Bayesian Approaches to Subgroup Analysis and Related Adaptive Clinical Trial Designs

CIARA NUGENT[1], WENTIAN GUO[2], PETER MÜLLER[1,⋆] , YUAN JI[3,⋆]

[1] University of Texas at Austin, Austin, TX, USA

[2] Laiya Consulting, Shanghai, China

[3] University of Chicago, Chicago, IL, USA

⋆ corresponding email yji@health.bsd.uchicago.edu

⋆ *PM and YJ contributed equally to the manuscript*

Abstract

PURPOSE. We review Bayesian and Bayesian decision theoretic approaches to subgroup analysis and applications to subgroup-based adaptive clinical trial designs. Subgroup analysis refers to inference about subpopulations with significantly distinct treatment effects.
METHODS. The discussion mainly focuses on inference for a benefiting subpopulation, that is, a characterization of a group of patients who benefit from the treatment under consideration more than the overall population. We introduce alternative approaches, and demonstrate them with a small simulation study. Then, we turn to clinical trial designs. When the selection of the interesting subpopulation is carried out as the trial proceeds, the design becomes an adaptive clinical trial design, using subgroup analysis to inform the randomization and assignment of treatments to patients. We briefly review some related designs.
RESULTS. We demonstrate how subgroup analysis can be carried out by different Bayesian methods, and discuss how they identify slightly different subpopulations.
CONCLUSION. There are a variety of approaches to Bayesian subgroup analysis. Practitioners should consider the type of subpopulations they are interested in, and choose their methods accordingly.

KEYWORDS: Bayesian decision problem; Set estimation; Subgroup analysis; Adaptive clinical trial design.

1 Introduction

Subgroup analysis is about understanding heterogeneity in patient populations and is typically focused on finding benefiting subgroups. That is, subgroup analysis is concerned with the question of whether a conclusion for the entire eligible patient population in a clinical trial remains valid for subpopulations of interest. To be specific, consider a two-arm randomized trial that is carried out to learn about the effectiveness of an experimental therapy versus control, and assume that no significant treatment effect can be established in the overall population. A natural follow-up question is whether the treatment might be effective in some subpopulation. The question is closely related to personalized optimal treatment selection as it is needed, for example, in precision oncology.

The investigation of hypotheses related to the notion that certain subgroups may benefit from certain therapies more than others necessitates a disciplined way to identify such subgroups, and to quantify the benefit to these groups. Dangers that investigators face in setting up subgroup analyses include: studies being under-powered to detect significant effects in subgroups, the issue of multiple testing, and the possibility for data dredging.

We focus on a Bayesian perspective, because it naturally addresses multiplicities through the use of priors, and because of its alignment with decision theoretic approaches [1], which provide an elegant way to address competing aims in subgroup analysis. First, we discuss in general terms some of the questions that arise in subgroup analysis, then present some approaches that have been developed in the Bayesian literature and compare them in a small simulation. We then continue with a discussion of subgroup-based designs in ongoing clinical trials, and end with a consideration of advantages and drawbacks of the presented approaches. The discussion is not intended to be a complete literature review. For a more technical recent review of subgroup analysis see [2].

2 Approaches to Subgroup Analysis

In the following review of Bayesian subgroup analysis methods we focus on three main questions. Firstly, how does an approach represent subgroups? A common way to represent subgroups is through the use of tree structures. This can be explicit, or implied by a restriction to dichotomized covariates. Tree structures provide a logical set of rules to determine whether a patient is in a subgroup. Many methods, including several approaches that we review below, use tree structures because of

their interpretability and ease of use. Secondly, we ask: what is the primary inference target? Finally, we look at how different approaches induce parsimony. Parsimony can be induced by restricting the model space. For example, the model space might only include models defined by one or two binary covariates, thereby only allowing subgroups defined by at most two covariates. Alternatively, parsimony can be induced through the use an explicit utility function. Table 1 briefly summarizes how the following methods compare in terms of these considerations. Although we introduce the methods as alternative approaches, they could also be thought of as case studies in different aspects of subgroup analysis. Elements of different methods could be combined as needed, particularly for the decision theoretic approaches.

