputation model. In this case, the interactions are created by multiplication of the imputed main effects (as described in Section 3.4.5.2). The results in the first row of Figure 3.36 show that the bias of the estimate of $\beta_{1,2}$ linearly increases in absolute terms for all three imputation methods as the proportion of missing data increases. While there is some bias for the estimate of β_1 when using PMM, which linearly increases for increasing proportions of missing data (see middle plot of Figure 3.36c), there is substantial bias for both, LR and PLR (middle plots of Figures 3.36a and 3.36b). This is similar for the estimate of β_2. While there is no visible bias for $\hat{\beta}_2$ when using PMM, there is a constant, substantial bias for all proportions of missing data when using LR or PLR.

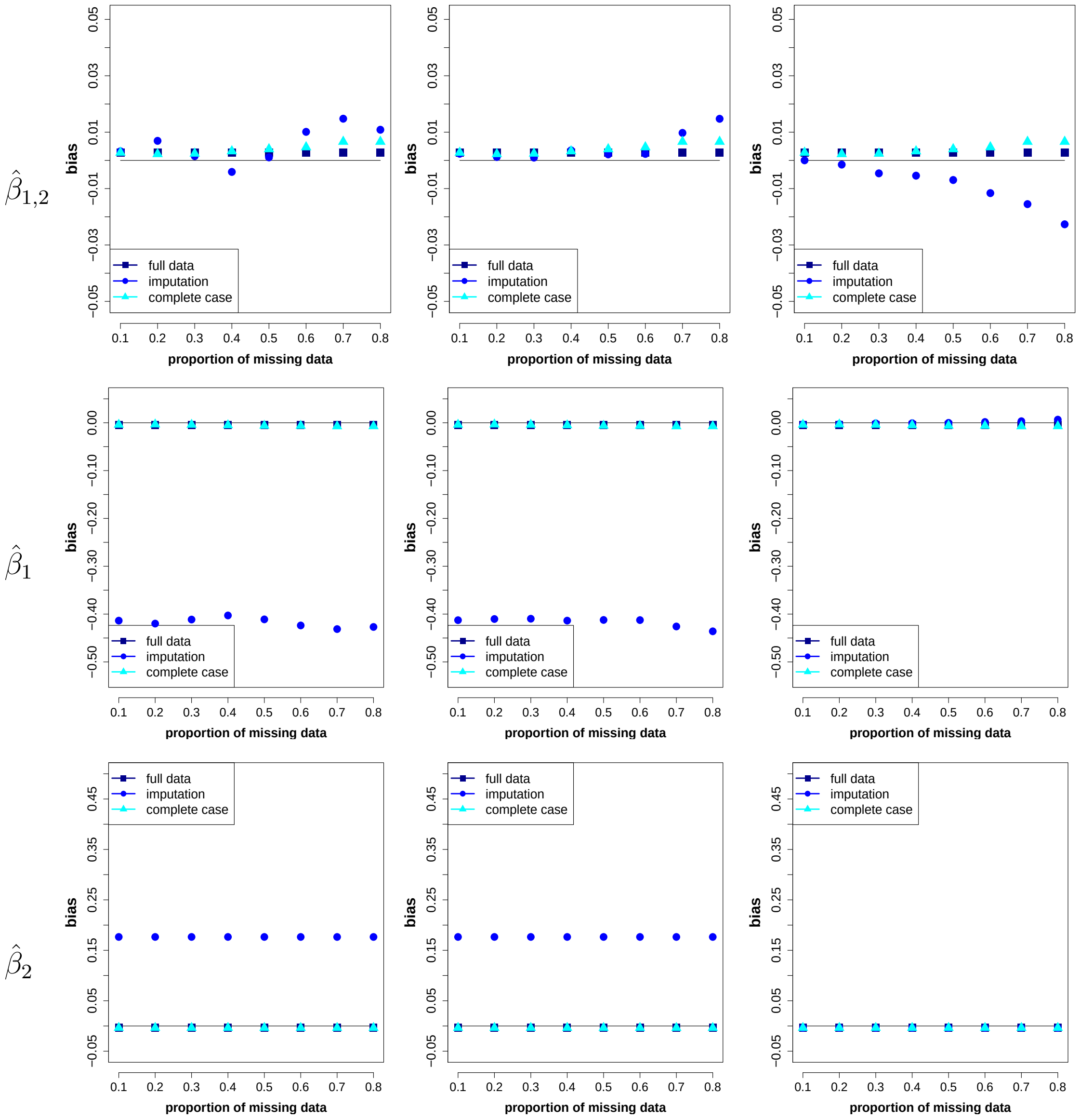
Subsequently, two imputation approaches for handling interactions terms were implemented and tested: the stratify-approach and the JAV-approach. When the stratify-approach is used, the bias of $\hat{\beta}_{1,2}$ is drastically reduced for all three methods. While there is only a small amount of bias for larger proportions of missing data when using LR and PLR, there is some bias for PMM which increases in absolute terms with increasing proportions of missing data (see first row of Figure 3.37). Note the different scaling of the y-axis for $\hat{\beta}_{1,2}$ compared to Figure 3.36. However, while PMM yields approximately unbiased estimates for β_1 and β_2, there is still a substantial bias for the estimates of β_1 and β_2 when using LR or PLR, and the bias is more or less constant over all proportions of missing data (see second and third row of Figures 3.37a and 3.37b).



(a) Logistic regression (b) Polytomous logistic regression (c) Predictive mean matching

Figure 3.36: Bias for imputation methods logistic regression, polytomous logistic regression, and predictive mean matching for $\hat{\beta}_{1,2}$, $\hat{\beta}_1$, and $\hat{\beta}_2$ when imputation model does not include interactions (naive imputation). B_1 is the prioritized biomarker in the allocation algorithm, data is MCAR, and m=10 imputations.

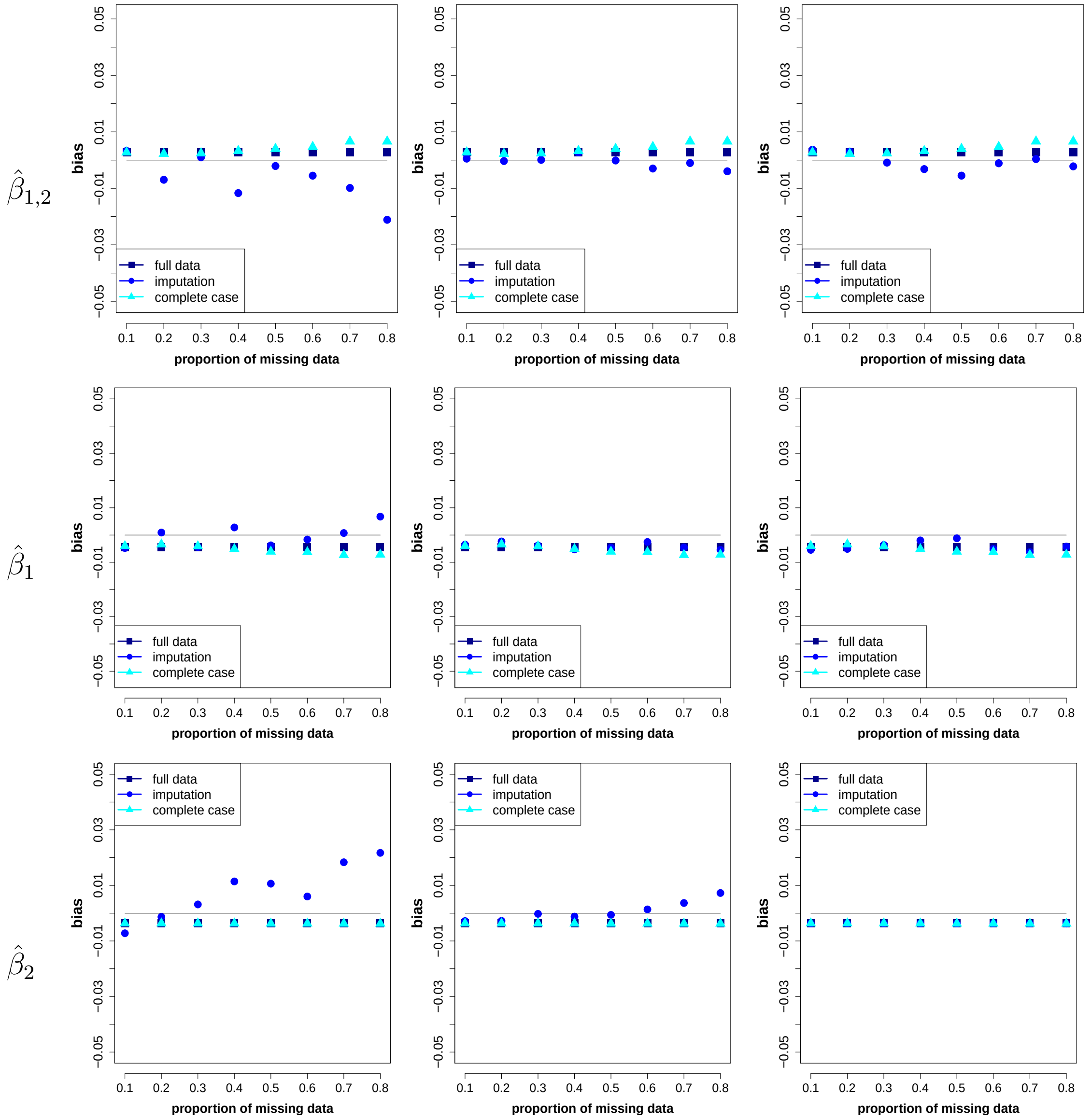
Using the JAV-approach results in considerable improvements for all three methods and all estimates compared to PI. There is now only a small bias left for $\hat{\beta}_{1,2}$ and $\hat{\beta}_2$ when using LR for larger proportions of missing data (see Figure 3.38). There is hardly any visible bias for $\hat{\beta}_1$ for either method. PLR appears to perform slightly better than regular LR for larger proportions of missing data. For PMM, there is no notable bias for any of the three estimates. Note the different scaling of the y-axes for $\hat{\beta}_1$ and $\hat{\beta}_2$ compared to the previous two figures.



(a) Logistic regression (b) Polytomous logistic regression (c) Predictive mean matching

Figure 3.37: Stratify: Bias for LR, PLR, and PMM for $\hat{\beta}_{1,2}$, $\hat{\beta}_1$, and $\hat{\beta}_2$ when data is stratified by treatment before imputation. B_1 is the prioritized biomarker in the allocation algorithm, data is MCAR, and m=10 imputations.

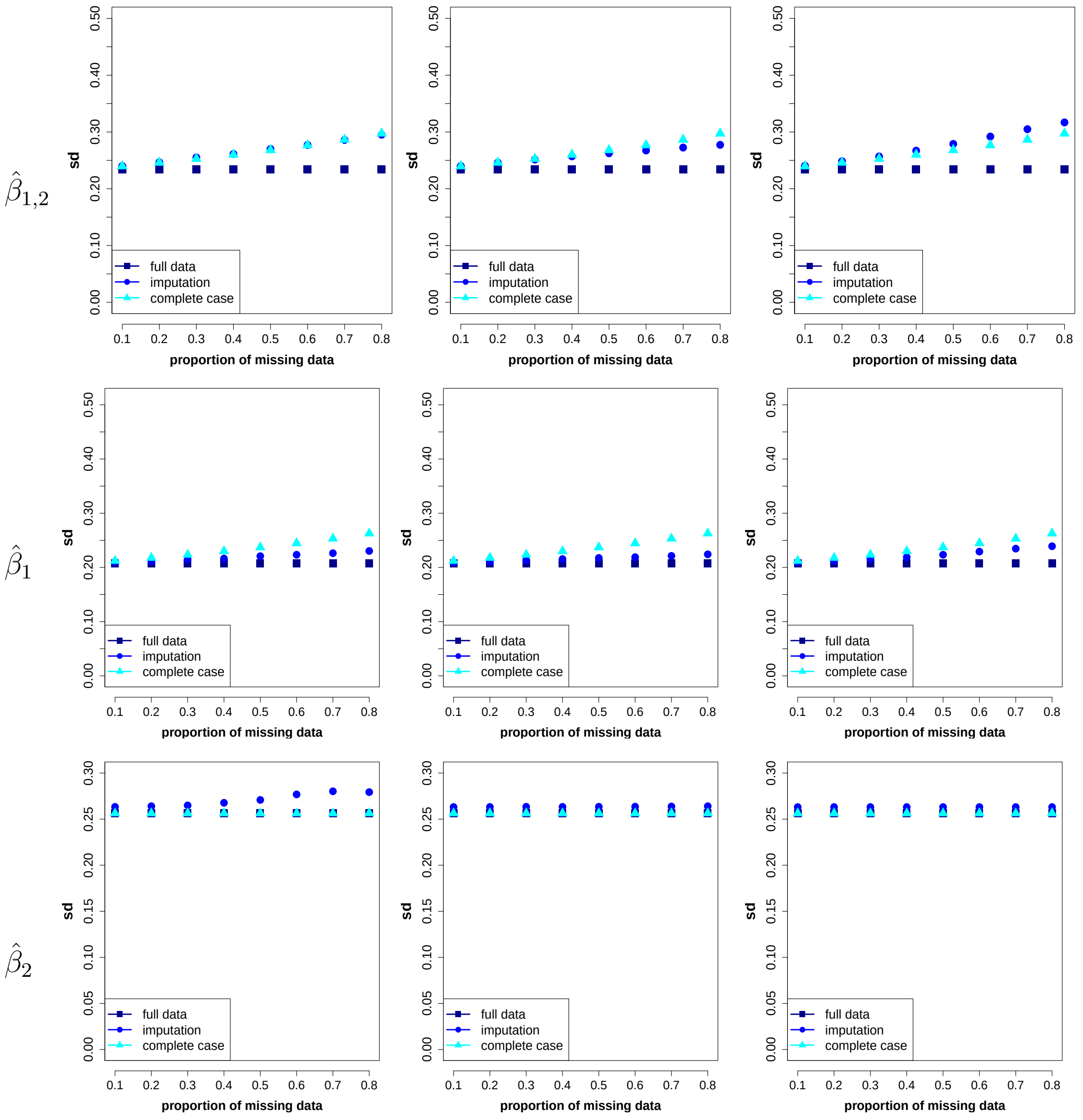
Regarding standard deviation, a small improvement can bee seen for $\hat{\beta}_1$ for all three methods compared to CC analysis (see Figure 3.39). For $\hat{\beta}_{1,2}$ and $\hat{\beta}_2$, only minor differences can be seen between imputation and CC analysis, and it varies which of the two yields a slightly smaller standard deviation.



(a) Logistic regression (b) Polytomous logistic regression (c) Predictive mean matching

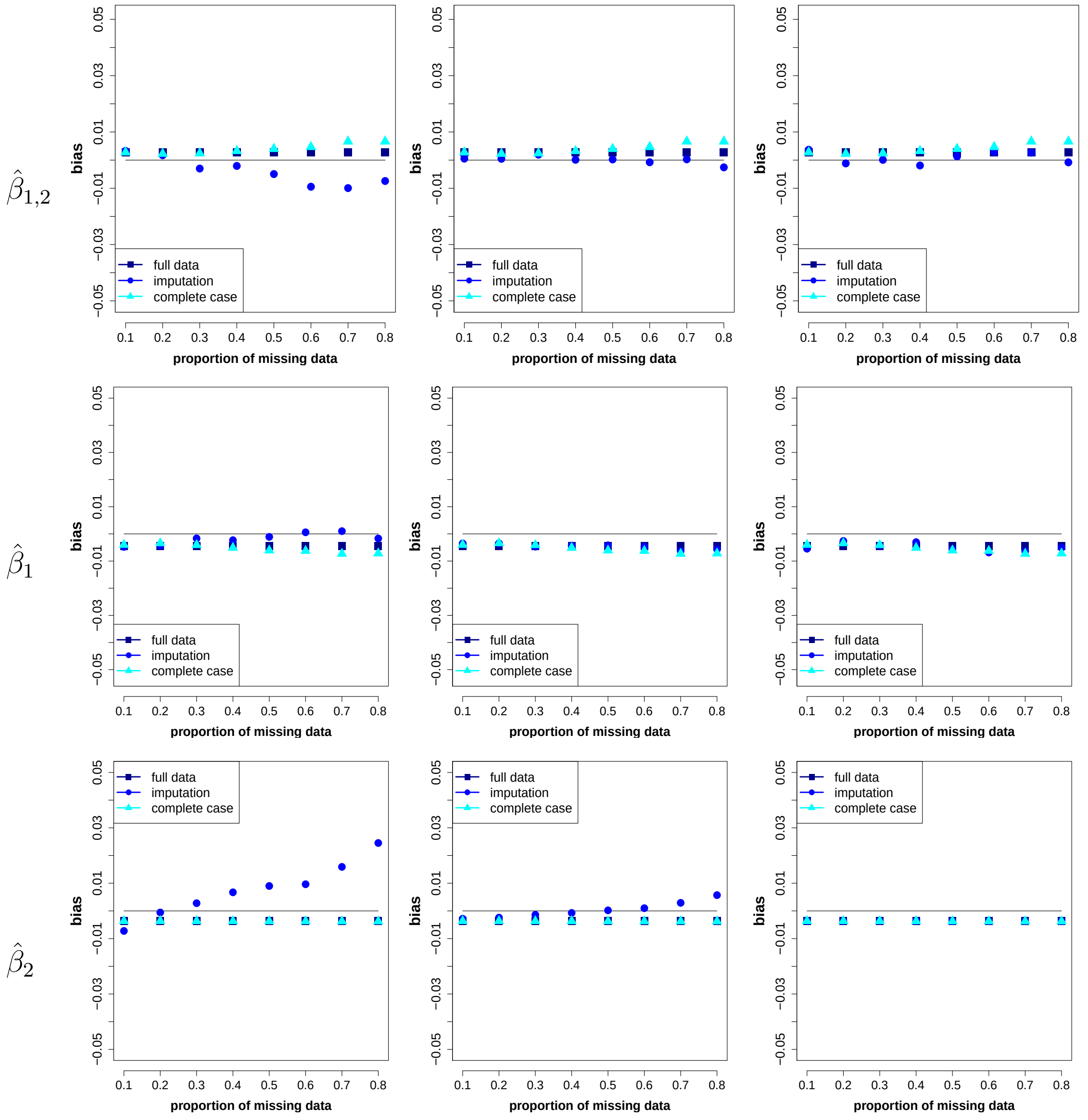
Figure 3.38: JAV: Bias for LR, PLR, and PMM for $\hat{\beta}_{1,2}$, $\hat{\beta}_1$, and $\hat{\beta}_2$ treating interactions as JAV in imputation model. B_1 is the prioritized biomarker in the allocation algorithm, data is MCAR, and m=10 imputations.

