JOURNAL OF BIOPHARMACEUTICAL STATISTICS
2016, VOL. 26, NO. 1, 99–119
http://dx.doi.org/10.1080/10543406.2015.1092034

Taylor & Francis
Taylor & Francis Group

Methods for identification and confirmation of targeted subgroups in clinical trials: A systematic review

Thomas Ondra[a], Alex Dmitrienko[b], Tim Friede[c], Alexandra Graf[a], Frank Miller[d], Nigel Stallard[e], and Martin Posch[a]

[a]Center for Medical Statistics and Informatics, Medizinische Universität Wien, Vienna, Austria; [b]Center for Statistics in Drug Development, Quintiles, Overland Park, Kansas, USA; [c]Department of Medical Statistics, Universitaetsmedizin Göttingen, Göttingen, Germany; [d]Statistiska institutionen, Stockholms Universitet, Stockholm, Sweden; [e]Department of Statistics and Epidemiology, University of Warwick, Coventry, UK

ABSTRACT

Important objectives in the development of stratified medicines include the identification and confirmation of subgroups of patients with a beneficial treatment effect and a positive benefit-risk balance. We report the results of a literature review on methodological approaches to the design and analysis of clinical trials investigating a potential heterogeneity of treatment effects across subgroups. The identified approaches are classified based on certain characteristics of the proposed trial designs and analysis methods. We distinguish between exploratory and confirmatory subgroup analysis, frequentist, Bayesian and decision-theoretic approaches and, last, fixed-sample, group-sequential, and adaptive designs and illustrate the available trial designs and analysis strategies with published case studies.

ARTICLE HISTORY
Received 14 August 2015
Accepted 14 August 2015

KEYWORDS
Enrichment design; personalized medicine; precision medicine; predictive biomarkers; subgroup identification

1. Introduction

A major challenge in the development of stratified medicines is the identification and confirmation of subgroups where a treatment is effective and has a positive benefit-risk balance. Many modern anti-cancer drugs are understood to act on specific genetic targets, and as a consequence it is expected that the treatment will be effective only in patients where the target is present. The identification and confirmation of targeted subgroups raises several statistical issues, such as the multiplicity problem when assessing multiple populations or the low power to detect treatment effects in subgroups with low prevalence. In recent years, an impressive amount of methodological research has been conducted to derive efficient trial designs and analysis strategies and to better understand the possibilities to obtain evidence on the heterogeneity of treatment effects across subgroups.

For the investigation of targeted therapies, patient subgroups are typically defined by genetic or proteomic biomarkers. In this review we restrict our attention to biomarkers that are measured prior to treatment and therefore cannot be affected by outcome. However, we also consider settings where the determination of a biomarker's status may affect the outcome, for example if the determination requires an invasive procedure or takes so long that it delays the start of treatment administration. Following the standard terminology, we define *prognostic biomarkers* as biomarkers that allow one to predict the outcome independently of any specific therapy and *predictive biomarkers* as biomarkers that predict the treatment effect of an experimental treatment in comparison to control (Ziegler et al., 2012; Beckman et al., 2011). The difference between prognostic and predictive biomarkers becomes most clear when

CONTACT Martin Posch martin.posch@meduniwien.ac.at Center for Medical Statistics and Informatics, Medizinische Universität Wien, Spitalgasse 23, 1090 Vienna, Austria.

considering a regression model with the outcome variable as dependent variable and the independent variables treatment, biomarker, and the interaction term of treatment and biomarker. Then, the prognostic effect of a biomarker is modeled with the main biomarker term in the model and the predictive effect with the interaction term. For the development of targeted therapies, predictive biomarkers are of main interest, and therefore we focus here on methods to identify and confirm a predictive biomarker. If the treatment effects differ between subgroups, we speak of a quantitative interaction, if the treatment effects have different signs; this is called a qualitative interaction.

We conducted a literature survey and give an overview on methodology for clinical trial designs and analysis methods investigating differential treatment effects in subpopulation(s) that address these challenges.

The article is structured as follows: In Section 2, we describe the literature search and classification strategy. In Sections 3 and 4, we report on the identified methods and trial designs. In Section 5, we discuss several case studies and conclude with a discussion.

2. Literature search

We conducted a literature search on the PubMed web site at http://www.ncbi.nlm.nih.gov/pubmed/advanced on April 5, 2015, using the following search strategy *enrichment OR subgroup selection OR subgroup analysis OR subgroup identification.*

The search was restricted to methodological journals given in Table 1. In addition, relevant papers from the list of references in the identified manuscripts as well as papers that were discovered via manual searches have been included in the review. After screening the abstracts of the identified manuscripts, we included only papers that focused on statistical methods for the design and analysis of clinical trials that investigate subgroup effects.

Table 1. Journals included in the literature search.

Included journals	
Biometrical Journal	*Biometrics*
Biostatistics	*BMC Medical Research Methodology*
Clinical Trials	*Contemporary Clinical Trials*
Controlled Clinical Trials	*Journal of Biopharmaceutical Statistics*
Journal of the American Statistical Association	*Journal of the Royal Statistical Society: Series B*
Journal of the Royal Statistical Society: Series C	*Pharmaceutical Statistics*
Statistics in Biopharmaceutical Research	*Statistics in Medicine*

Table 2. Classification criteria.

Classification criterion	Description
Confirmatory trial (CT)	Confirmatory clinical trials investigating up to three prespecified subgroups controlling Type I error rates.
Exploratory trial (ET)	Exploratory clinical trials investigating more than three prespecified subgroups or not controlling Type I error rates.
Frequentist method (FM)	Includes methods based on hypothesis testing as well as regression models if frequentist properties are considered.
Bayesian method (BM)	Approaches where inference is based on posterior distributions of parameters or the trial design is based on Bayesian techniques (as, e.g., in adaptive trials).
Decision-theoretic method (DM)	Inference or trial design is based on maximizing a utility function.
Trial design type	Classification into fixed sample designs (FD), group sequential designs (GS), adaptive designs based on conditional error rate approach (ADce), adaptive designs based on combination functions (ADcf), response adaptive designs (RA), and other adaptive designs (ADo).
Biomarker type	Continuous (C), categorical (Cat), binary (B).
Number of prespecified subgroups	The number of prespecified subgroups is classified into few (≤ 3), any number (Any), and no subgroups prespecified (None).
Endpoint type	Continuous (C), binary (B), time-to-event (TtE), categorical (Cat).
Class of exploratory methods	Global outcome modeling (GOM), global treatment effect modeling (GTEM), and local modeling (LM).

The identified manuscripts were classified according to the criteria defined in Table 2. We distinguished between *confirmatory settings,* where a small number (up to three) of prospectively defined subgroups of patients are investigated and the frequentist error rates such as the familywise error rate are explicitly protected, and *exploratory settings,* where multiple subgroups may be considered and error rate control may not be addressed.

The analysis methods used in confirmatory and exploratory settings were classified as follows:

- *Frequentist methods* (FM) that deal with assessing frequentist properties of parameter estimates and controlling Type I error rates in hypothesis testing problems;
- *Bayesian methods* (BM) that rely on inferences based on posterior distributions of parameters or the trial design is based on Bayesian techniques;
- *Decision-theoretic methods* (DM) that are based on utility functions that assign gains and costs to different decisions based on the clinical trial data.

Note that some of the proposed approaches fall into more than one of these categories because they combine frequentist, Bayesian, and decision-theoretic methods, e.g., by considering multiple testing procedures for hypothesis testing but a Bayesian decision theoretic approach to optimize trial designs. Furthermore, the methods were classified according to the trial endpoint type, i.e., continuous, binary, categorical, or time-to-event endpoints (no count-type endpoints were found), and biomarker type, i.e., binary, categorical, and continuous biomarkers, which define the subgroup(s) of interest.

Another classification factor was the number of prespecified patient subgroups (which is related to the exploratory/confirmatory classification criterion). While, by definition, binary biomarkers used in confirmatory studies define two subgroups, categorical or continuous biomarkers, or combinations of several biomarkers may define several subgroups. We distinguished methods that can be applied to any (fixed) set of subgroups from methods where no candidate subgroups are prespecified. Note that a method which controls the Type I error rate but is designed for an arbitrary number of predefined subgroups was classified as both confirmatory and exploratory since it can be applied to settings with a few subgroups as well as a large number of subgroups. Clinical trial designs were classified into fixed-sample designs, adaptive designs based on the combination function or conditional error approach, group-sequential designs, response adaptive designs, and other adaptive designs. Finally, exploratory subgroup analysis methods were further classified into three subcategories introduced in Lipkovich and Dmitrienko (2014a): *global outcome modeling* (GOM), *global treatment effect modeling* (GTEM), and *local modeling* (LM), see Section 4 for a detailed description of these categories.

We identified in total 239 papers of which 86 were classified as relevant for this survey (i.e., papers on novel methodology on the identification and confirmation of patient subgroups in clinical trials). The results of the literature search are summarized in Figure 1 and Tables 3, 4 and 5. A table with the list of all papers and their categorization can be found in the supplementary material.

3. Confirmatory methods

By definition, confirmatory approaches control the Type I error rate (familywise error rate) and hence fall into the category of frequentist methods. However, some of these methods use frequentist approaches for statistical inference but employ Bayesian or decision-theoretic methods in the trial design. These approaches will be discussed in Sections 3.3 and 3.4.

The most frequently investigated scenarios are parallel-group designs with an experimental treatment arm (T) that is compared to a control (C). In the simplest case, a single prespecified subgroup S (termed the *target subgroup*) defined by a binary biomarker or binary classifier derived from one or more continuous biomarkers is considered. We denote the complement of the subgroup by S' and the full population by F.