We use the following notation. For patient i, let y_i denote an outcome that is assumed to be some measure of efficacy, let t_i denote a treatment indicator, and let $\boldsymbol{x_i} = (x_{i1}, \ldots, x_{iJ})$ denote a set of covariates. The covariates could potentially define subgroups.

2.1 Regression

One of the earliest discussions of Bayesian subgroup analysis is by Dixon and Simon [3] who approach subgroup analysis as inference on regression coefficients, specifically the coefficient for the interaction between a treatment indicator (t_i) and a covariate (x_{ij}). A significant value of an interaction coefficient defines a subgroup, as illustrated in Figure 1. In their original approach, Dixon and Simon place shrinkage priors centered at zero on these coefficients. While this shrinks the covariates for variables with no effect towards zero, it does not introduce sparsity. Sparsity could be enforced by putting alternate priors on these coefficients, such as spike and slab [4, 5, 6, 7], or non-local priors [8].

2.2 Model Selection

Here, subgroups are defined by competing models, where the covariates included in the winning model define the subgroup. The general setup starts with a list of possible models, and then compares models in a principled manner. In general, Bayesian model selection induces sparsity through prior distributions that are specified for each model. One can argue that the use of prior probability models implicitly introduces a complexity penalty by spreading out probability mass for additional parameters [9, 10]. By including the model choice in the prior specification, the approach naturally considers the entire model space at once, making all models directly comparable.

We consider two methods that use model selection to perform subgroup analysis [11, 12]. In [11] the model space is restricted to models that can be written as a tree defined by one or two covariates. By construction, this only allows for subgroups defined by up to two covariates, enforcing a minimum level of sparsity. Rather than trees, [12] uses clustering to define a model space. They consider families of models, where each family is defined by a single covariate. Covariates must be categorical, see the appendix for more details. Since only models within a family are directly comparable, the approach introduces a reasonable, but ad-hoc, rule based on thresholds for selecting models to report as formal subgroups. The thresholds can be tuned to control desired frequentist properties of the procedure, such as the rate of falsely choosing a model with a common overall treatment effect over a subgroup model. Since the algorithm may select multiple subgroups corresponding to different families, it can potentially be difficult to interpret the results, especially since models in different families are not directly comparable. In particular, if there is a significant amount of correlation between covariates, the algorithm is not designed to pick a clear winner.

2.3 Potential Outcomes

While many approaches to subgroup analysis implicitly include a notion of potential outcomes, some explicitly use potential outcomes as the input to the search for subgroups. Such methods first model the response under treatment ($y_i^{(1)}$) and control ($y_i^{(0)}$) for each individual. Subgroup analysis is then based on a regression of the estimated differences ($\delta(\boldsymbol{x})$) between the expected potential response $y_i^{(1)}$ and $y_i^{(0)}$, on patient covariates.

The above strategy is introduced in [13] as the Virtual Twins method (VT). They first use a random forest to estimate the potential outcomes $(y_i^{(0)}, y_i^{(1)})$ for each patient, and then input this into another tree-based regression model to identify an estimate $\hat{B} = \{\boldsymbol{x}: \delta(\boldsymbol{x}) > 0\}$ of the true benefiting subset $B = \{\boldsymbol{x}: \Delta(\boldsymbol{x}) > 0\}$. Here $\Delta(\boldsymbol{x})$ is the unknown true difference in expected potential outcomes. Subgroups are identified as the paths that lead to leaves of the tree that have an estimated treatment effect above zero, or above a specified threshold.

2.4 Decision Problems

Some approaches focus on the nature of a subgroup report as a decision problem, and introduce a utility function $u(a, \Theta, y)$ to characterize the

preference for a decision a, in this case the benefiting subgroup, under data y and a hypothetical truth, represented by the parameters Θ, in an assumed sampling model. The utility usually includes a preference for parsimony, but can also account for other aims, such as a preference for patient populations that are not well served by currently available treatments. The optimal subgroup report is then the one that is expected to maximize $u(a, \Theta, y)$. While we here introduce a decision theoretic choice of a subgroup report as an alternative to other methods, it could also be argued that it can be included as an additional step on top of any of the earlier considered approaches.