Up to this point, only 10 imputations were used for the simulations. For Figure 3.40, the recommendation of Bodner (2008) and White et al. (2011) was followed to use the percentage of missing data as number of imputations. However, this only results in minor visible improvements compared to Figure 3.38.



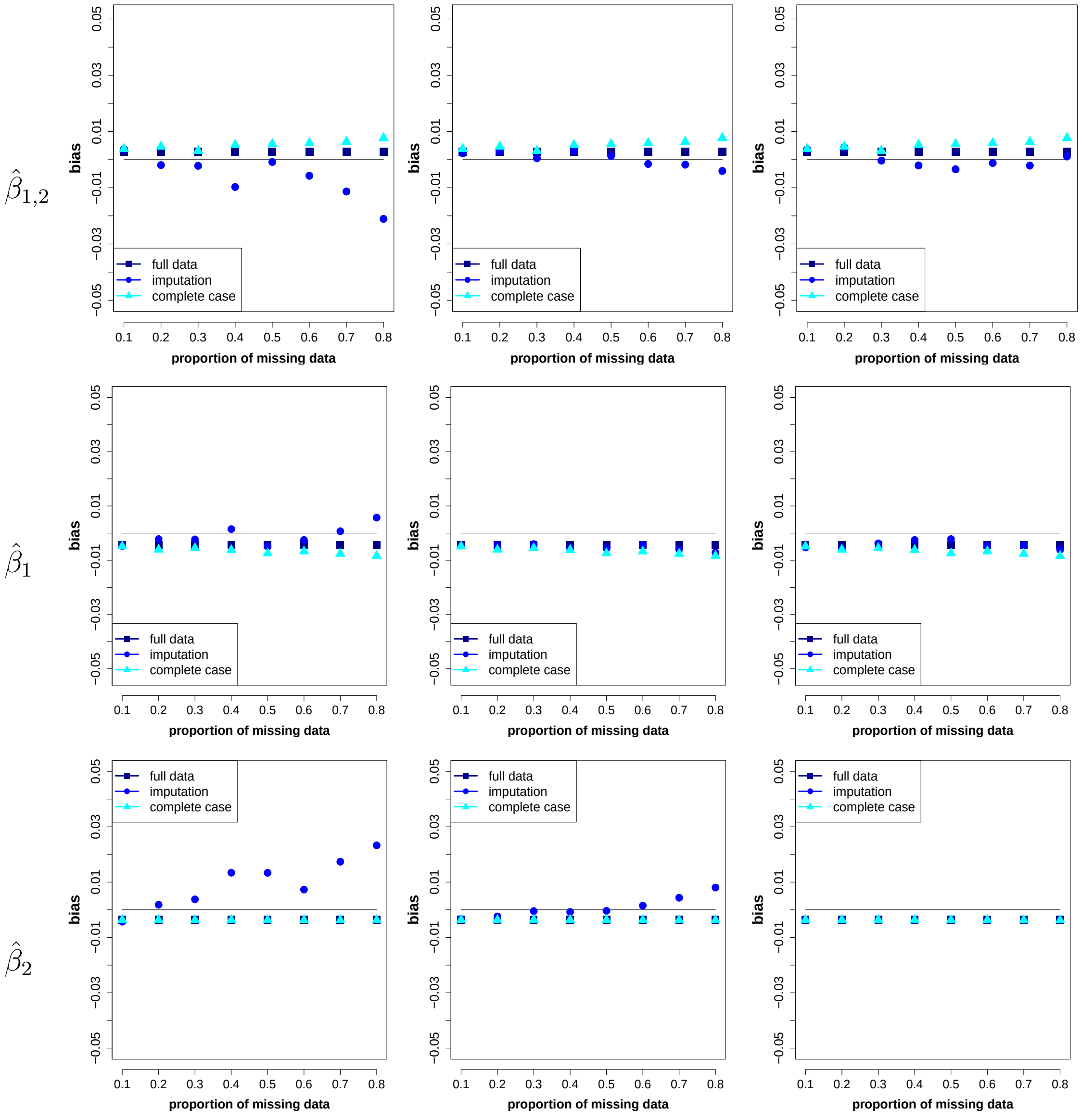
(a) Logistic regression (b) Polytomous logistic regression (c) Predictive mean matching

Figure 3.39: JAV: Standard deviation for LR, PLR, and PMM for $\hat{\beta}_{1,2}$, $\hat{\beta}_1$, and $\hat{\beta}_2$ treating interactions as JAV in imputation model. B_1 is the prioritized biomarker in the allocation algorithm, data is MCAR, and m=10 imputations.



(a) Logistic regression (b) Polytomous logistic regression (c) Predictive mean matching

Figure 3.40: Bias for LR, PLR, and PMM for $\hat{\beta}_{1,2}$, $\hat{\beta}_1$, and $\hat{\beta}_2$ treating interactions as JAV in imputation model when the number of imputations equal to percentage of missing data. B_1 is the prioritized biomarker in the allocation algorithm and data is MCAR.



(a) Logistic regression *(b) Polytomous logistic regression* *(c) Predictive mean matching*

Figure 3.41: Bias for LR, PLR, and PMM for $\hat{\beta}_{1,2}$, $\hat{\beta}_1$, and $\hat{\beta}_2$ treating interactions as JAV in imputation model when data is MAR. B_1 is the prioritized biomarker in the allocation algorithm and m=10 imputations.

Additional to data that is MCAR, a case with data that is MAR was considered, where the missingness of B_2 depends on entry time such that the probability of missingness is higher, the earlier a patient entered the study (cf. Section 3.4.6). Again, not much of a difference can be observed between the results for data that is MAR (Figure 3.41) and for data that is MCAR (Figure 3.40). Note that this may be different for other or more extreme scenarios of data that is MAR.

3.4.6.1 Sensitivity analysis

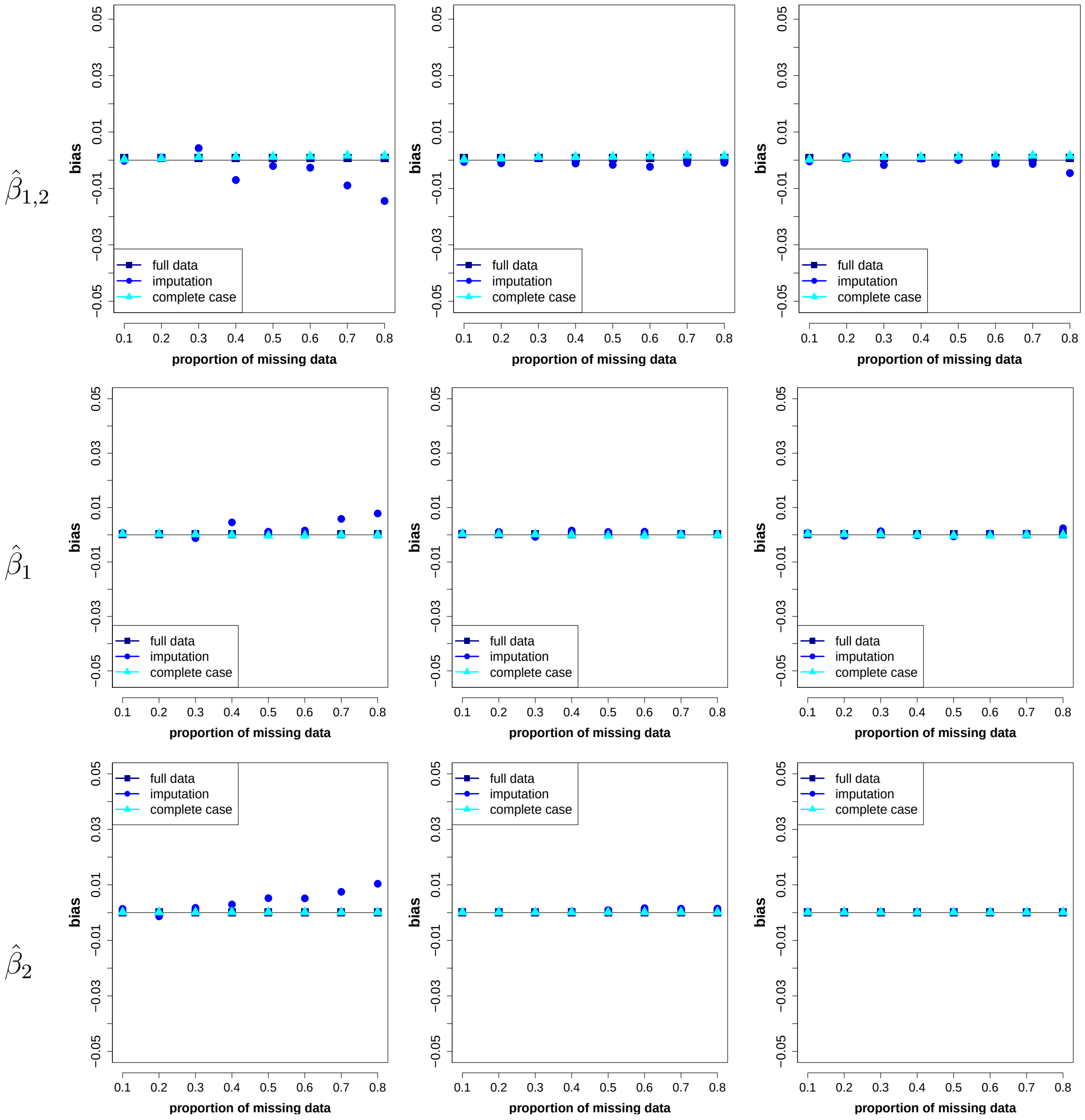
Additional to the scenario with data that is MAR, it was also investigated whether the results are sensitive to a change of parameters, focusing on the newly introduced parameters in Section 3.4.3: The time at which G_2 is added, t_b, the priority of the biomarkers, and the factor by which the hazard ratio for B_2-patients treated with Exp_1 differs from the hazard ratio for B_1-patients treated with Exp_1, i.e. the effect τ, and the priority of the biomarkers, i.e. to which biomarker-group patients are assigned to if they have both biomarkers.

For a scenario with $t_b = 6$ months, Figure 3.42 shows that for all imputation methods the bias of the estimates is reduced compared to Figure 3.38, for which $t_b = 12$ months. This could be expected, since altering t_b has an effect on the group sizes. Choosing t_b to be smaller means that G_2 is added earlier and there is more time to accrue patients to this group. Simultaneously, the number of patients in G_1 which were accrued before t_b gets smaller, causing the number of patients with potentially missing biomarker status for B_2 to be smaller. Hence, with more complete cases in the study data, the bias of the effect estimates obtained from the data imputation would be expected to be smaller.

For a scenario where B_2 is chosen as priority rather than B_1, Figure 3.43 shows that there is a larger bias for $\hat{\beta}_{1,2}$ and $\hat{\beta}_1$ for all methods compared to Figure 3.38. Again, this could be expected, since giving priority to B_2 would lead to a smaller number of patients in G_1 and especially to a smaller number of patients with both biomarkers in G_1, since patients with both biomarkers are allocated to G_2 after t_b.

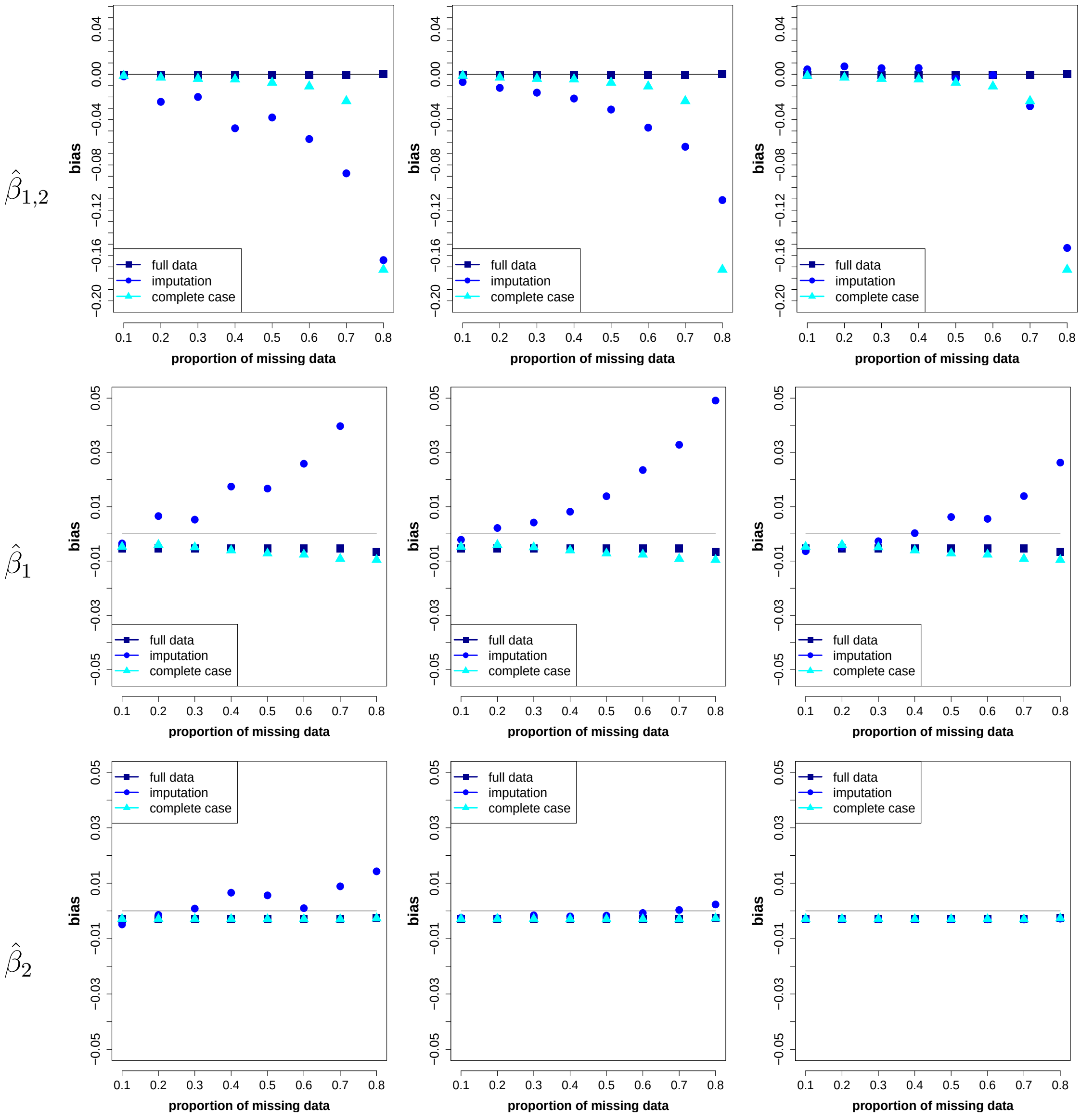
Finally, when the effect τ is set to 2 instead of 1.5, there are barely any noticeable differences compared to Figure 3.38. For $\tau = 0.75$, the bias of $\hat{\beta}_2$ again shows no visible differences. The bias of $\hat{\beta}_{1,2}$ and $\hat{\beta}_1$ behaves mostly similar for PLR and PMM and is reduced for LR for larger proportions of missing data. The results are shown in

Figures A.5 and A.6 in the Appendix.



(a) Logistic regression *(b) Polytomous logistic regression* *(c) Predictive mean matching*

Figure 3.42: t_b=6: Bias for LR, PLR, and PMM for $\hat{\beta}_{1,2}$, $\hat{\beta}_1$, and $\hat{\beta}_2$ treating interactions as JAV in imputation model when G_2 is added to the study after 6 months. B_1 is the prioritized biomarker in the allocation algorithm, data is MCAR, and m=10 imputations.



(a) Logistic regression (b) Polytomous logistic regression (c) Predictive mean matching

Figure 3.43: Priority=B_2, JAV: Bias for LR, PLR, and PMM for $\hat{\beta}_{1,2}$, $\hat{\beta}_1$, and $\hat{\beta}_2$ treating interactions as JAV in imputation model. B_2 is the prioritized biomarker in the allocation algorithm, data is MCAR, and m=10 imputations.

3.5 Choice of strategy for subgroup analyses

When analyzing survival data, there is often more than one hypothesis that is of interest. During the planning stage of a clinical trial, it is important to formulate all relevant hypotheses. Subsequently, it should be determined which one is the primary hypothesis and which hypotheses are secondary. For multiple biomarker trials, there are many potential hypotheses that could be investigated. In the following, the focus will be on hypotheses regarding effectiveness of treatment. Besides testing within each biomarker-group individually, an overall biomarker-guided treatment strategy could be evaluated, as previously discussed for the stratified randomize-all design in Section 3.2. Alternatively, a group of subgroups could be tested (e.g. all biomarker-positive patients), which could for example be used if not all biomarker-groups are randomized, as discussed for the multiple-biomarker hybrid design in Section 3.3. If it is preferred to test the biomarker-groups individually, but small sample sizes are expected, a hypothesis could be chosen which tests if a treatment benefit can be detected in at least one of the biomarker-groups. For designs like the stratified randomize-all design (see Figure 3.1, Design 3) or the multiple-biomarker hybrid design (see Figure 3.1, Design 4) and for a case with two biomarker-positive groups and one biomarker-negative group, hypotheses of interest could be, but are not limited to:

- H_1: Overall strategy: Experimental therapy vs. standard,
  i.e. $\beta = 0$ vs. $\beta \neq 0$
- H_2 - H_4: Individual hypotheses:

  H_2: Experimental therapy vs. standard within biomarker-group 1,
  i.e. $\beta_1 = 0$ vs. $\beta_1 \neq 0$

  H_3: Experimental therapy vs. standard within biomarker-group 2,
  i.e. $\beta_2 = 0$ vs. $\beta_2 \neq 0$

  H_4 (if applicable): Experimental therapy vs. standard within biomarker-negative group, i.e. $\beta_0 = 0$ vs. $\beta_0 \neq 0$
- H_5: Group of subgroups: E.g. experimental therapy vs. standard for biomarker-positive groups only, i.e. $\beta_+ = 0$ vs. $\beta_+ \neq 0$
- H_6: At least one: H_2 <u>or</u> H_3 <u>or</u> H_4

where β is the overall treatment effect, β_i is the treatment effect for biomarker-group i, and β_+ is the overall treatment effect for the biomarker-positive groups.