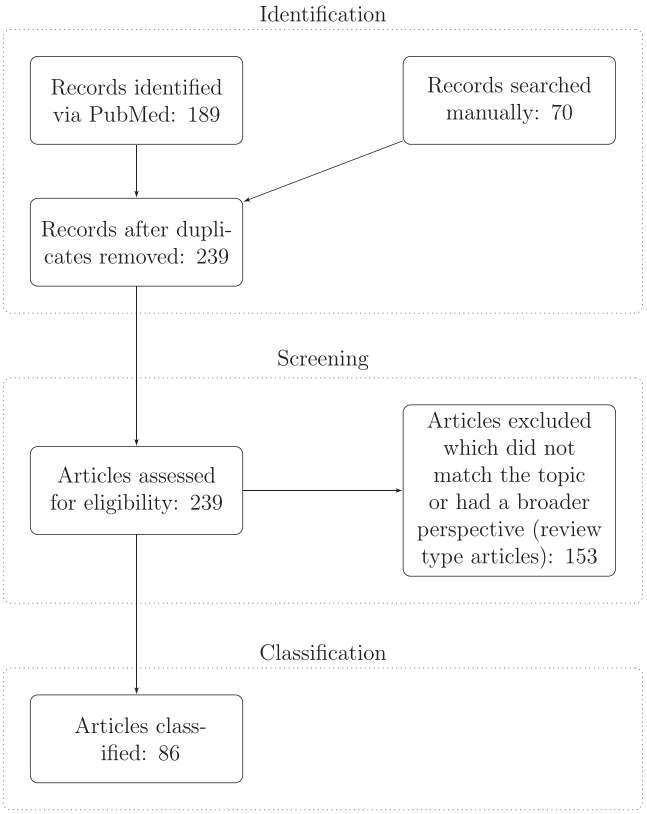

Figure 1. Flow diagram of selected manuscripts.

Table 3. Classification of the methodological approaches.

Classification criterion				
CT/ET	36 (CT)	36 (ET)	14 (CT & ET)	
Number of prespecified subgroups	35 (≤ 3)	35 (Any)	16 (None)	
Endpoint type	18 (B)	14 (TtE)	53 (C)	1 (Cat)
Biomarker type	34 (B)	29 (Cat)	23 (C)	
Class of exploratory method	25 (GOM)	6 (GTEM)	5 (LM)	

B, binary; C, continuous; Cat, categorical; CT, confirmatory trial; ET, exploratory trial; GOM, global outcome modeling; GTEM, global treatment effect modeling; LM, local modeling; TtE, time-to-event

Table 4. Exploratory and confirmatory approaches stratified by frequentist methods (FM), Bayesian methods (BM), and decision-theoretic methods (DM).

Exploratory / Confirmatory	FM	BM	DM
Confirmatory (CT)	36	5	2
Exploratory (ET)	18	19	8
Confirmatory & Exploratory (CT and ET)	14	0	0

Note. Papers using more than one method (e.g., frequentist and Bayesian) appear more than once in this table.

Table 5. Results for trial design type classification criterion.

Trial design type	ADce	ADcf	ADo	RA	GS	FD
Number	3	11	13	6	1	52

The following notation will be used throughout this section. With λ, we denote the population prevalence of biomarker-positive patients. Furthermore, Δ^j, $j \in \{+, -, F\}$, denotes the treatment effect in the respective populations, e.g., Δ^+ specifies the mean treatment difference in the biomarker-positive population under the assumption of a normally distributed outcome variable. For effect sizes defined as mean differences, it is easy to verify that the treatment effect in the full population is given by $\Delta^F = \lambda\Delta^+ + (1-\lambda)\Delta^-$. We assume that $\Delta^+ \geq \Delta^-$, reflecting settings where due to the mode of action of the treatment it is justified to assume that the treatment effect is more pronounced (or only present) in S.

A confirmatory clinical trial investigating treatment effects in this scenario can have the following three distinct objectives:

- Objective O1: Demonstrate the efficacy of the treatment in S only.
- Objective O2: Demonstrate the efficacy in F only.
- Objective O3: Demonstrate the efficacy in F and enhanced efficacy in S.

Accordingly, a multiple testing procedure can be used to control the familywise error rate for the null hypotheses of no-treatment effects in S and in F.

Millen et al. (2012) emphasized the importance of accounting for incorrect decisions related to Objectives O1 and O3 and introduced tools for facilitating the decision-making process in multi-population trials (known as the influence and interaction conditions). For example, a statistical test may show a significant effect in the full population F which is entirely driven by a strong treatment effect in S. The influence condition states that Objective O2 needs to be restricted to Objective O1 if no treatment benefit is established in S'. Similarly, when considering Objective O3, it is important to ensure that the treatment effect in the target subgroup S is meaningfully different from that in the complementary subgroup S'. If the appropriately defined interaction condition is met, Objective O3 is valid, and it needs to be replaced with Objective O2 otherwise.

Frequentist and Bayesian rules for evaluating the influence and interaction conditions were proposed in Millen et al. (2012, 2014). The Bayesian rules enable clinical trial researchers to account for available prior information and uncertainty around the estimated treatment effects. As an illustration, within a simple frequentist framework, the influence condition is met if $\Delta^- \geq \eta$, where η is a prespecified constant which defines a meaningful treatment effect in the complementary subgroup, which supports the conclusion that the treatment effect is homogeneous across the full population. Alternatively, η can be viewed as a tuning parameter and selected based on appropriate statistical criteria. Switching to a Bayesian framework, the influence condition is satisfied if the posterior probability of a meaningful effect in the complementary subgroup given the available data is high enough, i.e.,

$$P(\Delta^- \geq c_1 | D) \geq \gamma_1,$$

where $0<\gamma_1<1$ is a risk-tolerance parameter and D denotes the available data. Similarly, using a Bayesian rule, the interaction condition is met if

$$P(\Delta^+ \geq c_2\Delta^- | D) \geq \gamma_2.$$

Here $c_2>1$ is an application-specific threshold of clinical relevance, γ_2 is again a risk-tolerance parameter, and we assume that the prior on Δ^- gives no weight to negative effect sizes. A detailed mathematical description of the proposed assessment strategy for the the influence and interaction condition for binary, time-to-event, and continuous endpoints is provided in Millen et al. (2014).

A special class of treatment strategies need to be considered in settings where the process determination of a patient's biomarker status may impact the outcome, for example, because of the time needed to determine the biomarker status or because invasive examinations are involved. In these cases, it is of interest to assess whether patients treated based on their biomarker status have a better outcome than patients whose treatment does not account for the biomarker. This leads to the following objective:

- Objective O4. Demonstrate the superiority of a biomarker-guided treatment strategy compared to a treatment strategy that does not take the biomarker into account, for example, strategies that assign all patients to the control treatment or all patients to the experimental treatment.

Depending on the trial's objectives defined above, different strategies can be considered to utilize the available biomarker in the design or analysis of biomarker-driven clinical trials:

- Strategy B1. The biomarker status is used as part of the inclusion criteria.
- Strategy B2. Treatments are assigned based on the patient's biomarker status.
- Strategy B3. The biomarker status is used in the analysis as a stratification factor or identifies an important subgroup, which is included in the primary analysis.

Different proposals for applying Strategies B1, B2, and B3 in different trial designs to reach the design-specific objectives are discussed below.

3.1. *Single-stage designs*

Single-stage clinical trial designs (Mandrekar and Sargent, 2009; Freidlin et al., 2010b; Mandrekar and Sargent, 2011b; Freidlin et al., 2012; Ziegler et al., 2012) recruit patients from prespecified populations according to a prespecified sampling rule. The designs proposed in the literature differ in the way populations are investigated, the role of the biomarker in the trial design, and the way multiple hypothesis testing is implemented. Figure 2 gives an overview of five proposed designs (Designs F1 through F5) and associated sampling and treatment allocation rules. Below we discuss

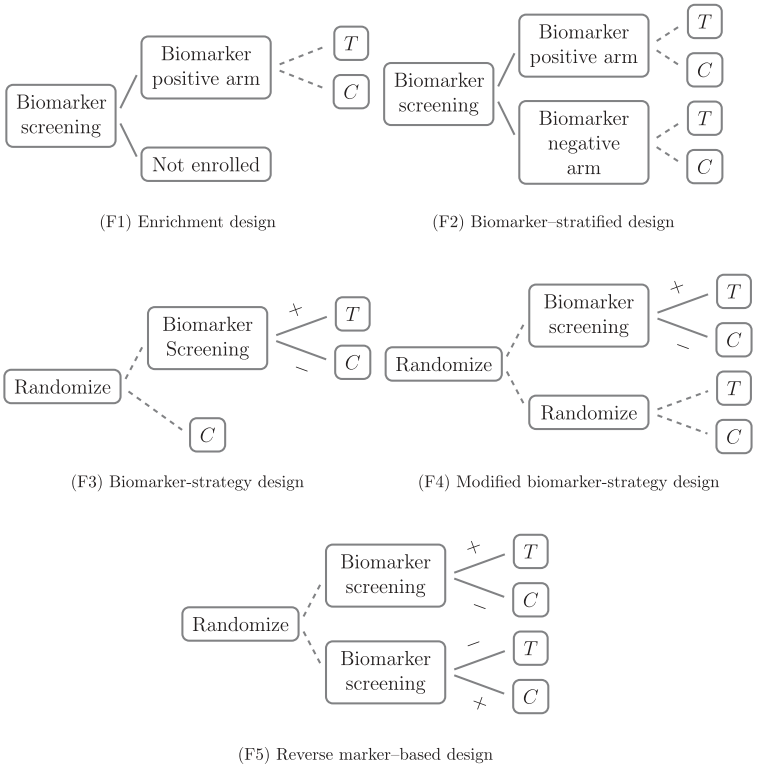

Figure 2. Overview of single-stage clinical trial designs. *T* is the experimental treatment, *C* is the control treatment. Solid lines indicate deterministic decisions whereat dashed lines indicate that a randomization procedure is involved.

their relation to Objectives O1 through O4 and identify the appropriate biomarker analysis strategies (Strategies B1 through B3) that were implemented in each design.