The approach taken in [14] defines a utility function that favors a subgroup with a large difference in treatment effect relative to the overall population; a large subgroup that includes many potential patients; and a subgroup that can be parsimoniously described by only a few covariates. Additionally, frequentist summaries are accounted for, and controlled, through tuning parameters in the utility function. Besides type-I and type-II errors related to selecting an overall null or an overall alternative hypothesis, meaning the absence of any treatment effect or the presence of a common treatment effect, several other frequentist summaries are of interest. For example, under a scenario with a true subgroup, [14] reports a false subgroup rate as the probability of reporting a subgroup effect that is different from the true subgroup effect, or a true subgroup rate as the probability of correctly reporting the true subgroup.

2.5 Subgroup as a Random Quantity

The nature of the true subgroup (B) as a function of the unknown parameters, i.e., as a random variable, is the focus in [15]. While B itself is unknown, bounds on B can be calculated. These bounds are represented by a credible pair of subsets (S, E). All patients with covariates described in S have a high probability of benefitting, and all patients with covariates that place them outside of E, i.e. in E^c, have a low probability of benefitting. In this way they define a credible region, bounded by S and E, that contains the true benefitting subgroup (B) with a prespecified probability. The approach directly handles the problem of multiplicities by using simultaneous inference across all $\boldsymbol{x}$. The discussion in [16] includes a straightforward algorithm to determine (S, E). A strength, and perhaps at the same time a limitation, is that S and E could be very complicated subsets of the covariate space that are not necessarily easy to communicate, and the joint report of (S, E)

could be difficult to interpret. The approach is generalized to multiple endpoints in [17] and to semiparametric models in [16].

3 A Simulation Example

We carry out a simulation study to compare implementations of the described methods. We refer to the approaches by the following acronyms: DS [3]; VT, Virtual Twins [13]; BW [11]; PU, Polya Urn [12]; BaPoFi, Bayesian Population Finding [14]; CS-I, Credible Subgroups (Inclusive Subset) [15]; and CS-E, Credible Subgroups (Exclusive Subset) [15].

3.1 Setup of the Simulation Study

Consider a patient population where three biomarkers of interest have been recorded for each patient. We assume that each biomarker in patient i is recorded as a binary variable $x_{ij} \in \{0,1\},\ j = 1,2,3$.

Simulation truth. We assume a sample size of $n = 300$. We generate the three biomarkers, assign patients to treatment, and generate a response based on a simulation truth. We consider two scenarios, one that has a predictive effect for biomarker 3, and another that has a predictive effect for biomarker 3 as well as a prognostic effect for biomarker 1. For both scenarios the true benefiting subgroup is $B = \{\boldsymbol{x}: x_3 = 1\}$. See the appendix for details. We carry out $M = 100$ repeat simulations, i.e., we repeatedly generate hypothetical data sets, go through the subgroup analyses, and record the results.

Summaries. To keep results comparable across methods we proceed as follows. Let $\boldsymbol{x} = (x_1, x_2, x_3)$ denote the covariates for a hypothetical future patient. For all possible values of $\boldsymbol{x}$ we record whether a patient would be include in the recommended subgroup $\hat{B}$. The results are summarized in table 2 and table 3. Under the simulation truth, patients with the following baseline covariates ($\boldsymbol{x} = (x_1, x_2, x_3)$) should be flagged: $\{(0,0,1), (1,0,1), (0,1,1), (1,1,1)\}$.

In addition to identifying subsets of benefiting patients, some methods allow a report of no treatment effect (H_0), meaning the treatment is not effective for anyone in the population; or an overall treatment effect (H_1), meaning that the treatment is equally effective for everyone in the population regardless of the presence or absence of a biomarker. Since H_1 was never chosen in this simulation, it is excluded from the table. See the appendix for more detail.