As main focus or primary hypothesis, one could either choose proof of concept/treatment strategy (H_1), or the individual biomarkers (H_2 - H_6). Decision criteria to choose a primary hypothesis could be the prevalence of the biomarkers, or the extent of similarities between the biomarkers. To maintain an overall α-level, one option to test these hypotheses is a multiple comparison procedure that adjusts for multiple testing. If the sample size calculation is only based on the primary hypothesis H_1, the other hypotheses may be under-powered, especially those hypotheses which only include a fraction of the patients included in the study. In this case, a classical method like Bonferroni, where each of m hypotheses is tested at α/m level, would be very restrictive. Instead, a procedure which allows reallocation of the unspent α could be considered. A potential approach is a so-called chain procedure (Millen and Dmitrienko 2012). Chain procedures are a class of closed testing procedures, which are based on the closure principle, meaning that a hypothesis is only rejected if and only if all intersection hypotheses that contain this hypothesis can be rejected. The advan-

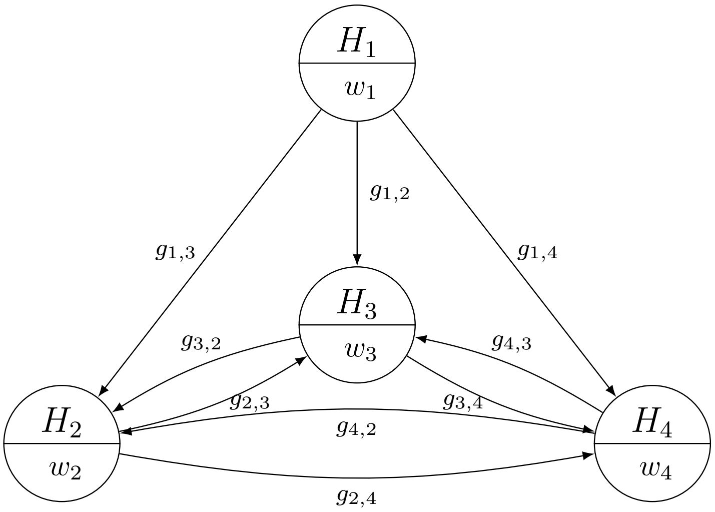


Figure 3.44: Cyclical chain procedure for testing four hypotheses incorporating logical relationships for an example where H_1 is the main hypothesis of interest. If $w_1 = 1$ is chosen, H_2-H_4 are only tested if H_1 is rejected. If $w_1 < 1$ is chosen, then H_2-H_4 can still be tested at the remaining α. (Adapted from Millen and Dmitrienko (2012)).

tage of a closed testing procedure is that, after a hypothesis is rejected, the unspent α can be carried over to the remaining hypotheses. For a chain procedure, initially all hypotheses are assigned weights w_c, such that in the beginning, hypothesis H_c is allocated a fraction of the overall α, $\alpha_c = w_c\alpha$. Additionally, transition parameters $g_{c,d}$ are determined which, after testing H_c, are used to reallocate the unspent α_c from hypothesis c to hypothesis d, i.e. α_d is updated such that $\alpha_{d_{new}} = \alpha_d + g_{c,d}\alpha_c$.

Millen and Dmitrienko (2012) describe different kinds of chain procedures. One type are serial chain procedures, which test a family of ordered hypotheses, i.e. the order in which the hypotheses are tested is prespecified. Unlike serial chain procedures, cyclical chain procedures do not assume an ordering of the hypotheses; the hypothesis tested first is the one with the most significant weighted p-value. If desired, a serial and a cyclical chain procedure can be combined to incorporate a logical relationship, e.g. to test the primary hypothesis first (see Fig. 3.44). If the weight of the primary hypothesis is chosen to be one, i.e. if the secondary hypotheses are only tested if the primary hypothesis can be rejected, this combination of cyclical and serial chain procedure can also be classified as a type of gatekeeping procedure. Here, the primary hypothesis would be the 'gatekeeper', which needs to be rejected in order to test the remaining hypotheses.

The following sections will suggest several testing strategies for testing multiple hypotheses. These strategies are based on the approaches of Millen and Dmitrienko (2012) and Bretz et al. (2009, 2011) for multiple comparison and closed testing procedures. Note that the hypothesis weights and α transition weights in the following sections were chosen to be equal among hypotheses of the same hierarchy (e.g. among all secondary hypotheses). These weights can be adjusted as needed. The figures shown for the different strategies were created using the gMCP package for R (Rohmeyer and Klinglmueller 2015), which was developed based on Bretz et al. (2011).

3.5.1 Options for main focus "Proof of concept"

If proof of concept or proof of treatment strategy is the main point of interest, H_1 could be chosen to be the primary hypothesis. A few of the many possible testing strategies are briefly presented in this section. A conservative approach would be the Bonferroni method, where α is split equally between the secondary hypotheses. Taking advantage of the closure principle, the unspent α can be reallocated to the secondary hypotheses if the primary hypothesis is rejected. The first option that is suggested below is a combination of serial and cyclical chain procedure. It is a sensible strategy if all secondary hypotheses are of similar interest and importance to the investigators. If, on the other hand, the investigators would prefer the secondary hypotheses to be tested in a specific order, a serial chain procedure could be used, where the hypotheses are tested in a prespecified order and the unspent α is split between the remaining hypotheses. The way both these options are shown in Figures 3.45 and 3.46, with $w_1 = 1$, these testing strategies can also be categorized as two-stage gatekeeping procedures, as mentioned in the previous section. If, additionally, the hypotheses can be classified as secondary, tertiary, and quaternary, a multi-stage serial gatekeeping procedure can be considered instead, where the unspent α from a hypothesis is only reallocated to the hypothesis that is next in line in the prespecified order.

Option 1: Mixed chain procedure

Step 1: Test H_1 (overall strategy) at level α.

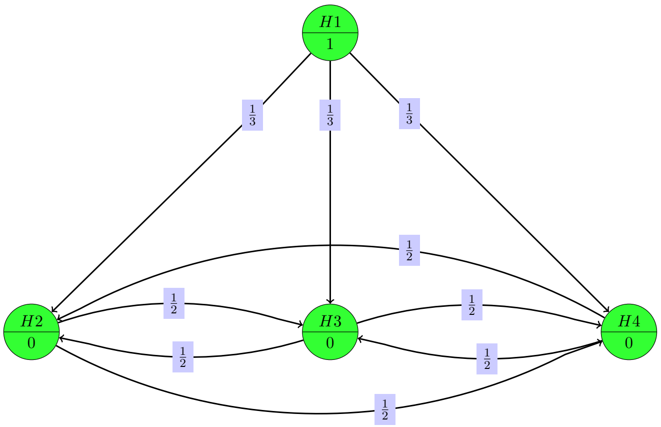


Figure 3.45: Mixed chain procedure for multiple testing when the main hypothesis is the overall treatment strategy.

Step 2: If H_1 is rejected, pass α on to H_2 - H_4 using the prespecified transition parameters. In case of a rejection of H_c (where $c = 2, 3, 4$), α_c can be split equally between the remaining hypotheses to be tested (see Figure 3.45).

If H_1 cannot be rejected, subgroup analyses could be performed on an exploratory basis.

Option 2: Serial chain procedure

Step 1: Define order for H_2 - H_4 (individual biomarkers), e.g. based on prevalence or expected outcome. Let H_2^*, H_3^*, H_4^* denote H_2 - H_4 ordered by this hierarchy.

Step 2: Test H_1 (overall strategy) at level α.

Step 3: If H_1 is rejected, pass α on to H_2^* - H_4^* (individual biomarkers). In case of a rejection of H_c^* (where $c = 2, 3, 4$), α_c can be split equally between the remaining hypotheses to be tested (see Figure 3.46).

If a hypothesis in the sequence cannot be rejected, the remaining subgroup analyses could be performed on an exploratory basis.

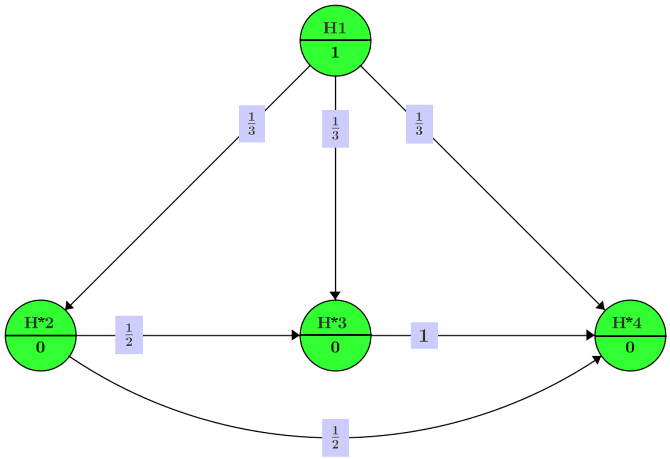


Figure 3.46: Serial chain procedure for multiple testing when the main hypothesis is the overall treatment strategy.

Option 3: Multi-stage serial gatekeeping procedure

Step 1: Define order for H_2 - H_4 (individual biomarkers), e.g. based on prevalence or expected outcome. Let H_2^*, H_3^*, H_4^* denote H_2 - H_4 ordered by this hierarchy.

Step 2: Test H_1 (overall strategy) at level α.

Step 3: If H_1 is rejected, pass α on to H_2^* (highest ranked individual biomarker, according to prespecified order).

Step 4: Test H_2^* at level α. If H_2^* cannot be rejected, stop. If H_2^* is rejected, pass α on to next hypothesis in the sequence. Proceed the same way for H_3^* and H_4^*.

If a hypothesis in the sequence cannot be rejected, the remaining subgroup analyses could be performed on an exploratory basis.

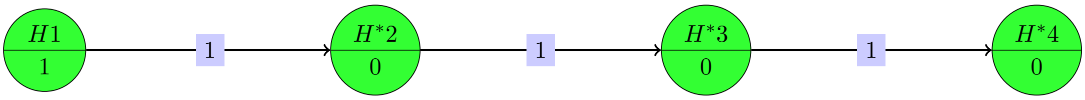


Figure 3.47: Multi-stage serial gatekeeping procedure for multiple testing when the main hypothesis is the overall treatment strategy.

If the individual biomarker-groups are expected to be small and testing the individual hypotheses seems unpromising, either H_5 (group of subgroups) or H_6 (at-least-one) could be tested instead.

Option 4: Group of subgroups

Step 1: Test H_1 (overall strategy) at level α.

Step 2: If H_1 is rejected, pass α on to H_5 (group of subgroups). If H_1 cannot be rejected, subgroup analyses could be performed on an exploratory basis.

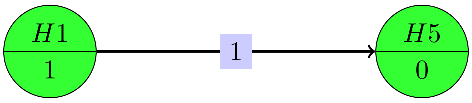


Figure 3.48: Testing strategy with hypothesis for group of subgroups instead of separate hypotheses for individual biomarkers.

Option 5: At least one

Step 1: Test H_1 (overall strategy) at level α.

Step 2: If H_1 is rejected, pass α on to H_6 (at least one). If H_1 cannot be rejected, subgroup analyses could be performed on an exploratory basis.

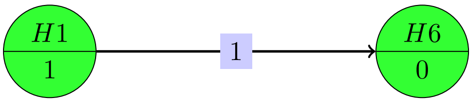


Figure 3.49: Testing strategy with "at-least-one" hypothesis instead of separate hypotheses for individual biomarkers.

Option Overview

Table 3.9: Overview of testing strategies, when the main focus of the analysis is a proof of concept.

	Focus: Overall treatment strategy				
Strategy	**Mixed chain procedure**	**Serial chain procedure**	**Serial gatekeeping**	**Group of subgroups**	**At least one**
Step 1	Test H_1 at level α				
Step 2	Split remaining α equally among H_2 - H_4	Define order for H_2 - H_4 and split remaining α equally among remaining hypotheses	Define order for H_2 - H_4 and pass remaining α down sequentially	Test H_5 at remaining α	Test H_6 at remaining α

3.5.2 Options for main focus "Individual biomarkers"

If investigators believe that they will have sufficiently large biomarker-groups, but would also like to evaluate the overall treatment strategy, they could choose the individual biomarkers (H_2 - H_4) as main hypotheses and the treatment strategy as secondary hypothesis. Similar to the previous section, a mixed chain procedure could be applied where the primary hypotheses are tested first and the unspent α is split between the remaining hypotheses to be tested. Or, if the hypotheses are to be tested in a prespecified order, a serial chain procedure or a multi-stage serial gatekeeping procedure could be chosen instead. Again, if $w_1 = 1$, as shown in Figures 3.50 and 3.51, both chain procedures can also be categorized as two-stage gatekeeping procedures. Finally, instead of testing H_2 - H_4, H_5 or H_6 could be chosen as primary hypotheses.

If some of the hypotheses cannot be tested, remaining analyses of interest could be performed on an exploratory basis.

Option 1: Mixed chain procedure

Step 1: Test H_2 - H_4 (individual biomarkers). In case of a rejection of H_c (where $c = 2, 3, 4$), its proportion of α can be split equally between the remaining hypotheses to be tested (including H_1; see Fig. 3.45).

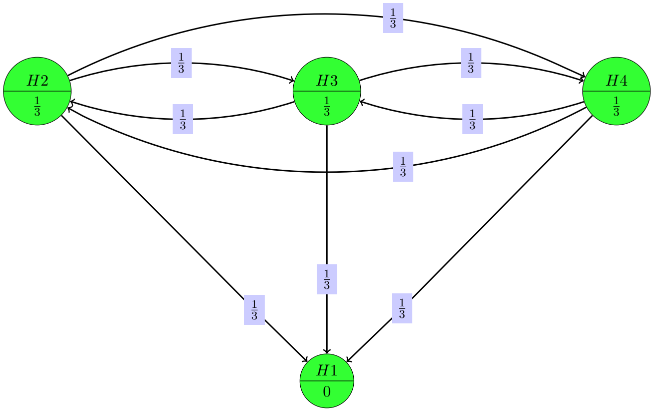


Figure 3.50: Mixed chain procedure for multiple testing when the main hypothesis are the individual biomarkers.

Step 2: Test H_1 (overall strategy) at the level resulting from *Step 1*, i.e. if one hypothesis was rejected, test at $\frac{1}{9}\alpha$, if two were rejected, test at $\frac{1}{3}\alpha$, and if all preceding hypotheses were rejected, test at full α.

Option 2: Serial chain procedure

Step 1: Define order for H_2 - H_4 (individual biomarkers), e.g. based on prevalence or expected outcome. Let H_2^*, H_3^*, H_4^* denote H_2 - H_4 ordered by this hierarchy.

Step 2: Test H_2^* at level α. If H_2^* cannot be rejected, stop. If H_2^* is rejected, pass α on to next hypothesis in the sequence. Proceed the same way for H_3^* and H_4^*.

Step 3: Test H_1 (overall strategy) at the level resulting from *Step 1*, i.e. if one hypothesis was rejected, test at $\frac{1}{9}\alpha$, if two were rejected, test at $\frac{1}{3}\alpha$, and if all preceding hypotheses were rejected, test at full α.

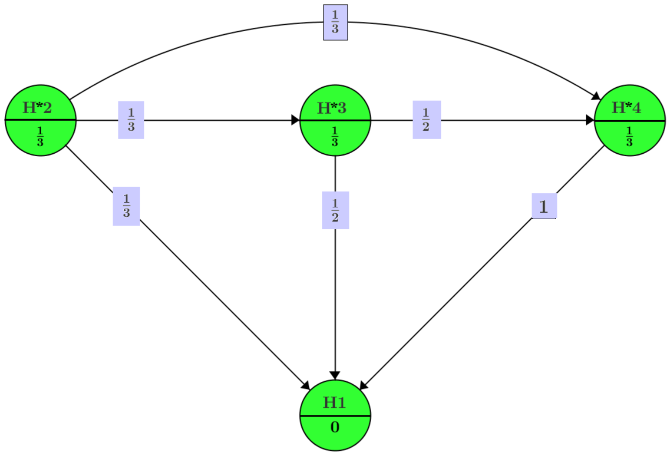


Figure 3.51: Serial chain procedure for multiple testing when the main hypothesis are the individual biomarkers.

Option 3: Multi-stage serial gatekeeping procedure

Step 1: Define order for H_2 - H_4 (individual biomarkers), e.g. based on prevalence or expected outcome. Let H_2^*, H_3^*, H_4^* denote H_2 - H_4 ordered by this hierarchy.

Step 2: Test H_2^* at level α. If H_2^* cannot be rejected, stop. If H_2^* is rejected, pass α on to next hypothesis in the sequence. Proceed the same way for H_3^* and H_4^*.

Step 3: If H_2^*, H_3^*, H_4^* were all rejected, test H_1 (overall strategy) at level α.

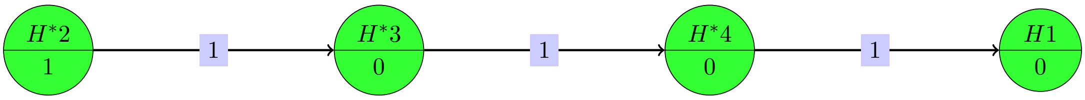


Figure 3.52: Multi-stage serial gatekeeping procedure for multiple testing when the main hypothesis are the individual biomarkers.

Option 4: Group of subgroups

Step 1: Test H_5 (group of subgroups) at level α.

Step 2: If H_5 is rejected, pass α on to H_1 (overall strategy).

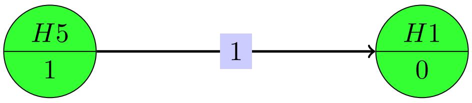


Figure 3.53: Testing strategy with hypothesis for group of subgroups instead of separate hypotheses for individual biomarkers when the main hypothesis are the individual biomarkers.

Option 5: At least one

Step 1: Test H_6 (at least one) at level α.

Step 2: If H_6 is rejected, pass α on to H_1 (overall strategy) .

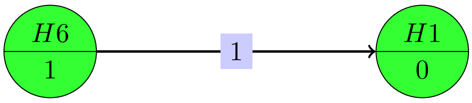


Figure 3.54: Testing strategy with "at-least-one" hypothesis instead of separate hypotheses for individual biomarkers when the main hypothesis are the individual biomarkers.

Option Overview

Table 3.10: Overview of testing strategies when the main focus of the analysis is on the individual biomarkers.