Design F1 is a simple *enrichment* or subpopulation-only design which has the following features. Before inclusion in the trial, patients are screened and selected by their biomarker status such that only biomarker-positive patients enter the trial and are randomized to the experimental treatment T or control C. The biomarker is utilized in this design using Strategy B1 and enables the trial's sponsor to test only one null hypothesis, namely, the null hypothesis of no-treatment effect in the biomarker-positive population, i.e., $H_+ : \Delta^+ \leq 0$. This means that the treatment effect can be studied in this population only (Objective O1).

To investigate whether a treatment is also effective in a larger patient population, including biomarker-negative patients, more complex trial designs are required. An example is the *biomarker-stratified* or multi-population design (Design F2). This design includes four arms, where patients are screened for biomarker status and randomization, stratified for the biomarker status, is performed. Biomarker-positive as well as biomarker-negative patients are randomized to the treatment T and control C, which means that Strategy B3 is applied in this design. Compared to Design F1, this design supports testing the null hypotheses of no effect in the biomarker-positive and full populations, i.e., $H_+ : \Delta^+ \leq 0$ and $H_F : \Delta^F \leq 0$, which means that it may be used to address Objectives O2 or O3 (the influence and interaction conditions need to be applied to determine the most relevant objective).

An important feature of Design F2 is that several null hypotheses are tested to examine the efficacy of the experimental treatment. This leads to Type I error rate inflation and a multiplicity adjustment must be applied to control the familywise error rate (FWER) in the strong sense. FWER is controlled strongly if the probability to commit at least one Type I error does not exceed the nominal level (e.g., one-sided $\alpha = 0.025$) regardless of how many and which null hypotheses are true (see, e.g., the tutorial (Dmitrienko and D'Agostino, 2013)). Suitable multiple testing methods include non-parametric procedures such as the Bonferroni procedure, Holm procedure, or more flexible Bonferroni-based stepwise procedures (Bretz et al., 2009). However, since the test statistics for the null hypotheses in the full and biomarker-positive populations are positively correlated, Bonferroni-based procedure may become conservative and more efficient approaches include semi-parametric procedures (e.g., Hochberg procedure) or parametric procedures (e.g., methods similar to the Dunnett procedure). These approaches gain a power advantage over non-parametric procedures by taking into account the correlation between the test statistics. Several authors noted that for many settings the test statistics follow an approximate multivariate normal distribution and computed critical values for single-step or stepwise parametric procedures (Song and Chi, 2007; Alosh and Huque, 2009; Spiessens and Debois, 2010; Millen and Dmitrienko, 2011; Alosh and Huque, 2013; Bretz et al., 2011). Zhao et al. (2010) considered a more general parametric test combining the test statistics with combination functions.

In Designs F1 and F2 all patients are screened and their biomarker status is determined prior to treatment. Therefore, these designs do not support the investigation of the impact of screening and determination of a patient's biomarker status on the outcome. For example, it may take time to ascertain the biomarker status, which will lead to a delay of the treatment start and may negatively impact the outcome. Design F3 is a trial design, known as the *biomarker-strategy design,* which addresses Objective O4. This design facilitates the comparison of a biomarker-guided therapy, where a patient's biomarker status is determined and only biomarker-positive patients receive the experimental treatment, to a treatment regimen without screening, where all patients are allocated to the control. Patients are randomized to either the control (without screening) or the biomarker-guided treatment strategy arm. Within the latter arm, the biomarker status is determined and all biomarker-positive patients receive the experimental treatment T, whereas the biomarker-negative patients receive the control C. Thus, Design F3 applies Strategy B2 for utilizing the biomarker information in the trial. Note that this trial design may require a large sample size to achieve adequate statistical power because biomarker-negative patients receive the control treatment in both arms (biomarker-guided arm and control arm) which results in a diluted effect size.

While Design F3 can demonstrate that the biomarker-guided treatment is superior to the strategy which relies on treating all patients with the control, it does not address the goal of assessing whether all patients (including biomarker-negative patients) would benefit from the experimental treatment. In other words, this design does not address Objective O2. The latter can be addressed (in addition to Objective O4) by the *modified biomarker-strategy design* (Design F4). It differs from Design F3 in that patients randomized to the non-biomarker strategy arm are again randomized between the experimental treatment and control. This design tests the impact of the biomarker-guided strategy against a random allocation procedure which does not take the biomarker into account (Objective O4). Furthermore, it allows the sponsor to test and estimate the treatment effect of the experimental treatment without a biomarker-guided treatment strategy Objective (O2). However, similar to Design F3, Design F4 may require a larger sample size because some of the biomarker-negative patients in the randomization arm also receive the control treatment and some of the biomarker-positive patients the experimental treatment. This leads to a diluted treatment effect and may result in lower statistical power.

Recently, Eng (2014) proposed a fixed-sample design termed the *reverse biomarker-based strategy*. This design was compared to Designs F3 and F4 in the case of binary outcomes. Specifically, patients are randomly assigned to one of the two treatment strategies. In the first arm, biomarker-positive patients receive the experimental treatment, whereas biomarker-negative patients are allocated to receive the control. By contrast, in the second arm, biomarker-positive patients receive the control and biomarker-negative patients receive the treatment. It is easy to see that the effect size for comparing the treatment strategy arms will be typically larger than in Designs F3 and F4. Furthermore, with this approach, the effect size of the experimental treatment compared to the control is easily estimated for each subgroup separately. However, the reverse biomarker-based design cannot address the question if a treatment strategy that does not require the determination of the biomarker status (which may, e.g., delay treatment) would be superior to the biomarker-guided treatment strategies.

Phase III trial designs are often based on the outcomes observed in Phase II trials. As an example, Freidlin et al. (2012) proposed an ad-hoc method for the selection of a biomarker-driven trial design, which utilizes a simplified version of the influence and interaction conditions (Millen et al., 2012). Considering a Phase II trial with a time-to-event endpoint, begin with the treatment effect test in the biomarker-positive subgroup S. If the null hypothesis of no effect is rejected in S at $\alpha = 0.1$, there is some evidence that the experimental treatment is superior to the control in the biomarker-positive subgroup. After that, an 80% confidence interval for the hazard ratio in the biomarker-negative population is constructed. If the interval lies entirely below 1.3, Design F1 is recommended for the subsequent Phase III trial since the treatment effect is unlikely to be beneficial in biomarker-negative patients. If the confidence interval includes 1.3 or 1.5, Design F2 should be used in the Phase III trial. If the confidence interval lies entirely above 1.5, a standard design in the full population should be conducted. On the other hand, if the null hypothesis of no effect is not rejected in S, the null hypothesis is tested in the full population at $\alpha = 0.05$. If this null hypothesis is rejected, a standard Phase III trial should be conducted and the development program should be terminated otherwise.

Further, Mandrekar and Sargent (2011b) and Mandrekar et al. (2013) suggested to base the choice of the trial design in a Phase III program on the preliminary evidence of efficacy in different subgroups, assay performance, marker prevalence and turnaround times. Here *assay performance* refers to how reliably the membership in a subgroup is defined in the presence of measurement errors in the determination of a patient's biomarker status. *Turnaround time* is defined as the time it takes to determine a patient's biomarker status.

3.2. *Multi-stage designs*

In multi-stage trials, patients are recruited in several stages and an interim analysis is performed after each stage. Based on the results of each interim analysis, the trial may be continued as initially planned,

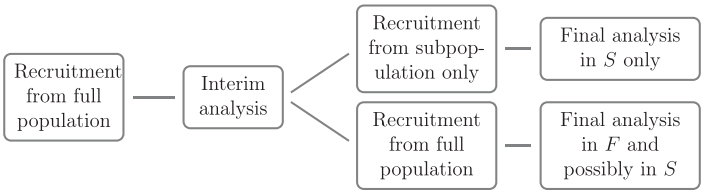

Figure 3. Adaptive multi-stage design.

stopped early, or the trial's design may be adaptively modified. For example, based on results of an interim analysis, patient recruitment in the second stage may be restricted to a predefined subgroup (see Figure 3). Multi-stage designs enable the trial's sponsor to change the biomarker's role in a clinical trial. In the first stage, the biomarker may be used as a stratification factor (as in Strategy B3) to test the treatment effect in the full and biomarker-positive populations (addressing Objectives O1, O2, or O3). In the second stage, an option to restrict the inclusion criteria to focus on patients with a certain biomarker status may be introduced as in Strategy B1. As a result, the trial's goal may be modified to focus on Objective O1. In addition to adapting the patient recruitment, the analysis strategy at the final analysis may be modified based on the outcome of an interim assessment. For example, a decision may be made to continue recruiting patients from the full population, but irrelevant null hypotheses may be dropped or the multiple testing strategy may be modified by updating the weights of the null hypotheses in the full and biomarker-positive populations.

In contrast to approaches where population selection/enrichment and evaluation of the treatment effect are addressed in separate trials, the data from different stages may be combined for the statistical analysis of multi-stage designs. This can be implemented using the combination function approach (Bauer and Koehne, 1994; Bauer and Kieser, 1999; Hommel, 2001; Wang et al., 2007; Brannath et al., 2009; Wang et al., 2009; Jenkins et al., 2011) or conditional error rate approach (Proschan and Hunsberger, 1995; Müller and Schäfer, 2004; Mehta and Gao, 2011; Friede et al., 2012; Irle and Schäfer, 2012; Mehta et al., 2014).

With both approaches, strong FWER control is achieved by the closure principle. For example, to test H_F and H_+ with FWER control at α, (local) level-α combination tests for the *intersection hypotheses,* i.e., the elementary hypotheses H_F and H_+ and global null hypothesis $H_F \cap H_+$, need to be defined. A closed testing procedure that rejects an elementary hypothesis H_i, $i \in \{F, +\}$ at multiple level α if H_i and the global intersection hypotheses can be rejected by their respective local level-α combination tests. If more than two hypotheses are tested, the closure principle requires to consider all intersection hypotheses (Marcus et al., 1976), although in many settings the so-called shortcuts can be identified that streamline the testing algorithm (Brannath and Bretz, 2010).

Testing procedures based on the combination function approach as well as conditional error rate approach in principle do not require the prespecification of the adaptation rule (as, e.g., population selection criteria or sample size adaptation algorithms) to guarantee FWER control in the strong sense. However, the proposed procedures require that the set of initially considered subgroups (from which subgroups may be selected at an interim analysis) is prespecified in advance rather than defined based on the first-stage data. An exception is the statistical test used in Mehta and Gao (2011) (see also Hommel (2001)). Even though the authors considered a setting with a small set of prespecified subgroups, the statistical test used provides FWER control even if any subgroup was defined based on the interim data. The additional flexibility comes at the price of having to use the second-stage data only if the trial is enriched. Similar two-stage procedures in a more exploratory context have been proposed in (Freidlin, 2005; Freidlin et al., 2010a) and are discussed in Section 4.1. Even though in most publications on adaptive enrichment designs specific adaptation rules are proposed, deviations from these adaptation rules in an actual trial do not lead to FWER inflation. Furthermore, the adaptation rule may depend on unblinded interim data in any way. For example, in addition to population selection, the second-stage sample size may be adapted at the interim

analysis. Special care, however, is required in trials with time-to-event endpoints, where some patients recruited before the interim analysis may not have experienced an event at the time of the interim analysis (Bauer and Posch, 2004). Several proposals have been made to guarantee FWER control in this setting, namely, either adaptations may depend only on the interim test statistics (and variables the analysis is stratified for) (Brannath et al., 2009) or it is guaranteed that the maximal observation time of patients recruited at the first stage is not adapted (Jenkins et al., 2011; Irle and Schäfer, 2012). See Magirr et al. (2014) for a detailed discussion.

Stallard et al. (2014) performed a simulation study comparing different adaptive designs with population selection from a prespecified set of subpopulations. They considered designs based on the combination function and conditional error rate approaches to account for adaptations as well as a design with two separate trials where the first trial is used for population selection only and the second stage is used for hypothesis testing. In the combination tests, multiplicity was accounted for using the Simes test and Spiessens–Debois test (Spiessens and Debois, 2010). Furthermore, they compared different adaptation rules to select the second-stage population and hypotheses tested at the final analysis. This included a decision rule that is based on separate thresholds for the treatment effect estimates in the full population and biomarker-positive population as in Jenkins et al. (2011) and, in addition, a decision rule that is based on the difference of treatment effects in the two populations proposed in Friede et al. (2012). The simulation study suggested that, especially if the prevalence of biomarker-positive patients is small, adaptive designs that support enrichment of the second-stage population are more powerful than adaptive designs where the adaptations are restricted to the selection of the hypotheses to be tested. Furthermore, as expected, adaptive designs that take into account the first-stage data for hypotheses testing are more efficient compared to the designs that only utilize second-stage data for hypothesis testing.

Wassmer and Dragalin (2014) generalized these designs and described adaptive population enrichment designs based on combination functions for an arbitrary number of prespecified subpopulations. Several possible intersection hypothesis tests as well as overall p-values and confidence intervals were discussed. Simulation studies for adaptive designs, where populations were selected based on the differences between the observed effect sizes in the different populations, were presented.

Magnusson and Turnbull (2013) proposed a group-sequential approach to implement population enrichment designs. These designs require that a set of subpopulations should be prespecified and can have two or more stages. However, subpopulations may be dropped after the first stage only. Two decision rules for subgroup selection were presented. With the first rule, all subgroups with the observed treatment effect below a certain threshold are discontinued. For the second selection rule, a hierarchy among the subgroups is predefined. If the treatment effect for a certain subgroup exceeds a predefined threshold, all higher-ordered subgroups are selected for the second stage. Stopping boundaries based on an α-spending approach for futility and efficacy controlling the FWER in the strong sense were calculated. Point and interval estimates for the treatment effect in the selected population as well as power and sample size calculations were provided. Note that the approach based on group-sequential designs guarantees FWER control only if the investigators follow the prespecified population selection rule. This holds because the rejection region of the trial depends on the prespecified rule, and a deviation from this rule may lead to FWER inflation. This is in contrast to adaptive designs based on combination functions and conditional error rates discussed above that do not require a prespecification of population selection rules.

Boessen et al. (2013) performed a simulation-based comparisons of fixed designs, group-sequential designs, and adaptive designs in trials with a prespecified subpopulation. In the fixed-design setting, the null hypotheses in the full population and sub-population, H_F and H_+, were tested using the Hochberg test. In the group-sequential design, an interim analysis was added with the option to stop for efficacy (and reject both null hypotheses) or futility (retain both null hypotheses) or to

continue to the second stage recruiting patients from the total population. For the adaptive design with population enrichment, they considered adaptation rules that, in addition to the stopping rules of the group-sequential design, support selection of the subgroup only for the second stage based on a version of the selection rule introduced in Friede et al. (2012). It was shown via simulations that, for a given overall significance level α and given power level, the group-sequential as well as adaptive designs have a lower sample size compared to the fixed-sample design. In addition, the adaptive designs can lead to a further reduction in the sample size compared to the group-sequential designs if the difference between the treatment effects in the prespecified subpopulation and its complement is large.

3.3. *Multi-stage designs with Bayesian rules*

Bayesian decision tools can be useful to guide subgroup selection as well as futility stopping decisions in multi-stage trials. It is important to note that, while the adaptations are based on Bayesian principles, the hypothesis tests at the final analysis are conducted using frequentist approaches to control the FWER in the strong sense.

Song (2014) considered adaptive population enrichment designs in a trial with a time-to-event endpoint that requires a long-term follow-up, which is common in oncology trials. Since a small number of events of interest is expected to be accrued at an interim analysis, the interim effect size is estimated using Bayesian tools that incorporate information from a short-term binary endpoint (surrogate) that is already available at the time of the interim analysis (Huang et al., 2009). This information synthesis can increase the precision of the interim estimates. It was pointed out that the use of the surrogate leads to a substantial increase in precision only if it is strongly related to the primary time-to-event endpoint. Moreover, the prior distribution reflecting the available information on the surrogate endpoint should be carefully selected by taking into account historical data. For the frequentist test of the respective null hypotheses, Song (2014) considered the testing framework proposed in Wang et al. (2007). However, the specific complexities associated with FWER control in adaptive survival trials discussed above (Bauer and Posch, 2004) were not addressed.

Brannath et al. (2009) also proposed to use Bayesian tools for interim decision-making in an adaptive enrichment design with a time-to-event endpoint. Based on predictive probabilities for rejecting certain null hypothesis, a decision tool was developed to determine which of the two populations (full population or subpopulation) should be further investigated in the second stage. Possible decisions included options to continue patient enrollment in the full population, subpopulation only, or stop for futility.

3.4. *Decision-theoretic approaches*

Several authors have proposed decision-theoretic methods to derive optimized trial designs with prespecified subpopulations. These are based on utility functions that assign a certain utility to every trial outcome (e.g., rejection of the null hypothesis of no-treatment effect in the full population). The utility may depend, e.g., on the trial outcome (which hypotheses are rejected), sample size, treatment effect estimate, and, in addition, on the (typically unknown) true value of the efficacy parameters in different subgroups. Because the true parameter values and the outcome of the trial are unknown at the planning stage of a trial, the utility of a trial design is unknown and cannot be calculated a priori. However, one can compute the expected utility by averaging over the efficacy parameters based on a prior distribution and averaging over the data distribution given the values of the efficacy parameters. Optimized trial designs can then be derived by maximizing the expected utility. Instead of maximizing a utility function, one can equivalently specify a loss function and minimize the expected loss, which is termed the Bayes risk. These decision-theoretic approaches can be used to optimize single-stage as well as adaptive multi-stage

designs. In the latter case, the decision-theoretic framework can be applied to optimally select adaptation rules at interim looks.

Beckman et al. (2011) developed a Bayesian decision-analytic approach to decide if a subsequent Phase III trial should be enriched, stratified in the full population, adaptive or not conducted at all based on the available Phase II data.

Krisam and Kieser (2014) considered a decision-theoretic approach for single-stage designs minimizing a quadratic loss function that assigns losses if the full population is selected, although the treatment effect in the subgroup is substantially larger than in the complement as well as in the opposite case if the subgroup is selected, while the treatment effect in the complement is similar to (or smaller than) then the effect in the subgroup. They derived optimal decision functions to select either a prespecified subpopulation or the overall population. In addition, they also investigated the impact of errors in the determination of a patient's biomarker status.

Using a similar approach, Götte et al. (2014) derived optimal rules for selecting a patient population at an interim analysis or futility assessment. They aimed to maximize a utility function defined as the expected probability of a correct selection and considered a simple three-point prior over the three scenarios of no-treatment effect in the full population or subpopulation, treatment effect in the subpopulation only, and a homogeneous treatment effect in the full population. The adaptation rules were optimized by optimally selecting thresholds for three different classes of adaptation strategies that are all based on the estimated effect sizes in the subpopulation and its complement. Besides a decision rule, called the "simple rule", which is based on the signs of the estimated effect sizes in the population chosen in the second stage, they proposed more general rules that take into account weighted averages of either the effect size or the conditional power in the subpopulation and its complement. Optimal thresholds were derived to maximize the expected probability of a correct selection for these rules.