	Focus: Individual biomarkers				
Strategy	**Mixed chain procedure**	**Serial chain procedure**	**Multi-stage serial gatekeeping**	**Group of subgroup**	**At least one**
Step 1	Test H_2 - H_4	Test H_2 - H_4 sequentially, split α equally among remaining hypotheses	Test H_2 - H_4 sequentially, allocate unspent α to next hypothesis	Test H_5 at level α	Test H_6 at level α
Step 2	Test H_1 at remaining α	Test H_1 at remaining α	Test H_1 at remaining α	Test H_1 at remaining α (H_2 - H_4 exploratory?)	Test H_1 at remaining α (H_2 - H_4 exploratory?)

The are many strategies to choose from for testing multiple hypotheses. The options suggested here are just a selection and were chosen in consideration of the study designs discussed in this thesis. The strategies in Section 3.5.1 can be utilized for the stratified randomize-all design as discussed in Section 3.2, while Section 3.5.2 could be useful for the multi-biomarker hybrid design as discussed in Section 3.3. Especially for designs that evaluate an overall treatment strategy, secondary hypotheses that test the individual biomarkers should be included, to avoid false conclusions. For further reading on multiple comparison procedures refer to e.g. Dmitrienko et al. (2009).

Chapter 4

Discussion

Planning and analyzing a multiple biomarker trial is a challenging task comprising various factors which have to be considered. It is an area of ongoing research and only a limited number of multiple biomarker trials have already been completed and their results published. Learning from these completed trials is an important part of the planning process, which can help to avoid issues and pitfalls that these trials may have encountered. Some of the issues which were reported by completed trials, such as low prevalence of the biomarkers and not being able to react to the latest developments regarding biomarkers and treatments, have been addressed in this thesis. Sample size calculation and data analysis methods for testing an overall treatment strategy were investigated for situations where biomarker prevalences make it unfeasible to test within the individual biomarker-groups. The results will be discussed in Section 4.1. Additionally, the issue of a large number of biomarker-negative patients was addressed, which is a side effect in trials that investigate lower prevalence biomarkers. Different analysis approaches for a trial that includes biomarker-negative patients were compared and it was examined whether inclusion of biomarker-negative patients in the analysis can improve bias and standard deviation of the treatment effect estimates. Section 4.2 will discuss the results in more detail. Finally, a flexible study design was considered that allows a new biomarker-group with corresponding experimental treatment to be added to the study after accrual has already begun. Different aspects of study design modification were discussed and different models for analysis of such a study were compared. Furthermore, the issue of missing biomarker data was addressed. If the initial biomarker screening did not include the new biomarker before it was added to the study, the biomarker status regarding this biomarker has to be determined retrospectively for patients that were included in the study before adding the new biomarker. This may lead to missing biomarker data for some or all of the patients. For cases where data is only partially missing, different methods for missing data imputation for models with interaction terms were investigated and compared. The results for the different analysis models and the missing data will be discussed in Section 4.3.

4.1 Sample size calculation and evaluation of overall treatment strategy

Section 4.1 is largely part of a paper that has already been published. The relevant passages have been taken verbatim from Beisel et al. (2017).

With regard to sample size calculation, the Schoenfeld method is acceptable to use in situations with only minor heterogeneity of treatment effects. But for more heterogeneous cases, which was the focus of Section 3.2, the formula by Palta and Amini provides the most adequate sample size, in the sense that it delivers the level of power which was aimed for at the planning stage. This is true not just for the case of equal hazard ratios, but also for minorly to moderately heterogeneous hazard ratios. If the hazard ratios are extremely heterogeneous, the adherence to the level of power depends on the analysis method used. It seems that in this case, the formula by Palta and Amini cannot compensate for the loss of power of the stratified Cox PH model in spite of taking into consideration the heterogeneous hazard ratios.

However, comparing the asymptotic and the exact log-rank test for these particular scenarios, and taking into consideration the calculated sample sizes, an obvious question is whether the reason for the poorer performance of the log-rank test and the stratified Cox PH model is that the assumptions regarding asymptotic properties are not met. For the most extreme scenarios, the calculated sample size is below 300, which, with prevalences of 0.5, 0.25, and 0.25 for G_0, G_1, and G_2 respectively, results in stratum sizes of less than 75 patients for G_1 and G_2. Hence, asymptotic assumptions are problematic in these cases. A simulation study with equal hazard ratios across strata but smaller sample sizes in Section 3.2.3.4 showed that there is indeed some loss of power, but for an overall sample size of 115 patients (which is very close to the 113 patients in the most extreme case considered), the loss of power is minor (see Table 3.4). For control of the type-I-error rate, the exact log-rank test should be considered instead if the strata sizes are expected to be small, i.e. in the double digits, and there is too little data available for reliable approximations. The shared frailty model also appears to be an appropriate choice and additionally offers the possibility to include covariates in the model, which is an advantage over the exact log-rank test.

Note that the shared frailty model performs superior to the stratified Cox PH model for strata with different treatment effects, even if there is no stratum-specific random effect present in the data, which is assumed by the shared frailty model. So, even though it is meant to model heterogeneity in baseline hazards, not hazard ratios, it may be that it can handle heterogeneity better in general. Note that if one tries to estimate a stratum effect that is not there, the less degrees of freedom are spent on this attempt, the better. Hence the shared frailty model would perform better in this case than the stratified model (with the simple non-stratified model being the best). Additionally, since the shared frailty model assumes a certain functional shape causing heterogeneity in treatment effects between strata, it is likely to perform better in terms of power than a method making lesser assumptions, like the stratified Cox PH model, which allows more general forms of baseline hazards than the constant baseline hazards that are considered here.

Another observation worth mentioning is that the formula by Palta and Amini provides a reasonable sample size for the lognormal shared frailty model in the case of minor random effects. For stronger random effects, the shared frailty model does suffer some loss of power, but less than the other methods (see Figures in Section 3.2.3.4). In those cases, it may be that the shared frailty model is not able to compensate for both, heterogeneity of baseline hazards and heterogeneity of treatment effects. If a strong random effect is expected, an empirical calculation of the required sample size should be considered instead (see Section 3.2.2).

With regard to the different analysis methods compared, the exact log-rank test, Mehrotra's two-step approach, and the shared frailty model yield at least the level of power which was aimed for, regardless of the heterogeneity of the treatment effects. The exact log-rank test is only of use if one is not interested in including additional covariates in the model. The two-step approach, as well as the shared frailty model, does facilitate this option.

Note that using an overall analysis method as suggested should be understood as assessing the benefit of a certain treatment strategy for the overall population and should be performed in conjecture with subsequent subgroup analyses to confirm the findings and to avoid false conclusions about treatments and subpopulations.

In consideration of the results in Section 3.2, the asymptotic log-rank test and the stratified Cox PH model should not be used in case of small sample sizes and heterogeneous treatment effects. In this case, the exact log-rank test, Mehrotra's two-step approach, or the shared frailty model appear to be suitable alternatives. For all models, these results should be used with caution if one wishes to model a greater magnitude of stratification or include a strong random effect in addition to heterogeneous treatment effects. With respect to sample size calculation for a situation with stratification and heterogeneous treatment effects, it is suggested to use the not widely recognized extension of Schoenfeld's formula by Palta and Amini rather than the formula by Lachin and Foulkes.

4.2 Inclusion of biomarker-negative patients and small sample size bias

Section 4.2 is largely part of a paper that has already been published. The relevant passages have been taken verbatim from Habermehl et al. (2017).

While there already exists some literature on biomarker designs, the inclusion of biomarker-negative patients is rarely addressed. Therefore, the focus of Section 3.3 was a study design including biomarker-negative patients and possible analysis approaches for the resulting data.

If it is contemplated to include biomarker-negative patients in a trial, the pros and cons should be carefully considered. The added expense of inclusion of biomarker-negative patients should be weighed against potential gains in efficiency for treatment effect estimation. If it is possible, it might be a better option to recruit a larger number of biomarker-positive patients instead of spending resources on following biomarker-negative patients. If there is strong interest in gaining information about prognostic effects, then it would be necessary to study biomarker-positive and -negative patients under standard therapy. But, depending on availability, this might also be done retrospectively by subdividing a patient cohort from an already existing study using stored specimens. For each trial it should be considered individually which goals are most important and then resources can be allocated accordingly.

For the situation when biomarker-negative patients are to be included in the study, the results of the simulation study in Section 3.3 show that for smaller sample sizes, and especially for biomarkers indicative of a poor prognosis, using a combined model and including the biomarker-negative patients in the analysis can reduce bias and standard deviation of the estimates of the regression coefficients compared to excluding biomarker-negative patients from the analysis and to performing separate analyses for the biomarkers, assuming proportional hazards. For larger sample sizes ($n = 1,000$) no noticeable reduction in bias and standard deviation can be observed. This leads to the conclusion that the observed benefit with respect to bias reduction could be due to the reduction of the small sample size bias of the Cox PH model. The different results for biomarkers with different prognoses are likely due to the smaller number of events in the biomarker-group indicative of a favorable prognosis, which seems to cause additional bias.

The increase of bias and standard deviation in absolute terms as the treatment effect gets larger shows that the bias not only depends on the sample size but also on the number of events. Hence, other factors influencing the bias of the treatment effect estimates appear to be baseline hazard and hazard ratio, which agrees with the results reported by Langner et al. (2003). The greatest benefit with respect to reduction of bias was observed for large treatment effects and for the biomarker indicative of a poor prognosis relative to the biomarker-negative population. These results indicate that, for small sample situations such as low prevalence biomarkers, using a Cox PH model with the biomarker status as factor variable and one variable for each biomarker-specific treatment can help reduce bias and standard deviation of the estimates. Additionally, this modeling approach facilitates the opportunity to estimate the treatment effects and prognostic effects for all biomarkers in one model.

The simulation study also demonstrates the benefit of the Firth correction in small to moderate sample size situations with respect to reduction of bias and also standard deviation, which increases as the treatment effect for the biomarker-group increases. This was observed for both biomarkers. The upwards bias that can be seen for $\hat{\beta}_1$ for small sample sizes when the Firth correction is used could be caused by the small number of events, a behavior which Elgmati et al. (2015) described for logistic regression.

As with all simulation studies, there are some limitations because of model assumptions. The model assumptions, such as constant baseline hazards and proportional hazards, were fulfilled by the simulated data set but that is not necessarily the case for real data sets. The simulation results in Section 3.3.2.3 suggest that, while using different biomarker prevalence and time-dependent baseline hazards did not have much of an impact on bias and standard deviation, one needs to be careful in the case of very small numbers of events. Furthermore, the assumption of proportional hazards is a key assumption of the Cox PH model. Therefore, it should always be checked if this assumption is reasonable before applying the model. For mid to late phase trials, it can be assumed that there is some knowledge available prior to the study that can give guidance on whether proportional hazards are a reasonable assumption. If non-proportional hazards are expected, the model(s) suggested in Section 3.3 may not be an appropriate choice. A simulation study on the extent of bias of the hazard ratio caused by non-proportional hazards, as well as censoring rate, type of censoring, and sample size can be found in Persson and Khamis (2005). Depending on the individual situation, alternatives or extensions to the Cox PH model that can handle non-proportional hazards should be considered. Some literature on Cox regression under non-proportional hazards can, e.g. be found in Schemper (1992). Alternative models to the Cox PH model are, e.g. additive hazards models or accelerated failure time models, which are, in the context of causal inference, suggested by Aalen et al. (2015).

Overall, based on the results of the simulation study, it can be concluded that the small sample size bias of the Cox PH model should not be neglected. Additionally, attention should be paid to the expected number of events, which also appears to influence the bias. In the analysis of a multiple-biomarker trial with low prevalence biomarkers and hence small sample sizes, a bias correction should be applied, such as the well-known Firth correction. Moreover, the results of the simulation study demonstrate that for biomarkers indicative of a poor prognosis, the inclusion of biomarker-negative patients in the analysis can help to further reduce bias of the effect estimates and can additionally lead to small improvements of the standard deviation of the estimates.

4.3 Adding a new biomarker-group: Interaction effects and missing data

When it is intended to add a new biomarker-group to an ongoing clinical trial, many factors need to be considered. The study protocol should facilitate this option from the beginning and explicitly state the conditions under which a new biomarker-group can be added throughout the trial. This should include the time point up to which it is possible to add the new biomarker-group to the study, as well as guidelines regarding allocation to the biomarker-groups for new patients with both biomarkers after the new group is added, e.g. by assigning priorities to the biomarkers. Additionally, it should be discussed how to adjust the final analysis for the belatedly added group and how to address the potential overlap of biomarkers within the groups. While simply excluding the patients with both biomarkers would be an acceptable option, with respect to bias of the treatment effect estimates, the overlap of the biomarkers and a potential interaction between therapy and additional biomarker could be accounted for by including an interaction term in the model. Both approaches can help avoiding biased treatment effect estimates but the interaction term offers the additional benefit of quantifying the interaction effect. Similar to Section 3.3, a combined model, which estimates the treatment effect estimates for both biomarkers and additionally the interaction term could be applied.

If the biomarker status with respect to the newly added biomarker cannot be determined retrospectively for all patients which were included prior to adding the new biomarker-group, multiple imputation can be utilized. However, data imputation for models that include interaction terms can be difficult and the simulation study in Section 3.4.6 shows that it is rather sensitive to model misspecification, such as omission of interaction terms from the imputation model.

When the interaction terms are imputed passively, i.e. calculated from the imputed main effects, the interaction terms are omitted from the imputation model. This results in moderately to severely biased estimates for all considered imputation methods, which were predictive mean matching, logistic regression, and polytomous logistic regression. When the stratify-approach is used instead of PI, the bias of $\hat{\beta}_{1,2}$ is reduced, but both logistic regression methods still yield biased estimates for β_1 and β_2. This may be due to the specific model considered here, which only contains one of the main effects. Since the two biomarker-groups receive a different experimental treatment, a main effect for treatment would not be meaningful in this context. Another aspect making this a special case is that one of the three data sets into which the data is split for the stratify-approach (one for each treatment group) does not contain any missing data. This is due to the fact that Exp_2 is added to the study at the same time when G_2 is added. Starting at this point, B_2 is determined for all patients upon entering the study, so the biomarker status for B_2 can only be missing for patients treated with Exp_1 or with standard of care who entered the study prior to adding G_2. It is possible that these circumstances lead to the biased estimates for the stratify-approach. Hence, while the stratify-approach may work well in general, it does not seem to be an appropriate choice for the situation considered here.

The JAV-approach on the other hand is able to eliminate most of the bias of $\hat{\beta}_{1,2}$, $\hat{\beta}_1$ and $\hat{\beta}_2$ for all three imputation methods, even for the MAR scenario considered here. Thus, for this particular situation, the JAV-approach seems to be the most appropriate choice. Using the percentage of missing data as number of imputations, as suggested by Bodner (2008), only resulted in minimal improvements in bias. This may not always be in proportion to the added computation time. PMM and PLR perform quite well even for larger proportions of missing data. With regard to standard deviation, PLR performs slightly better than PMM. The observed differences between the two logistic regression based methods, LR and PLR, could be due to differences in implementation of these methods in the used R package mice. While the imputation function for LR uses the function glm (generalized linear models) to fit a regression model, the function for PLR uses multinom (multinomial log-linear models). These two functions use different methods for model fitting. While glm uses iteratively reweighted least squares to fit the model, multinom fits a single-hidden-layer neural network via the function nnet, which is a machine learning technique for generalization of linear regression functions.

As usual, the conclusions drawn from the simulation study may not necessarily apply to other situations and different models. The results may also differ for more complex scenarios of data that is MAR or for violations of model assumptions. For the different parameter settings considered in the sensitivity analysis, no unexpected behavior could be observed.

Overall, an overlap in patient population and interaction of biomarker and treatment could be accounted for by including an interaction term in the model. Furthermore, if the biomarker status regarding the new biomarker is missing for some of the patients, using the interaction terms as variables in the imputation model appears to be the best way to avoid biased estimates for the situation considered here. PI should not be used for imputation of interaction terms. The simulation study suggests that, compared to PI and the stratify-approach, using the JAV-approach together with PLR yields the least biased estimates along with the smallest standard deviation.

4.4 Overall conclusion and outlook

This thesis aimed to address several issues which can arise when planning and analyzing a multiple biomarker trial, leading to several main conclusions that can be drawn from the results. The first issue which was addressed in this thesis was low prevalence of the biomarkers. It was aimed to investigate the evaluation of the biomarker-guided treatment strategy for situations with lower prevalence biomarkers. For a study which tests an overall biomarker-guided treatment strategy, the sample size calculation method by Palta and Amini appears to be the most appropriate choice when heterogeneous treatment effects are expected. The results from the simulation study suggest that the subsequent data analysis could be performed using the two-step approach suggested by Mehrotra or a shared frailty model. If no other covariates are included in the model, an exact log-rank test could also be used. The asymptotic log-rank test and the stratified Cox PH model suffered loss of power in the simulation study and therefore should not be used for heterogeneous treatment effects. To test the individual biomarker-groups as secondary hypotheses after testing the overall treatment strategy, some strategies for multiple testing were suggested.

The second issue that was addressed was a large expected number of biomarker-negative patients at the screening stage. It was aimed to investigate whether inclusion of biomarker-negative patients in the trial and the analysis provides additional benefit, such as improvement of power or reduction of bias. For a situation where an overall biomarker-guided treatment strategy is not desirable, a combined analysis model using the data from the entire study, including biomarker-negative patients, was investigated. This combined model estimates the treatment effects for both individual biomarkers. Application of the Firth correction appeared to be a good method for reduction of small sample size bias, which is likely to occur for low prevalence biomarkers. The inclusion of biomarker-negative patients in the model can provide a small additional benefit with respect to reduction of bias and standard deviation.

The third issue considered were the rapid developments in the field of biomarker research. It was aimed to be able to react to these continuous developments by investigating options to add new biomarkers and corresponding therapies to an ongoing study. Different models for data analysis were compared for a situation with a belatedly added biomarker, an overlap of biomarkers within the population, and an effect of the new biomarker on the response to the experimental treatment of the already existing biomarker-group. Adding an interaction term to the combined analysis model can help avoiding biased treatment effect estimates when there is overlap of the biomarkers within the patient population, and when patients with both biomarkers respond differently to the experimental therapy than patients with only one of the biomarkers. If there is missing data regarding the biomarker status of the belatedly added biomarker, data imputation can be utilized. However, the correct model specification is crucial to avoid biased estimates when interaction terms are part of the model for the final analysis. These interaction terms should already be included in the imputation model rather than imputing them passively. The simulation study suggests that for the considered scenario, the JAV-approach with PLR is the best option to avoid obtaining biased estimates after data imputation, and to reduce standard deviation compared to CC analysis.