Graf et al. (2015) used a decision-theoretic approach to evaluate fixed-sample and adaptive population enrichment designs that control the FWER. The utility functions considered assign utilities to the different outcomes of the hypothesis tests in the trial. Utility functions for different objectives were defined, representing the sponsor's as well as public health-policy maker's views. Considering adaptation strategies that depend on the interim treatment effect estimates in different populations, adaptation rules and stage-wise sample sizes were optimally selected. Settings where single-stage enrichment designs or trials in the full population are preferable to adaptive enrichment designs were identified.

4. Exploratory methods

In this section, we discuss methods for the investigation of patient subgroups with a beneficial treatment effect in an exploratory setting. Recall from Section 2 that the exploratory setting deals either with large sets of subgroups and/or designs without formal frequentist error rate control. This includes exploratory prospective trial designs based on adaptive Bayesian treatment allocation rules and methods aimed at post-hoc identification of patient subgroups with desirable properties (e.g., improved benefit). In response-adaptive trials, treatment allocation for newly recruited patients depends on biomarker status specific treatment effect estimates of the treatments investigated. These treatment effect estimates are updated after every patient based on Bayesian posterior distributions. The topic of subgroup identification/subgroup search has attracted much attention in the literature and key subgroup search methods are presented in Section 4.1. A review of recent developments in response-adaptive designs with applications to subgroup analysis is provided in Section 4.2.

4.1. *Subgroup identification methods*

A classification scheme for different approaches to investigation of treatment heterogeneity across patients subgroups (subgroup search methods) that are defined by biomarker profiles was proposed in Lipkovich and Dmitrienko (2014a) and Lipkovich et al. (2015). For a biomarker profile $x = (x_1, \ldots, x_k) \in X$, we denote the expected response (outcome) by

$$E(y|x,t) = f(x,t)$$

where $t \in \{C, T\}$ indicates that the patient received the control or experimental treatment, respectively. The outcome function can be modeled as $f(x,t) = h(x) + 1_{\{t=T\}}g(x)$, where h is the prognostic and g the predictive component, and $1_{\{\cdot\}}$ denotes the indicator function. The prognostic component helps evaluate a patient's outcome regardless of the treatment received and the predictive component helps investigate treatment-modification properties of a given biomarker. Depending on the goal of a subgroup search method, Lipkovich and Dmitrienko (2014a) distinguished between *GOM, GTEM,* and *LM* methods. GOM methods rely on modeling the outcome function *f*. By contrast, global treatment effect modeling methods focus on the treatment contrast $z(x)$, which is defined, for example, as $z(x) = f(x,T) - f(x,C)$. If it is assumed that f can be decomposed in an prognostic and a predictive term, as shown above, it follows that $z(x) = g(x)$ and the goal is to model the predictive component of the expected patient's response. Note that the treatment contrast can be defined on other scales, e.g., as the log odds ratio $\log(f(x,T)/(1-f(x,T)) - \log(f(x,C)/(1-f(x,C))$.

For single-arm trials, outcome and treatment effect modeling are equivalent and we categorized the respective papers to the treatment effect modeling group. Examples include oncology trials where the interest lies in comparing the response rates in biomarker-positive and biomarker-negative patients who are assigned the same treatment.

Finally, the LM approach aims at a direct identification of patient subgroups with an enhanced treatment effect, considering each subgroup separately to estimate treatment effects. Methods in this class do not model the outcome function over the entire covariate space but construct treatment effect estimates for individual subsets of the covariate space.

Beginning with GOM approaches, Chen et al. (2012) considered a Bayesian approach to search for qualitative interactions in a regression setting with adaptive decision rules. Qualitative interactions correspond to settings with subgroups where the treatment effect is reversed. The authors investigated several model selection algorithms addressing the multiplicity problem inherent in the selection of subgroups and consider continuous, binary, and time-to-event endpoints.

For the analysis of randomized single-stage clinical trials with longitudinal measurements, Moineddin et al. (2008) investigated multi-level models including random effects to identify subpopulations with differential treatment effects and applied the approach in a case study of a treatment of postmenopausal women experiencing hot flashes.

Cai et al. (2011) considered generalized linear models to estimate the mean outcome given a biomarker configuration x and treatment t with a two-stage estimation procedure. In the first stage, the model

$$E(y|x,t) = \varphi_t(\beta_t^\top u(x)), \;\; t \in \{C, T\},$$

where y denotes the patient's response, x the biomarker configuration, $u(x)$ a prognostic component, and φ_t a smooth, strictly increasing link function, was fitted to estimate the unknown parameter vector β_t. Based on these estimates, patients were grouped into subgroups $\{x : z(x) = v\}$, where $z(x) = \varphi_T\left(\hat{\beta}_T^\top u(x)\right) - \varphi_C\left(\hat{\beta}_C^\top u(x)\right)$ denotes the estimated treatment effects (to obtain subgroups of sufficient size, one can create strata by segmenting all possible values of $z(x)$ in intervals instead). In the second stage, the average treatment difference for each subgroup was estimated via a nonparametric function estimation method based on a local likelihood approach. It was pointed out that, if the parametric model fails to hold, inference based on the first stage of the method might be invalid.

However, the estimator constructed in the second stage is always a consistent estimator of the average treatment effect in the selected subgroups, regardless of the adequacy of the first-stage model.

Altstein et al. (2011) considered a parametric accelerated failure-time model for latent subgroup analysis of a right-censored time-to-event endpoint. Latent subgroup analysis can be used whenever subgroup membership is only observable in one arm of the trial. This occurs, for example, in oncology trials when a biopsy is performed only on patients allocated to the experimental treatment arm. For this setting, a general framework to estimate treatment effects in the latent subgroup was developed.

Foster et al. (2011) proposed a two-stage method for trials with binary outcomes, called *Virtual Twins*. In the first stage, random forests were used to estimate the patient-specific probabilities $P(y = 1|x, t), t \in \{C, T\}$, defined as the probabilities of response for the treatment-control "twins" with the biomarker configuration x. In the second stage, two alternative methods based on regression trees and classification techniques were considered to define the subpopulation S of patients who experienced enhanced treatment benefit. The authors defined a measure Q to quantify an enhanced treatment effect in the subgroup S as follows

$$Q(S) = (P(y = 1|t = T, x \in S) - P(y = 1|t = C, x \in S)) - p$$

where $p = P(y = 1|t = T) - P(y = 1|t = C)$ is a measure of the average treatment effect in the overall population. Several methods to estimate $Q(S)$ were proposed.

Kovalchik et al. (2013) derived a framework based on a proportional interactions model, where all treatment–biomarker interaction terms were assumed to be proportional to the main effects. To avoid model misspecification, a selection strategy taking all possible proportional interaction models into account was investigated. A modified Bonferroni correction for multiple testing was introduced. Zhao et al. (2013) proposed a parametric scoring system based on a patient's biomarker profile to estimate patient-specific treatment differences. Subgroups with an enhanced treatment effect consisted of patients whose estimated scores exceed a clinically relevant threshold.

Morita et al. (2014) compared two Bayesian trial designs, one based on subgroup analysis and the other on regression models for analyzing progression-free survival time. Both methods estimate Bayesian posterior probabilities of progression-free survival hazard ratios in the prespecified subgroups. For a setting where subgroups may be defined by several covariates, Varadhan and Wang (2014) proposed standardized marginal interaction models. With this approach, separate regression models for each covariate are fitted and, to account for confounding due to other covariates, the observations are appropriately re-weighted.

Freidlin (2005) suggested a two-step approach including a subgroup identification procedure and a subgroup evaluation step (adaptive signature design). Based on a logistic regression model, a biomarker signature is identified to define a targeted subgroup to be investigated in the second step. The design also allows for testing the treatment effect in the full population using data from both stages; however, the subgroup constructed in the first step is only tested with data from the second step. In Freidlin et al. (2010a) a cross-validated adaptive signature design was introduced as an extension to the adaptive signature design. The extension optimizes the efficiency of both the classifier development and the validation components of the design described above. The adaptive signature design uses two non-overlapping subsamples and the cross validation procedure allows a more efficient use of the trial population.

Several publications explored GTEM approaches in the context of subgroup identification. Jones et al. (2011) investigated several proposals to apply Bayesian regression models and shrinkage methods that directly estimate the treatment effect contrast. For a case study with eight subgroups arising from three binary biomarkers, they compared a model with no subgroup effect, a fully stratified model including all interaction terms, a simple regression model with first-order interactions only, a simple regression model including shrinkage, as well as proposals of Dixon and Simon (Dixon and Simon, 1991, 1992; Simon et al., 1996), where first-order interaction effects were

assumed to be exchangeable (and no higher-order interaction terms exist) and an extension including higher-order interaction terms.

Dusseldorp and Van Mechelen (2014) considered partitioning algorithms, termed qualitative interaction trees, resulting in a binary tree to identify qualitative treatment-by-biomarker interactions in the entire covariate space. At each partitioning step, a criterion which incorporates the treatment effect difference between subgroups as well as the subgroup sizes was considered to refine the subsets. A bootstrap algorithm was applied to prune the tree. Bonetti (2004) constructed simultaneous confidence intervals for treatment effects in subgroups defined by a single continuous covariate $x \in [x_{\min}, x_{\max}]$. The subgroups were defined by all patients for which the covariate exceeded or fell below a certain subgroup-specific threshold, i.e., $S_i = [x_i, x_{\max}]$ or $S_i = [x_{\min}, x_i]$, where x_i were pre-defined cutoffs. Alternatively, a "sliding window pattern" was proposed where the subgroups were defined by intervals of the covariate. The treatment effect estimators $\hat{\theta}_i$ and simultaneous confidence intervals were calculated within each subgroup. The authors developed plots of the estimates $\hat{\theta}_i$, termed subpopulation treatment effect pattern plots, together with their confidence bands versus the mean of the covariate x in the subpopulation S_i.

LM approaches were developed in Sivaganesan et al. (2011) who applied a Bayesian model selection approach to investigate treatment-by-subgroup interactions. For each covariate, a separate class of models was defined. Based on posterior probabilities computed for each model, inference on subgroup effects was made. Frequentist Type I and Type II error rates were controlled by adjusting the thresholds for the posterior probabilities.

Lipkovich et al. (2011) proposed a recursive partitioning algorithm called SIDES (Subgroup Identification based on Differential Effect Search) to identify patient subgroups with a differential treatment effect. The subgroups were defined using a step-wise procedure that started from the full population and partitioned the subgroups into increasingly smaller sets. For each parent subgroup S_P and each covariate x_i, two child subgroups, namely,

$$S_L(x_i, c_i) = \{x \in S_P : x_i \leq c_i\} \text{ and } S_H(x_i, c_i) = S_P \backslash S_L(x_i, c_i)$$

were constructed and the cutoff c_i was chosen by minimizing a prespecified splitting criterion. Several types of splitting criteria were proposed, for example, a differential effect between the two subgroups. The subgroup with the larger treatment effect was chosen as a candidate parent subgroup in the subsequent step if it satisfied additional restrictions, including restrictions on the number of child subgroups for a given parent, subgroup size, and magnitude of the treatment effect within the subgroup (complexity control). Tuning parameters of the subgroup search procedure were selected by cross-validation, and treatment effect p-values within the selected subgroups were adjusted to control the probability of incorrect subgroup discovery. A two-stage version of the SIDES procedure, known as the SIDEScreen procedure, was developed in Lipkovich and Dmitrienko (2014b). The first stage of this procedure induced a biomarker screen which selected the most promising biomarkers with high "variable importance", defined as a summary measure of a biomarker's predictive strength, and the regular SIDES procedure was applied in the second stage to the restricted set of biomarkers.

A Bayesian tree-based approach was developed by Berger et al. (2014) to identify subgroup effects. The subgroups were defined by the terminal nodes of the trees used to construct models for treatment effects and baseline covariates (the latter used for modeling prognostic effects). The model space consisted of all possible treatment and baseline regression models. Multiplicity adjustment was implemented by selecting prior distributions for the models.

4.2. *Response-adaptive designs*

Several authors (Zhou et al., 2008; Lee et al., 2010; Eickhoff et al., 2010; Zhong et al., 2013) investigated Bayesian adaptive designs where a global outcome model was applied to implement adaptations of treatment allocation proportions based on each patient's biomarker profiles. Zhou

et al. (2008), for example, considered a sequential multi-arm trial where after each patient response probabilities are estimated for each subgroup using a hierarchical Bayesian probit model. The allocation ratios in the randomization procedure are continuously updated and set to be proportional to the resulting current estimate of the response probabilities.

Berry et al. (2013) proposed a Bayesian hierarchical single-arm adaptive design for Phase II oncology trials to identify subgroups with an enhanced treatment effect. For every prespecified subgroup, they examined the hypothesis $H_0 : p \leq p_l$ versus $H_1 : p \geq p_h$, where p denotes the rate of tumor response and p_l and p_h are clinically relevant thresholds. They modelled the treatment effect assuming the difference between the log-odds of response and log-odds of the targeted threshold rate p_h for each subgroup to be normally distributed with unknown mean μ and variance σ^2. Following a hierarchical model approach, the parameters μ and σ^2 were assumed to be normally distributed. The proposed adaptive design included frequent interim analyses and stops for futility in a specific subgroup as soon as the posterior probability of a response rate larger than $(p_l + p_h)/2$ fell below a prespecified threshold. The design was compared to an adaptive design based on a (non-hierarchical) Bayesian model that treats each subgroup separately and is similar to the Simon's optimal two-stage design.

Simon and Simon (2013) proposed statistical tests for response adaptive enrichment designs that consist of the first stage where recruitment is not restricted and the second stage where the design allows one to continuously adapt the subpopulation under investigation. The procedure controls the Type I error rate for the test of the overall null hypothesis that no subpopulation benefits more from treatment than control. Applications of the test procedure to adaptive threshold enrichment designs, group sequential designs and several types of endpoints were discussed.

Gu et al. (2014) developed a Bayesian two-stage biomarker-based adaptive randomization design in the setting of the BATTLE-2 trial in non-small cell lung cancer. Four treatment groups were compared on the primary endpoint. In the first stage of the design, response adaptive randomization based on available outcome data and the biomarker KRAS was performed. Then an interim analysis with the option to stop treatment arms for futility was performed. Based on the first stage as well as external data, a predictive model combining several biomarkers and other predictive variables was derived. This model was then used in the second stage for a refined adaptive randomization algorithm. Finally, treatment effects, marker effects and the interactions were estimated and tested using data from both stages.

A further Bayesian approach for a subgroup based adaptive design, which utilized individual biomarker profiles and clinical outcome as they become available, was described in Xu et al. (2014). The main features of this design included the continuous re-classification of patient subgroups based on a random partition model and the random allocation of patients to the best treatment arm based on posterior predictive probabilities.

5. Clinical trial examples

In this section, we discuss selected case studies that were referred to in the methodological literature reviewed above.

An example for an enrichment trial design (F1) is the ToGA trial (trastuzumab for gastric cancer) (Bang et al., 2010), a multi–enter Phase III trial comparing trastuzumab (amonoclonal antibody against human epidermal growth factor receptor) in combination with chemotherapy with chemotherapy alone. Only patients with tumors showing overexpression of HER2 protein were considered for inclusion.

Examples for biomarker-stratified trials (F2) include the marker validation for erlotinib in the lung cancer (MARVEL) trial and cancer and leukemia group B (CALGB)-30506 trial (Freidlin et al., 2010b). In the MARVEL trial, patients were stratified by epidermal growth factor receptor gene (EGFR) status as measured by fluorescent in situ hybridization (FISH). After stratification, patients

were randomly assigned to erlotinib or pemetrexed. In the CALGB trial, patients were stratified by the lung metagene score and randomly assigned to either chemotherapy or present standard of care.

The ERCC1 trial (Cobo et al., 2007) and TCA (Cree et al., 2007) ovarian cancer trial provide examples for biomarker strategy trials (F3). In the ERCC1 trial, patients were randomized to control or an biomarker strategy arm. Patients in the control arm received docetaxel and cisplatin (standard of care). Participants in the biomarker strategy arm with low ERCC1 levels received docetaxel and cisplatin, whereas patients with high ERCC1 levels received docetaxel and gemcitabine. In the TCA (The Tumor Chemosensitivity Assay), ovarian cancer trial patients were randomized to a biomarker-guided arm using a chemosensitivity assay that measured ATP levels in drug-treated cancer cells.

The BATTLE trial is an example for a Bayesian response adaptive trial (Zhou et al., 2008). BATTLE is an umbrella trial consisting of four parallel Phase II studies. After a run-in phase, patients were adaptively randomized based on their biomarker status to one of the investigated treatments (Mandrekar and Sargent, 2011a).

Beckman et al. (2011) pointed out the importance of integrating biomarkers into clinical trials by, among others, the following two examples. Gefitinib is an example of a treatment that initially failed in the full population but was later shown to have a strong beneficial effect in patients with certain mutations. An example where the target population was chosen too small is cetuximab, an antibody directed against EGFR, which was expected to be effective only in patients with a certain biomarker profile. It was later discovered that the eligibility criteria were too restrictive and cetuximab was actually effective in a larger population. If however biomarkers are not predictive, including them in clinical trials might increase the costs, complexity and duration of trials (Beckman et al., 2011). To guarantee an efficient use of resources, exploratory clinical trials with analysis controlling false-positive rates may be useful to choose promising candidate biomarkers which are then further investigated in confirmatory clinical trials.

6. Discussion

In this review, we survey methodological papers investigating novel procedures to analyze the heterogeneity of treatment effects across patient subgroups in clinical trials. The systematic search was performed using medical statistics journals and, more generally, journals on clinical trial methodology indexed in PubMed. While PubMed has only partial coverage of some of the statistical journals considered, the choice of the PubMed database focused the search on methodological literature relevant for medical applications. Since some relevant papers of interest may not be listed in PubMed, we augmented the automated database search by including manuscripts from the list of references in the identified manuscripts as well as papers discovered via manual searches in the review.

Overall, we note that a broad range of methods have been developed for exploratory as well as confirmatory subgroup analysis methods. Many different statistical tools and approaches have demonstrated their utility, including multiple testing procedures, group-sequential and adaptive clinical trial designs, regression models, Bayesian hierarchical models, decision-theoretic models, machine learning algorithms (as, e.g., recursive partitioning), model selection algorithms, and shrinkage estimation.

In the confirmatory setting, the requirement of FWER control gives a clear frame work and facilitates the meaningful comparison of different approaches to trial design and analysis. However, there is a diversity of objectives that may be addressed in confirmatory trials with respect to the evaluation of treatment effects in relevant patient subgroups. There are several open issues where no agreement on the required level of evidence has been reached. Consider, for example, the setting where a treatment effect is claimed in the full population. Even though there is a broad consensus that the influence condition needs to be addressed to ensure that the overall treatment effect is not driven by a highly significant effect in a small subgroup, it is not clear how much evidence is required to claim efficacy in the full population.

In the exploratory setting, comparisons of available subgroup analysis/identification procedures are quite challenging, since the general goals of subgroup search vary from one class of methods to another. As an example, when considering a large number of potential subgroups in a typical subgroup search problem, tackling the multiplicity problem is important to improve the reproducibility of the results. However, strong FWER control may not be a useful concept in subgroup exploration as it leads to overly conservative procedures. Bayesian adjustments based on the concept of shrinkage present a viable alternative to traditional multiple testing procedures and have proved to be efficient tools in the subgroup identification setting, see Jones et al. (2011).