Points of future research based on the results of this thesis could be a comparison of the performance of the suggested multiple comparison strategies when they are used for the stratified randomize-all design. Furthermore, the Firth correction could be applied to the models investigated in Section 3.4 when smaller sample sizes are used. For all models considered, the behavior when covariates are added to the model could be investigated, as well as different parameter settings. For the flexible study design, methods for sample size recalculation after adding the new biomarker group could be investigated.

Due to the heterogeneity of biomarkers and treatments and the rapid developments in this field, the planning phase of a multiple-biomarker trial is a complex process and each trial has to be adjusted to the individual situation. This thesis can give guidance in some of the aspects that need to be considered, but of course there are many more aspects that need to be addressed. The study designs which were discussed could, for example, be extended to include and interim analysis strategy to facilitate sample size recalculation, early stopping, or stopping for futility.

Chapter 5

Summary

Planning and analyzing a multiple biomarker trial is a challenging task comprising various factors which have to be considered. It is an area of ongoing research and only a limited number of multiple biomarker trials have already been completed and their results published. Learning from these completed trials is an important part of the planning process, which can help to avoid issues and pitfalls that these trials may have encountered. Some of the issues which were reported by completed trials, such as low prevalence of the biomarkers and not being able to react to the latest developments regarding biomarkers and treatments, are addressed in this thesis.

Sample size calculation and data analysis methods for testing an overall treatment strategy are investigated for situations where biomarker prevalences make it unfeasible to test within the individual biomarker-groups. Additionally, the issue of a large number of biomarker-negative patients is addressed, which is a side effect in trials that investigate lower prevalence biomarkers. Different analysis approaches for a trial that includes biomarker-negative patients are compared and it is examined whether inclusion of biomarker-negative patients in the analysis can improve bias and standard deviation of the treatment effect estimates. Finally, a flexible study design is considered that allows a new biomarker-group with corresponding experimental treatment to be included in the study after accrual has already begun. Different aspects of study design modification are discussed and different models for analysis of such a study are compared. Furthermore, the issue of missing biomarker data is addressed. If the initial biomarker screening did not include the new biomarker before it was added to the study, the biomarker status regarding this biomarker has to be determined retrospectively for patients that are included in the study before adding the new biomarker. This may lead to missing data for some or all of the patients. For cases where data is only partially missing, different methods for missing data imputation for models with interaction terms are investigated and compared.

The first issue of three issues which are addressed in this thesis is low prevalence of the biomarkers. For a study which tests an overall biomarker-guided treatment strategy, the sample size calculation method by Palta and Amini appears to be the most appropriate choice when heterogeneous treatment effects are expected. The results from the simulation study suggest that the subsequent data analysis could be performed using the two-step approach suggested by Mehrotra or a shared frailty model. If no other covariates are included in the model, an exact log-rank test could also be used. The asymptotic log-rank test and the stratified Cox PH model suffers loss of power in the simulation study and therefore should not be used for heterogeneous treatment effects. To test the individual biomarker-groups as secondary hypotheses after testing the overall treatment strategy, some strategies for multiple testing are suggested.

The second issue that is addressed is a large expected number of biomarker-negative patients at the screening stage. For a situation where an overall biomarker-guided treatment strategy is not desirable, a combined analysis model using the data from the entire study, including biomarker-negative patients, is investigated. This combined model estimates the treatment effects for the individual biomarkers. Application of the Firth correction appeared to be a good method for reduction of small sample size bias, which is likely to occur for low prevalence biomarkers. The inclusion of biomarker-negative patients in the model can provide a small additional benefit with respect to reduction of bias and standard deviation.

The third issue considered is the constant discovery of new biomarkers and corresponding biomarker-guided experimental therapies. It is desirable for a clinical trial to be able to react to these continuous developments by investigating options to add new biomarkers and corresponding therapies to an ongoing study. Different models for data analysis are compared for a situation with a belatedly added biomarker, an overlap of biomarkers within the population, and an effect of the new biomarker on the response to the experimental treatment of an already existing biomarker-group. Adding an interaction term to the combined analysis model can help avoiding biased treatment effect estimates when there is overlap of the biomarkers within the patient population, and when patients with both biomarkers respond differently to the experimental therapy than patients with only one of the biomarkers. If there is missing data regarding the biomarker status of the belatedly added biomarker, data imputation can be utilized. However, the correct model specification is crucial to avoid biased estimates when interaction terms are part of the model for the final analysis. These interaction terms should already be included in the imputation model rather than imputing them passively. The simulation study suggests that for the considered scenario, the 'just-another-variable'-approach with polytomous logistic regression is the best option to avoid obtaining biased estimates after data imputation.

Due to the heterogeneity of biomarkers and treatments and the rapid developments in this field, the planning phase of a multiple-biomarker trial is a complex process and each trial has to be adjusted to the individual situation. This thesis can give guidance in some of the aspects that need to be considered, but of course there are many more aspects that need to be addressed.

References

Aalen O (1978) Nonparametric inference for a family of counting processes. Ann Appl Stat 701–726

Aalen OO, Cook RJ, Røysland K (2015) Does Cox analysis of a randomized survival study yield a causal treatment effect? Lifetime Data Anal 21: 579–593

Agresti A (2007) An introduction to categorical data analysis. Wiley Series in Probability and Statistics. Wiley and Sons, New York

An MW, Mandrekar SJ, Sargent DJ (2012) A 2-stage phase II design with direct assignment option in stage II for initial marker validation. Clin Cancer Res 18: 4225–4233

Beisel C, Benner A, Kunz C, Kopp-Schneider A (2017) Heterogeneous treatment effects in stratified clinical trials with time-to-event endpoints. Biom J 59: 511–530

Bender R, Augustin T, Blettner M (2005) Generating survival times to simulate Cox proportional hazards models. Stat Med 24: 1713–1723

Bodner TE (2008) What improves with increased missing data imputations? Struct Equ Modeling 15: 651–675

Brannath W, König F, Bauer P (2006) Estimation in flexible two stage designs. Stat Med 25: 3366–3381

Bretz F, Maurer W, Brannath W, Posch M (2009) A graphical approach to sequentially rejective multiple test procedures. Stat Med 28: 586–604

Bretz F, Posch M, Glimm E, Klinglmueller F, Maurer W, Rohmeyer K (2011) Graphical approaches for multiple comparison procedures using weighted Bonferroni, Simes, or parametric tests. Biom J 53: 894–913

Cordeiro GM, McCullagh P (1991) Bias correction in generalized linear models. J R Stat Soc Series B Stat Methodol 53: 629–643

Cox DR (1972) Regression models and life-tables. J R Stat Soc Series B Stat Methodol 34: 187–220

Dmitrienko A, Tamhane AC, Bretz F (2009) Multiple testing problems in pharmaceutical statistics. CRC Press, Boca Raton, FL

Do K, O'Sullivan Coyne G, Chen AP (2015) An overview of the NCI precision medicine trial - NCI MATCH and MPACT. Chin Clin Oncol 4: 31

Downing G (2001) Biomarkers Definitions Working Group. Biomarkers and Surrogate Endpoints. Clin Pharmacol Ther 69: 89–95

Duchateau L, Janssen P (2007) The frailty model. Springer Springer & Business Media, New York, NY

Elgmati E, Fiaccone RL, Henderson R, Matthews JN (2015) Penalised logistic regression and dynamic prediction for discrete-time recurrent event data. Lifetime Data Anal 21: 542–560

Fergusson D, Aaron SD, Guyatt G, Hébert P (2002) Post-randomisation exclusions: the intention to treat principle and excluding patients from analysis. BMJ 325: 652–654

Ferrarotto R, Redman MW, Gandara DR, Herbst RS, Papadimitrakopoulou V (2015) Lung-MAP - framework, overview, and design principles. Chin Clin Oncol 4: 36

Firth D (1993) Bias reduction of maximum likelihood estimates. Biometrika 80: 27–38

Gerber DE, Oxnard GR, Govindan R (2015) ALCHEMIST: Bringing genomic discovery and targeted therapies to early-stage lung cancer. Clin Pharmacol Ther 97: 447–450

Gosho M, Nagashima K, Sato Y (2012) Study designs and statistical analyses for biomarker research. Sensors 12: 8966–8968

Habermehl C, Benner A, Kopp-Schneider A (2017) Addressing small sample size bias in multiple-biomarker trials: Inclusion of biomarker negative patients and Firth correction. Biom J Advance online publication: https://doi.org/10.1002/bimj.201600226

Heinze G, Schemper M (2001) A solution to the problem of monotone likelihood in Cox regression. Biometrics 57: 114–119

Kalbfleisch JD, Prentice RL (1980) The statistical analysis of failure time data. Wiley series in probability and mathematical statistics. Wiley and Sons, New York

Kang BP, Slosberg E, Snodgrass S, Lebedinsky C, Berry DA, Corless CL, Stein S, Salvado A (2015) The signature program: bringing the protocol to the patient. Clin Pharmacol Ther 98: 124–126

Kaplan R, Maughan T, Crook A, Fisher D, Wilson R, Brown L, Parmar M (2013) Evaluating many treatments and biomarkers in oncology: A new design. J Clin Oncol 31: 4562–4568

Kleinbaum DG, Klein M (1996) Survival Analysis. Springer, New York

Lachin JM (1981) Introduction to sample size determination and power analysis for clinical trials. Control Clin Trials 2: 93–113

Lachin JM, Foulkes MA (1986) Evaluation of sample size and power for analyses of survival with allowance for nonuniform patient entry, losses to follow-up, noncompliance, and stratification. Biometrics 42: 507–519

Langner I, Bender R, Lenz-Tönjes R, Küchenhoff H, Blettner M (2003) Bias of maximum-likelihood estimates in logistic and Cox regression models: A comparative simulation study. Sonderforschungsbereich 386, Paper 362, https://www.statistik.lmu.de/sfb386/papers/dsp/paper362.pdf

Le Tourneau C, Delord JP, Gonçalves A, Gavoille C, Dubot C, Isambert N, Campone M, Trédan O, Massiani MA, Mauborgne C, Armanet S (2015) Molecularly targeted therapy based on tumour molecular profiling versus conventional therapy for advanced cancer (SHIVA): a multicentre, open-label, proof-of-concept, randomised, controlled phase 2 trial. Lancet Oncol 16: 1324–1334

Leuchs AK, Neuhäuser M (2013) A modified combination test for the analysis of a clinical trial when a protocol amendment changed the inclusion criteria. J Stat Comput Simul 83: 825–836

Lin IF, Chang WP, Liao YN (2013) Shrinkage methods enhanced the accuracy of parameter estimation using Cox models with small number of events. J Clin Epidemiol 66: 743–751

Liu S, Lee JJ (2015) An overview of the design and conduct of the BATTLE trials. Chin Clin Oncol 4: 33

Lopez-Chavez A, Thomas A, Rajan A, Raffeld M, Morrow B, Kelly R, Carter CA, Guha U, Killian K, Lau CC, Abdullaev Z (2015) Molecular profiling and targeted therapy for advanced thoracic malignancies: a biomarker-derived, multiarm, multihistology phase II basket trial. J Clin Oncol 33: 1000–1007

Lösch C, Neuhäuser M (2008) The statistical analysis of a clinical trial when a protocol amendment changed the inclusion criteria. BMC Med Res Methodol 8: 16

Mandrekar SJ, Sargent DJ (2009a) Clinical trial designs for predictive biomakrer validation: One size does not fit all. J Biopharm Stat 19: 530–542

Mandrekar SJ, Sargent DJ (2009b) Clinical trial designs for predictive biomarker validation: Theoretical considerations and practical challenges. J Clin Oncol 27: 4027–4034

Mandrekar SJ, Sargent DJ (2010) Randomized phase II trials: Time for a new era in clinical trial design. J Thorac Oncol 5: 932–934

Mandrekar SJ, Dahlberg SE, Simon R (2015) Improving clinical trial efficiency: thinking outside the box. Am Soc Clin Oncol Educ Book e141–e147

Mehrotra D, Su SC, Li X (2012) An efficient alternative to the stratified Cox model analysis. Stat Med 31: 1849–1856

Mehta CR, Nitin P, Senchaudhri P (1992) Exact stratified linear rank tests for ordered categorical and binary data. J Comput Graph Stat 1: 21–40

Millen BA, Dmitrienko A (2012) Chain procedures: A class of flexible closed testing procedures with clinical trial applications. Stat Biopharm Res 3: 14–30

Mullard A (2015) NCI-MATCH trial pushes cancer umbrella trial paradigm. Nat Rev Drug Discovery 14: 513–515

Palta M, Amini SB (1985) Consideration of covariates and stratification in sample size determination for survival time studies. J Chronic Dis 38: 801–809

Park JW, Liu MC, Yee D, Yau C, van't Veer LJ, Symmans WF, Paoloni M, Perlmutter J, Hylton NM, Hogarth M, DeMichele A (2016) Adaptive randomization of neratinib in early breast cancer. N Engl J Med 375: 11–22

Persson I, Khamis H (2005) Bias of the Cox model hazard ratio. J Mod Appl Stat Methods 4: 90–99

Rao CR (1948) Large sample tests of statistical hypotheses concerning several parameters with applications to problems of estimation. Math Proc Cambridge Philos Soc 44: 50–57

Renfro LA, Sargent DJ (2016) Statistical controversies in clinical research: basket trials, umbrella trials, and other master protocols: a review and examples. Ann Oncol 28: 34–43

Rohmeyer K, Klinglmueller F (2015) gMCP: Graph based multiple test procedures R package version 0.8-10. URL http://CRAN.R-project.org/package=gMCP

Rubin DB (1976) Inference and missing data. Biometrika 63: 581–592

Rubin DB (1987) Multiple Imputation for Nonresponse in Surveys. Wiley Series in Probability and Statistics. Wiley & Sons, New York

Schemper M (1992) Cox analysis of survival data with non-proportional hazard functions. Statistician 41: 455–465

Schlenk R, Froehling S, Hartmann F, Fischer J, Glasmacher A, del Valle F, Grimminger W, Goetze K, Waterhouse C, Schoch R, Pralle H (2004) Phase III study of all trans retinoic acid in previously untreated patients 61 years or older with acute myeloid leukemia. Leukemia 18: 1798–1803

Schlenk R, Lübbert M, Benner A, Lamparter A, Krauter J, Herr W, Martin H, Salih HR, Kündgen A, Horst HA, Brossart P (2016) All-trans retinoic acid as adjunct to intensive treatment in younger adult patients with acute myeloid leukemia: results of the randomized AMLSG 07-04 study. Ann Hematol 95: 1931–42

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

Seaman SR, Bartlett JW, White IR (2012) Multiple imputation of missing covariates with non-linear effects and interactions: an evaluation of statistical methods. BMC Med Res Methodol 12: 46

Strasser H, Weber C (1999) On the asymptotic theory of permutation statistics. Math Meth Stat 8: 220–250

Strimbu K, Tavel J (2010) What are biomarkers. Curr Opin HIV AIDS 5: 463

Tassara M, Döhner K, Brossart P, Held G, Götze K, Horst HA, Ringhoffer M, Köhne CH, Kremers S, Raghavachar A, Wulf G (2014) Alproic acid in combination with all-trans retinoic acid and intensive therapy for acute myeloid leukemia in older patients. Blood 123: 4027–4036

Therneau T, Grambsch P (2000) Modeling Survival Data: Extending the Cox Model. Springer & Business Media, New York, NY

Van Buuren S (2012) Flexible imputation of missing data. CRC press, Boca Raton, FL

Von Hippel PT (2009) How to impute interactions, squares, and other transformed variables. Sociol Methodol 39: 265–291

White IR, Royston P (2009) Imputing missing covariate values for the Cox model. Stat Med 28: 1982–1998

White IR, Royston P, Wood AM (2011) Multiple imputation using chained equations: issues and guidance for practice. Stat Med 30: 377–399

Wienke A (2010) Frailty models in survival analysis. CRC Press, Boca Raton, FL

Own Publications Related to this Dissertation

1. **Beisel[1] C** , Benner A, Kunz C and Kopp-Schneider A (2017) Heterogeneous treatment effects in stratified clinical trials. Biometrical Journal 59(3):511-530.

2. **Habermehl C** , Benner A and Kopp-Schneider A (2017) Addressing small sample size bias in multiple-biomarker trials: Inclusion of biomarker negative patients and Firth correction. Biometrical Journal. Advance online publication: https://doi.org/10.1002/bimj.201600226.

Further Own Publications

3. **Beisel[1] CB** (2014) Iterated conditional modes algorithm for medical image denoising. WESTERN ILLINOIS UNIVERSITY. (Master Thesis)

[1] Last name changed from Beisel to Habermehl in 2017

Chapter 6

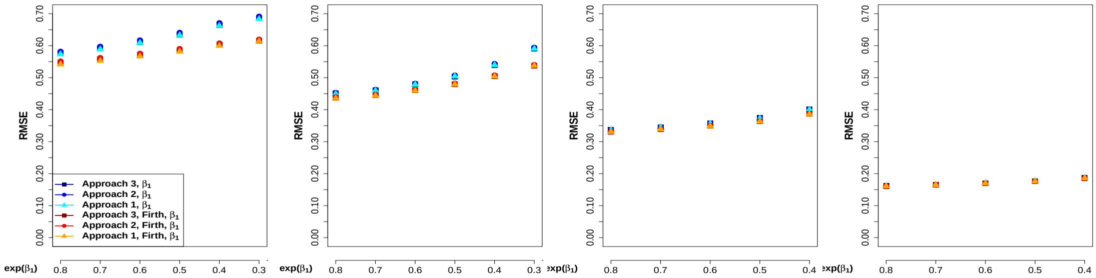
Appendix

A.1. Supplementary figures

Figures for Section 3.3

A simulation study was conducted to compare the performance of the three approaches discussed in Section 3.3.1 with respect to bias, standard deviation and RMSE of the different parameter estimates. It was investigated whether there is a benefit of including biomarker-negative patients in the study. Additionally it was investigated, whether application of the Firth correction can further reduce the bias.

Figures A.1 and A.2 show the RMSE of $\hat{\beta}_1$ and $\hat{\beta}_2$ respectively. Figure A.3 shows bias, standard deviation, and RMSE of $\hat{\gamma}_1$ and $\hat{\gamma}_2$. Figure A.4 shows the behavior of the standard deviation of $\hat{\beta}_1$ and $\hat{\beta}_2$ for different baseline hazards.