In this review, we focused on subgroups identified by biomarkers. Notably, the term enrichment designs has been used also for similar types of designs applied in chronic pain studies. Recently, Moore et al. (2015) identified 25 trials with such a design in chronic non-cancer pain. In these studies, subgroups are not identified based on a baseline biomarker but based on a pain rating (a surrogate variable) obtained during an initial open-label treatment with the drug under investigation. Because this surrogate variable is believed to be predictive for the treatment effect, only patients responding to drug (and tolerating it) are included in the second part of the study and randomized to the experimental drug or control. Strictly speaking, these so-called enriched enrollment randomized withdrawal designs (McQuay et al., 2008; Straube et al., 2008) test the effect of withdrawal of treatment. Alternatively, it has been proposed to treat all patients in an enrollment period with the intended control treatment (placebo or active control). Non-responders to control treatment are then identified, and this subpopulation is selected for randomization to new treatment or control, see e.g. FDA (2012). Here the non-responsiveness to control is used as surrogate variable that defines the study population of the randomized part of the study. An extension of this approach is provided by Sequential Multiple Assignment Randomized Trials (SMART) that investigate adaptive intervention schemes where treatment allocations at different treatment periods for an individual patient depend on outcome variables measured during previous periods of the same patient, see e.g. Almirall et al. (2012). Optimal treatment regimes and individualized treatment rules were developed in Zhang et al. (2012) and Zhao et al. (2012).

Funding

This work was conducted as part of the InSPiRe (Innovative methodology for small populations research) project funded by the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement number FP7 HEALTH 2013-602144.

References

Almirall, D., Compton, S.N., Gunlicks-Stoessel, M., Duan, N., Murphy, S.A. (2012). Designing a pilot sequential multiple assignment randomized trial for developing an adaptive treatment strategy. *Statistics in Medicine* 31:1887–1902.

Alosh, M., Huque, M.F. (2009). A flexible strategy for testing subgroups and overall population. *Statistics in Medicine* 28:3–23.

Alosh, M., Huque, M.F. (2013). Multiplicity considerations for subgroup analysis subject to consistency constraint: Multiplicity considerations for subgroup analysis. *Biometrical Journal* 55:444–462.

Altstein, L.L., Li, G., Elashoff, R.M. (2011). A method to estimate treatment efficacy among latent subgroups of a randomized clinical trial. *Statistics in Medicine* 30:709–717.

Bang, Y.J., Van Cutsem, E., Feyereislova, A., Chung, H.C., Shen, L., Sawaki, A., Lordick, F., Ohtsu, A., Omuro, Y., Satoh, T., Aprile, G., Kulikov, E., Hill, J., Lehle, M., Rüschoff, J., Kang, Y.K. (2010). Trastuzumab in combination with chemotherapy versus chemotherapy alone for treatment of her2-positive advanced gastric or gastro-oesophageal junction cancer (toga): a phase 3, open-label, randomised controlled trial. *The Lancet* 367:687–698.

Bauer, P., Kieser, M. (1999). Combining different phases in the development of medical treatments within a single trial. *Statistics in Medicine* 18:1833–1848.

Bauer, P., Koehne, K. (1994). Evaluation of experiments with adaptive interim analyses. *Biometrics* 50:1029–1041.

Bauer, P., Posch, M. (2004). Letter to the editor. Modification of the sample size and the schedule of interim analyses in survival trials based on data inspections by H. Schäfer and H.-H. Müller, *Statistics in Medicine* 2001; 20: 3741–3751. *Statistics in Medicine* 23:1333–1334.

Beckman, R.A., Clark, J., Chen, C. (2011). Integrating predictive biomarkers and classifiers into oncology clinical development programmes. *Nature Reviews Drug Discovery* 10:735–748.

Berger, J.O., Wang, X., Shen, L. (2014). A Bayesian approach to subgroup identification. *Journal of Biopharmaceutical Statistics* 24:110–129.

Berry, S.M., Broglio, K.R., Groshen, S., Berry, D.A. (2013). Bayesian hierarchical modeling of patient subpopulations: Efficient designs of phase II oncology clinical trials. *Clinical Trials* 10:720–734.

Boessen, R., van der Baan, F., Groenwold, R., Egberts, A., Klungel, O., Grobbee, D., Knol, M., Roes, K. (2013). Optimizing trial design in pharmacogenetics research: comparing a fixed parallel group, group sequential, and adaptive selection design on sample size requirements. *Pharmaceutical Statistics* 12:366–374.

Bonetti, M. (2004). Patterns of treatment effects in subsets of patients in clinical trials. *Biostatistics* 5:465–481.

Brannath, W., Bretz, F. (2010). Shortcuts for locally consonant closed test procedures. *Journal of the American Statistical Association* 105:660–669.

Brannath, W., Zuber, E., Branson, M., Bretz, F., Gallo, P., Posch, M., Racine-Poon, A. (2009). Confirmatory adaptive designs with Bayesian decision tools for a targeted therapy in oncology. *Statistics in Medicine* 28:1445–1463.

Bretz, F., Maurer, W., Brannath, W., Posch, M. (2009). A graphical approach to sequentially rejective multiple test procedures. *Statistics in Medicine* 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. *Biometrical Journal* 53:894–913.

Cai, T., Tian, L., Wong, P.H., Wei, L.J. (2011). Analysis of randomized comparative clinical trial data for personalized treatment selections. *Biostatistics* 12:270–282.

Chen, W., Ghosh, D., Raghunathan, T.E., Norkin, M., Sargent, D.J., Bepler, G. (2012). On Bayesian methods of exploring qualitative interactions for targeted treatment. *Statistics in Medicine* 31:3693–3707.

Cobo, M., Isla, D., Massuti, B., Montes, A., Sanchez, J.M., Provencio, M., Viñolas, N., Paz-Ares, L., Lopez-Vivanco, G., Muñoz, M.A., Felip, E., Alberola, V., Camps, C., Domine, M., Sanchez, J.J., Sanchez-Ronco, M., Danenberg, K., Taron, M., Gandara, D., Rosell, R. (2007). Customizing cisplatin based on quantitative excision repair cross-complementing 1 mRNA expression: A phase III trial in non-small-cell lung cancer. *Journal of clinical oncology* 25:2747–2754.

Cree, I.A., Kurbacher, C.M., Lamont, A., Hindley, A.C., Love, S. (2007). A prospective randomized controlled trial of tumour chemosensitivity assay directed chemotherapy versus physician's choice in patients with recurrent platinum-resistant ovarian cancer. *Anticancer drugs* 18:1093–1101.

Dixon, D.O., Simon, R. (1991). Bayesian subset analysis. *Biometrics* 47:871–881.

Dixon, D.O., Simon, R. (1992). Bayesian subset analysis in a colorectal cancer clinical trial. *Statistics in Medicine* 11:13–22.

Dmitrienko, A., D'Agostino, R.B. (2013). Tutorial in biostatistics: Traditional multiplicity adjustment methods in clinical trials. *Statistics in Medicine* 32:5172–5218.

Dusseldorp, E., Van Mechelen, I. (2014). Qualitative interaction trees: a tool to identify qualitative treatment-subgroup interactions. *Statistics in Medicine* 33:219–237.

Eickhoff, J.C., Kim, K., Beach, J., Kolesar, J.M., Gee, J.R. (2010). A Bayesian adaptive design with biomarkers for targeted therapies. *Clinical Trials* 7:546–556.

Eng, K.H. (2014). Randomized reverse marker strategy design for prospective biomarker validation. *Statistics in Medicine* 33:3089–3099.

FDA (2012). Guidance for industry: Enrichment strategies for clinical trials to support approval of human drugs and biological products. http://www.fda.gov.

Foster, J., Taylor, J., Ruberg, S. (2011). Subgroup identification from randomized clinical trial data. *Statistics in Medicine* 30:2867–2880.

Freidlin, B. (2005). Adaptive signature design: An adaptive clinical trial design for generating and prospectively testing a gene expression signature for sensitive patients. *Clinical Cancer Research* 11:7872–7878.

Freidlin, B., Jiang, W., Simon, R. (2010a). The cross-validated adaptive signature design. *Clinical Cancer Research* 16:691–698.

Freidlin, B., McShane, L.M., Korn, E.L. (2010b). Randomized clinical trials with biomarkers: Design issues. *JNCI Journal of the National Cancer Institute* 102:152–160.

Freidlin, B., McShane, L.M., Polley, M.Y.C., Korn, E.L. (2012). Randomized Phase II trial designs with biomarkers. *Journal of Clinical Oncology* 30:3304–3309.

Friede, T., Parsons, N., Stallard, N. (2012). A conditional error function approach for subgroup selection in adaptive clinical trials. *Statistics in Medicine* 31:4309–4320.

Graf, A.C., Posch, M., Koenig, F. (2015). Adaptive designs for subpopulation analysis optimizing utility functions. *Biometrical Journal* 57:76–89.

Gu, X., Chen, N., Wei, C., Liu, S., Papadimitrakopoulou, V.A., Herbst, R.S., Lee, J.J. (2014). Bayesian two-stage biomarker-based adaptive design for targeted therapy development. *Statistics in Biosciences*. Published online, doi:0.1007/s12561–014–9124–2.

Götte, H., Donica, M., Mordenti, G. (2014). Improving probabilities of correct interim decision in population enrichment designs. *Journal of Biopharmaceutical Statistics* 25:1020–1038.

Hommel, G. (2001). Adaptive modifications of hypotheses after an interim analysis. *Biometrical Journal* 43:581–589.