$n = 100$ $n = 150$ $n = 250$ $n = 1000$

Figure A.1: The RMSE of log hazard ratio β_1, the treatment effect estimate for the biomarker indicative of a favorable prognosis (B_1), without and with Firth correction for sample sizes 100, 150, 250, and 1,000.

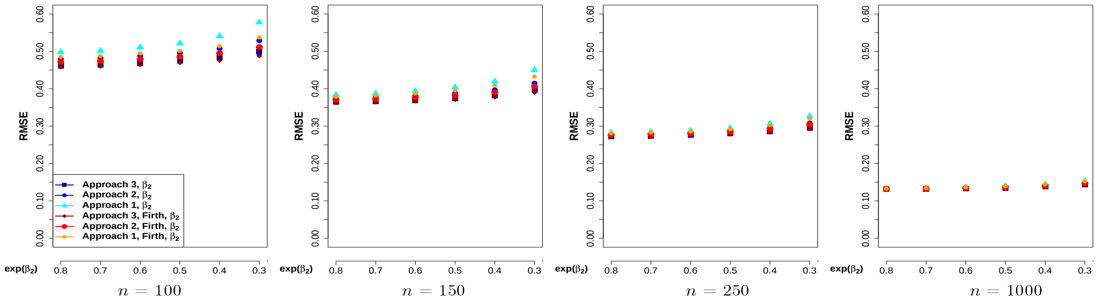


Figure A.2: The RMSE of log hazard ratio β_2 without and with Firth correction for sample sizes 100, 150, 250, and 1,000.

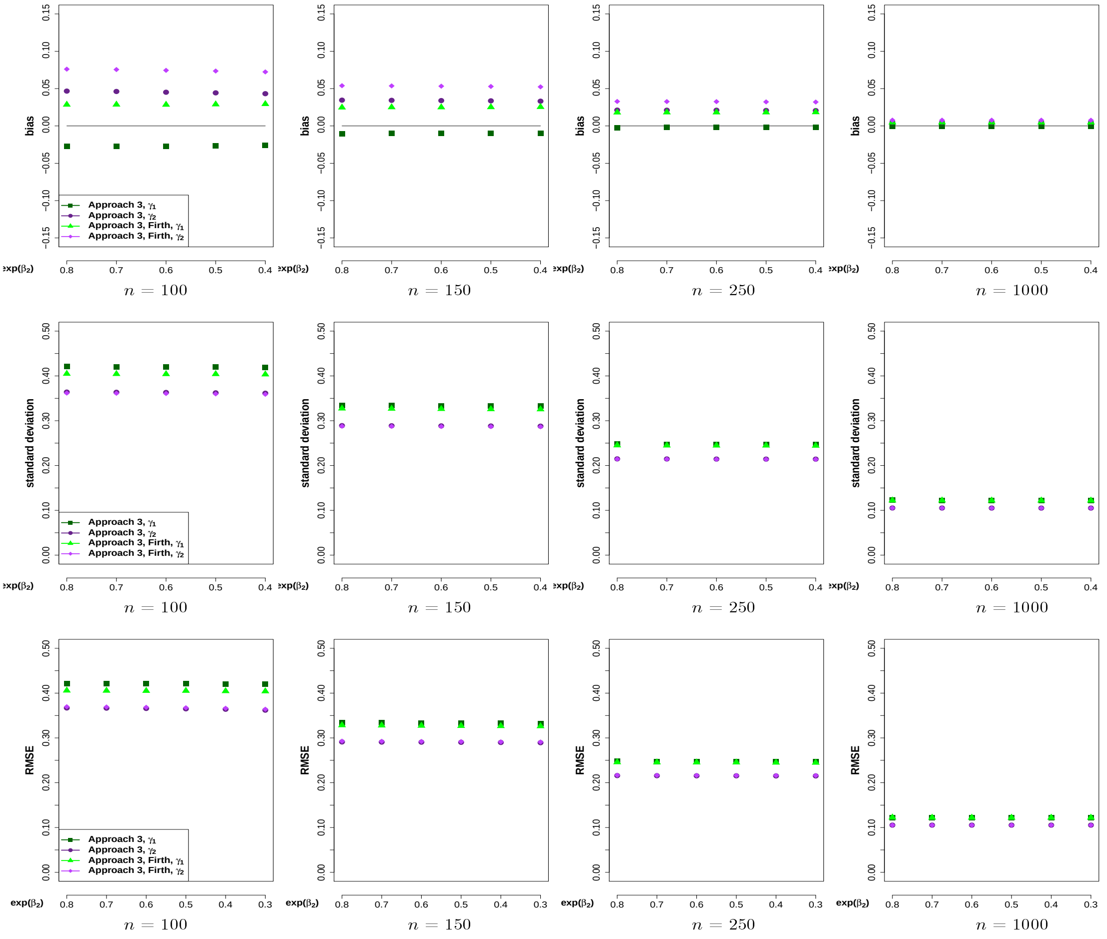
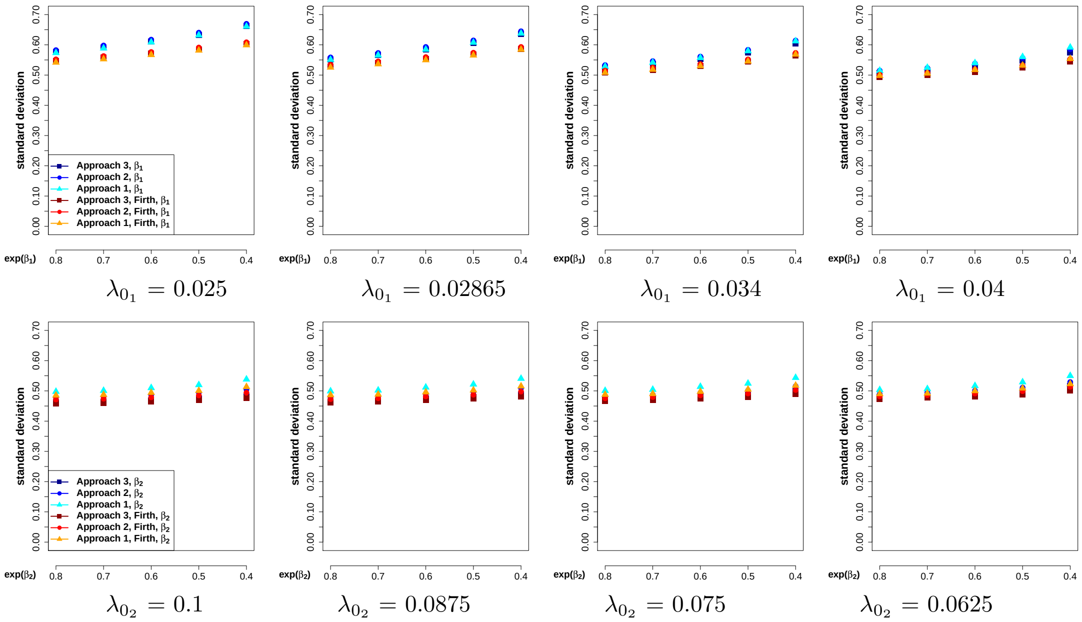


Figure A.3: Bias and standard deviation of the estimates of γ_1 and γ_2, the prognostic effects for biomarkers 1 and 2, respectively, using a Cox model without and with Firth correction for Approach 3 and sample sizes 100, 150, 250, and 1,000.



$\lambda_{0_1} = 0.025$ $\lambda_{0_1} = 0.02865$ $\lambda_{0_1} = 0.034$ $\lambda_{0_1} = 0.04$

$\lambda_{0_2} = 0.1$ $\lambda_{0_2} = 0.0875$ $\lambda_{0_2} = 0.075$ $\lambda_{0_2} = 0.0625$

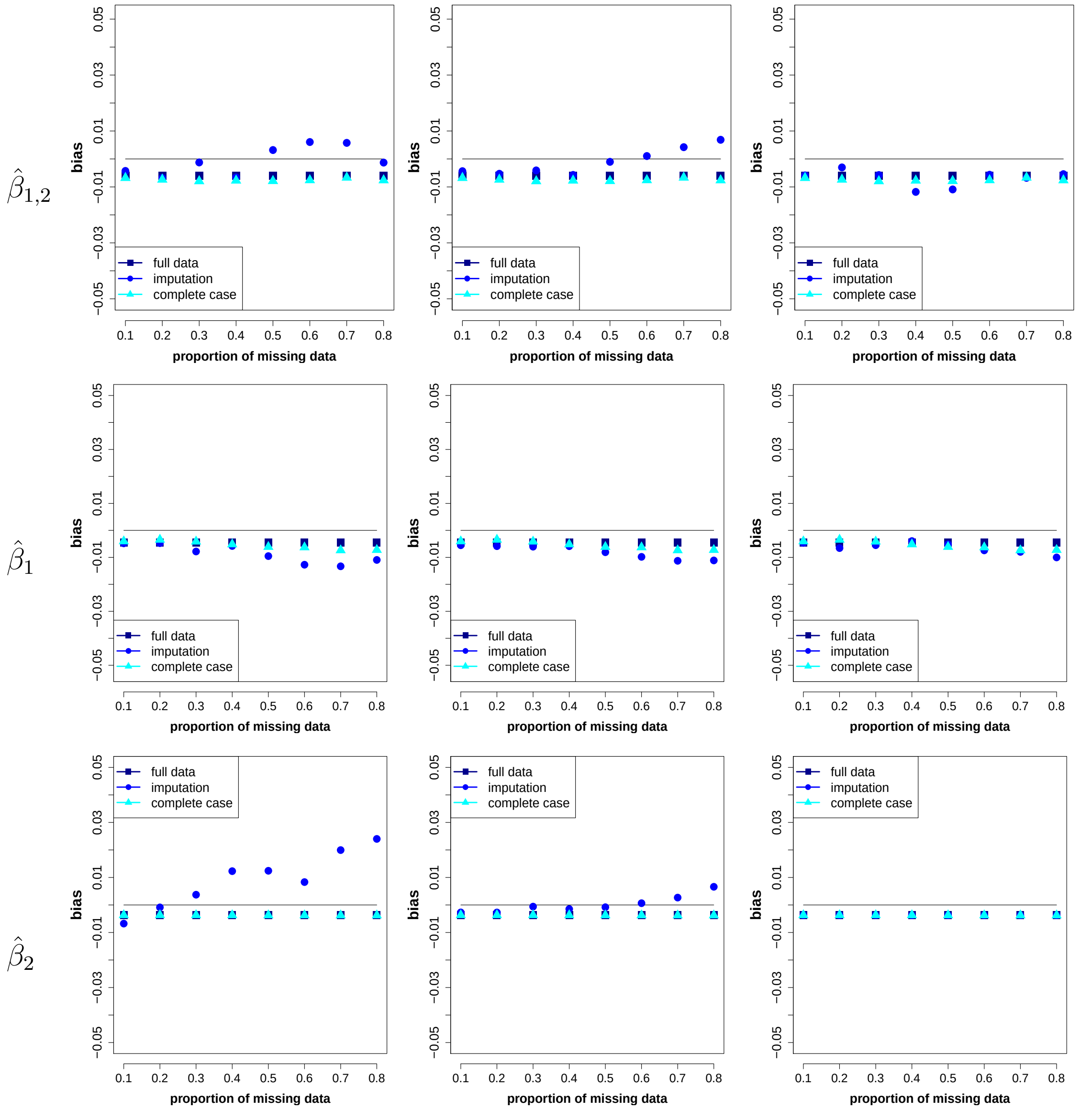
Figure A.4: Comparison of standard deviation for different baseline hazards for $n = 100$: Standard deviation of the estimates of log hazard ratio β_1 and log hazard ratio β_2, using a Cox model without and with Firth correction given baseline hazards λ_{0_1}: 0.025, 0.02865, 0.034, 0.04 and λ_{0_2}: 0.1, 0.0875, 0.075, 0.0625, respectively. The baseline hazards for B$_1$ correspond to multiplying λ_{0_0} with $\exp(\gamma_1)$, where $\gamma_1 = \log(\frac{1}{2})$, $\gamma_1 = \log(4/7)$, $\gamma_1 = \log(2/3)$, and $\gamma_1 = \log(4/5)$, respectively. The baseline hazards for B$_2$ correspond to multiplying $\lambda_{0_0} = 0.05$ with $\exp(\gamma_2)$, where $\gamma_2 = \log(2)$, $\gamma_2 = \log(7/4)$, $\gamma_2 = \log(3/2)$, and $\gamma_2 = \log(5/4)$, respectively.

For all Figures, bias, standard deviation, and RMSE were calculated from 10,000 Simulations runs. Fixed parameters: $\exp(\beta_2) = 0.8$, $\lambda_{0_0} = 0.05$, $\gamma_1 = \log(0.5)$ and $\gamma_2 = \log(2)$. The patient proportions in biomarker groups B$_0$, B$_1$, and B$_2$ are 0.5, 0.25, and 0.25, respectively, and treatment allocation ratio between treatments for B$_1$, and B$_2$ is 1:1.

Figures for Section 3.4

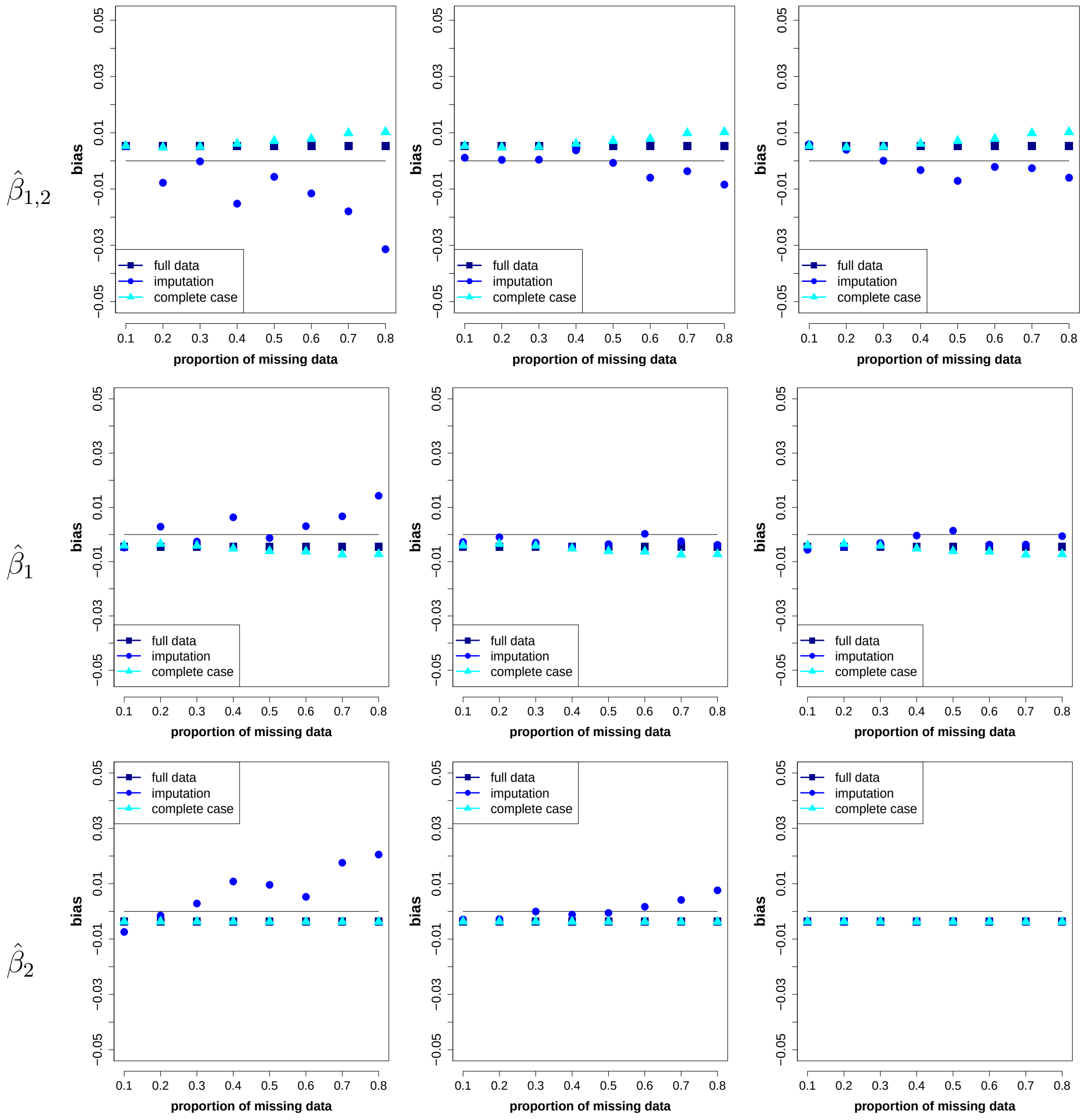
The bias of the estimates of β_1, β_2, and $\beta_{1,2}$ when using PI, the JAV-approach or the stratify-approach was compared to the bias resulting from CC analysis and from the analysis of the full data set. Three imputation methods to predict the missing values were used and compared for the three imputation approaches: logistic regression (LR), polytomous logistic regression (PLR), and predictive mean matching (PMM). Unless otherwise indicated, missing data are MCAR.

It was investigated whether the results of the simulation study are sensitive to a change of parameters, e.g. for a different effect τ, i.e. the factor by which the hazard ratio for B$_2$-patients treated with Exp$_1$ differs from the hazard ratio for B$_1$-patients treated with Exp$_1$. Figures A.5 and A.6 show the results for $\tau = 0.75$ and $\tau = 2$, respectively.



(a) Logistic regression *(b) Polytomous logistic regression* *(c) Predictive mean matching*

Figure A.5: Effect τ=0.75: Bias for LR, PLR, and PMM for $\hat{\beta}_{1,2}$, $\hat{\beta}_1$, and $\hat{\beta}_2$ treating interactions as JAV in imputation model. B$_1$ is the prioritized biomarker in the allocation algorithm, data is MCAR, and m=10 imputations.



(a) *Logistic regression*   (b) *Polytomous logistic regression*   (c) *Predictive mean matching*

Figure A.6: Effect $\tau = 2$: Bias for LR, PLR, and PMM for $\hat{\beta}_{1,2}$, $\hat{\beta}_1$, and $\hat{\beta}_2$ treating interactions as JAV in imputation model. B_1 is the prioritized biomarker in the allocation algorithm, data is MCAR, and m=10 imputations.