Huang, X., Ning, J., Li, Y., Estey, E., Issa, J.P., Berry, D.A. (2009). Using short-term response information to facilitate adaptive randomization for survival clinical trials. *Statistics in Medicine* 28:1680–1689.

Irle, S., Schäfer, H. (2012). Interim design modifications in time-to-event studies. *Journal of the American Statistical Association* 107:341–348.

Jenkins, M., Stone, A., Jennison, C. (2011). An adaptive seamless phase II/III design for oncology trials with subpopulation selection using correlated survival end-points†. *Pharmaceutical Statistics* 10:347–356.

Jones, H.E., Ohlssen, D.I., Neuenschwander, B., Racine, A., Branson, M. (2011). Bayesian models for subgroup analysis in clinical trials. *Clinical Trials* 8:129–143.

Kovalchik, S.A., Varadhan, R., Weiss, C.O. (2013). Assessing heterogeneity of treatment effect in a clinical trial with the proportional interactions model. *Statistics in Medicine* 32:4906–4923.

Krisam, J., Kieser, M. (2014). Decision rules for subgroup selection based on a predictive biomarker. *Journal of Biopharmaceutical Statistics* 24:188–202.

Lee, J.J., Gu, X., Liu, S. (2010). Bayesian adaptive randomization designs for targeted agent development. *Clinical Trials* 7:584–596.

Lipkovich, I., Dmitrienko, A. (2014a). Biomarker identification in clinical trials. In C. Carini, S. Menon, M. Chang, editors, *Clinical and Statistical Considerations in Personalized Medicine*. New York: Chapman and Hall/CRC Press.

Lipkovich, I., Dmitrienko, A. (2014b). Strategies for identifying predictive biomarkers and subgroups with enhanced treatment effect in clinical trials using sides. *Journal of Biopharmaceutical Statistics* 24:130–153.

Lipkovich, I., Dmitrienko, A., D'Agostino, R.B. (2015). Tutorial in biostatistics: Exploratory subgroup analysis in clinical trials. *Statistics in Medicine* 34:To appear.

Lipkovich, I., Dmitrienko, A., Denne, J., Enas, G. (2011). Subgroup identification based on differential effect search-a recursive partitioning method for establishing response to treatment in patient subpopulations. *Statistics in Medicine* 30:2601–2621.

Magirr, D., Jaki, T., Koenig, F., Posch, M. (2014). Adaptive survival trials. arXiv:1405.1569 [stat.AP]. http://arxiv.org/abs/1405.1569

Magnusson, B.P., Turnbull, B.W. (2013). Group sequential enrichment design incorporating subgroup selection. *Statistics in Medicine* 32:2695–2714.

Mandrekar, S.J., An, M.W., Sargent, D.J. (2013). A review of phase II trial designs for initial marker validation. *Contemporary Clinical Trials* 36:597–604.

Mandrekar, S.J., Sargent, D.J. (2009). Clinical trial designs for predictive biomarker validation: One size does not fit all. *Journal of Biopharmaceutical Statistics* 19:530–542.

Mandrekar, S.J., Sargent, D.J. (2011a). All-comers versus enrichment design strategy in phase II trials. *Journal of thoracic oncology: official publication of the International Association for the Study of Lung Cancer* 6:658.

Mandrekar, S.J., Sargent, D.J. (2011b). Design of clinical trials for biomarker research in oncology. *Clinical Investigation* 1:1627–1636.

Marcus, R., Peritz, E., Gabriel, K.R. (1976). On closed testing procedures with special reference to ordered analysis of variance. *Biometrika* 63:655–660.

McQuay, H.J., Derry, S., Moore, R.A., Poulain, P., Legout, V. (2008). Enriched enrolment with randomised withdrawal (eerw): time for a new look at clinical trial design in chronic pain. *Pain* 135:217–220.

Mehta, C., Schäfer, H., Daniel, H., Irle, S. (2014). Biomarker driven population enrichment for adaptive oncology trials with time-to-event endpoints. *Statistics in Medicine* 33:4515–4531.

Mehta, C.R., Gao, P. (2011). Population enrichment designs: case study of a large multinational trial. *Journal of Biopharmaceutical Statistics* 21:831–845.

Millen, B.A., Dmitrienko, A. (2011). Chain procedures: A class of flexible closed testing procedures with clinical trial applications. *Statistics in Biopharmaceutical Research* 3:14–30.

Millen, B.A., Dmitrienko, A., Ruberg, S., Shen, L. (2012). A statistical framework for decision making in confirmatory multipopulation tailoring clinical trials. *Therapeutic Innovation & Regulatory Science* 46:647–656.

Millen, B.A., Dmitrienko, A., Song, G. (2014). Bayesian assessment of the influence and interaction conditions in multipopulation tailoring clinical trials. *Journal of Biopharmaceutical Statistics* 24:94–109.

Moineddin, R., Butt, D.A., Tomlinson, G., Beyene, J. (2008). Identifying subpopulations for subgroup analysis in a longitudinal clinical trial. *Contemporary Clinical Trials* 29:817–822.

Moore, R.A., Wiffen, P.J., Eccleston, C., Dery, S.., Baron, R., Bell, R.F., Furlan, A.D., Gilron, I., Haroutounian, S., Katz, N.P., Lipman, A.G., Morley, S., Peloso, P.M., Quessy, S.N., Seers, K., Strassels, S.A., Straube, S. (2015). Systematic review of enriched-enrolment randomised-withdrawal trial designs in chronic pain: A new framework for design and reporting. *Pain* 156:1382–1395.

Morita, S., Yamamoto, H., Sugitani, Y. (2014). Biomarker-based Bayesian randomized phase II clinical trial design to identify a sensitive patient subpopulation. *Statistics in Medicine* 33:4008–4016.

Müller, H.H., Schäfer, H. (2004). A general statistical principle for changing a design any time during the course of a trial. *Statistics in Medicine* 23:2497–2508.

Proschan, M.A., Hunsberger, S.A. (1995). Designed extension of studies based on conditional power. *Biometrics* 51:1315–1324.

Simon, N., Simon, R. (2013). Adaptive enrichment designs for clinical trials. *Biostatistics* 14:613–625.

Simon, R., Dixon, D.O., Freidlin, B. (1996). Bayesian subset analysis of a clinical trial for the treatment of hiv infections. In D. A. Berry, D. K. Stangl, editors, *Bayesian Biostatistics*. New York: CRC Press.

Sivaganesan, S., Laud, P.W., Müller, P. (2011). A Bayesian subgroup analysis with a zero-enriched Polya urn scheme. *Statistics in Medicine* 30:312–323.

Song, J.X. (2014). A two-stage patient enrichment adaptive design in phase II oncology trials. *Contemporary Clinical Trials* 37:148–154.

Song, Y., Chi, G.Y.H. (2007). A method for testing a prespecified subgroup in clinical trials. *Statistics in Medicine* 26:3535–3549.

Spiessens, B., Debois, M. (2010). Adjusted significance levels for subgroup analyses in clinical trials. *Contemporary Clinical Trials* 31:647–656.

Stallard, N., Hamborg, T., Parsons, N., Friede, T. (2014). Adaptive designs for confirmatory clinical trials with subgroup selection. *Journal of Biopharmaceutical Statistics* 24:168–187.

Straube, S., Derry, S., McQuay, H.J., Moore, A.M. (2008). Enriched enrolment: definition of effects of enrichment and dose in trials of pregabalin and gabapentin in neuropathic pain. a systematic review. *British Journal of Clinical Pharmacology* 66:266–275.

Varadhan, R., Wang, S.J. (2014). Standardization for subgroup analysis in randomized controlled trials. *Journal of Biopharmaceutical Statistics* 24:154–167.

Wang, S.J., James Hung, H.M., O'Neill, R.T. (2009). Adaptive patient enrichment designs in therapeutic trials. *Biometrical Journal* 51:358–374.

Wang, S.J., O'Neill, R.T., Hung, H.M.J. (2007). Approaches to evaluation of treatment effect in randomized clinical trials with genomic subset. *Pharmaceutical Statistics* 6:227–244.

Wassmer, G., Dragalin, V. (2014). Designing issues in confirmatory adaptive population enrichment trials. *Journal of Biopharmaceutical Statistics* 25:651–669.

Xu, Y., Trippa, L., Müller, P., Ji, Y. (2014). Subgroup-based adaptive (suba) designs for multi-arm biomarker trials. *Statistics in Biosciences*, 1–2.

Zhang, B., Tsiatis, A.A., Davidian, M., Zhang, M., Laber, E.B. (2012). Estimating optimal treatment regimes from a classification perspective. *Statistics* 1:103–114.

Zhao, L., Tian, L., Cai, T., Claggett, B., Wei, L.J. (2013). Effectively selecting a target population for a future comparative study. *Journal of the American Statistical Association* 108:527–539.

Zhao, Y., Zeng, D., Rush, A.J., Kosorok, M.R. (2012). Estimating individualized treatment rules using outcome weighted learning. *Journal of the American Statistical Association* 107:1106–1118.

Zhao, Y.D., Dmitrienko, A., Tamura, R. (2010). Design and analysis considerations in clinical trials with a sensitive subpopulation. *Statistics in Biopharmaceutical Research* 2:72–83.

Zhong, W., Koopmeiners, J.S., Carlin, B.P. (2013). A two-stage Bayesian design with sample size reestimation and subgroup analysis for phase II binary response trials. *Contemporary Clinical Trials* 36:587–596.

Zhou, X., Liu, S., Kim, E.S., Herbst, R.S., Lee, J.J. (2008). Bayesian adaptive design for targeted therapy development in lung cancer – a step toward personalized medicine. *Clinical Trials* 5:181–193.

Ziegler, A., Koch, A., Krockenberger, K., Großhennig, A. (2012). Personalized medicine using DNA biomarkers: A review. *Human Genetics* 131:1627–1638.