A.2. Supplementary tables

Tables for Section 3.2

The formulas for sample size calculation by Schoenfeld, Palta and Amini, and Lachin and Foulkes were compared (Table A.1). For each hazard ratio constellation of the simulation study, the required sample size for a power of 0.8 is determined with each of the sample size formulas.

Afterwards, the different analysis methods were evaluated with respect to the power to detect a significant overall treatment effect, given specific hazard ratio scenarios. The numerical results for the sample size obtained from the formulas by Schoenfeld and Lachin and Foulkes are shown in Tables A.2-A.4.

For all following tables, the hazard ratios of B_1 and B_2 are varied between 0.8 and 0.4, and 0.8 and 0.3, respectively; the hazard ratio for B_0 is held constant at 0.8. 10,000 simulations.

Table A.1: Comparison of sample sizes calculated from formulas by Schoenfeld, Palta and Amini, and Lachin and Foulkes for different scenarios.

HR	**Sample size**		
B_0, B_1, B_2	Schoenfeld	Palta Amini	Lachin Foulkes
0.8 0.8 0.8	831	763	686
0.8 0.8 0.7	652	574	540
0.8 0.8 0.6	495	437	417
0.8 0.8 0.5	365	336	316
0.8 0.8 0.4	263	259	236
0.8 0.8 0.3	185	200	172
0.8 0.7 0.7	492	461	404
0.8 0.7 0.6	390	360	325
0.8 0.7 0.5	300	282	256
0.8 0.7 0.4	224	222	197
0.8 0.7 0.3	163	174	149
0.8 0.6 0.6	305	297	249
0.8 0.6 0.5	243	238	204
0.8 0.6 0.4	188	190	163
0.8 0.6 0.3	141	152	127
0.8 0.5 0.5	195	200	160
0.8 0.5 0.4	156	163	132
0.8 0.5 0.3	121	131	106
0.8 0.4 0.4	128	138	105
0.8 0.4 0.3	102	113	87

Table A.2: Numerical results for the power comparison of the stratified Cox model, Mehrotra's two-step approach, the lognormal shared frailty model, and stratified exact and asymptotic log-rank test to detect a true treatment effect when using Schoenfeld's sample size formula under different scenarios.

Input		**Power**				
HR	Sample	Strat.Cox	Twostep	Frailty	Exact	Asympt.
B_0, B_1, B_2	Size	Wald	Wald	lognorm	log-rank	log-rank
0.8 0.8 0.8	831	0.833	0.833	0.832	0.828	0.833
0.8 0.8 0.7	652	0.841	0.839	0.841	0.831	0.842
0.8 0.8 0.6	495	0.842	0.843	0.842	0.834	0.842
0.8 0.8 0.5	365	0.818	0.830	0.826	0.826	0.819
0.8 0.8 0.4	263	0.781	0.812	0.808	0.809	0.783
0.8 0.8 0.3	185	0.722	0.789	0.773	0.787	0.725
0.8 0.7 0.7	492	0.817	0.823	0.822	0.822	0.817
0.8 0.7 0.6	390	0.824	0.830	0.829	0.828	0.825
0.8 0.7 0.5	300	0.815	0.834	0.827	0.828	0.816
0.8 0.7 0.4	224	0.774	0.807	0.798	0.806	0.776
0.8 0.7 0.3	163	0.725	0.791	0.774	0.790	0.727
0.8 0.6 0.6	305	0.809	0.824	0.817	0.823	0.811
0.8 0.6 0.5	243	0.795	0.816	0.810	0.814	0.796
0.8 0.6 0.4	188	0.767	0.804	0.798	0.804	0.769
0.8 0.6 0.3	141	0.730	0.795	0.778	0.796	0.735
0.8 0.5 0.5	195	0.770	0.805	0.792	0.809	0.772
0.8 0.5 0.4	156	0.750	0.795	0.780	0.796	0.753
0.8 0.5 0.3	121	0.713	0.778	0.765	0.785	0.717
0.8 0.4 0.4	128	0.723	0.787	0.762	0.790	0.727
0.8 0.4 0.3	102	0.689	0.764	0.747	0.773	0.695

Table A.3: Numerical results for the power comparison of the stratified Cox model, Mehrotra's two-step approach, the lognormal shared frailty model, and stratified exact and asymptotic log-rank test to detect a true treatment effect when using Lachin and Foulkes' sample size formula under different scenarios.

Input		**Power**				
HR	Sample	Strat.Cox	Twostep	Frailty	Exact	Asympt.
B_0, B_1, B_2	Size	Wald	Wald	lognorm	log-rank	log-rank
0.8 0.8 0.8	686	0.760	0.761	0.762	0.759	0.761
0.8 0.8 0.7	540	0.768	0.767	0.769	0.760	0.769
0.8 0.8 0.6	417	0.775	0.777	0.778	0.767	0.776
0.8 0.8 0.5	316	0.761	0.774	0.773	0.767	0.763
0.8 0.8 0.4	236	0.734	0.767	0.759	0.764	0.736
0.8 0.8 0.3	172	0.680	0.751	0.733	0.751	0.683
0.8 0.7 0.7	404	0.735	0.744	0.741	0.740	0.737
0.8 0.7 0.6	325	0.751	0.760	0.756	0.754	0.752
0.8 0.7 0.5	256	0.742	0.759	0.754	0.754	0.744
0.8 0.7 0.4	197	0.726	0.762	0.757	0.763	0.728
0.8 0.7 0.3	149	0.681	0.748	0.731	0.748	0.684
0.8 0.6 0.6	249	0.717	0.735	0.729	0.733	0.719
0.8 0.6 0.5	204	0.712	0.739	0.733	0.739	0.714
0.8 0.6 0.4	163	0.700	0.741	0.734	0.744	0.703
0.8 0.6 0.3	127	0.675	0.741	0.727	0.740	0.678
0.8 0.5 0.5	160	0.677	0.717	0.707	0.720	0.680
0.8 0.5 0.4	132	0.670	0.726	0.713	0.725	0.674
0.8 0.5 0.3	106	0.649	0.725	0.710	0.731	0.655
0.8 0.4 0.4	105	0.629	0.698	0.679	0.700	0.635
0.8 0.4 0.3	87	0.627	0.695	0.691	0.714	0.634

Table A.4: Numerical results for the power comparison of the stratified Cox model, Mehrotra's two-step approach, the lognormal shared frailty model, and stratified exact and asymptotic log-rank test to detect a true treatment effect when using Lachin and Foulkes' sample size formula (with sample proportion weights) under different scenarios.

Input		**Power**				
HR	Sample	Strat.Cox	Twostep	Frailty	Exact	Asympt.
B_0, B_1, B_2	Size	Wald	Wald	lognorm	log-rank	log-rank
0.8 0.8 0.8	632	0.721	0.719	0.720	0.714	0.722
0.8 0.8 0.7	469	0.711	0.712	0.713	0.703	0.712
0.8 0.8 0.6	360	0.715	0.718	0.716	0.706	0.716
0.8 0.8 0.5	282	0.701	0.717	0.715	0.710	0.703
0.8 0.8 0.4	226	0.705	0.739	0.733	0.740	0.707
0.8 0.8 0.3	184	0.722	0.786	0.770	0.788	0.724
0.8 0.7 0.7	381	0.716	0.724	0.724	0.725	0.718
0.8 0.7 0.6	298	0.712	0.720	0.719	0.711	0.713
0.8 0.7 0.5	238	0.706	0.723	0.723	0.720	0.708
0.8 0.7 0.4	193	0.711	0.746	0.739	0.745	0.713
0.8 0.7 0.3	159	0.714	0.779	0.762	0.780	0.718
0.8 0.6 0.6	249	0.717	0.735	0.729	0.733	0.719
0.8 0.6 0.5	201	0.707	0.735	0.730	0.733	0.710
0.8 0.6 0.4	165	0.705	0.748	0.739	0.745	0.708
0.8 0.6 0.3	137	0.707	0.772	0.756	0.771	0.710
0.8 0.5 0.5	171	0.706	0.743	0.733	0.746	0.709
0.8 0.5 0.4	142	0.708	0.760	0.746	0.761	0.712
0.8 0.5 0.3	119	0.703	0.772	0.757	0.779	0.708
0.8 0.4 0.4	122	0.699	0.765	0.740	0.768	0.703
0.8 0.4 0.3	103	0.695	0.769	0.752	0.782	0.701

Tables for Section 3.3

Simulations were run to investigate the robustness of the estimators for the approaches in Section 3.3 against violations of model assumptions and change of parameters. The bias and standard deviation were investigated for different biomarker prevalences and for a non-constant (Weibull distributed) hazard function, with Weibull shape parameters 0.4, 1 and 5. For better comparability, λ_0 was adjusted such that the same number of events is reached at 60 months for all shape parameters, i.e. $\lambda_0 = 0.05 \cdot 60/60^{\text{shape}}$. Tables A.5 and A.6 show the type I error rates for these scenarios for all three approaches for β_1 and β_2 without and with Firth correction.

Table A.5: Type I error for the different approaches and different scenarios. Simulations: 10,000. Sample size: 10,000.

Approach 1		**Approach 2**		**Approach 3**		**Mean no.**	**Scenario**
β_1	β_2	β_1	β_2	β_1	β_2	events	
0.050	0.048	0.050	0.048	0.050	0.048	8371	
0.048	0.049	0.048	0.048	0.048	0.048	8176	different biomarker prevalences
0.052	0.050	0.052	0.051	0.052	0.050	9189	Weibull shape 0.4 (no censoring)
0.050	0.048	0.050	0.048	0.050	0.048	9189	Weibull shape 1 (no censoring)
0.052	0.049	0.051	0.049	0.052	0.049	9189	Weibull shape 5 (no censoring)
0.052	0.050	0.052	0.051	0.052	0.050	7848	Weibull shape 0.4 (with censoring)
0.052	0.049	0.051	0.049	0.052	0.049	6053	Weibull shape 5 (with censoring)

Table A.6: Type I error for the different approaches and different scenarios with Firth correction. Simulations: 10,000. Sample size: 10,000.

Approach 1		**Approach 2**		**Approach 3**		**Mean no.**	**Scenario**
β_1	β_2	β_1	β_2	β_1	β_2	events	
0.050	0.048	0.050	0.048	0.050	0.048	8371	
0.048	0.049	0.048	0.048	0.048	0.048	8176	different biomarker prevalences
0.052	0.050	0.052	0.051	0.052	0.050	9189	Weibull shape 0.4 (no censoring)
0.050	0.048	0.050	0.048	0.050	0.048	9189	Weibull shape 1 (no censoring)
0.052	0.049	0.051	0.049	0.052	0.049	9189	Weibull shape 5 (no censoring)
0.052	0.050	0.052	0.051	0.052	0.050	7848	Weibull shape 0.4 (with censoring)
0.052	0.049	0.051	0.049	0.052	0.049	6053	Weibull shape 5 (with censoring)

A.3. Selected R-Code

Sample size calculation for 3.2.

```
########################################################################################
### estimate survival probabilities ###

t <- (accrual / 2) + followup
study_length <- accrual + followup

surv_prob_std <- sum((bio_prop * (1-rand.prob) / p_std) * exp(-lambda0 * t))
surv_prob_exp <- sum((bio_prop *    rand.prob  / p_exp) * exp(-lambda1 * t))

########################################################################################
### estimate average hazard ratio and hazard rates ###

### for Z=0 ###
lambda_std <- -log(surv_prob_std) / t

### for Z=1 ###

HR_avg <- log(surv_prob_exp) / log(surv_prob_std)
lambda_exp <- HR_avg * lambda_std

lambda_star_std <- lambda_cens + lambda_std
lambda_star_exp <- lambda_cens + lambda_exp

eventrate_std <- (lambda_std/lambda_star_std) * (1 + (1/(accrual * lambda_star_std)) *
                  exp(-lambda_star_std * (study_length)) -
                  (1/(accrual * lambda_star_std)) * exp(-lambda_star_std * followup))

eventrate_exp <- (lambda_exp/lambda_star_exp) * (1 + (1/(accrual * lambda_star_exp)) *
                  exp(-lambda_star_exp * (study_length)) -
                  (1/(accrual * lambda_star_exp))*exp(-lambda_star_exp*followup))

eventrate <- p_std * eventrate_std + p_exp * eventrate_exp

########################################################################################
### Schoenfeld ###

n_schoenfeld <- ceiling((qnorm(1 - beta) + qnorm(1 - alpha/2))^2 /
                        (p_std * p_exp * log(HR)^2 * eventrate))

########################################################################################
### Palta & Amini ###

V_s <- rand.prob * (1 - (exp(-lambda1 * followup) -
       exp(-lambda1 * study_length)) / (lambda1 * accrual)) +
       (1-rand.prob) * (1-(exp(-lambda0 * followup) -
       exp(-lambda0 * study_length)) / (lambda0 * accrual))

mu <- (sum(log(HR) * bio_prop * rand.prob * (1 - rand.prob) * V_s)) /
      (sqrt(sum(bio_prop * rand.prob) * (1 - rand.prob) * V_s)))

n_palta <- ceiling((qnorm(1 - beta) + qnorm(1 - alpha/2))^2 / mu^2)

########################################################################################
### Lachin & Foulkes ###

Phi <- function(lambda){
  phi <- lambda^2 * ((lambda + lambda_cens * (1-(1/(accrual * (lambda+lambda_cens))) *
               exp(-(lambda + lambda_cens) * followup) -
               exp(-(lambda + lambda_cens) * (accrual + followup)))) / lambda)
return(phi)
}

lambda_bar <- rand.prob * lambda1 + (1 - rand.prob) * lambda0
lambda0_bar <- sum(lambda0 * bio_prop)
lambda1_bar <- sum(lambda1 * bio_prop)
lambda_bar_mean <- sum(lambda_bar * bio_prop)

### initial sample size for sigma weights from Lachin formula ###

n_init <- ceiling(((qnorm(1-alpha/2) * sqrt(Phi(lambda_bar_mean) * ((1/rand.prob) +
                   (1/(1-rand.prob)))) + qnorm(1-beta) * sqrt(Phi(lambda1_bar) *
                   (1/rand.prob) + Phi(lambda0_bar) *  (1/(1-rand.prob)))) /
                   mean(lambda1-lambda0))^2)

Psi0 <- Phi(lambda_bar) * (1/rand.prob + 1/(1 - rand.prob))
Psi1 <- Phi(lambda1) / rand.prob + Phi(lambda0) / (1 - rand.prob)
Sigma <- sum((bio_prop * Psi1) / (Psi0^2))
Omega <- sum(bio_prop / Psi0)

###sigma weights###

sigma0 <- Psi0 / (n_init * bio_prop)
w <- (1/sigma0) / ((1/sigma0[1]) + (1/sigma0[2]) + (1/sigma0[3]))

lambda_diff <- sum(w * (lambda1 - lambda0))

n_lachinfoulkes <- ceiling(((qnorm(1-alpha/2) * sqrt(Omega^(-1)) + qnorm(1-beta) *
                             sqrt(Omega^(-2) * Sigma)) / lambda_diff)^2)

########################################################################################
###Lachin & Foulkes with sample size weigths###

w_ssize <- bio_prop
lambda_diff_ssize <- sum(w_ssize * (lambda1 - lambda0))

n_lachinfoulkes_ssize <- ceiling(((qnorm(1 - alpha/2) * sqrt(Omega^(-1)) +
                                   qnorm(1 - beta) * sqrt(Omega^(-2) * Sigma)) /
                                   lambda_diff_ssize)^2)
```

Data generation for Sections 3.2 and 3.3.

```
data <- data.frame(patientID=rep(NA, n), adm_cens_time=rep(NA, n),
                   biomarker_status=rep(NA, n), biomarker=rep(NA, n),
                   survival_time=rep(NA, n), cens_time=rep(NA, n), group=rep(NA, n),
                   arm=rep(NA, n), status=rep(NA, n), os=rep(NA, n))

### patients ###
data$patientID <- c(1:n)

########################################################################################
### Draw biomarker status for each patient ###

data$biomarker_status <- sample(c(0:2), n, replace = T, prob = bio_prop)

nB0 <- length(data$biomarker_status[data$biomarker_status == 0])
nB1 <- length(data$biomarker_status[data$biomarker_status == 1])
nB2 <- length(data$biomarker_status[data$biomarker_status == 2])

########################################################################################
```

```
### Randomize within arms ###

Z0 <- rbinom(nB0, 1, rand.prob[[1]])
Z1 <- rbinom(nB1, 1, rand.prob[[2]])
Z2 <- rbinom(nB2, 1, rand.prob[[3]])

data$biomarker[data$biomarker_status == 0] <- ’B0’
data$group[data$biomarker_status == 0]     <-  Z0
data$biomarker[data$biomarker_status == 1] <- ’B1’
data$group[data$biomarker_status == 1]     <-  Z1
data$biomarker[data$biomarker_status == 2] <- ’B2’
data$group[data$biomarker_status == 2]     <-  Z2

data$arm[data$biomarker_status == 0 & data$group == 0] <- ’B00’
data$arm[data$biomarker_status == 0 & data$group == 1] <- ’B01’
data$arm[data$biomarker_status == 1 & data$group == 1] <- ’B11’
data$arm[data$biomarker_status == 2 & data$group == 0] <- ’B20’
data$arm[data$biomarker_status == 2 & data$group == 1] <- ’B21’

########################################################################################
### Survival times ###

nB10 <- length(data$arm[data$arm == "B10"])
nB11 <- length(data$arm[data$arm == "B11"])
nB20 <- length(data$arm[data$arm == "B20"])
nB21 <- length(data$arm[data$arm == "B21"])

U1 <- runif(nB0,  min = 0, max = 1)
U2 <- runif(nB10, min = 0, max = 1)
U3 <- runif(nB11, min = 0, max = 1)
U4 <- runif(nB20, min = 0, max = 1)
U5 <- runif(nB21, min = 0, max = 1)


data$survival_time[data$arm == ’B00’] <- (-(log(U1)/(lambda0[[1]])))^(1/shape)
data$survival_time[data$arm == ’B01’] <- (-(log(U2)/(lambda1[[1]])))^(1/shape)
data$survival_time[data$arm == ’B10’] <- (-(log(U2)/(lambda0[[2]])))^(1/shape)
data$survival_time[data$arm == ’B11’] <- (-(log(U3)/(lambda1[[2]])))^(1/shape)
data$survival_time[data$arm == ’B20’] <- (-(log(U4)/(lambda0[[3]])))^(1/shape)
data$survival_time[data$arm == ’B21’] <- (-(log(U5)/(lambda1[[3]])))^(1/shape)

########################################################################################
### random censoring ###

if(p.rand.cens > 0 & p.rand.cens <= 1){
   lambda.cens    <- (p.rand.cens*sum(bio_prop*((lambda0+lambda1)/2)))/
                     (1-p.rand.cens)
   data$cens_time <- rexp(n, rate = lambda.cens)
} else{
   data$cens_time <- rep(Inf, n)
}

########################################################################################
### staggered entry/administrative censoring ###

if(accrual > 0){
data$adm_cens_time <- runif(n, min = followup, max = accrual+followup)
} else{
data$adm_cens_time <- rep(Inf, n)
}


########################################################################################
### overall survival ###
```

```
data$os <- pmin(data$survival_time, data$cens_time, data$adm_cens_time)

#######################################################################################
### status 1=dead 0=alive ###

data$status[data$survival_time <= data$os] <- 1
data$status[data$survival_time >  data$os] <- 0
```

Data analysis for Section 3.2

```
### Surv objects ###

Surv   <- Surv(data$os, data$status)
SurvB0 <- Surv(data$os[data$biomarker_status==0], data$status[data$biomarker_status==0])
SurvB1 <- Surv(data$os[data$biomarker_status==1], data$status[data$biomarker_status==1])
SurvB2 <- Surv(data$os[data$biomarker_status==2], data$status[data$biomarker_status==2])

#######################################################################################
### (Stratified) asymptotic log-rank test ###

diff   <- survdiff(Surv   ~ group + strata(biomarker_status), data=data)
diffB0 <- survdiff(SurvB0 ~ group[data$biomarker_status==0],  data=data)
diffB1 <- survdiff(SurvB1 ~ group[data$biomarker_status==1],  data=data)
diffB2 <- survdiff(SurvB2 ~ group[data$biomarker_status==2],  data=data)

#######################################################################################
### Log-rank pvalues ###

p.val   <- (1 - pchisq(diff$chisq,   length(diff$n) - 1))
p.valB0 <- (1 - pchisq(diffB0$chisq, length(diffB0$n) - 1))
p.valB1 <- (1 - pchisq(diffB1$chisq, length(diffB1$n) - 1))
p.valB2 <- (1 - pchisq(diffB2$chisq, length(diffB2$n) - 1))

nB0 <- length(data$biomarker_status[data$biomarker_status==0])
nB1 <- length(data$biomarker_status[data$biomarker_status==1])
nB2 <- length(data$biomarker_status[data$biomarker_status==2])
n   <- nB0 + nB1 + nB2

#######################################################################################
### Stratified exact/approximate log-rank test ###

diff.exactLR <- logrank_test(Surv ~ factor(group)|factor(biomarker_status),
                             data=data, distribution=approximate(B=10000))

p.val.exactLR <- pvalue(diff.exactLR)[1]

#######################################################################################
### Frailty coxph (Gamma) ###

frail.coxph       <- coxph(Surv ~ group + frailty.gamma(biomarker_status), data=data)

p.val.frail.coxph <- coef(summary(frail.coxph))[1,6]

#######################################################################################
### Frailty coxme (Gaussian) ###

frail.coxme       <- coxme(Surv ~ group + (1|biomarker_status), data=data)
chisq.frail.coxme <- (fixef(frail.coxme))^2 / vcov(frail.coxme)
p.val.frail.coxme <- 1 - pchisq(chisq.frail.coxme, 1)
```

```
#######################################################################################
### Stratified Cox ###

strat.cox       <- coxph(Surv ~ group + strata(biomarker_status), data=data)
p.val.strat.cox <- coef(summary(strat.cox))[,5]

#######################################################################################
### Two-step procedure ###

fitB0 <- coxph(SurvB0 ~ group[data$biomarker_status==0], data=data)
fitB1 <- coxph(SurvB1 ~ group[data$biomarker_status==1], data=data)
fitB2 <- coxph(SurvB2 ~ group[data$biomarker_status==2], data=data)

fit_vec <- c(unname(fitB0$coefficients), unname(fitB1$coefficients),
             unname(fitB2$coefficients))
HR_vec  <- exp(fit_vec)
var_vec <- c(fitB0$var, fitB1$var, fitB2$var)

### ssize weigths ###

weights_ssize <- c(nB0/n, nB1/n, nB2/n)

### test statistic ###

twostep_HR    <- exp(sum(fit_vec * weights_ssize))
twostep_coef  <- sum(fit_vec * weights_ssize)
twostep_var   <- sum(var_vec * weights_ssize^2)
twostep_wald  <- (twostep_coef)^2 / twostep_var
p.val.twostep <- 1 - pchisq(twostep_wald, 1)
```

Data analysis for Section 3.3

```
### Surv objects ###

Surv   <- Surv(data$os, data$status)
SurvB1 <- Surv(data$os[data$biomarker_status==1], data$status[data$biomarker_status==1])
SurvB2 <- Surv(data$os[data$biomarker_status==2], data$status[data$biomarker_status==2])

SurvB1B2 <- Surv(data$os[data$biomarker_status!=0],
                 data$status[data$biomarker_status!=0])

data$treatment  <- data$biomarker_status * data$group
data$treatment1 <- as.numeric(data$treatment == 1)
data$treatment2 <- as.numeric(data$treatment == 2)

#######################################################################################
### Approach 3 (full model) ###

cox_full <- coxph(Surv ~ factor(biomarker_status) + treatment1 + treatment2, data=data)

beta.cox_full_gamma1 <- coef(summary(cox_full))[,1][1]
beta.cox_full_gamma2 <- coef(summary(cox_full))[,1][2]
beta.cox_full_trt1   <- coef(summary(cox_full))[,1][3]
beta.cox_full_trt2   <- coef(summary(cox_full))[,1][4]

#######################################################################################
### Approach 2 (no biomarker-negative patients) ###

cox_B1B2 <- coxph(SurvB1B2 ~ factor(biomarker_status) +treatment1 +treatment2,
                  data=data[data$biomarker_status!=0,])
```

```
beta.cox_B1B2_gamma2 <- unname(coef(summary(cox_B1B2))[,1][1])
beta.cox_B1B2_trt1 <- unname(coef(summary(cox_B1B2))[,1][2])
beta.cox_B1B2_trt2 <- unname(coef(summary(cox_B1B2))[,1][3])

#######################################################################################
### Approach 1 (individual models) ###

cox_B1 <- coxph(SurvB1 ~ group, data=data[data$biomarker_status==1,])

beta.cox_B1 <- coef(summary(cox_B1))[,1]


cox_B2 <- coxph(SurvB2 ~ group, data=data[data$biomarker_status==2,])

beta.cox_B2 <- coef(summary(cox_B2))[,1]

#######################################################################################
### Approach 3 with Firth (full model) ###

cox_full_firth <- coxphf(Surv ~ factor(biomarker_status) + treatment1 + treatment2,
                         data=data, firth=T)

beta.cox_full_firth_gamma1 <- unname(cox_full_firth$coefficients[1])
beta.cox_full_firth_gamma2 <- unname(cox_full_firth$coefficients[2])
beta.cox_full_firth_trt1 <- unname(cox_full_firth$coefficients[3])
beta.cox_full_firth_trt2 <- unname(cox_full_firth$coefficients[4])

#######################################################################################
### Approach 2 with Firth (no biomarker-negatives) ###

cox_B1B2_firth <- coxphf(SurvB1B2 ~ factor(biomarker_status) + treatment1 + treatment2,
                         data=data[data$biomarker_status!=0,], firth=T)

beta.cox_B1B2_firth_gamma2 <- unname(cox_B1B2_firth$coefficients[1])
beta.cox_B1B2_firth_trt1 <- unname(cox_B1B2_firth$coefficients[2])
beta.cox_B1B2_firth_trt2 <- unname(cox_B1B2_firth$coefficients[3])

#######################################################################################
### Approach 1 with Firth (individual models) ###

cox_B1_firth <- coxphf(SurvB1 ~ group, data=data[data$biomarker_status==1,], firth=T)

beta.cox_B1_firth <- unname(coef(cox_B1_firth))


cox_B2_firth <- coxphf(SurvB2 ~ group, data=data[data$biomarker_status==2,], firth=T)

beta.cox_B2_firth <- unname(coef(cox_B2_firth))
```

Data imputation and analysis for Section 3.4

```
### Surv object ###
Surv <- Surv(sim.data$os, sim.data$status)

### additional variables ###
sim.data$biomarkergroup[sim.data$biomarker=="B0"] <- 0
sim.data$biomarkergroup[sim.data$biomarker=="B1"] <- 1
sim.data$biomarkergroup[sim.data$biomarker=="B2"] <- 2

sim.data$treatment  <- sim.data$biomarkergroup*sim.data$group
sim.data$treatment1 <- as.numeric(sim.data$treatment==1)
sim.data$treatment2 <- as.numeric(sim.data$treatment==2)

#########################################################################################
### data analysis for Section 3.4.4.2 ###

sim.dataB1 <- sim.data$os[sim.data$biomarker == ’B1’],
              sim.data$status[sim.data$biomarker == ’B1’]

sim.dataB1_noB2 <- sim.data$os[sim.data$biomarker == ’B1’ & sim.data$B2= = 0],
                   sim.data$status[sim.data$biomarker == ’B1’  & sim.data$B2 == 0]

SurvB1      <- Surv(sim.dataB1$os, sim.dataB1$status)
SurvB1_noB2 <- Surv(sim.dataB1_noB2$os, sim.dataB1_noB2$status)

#########################################################################################
### Model 1 ###

cox_B1_noB2 <- coxph(SurvB1_noB2 ~ group, data=sim.dataB1_noB2)

beta.cox_B1_noB2 <- coef(summary(cox_B1_noB2))[1,1]

#########################################################################################
### Model 2 ###

cox_B1 <- coxph(SurvB1 ~ group + B2:group, data=sim.dataB1)

beta.cox_B1     <- coef(summary(cox_B1))[1,1]
beta.cox_B1_int <- coef(summary(cox_B1))[2,1]

#########################################################################################
### Model 3 ###

cox_full_true <- coxph(Surv(os, status) ~ factor(biomarkergroup) + B1:treatment1 +
                                          B2:treatment2 + B2:treatment1, data=sim.data)

beta.cox_full_true_gamma1 <- coef(summary(cox_full_true))[,1][1]
beta.cox_full_true_gamma2 <- coef(summary(cox_full_true))[,1][2]
beta.cox_full_true_trt1   <- coef(summary(cox_full_true))[,1][3]
beta.cox_full_true_trt2   <- coef(summary(cox_full_true))[,1][4]
beta.cox_full_true_int    <- coef(summary(cox_full_true))[,1][5]

#########################################################################################
### Data imputation for Section 3.4.5.3

### data frames ###
data_true     <- subset(sim.data, select=c(-B2_true))
data_true$B2  <- sim.data$B2_true
sim.data_miss <- subset(sim.data, select=c(-B2_true))
sim.data_imp  <- subset(sim.data, select=c(-B2_true))

model  <- imp.method[1]
method <- imp.method[2]
prop.miss <- as.numeric(imp.method[3])

sim.data_imp$nelsaal <- nelsonaalen(sim.data_imp, os, status)

#########################################################################################
### naive imputation ###
if(model == ’naive’){
        sim.data_imp$B2 <- as.factor(sim.data_imp$B2)

        ini  <- mice(sim.data_imp, max=0, print=FALSE)
        pred <- ini$pred
        meth <- ini$meth
        pred[c("B2"), c("treatment","os")] <- 0
        pred[c("B2"), c("nelsaal")]        <- 1
        meth["B2"] <- method

        imp <- mice(sim.data_imp, meth=meth, predictorMatrix=pred, seed=1, m=prop.miss,
                    print=F, maxit=10)

        cox_full_imp <- with(imp, {coxph(Surv(os, status) ~ factor(biomarkergroup) +
                             B1:treatment1 + as.numeric(B2):treatment2 +
                             as.numeric(B2):treatment1)})

        cox_full_imputed <- summary(pool(cox_full_imp))

########################################################################################
### JAV apporach ###
}else if(model == ’JAV’){
        ## create variables for interactions  ###
        sim.data_imp$B1trt1 <- as.factor(sim.data_imp$B1*sim.data_imp$treatment1)
        sim.data_imp$B2trt2 <- as.factor(sim.data_imp$B2*sim.data_imp$treatment2)
        sim.data_imp$B2trt1 <- as.factor(sim.data_imp$B2*sim.data_imp$treatment1)

        sim.data_imp$B2 <- as.factor(sim.data_imp$B2)

        ini <- mice(sim.data_imp, max=0, print=FALSE)
        pred <- ini$pred
        meth <- ini$meth
        pred[c("B2", "B2trt1", "B2trt2"),c("treatment","os")] <- 0
        pred[c("B2", "B2trt1", "B2trt2"),c("nelsaal")] <- 1
        pred[c("B2trt2"),] <- pred[c("B2trt1"),]
        pred[c("B2"), c("B2trt1")] <- 0
        pred[c("B2trt1"), c("B2")] <- 0
        pred[c("B2trt2"), c("B2")] <- 0
        meth[c("B2", "B2trt1", "B2trt2")] <- method

        imp <- mice(sim.data_imp, meth=meth, predictorMatrix=pred, seed=1, m=prop.miss,
                    print=F, maxit=10)

        cox_full_imp <- with(imp, {coxph(Surv(os, status) ~ factor(biomarkergroup) +
                             as.numeric(B1trt1)+as.numeric(B2trt2)+as.numeric(B2trt1))})

        cox_full_imputed <- summary(pool(cox_full_imp))

########################################################################################
### stratify approach ###
}else if(model == ’stratify’){

        sim.data_imp$B2 <- as.factor(sim.data_imp$B2)

        ### stratify data set variable ’treatment’ ###
        sim.data_imp_trt <- split(sim.data_imp, sim.data_imp$treatment)
        sim.data_imp_trt0 <- sim.data_imp_trt[[1]]
        sim.data_imp_trt1 <- sim.data_imp_trt[[2]]
        sim.data_imp_trt2 <- sim.data_imp_trt[[3]]

        ini <- mice(sim.data_imp, max=0, print=FALSE)
        pred <- ini$pred
        pred[c("B2"),c("treatment","os")] <- 0
        pred[c("B2"),c("nelsaal")] <- 1
        pred[c("B2"),c("B1")] <- 1
        meth <- ini$meth
        meth["B2"] <- method

        imp1 <- mice(sim.data_imp_trt1, meth = meth, predictorMatrix = pred, seed = 1,
                     m = prop.miss, print = F, maxit = 10)
        imp0 <- mice(sim.data_imp_trt0, meth = meth, predictorMatrix = pred, seed = 1,
                     m = prop.miss, print = F, maxit = 10)
        imp  <- rbind(imp0, imp1)

        ### load modified ’with’ function to add complete data set (sim.data_imp_trt2)
        source(’with_addcomplete.R’)
        cox_full_imp <- with_addcomplete(imp, {coxph(Surv(os, status) ~
                                         factor(biomarkergroup) + B1:treatment1 +
                                         as.numeric(B2):treatment2 +
                                         as.numeric(B2):treatment1)},
                                         data.complete = sim.data_imp_trt2)

        cox_full_imputed <- summary(pool(cox_full_imp))
}

beta.cox_full_imputed_gamma1 <- cox_full_imputed[1, 1]
vamd.cox_full_imputed_gamma1 <- cox_full_imputed[1,10]

beta.cox_full_imputed_gamma2 <- cox_full_imputed[2, 1]
vamd.cox_full_imputed_gamma2 <- cox_full_imputed[2,10]

beta.cox_full_imputed_trt1   <- cox_full_imputed[3, 1]
vamd.cox_full_imputed_trt1   <- cox_full_imputed[3,10]

beta.cox_full_imputed_trt2   <- cox_full_imputed[4, 1]
vamd.cox_full_imputed_trt2   <- cox_full_imputed[4,10]

beta.cox_full_imputed_int    <- cox_full_imputed[5, 1]
vamd.cox_full_imputed_int    <- cox_full_imputed[5,10]
```

Curriculum Vitae

PERSONAL INFORMATION

Family name	Habermehl née Beisel
First name	Christina
Date of birth	January 13th 1989
Place of birth	Darmstadt
Marital status	married

EDUCATION

09/2014 – present	**Doctoral student**, Division of Biostatistics, German Cancer Research Center (DKFZ), Heidelberg
08/2012 – 05/2014	**Master of Science**, Mathematics, Western Illinois University, Macomb, IL, USA
06/2012 – 08/2012	**Summer School**, Harvard University Summer School, Cambridge, MA, USA
01/2010 – 12/2010	**Exchange Year**, University of Wisconsin - Platteville, Platteville, WI, USA
09/2008 – 01/2012	**Bachelor of Science**, Applied Mathematics, Hochschule Darmstadt, Darmstadt
June 2008	**Abitur**, Bachgauschule Babenhausen

WORK EXPERIENCE

08/2013 – 12/2013	**Teaching Assistant**, Western Illinois University, Macomb, IL, USA
01/2013 – 05/2013	**Research Assistant**, Illinois Institute for Rural Affairs, Macomb, IL, USA
08/2012 – 12/2012	**Teaching Support Assistant**, Western Illinois University, Macomb, IL, USA

Acknowledgements

I am deeply indebted to my supervisor Prof. Dr. Annette Kopp-Schneider whose help, valuable suggestions, guidance and encouragement were most helpful for doing research for this thesis.

Furthermore, I am obliged to my advisors Axel Benner and Dr. Christina Kunz for all their help, time, support, advice and shared knowledge.

I would like to thank my Thesis Advisory Committee members Prof. Dr. Hanno Glimm and Dr. Stefan Michiels for three productive TAC meetings and their valuable suggestions and insights.

I would also like to thank my colleagues at the Biostatistics department for creating a pleasant and relaxed working environment. Special thanks goes to my office 'roommate' Julia, who makes our office a brighter place.

I am deeply grateful to my husband Kai for his never-ending love, support, and encouragement and for always being understanding. I would also like to thank him for proofreading the thesis and offering suggestions for improvement.

I wish to express my deepest gratitude to my mother Brigitte for her continuous support, for encouraging me in all my endeavors, and for always believing in me.

The data for the data example was kindly provided by the AMLSG study group.

Lastly, I would like to thank the DKFZ and the Helmholtz International Graduate School for Cancer Research for providing me with financial support through their fellowship program.