3.2 Interpretation

The results in tables 2 and 3 show that BaPoFi and CS-I performed the best in both scenarios. BaPoFi, in particular, correctly identified the benefiting subgroup, and only the benefiting subgroup, 100% of the time for both scenarios. All methods correctly identified the benefiting subgroup the vast majority of the time in both scenarios. However, other than BaPoFi and CS-I, they falsely identified non-benefiting patients as being part of the benefiting subgroup some percentage of the time. Additionally, all methods did slightly better in scenario 1 than scenario 2, as expected.

Some of the methods were more prone to identifying a larger subset. In particular, VT reported slightly inflated subgroups. This is probably because it allows all possible combinations, and does not include any penalty for identifying small or awkwardly described groups, or for lack of parsimony.

As noted, some of the methods explicitly only allow for subgroups defined by one covariate (BW), combinations of subgroups defined by one covariate (PU), or two covariates (BaPoFi). All these methods successfully chose a group defined by biomarker 3. When evaluating their performance, keep in mind that the simulation truth for both scenarios happens to fall within the restricted scope of these methods, with the exception of BW for scenario 2; the version of BW implemented in this simulation does not take into account prognostic effects, although it could be set up to do so.

Due to the nature of the models that are considered in BW, the model that usually wins is the model that finds an enhanced treatment effect for both patients with $x_{13} = 1$ as well as for patients with $x_{13} = 0$. This is not the same as an overall treatment effect, but rather a treatment effect for each group. In our scenarios, the effect for $x_{13} = 0$ was always close to zero, and the effect for $x_{13} = 1$ was much larger. BW considered a Bayesian Information Criterion (BIC) approximation to adjust for this. BaPoFi avoided this problem by explicitly including a penalty for complexity.

All of these methods can easily be extended to include more covariates, and can be modified to include interaction structures. However, because each method is geared towards a slightly different aim, it becomes difficult to directly compare them when the data becomes more complicated. Because of this, it becomes more important to specify the aim of the analysis. DS inference loses power due to multiplicity issues when many variables and interactions are included. The tree based methods in VT, in contrast, can easily incorporate many

variables and account for multiple interactions. However, results become harder to interpret the more variables one includes, and can easily become meaningless. A similar limitation applies to the CS method, with the addition that it is more computationally intensive than the other methods presented. The BW, PU and BaPoFi methods all have well defined model spaces, but limit the number of interactions that can be considered, if any; it may be necessary to define derived covariates to represent interactions of interest.

4 Adaptive Subgroup Enrichment Designs

Subgroup analysis is often carried out to inform the design of a future trial, or to inform decisions in an on-going trial. We briefly discuss such subgroup-based designs. It is useful to distinguish different types of subgroup-based designs, as shown in Figure 2.

4.1 Traditional Two-Step Process

The standard subgroup-driven clinical development consists of at least two clinical trials: an exploratory trial for subgroup finding, and a confirmatory trial for efficacy confirmation in the identified subgroup. If needed, more than two trials for further exploration or early confirmation may be included. At the end of each exploratory trial, any of the earlier discussed methods can be applied to identify subgroups. If evidence of the subgroup of interest is persuasive, the estimated subgroup from these completed exploratory clinical studies can be confirmed in the subsequent confirmatory trials.

Two different designs for the confirmatory trial may be considered depending on the different goals of different stake holders. The first is a one-stage subgroup enrichment design [18], in which one would prospectively specify the subgroups that the new therapy is expected to benefit. Such subgroup enrichment designs recruit patients from prespecified subpopulations and do not modify the inclusion criteria during the trial. The objective of such designs is to confirm the efficacy of the investigational therapy in the enriched subpopulation. On the other hand, a sponsor may be primarily interested in a large at-risk population, and may consider a biomarker-stratified design that recruits patients from a larger subpopulation, or even the whole patient population (denoted as F), and then includes patient strata (denoted as S). The objective of the biomarker-stratified designs is to demonstrate the efficacy of the treatment in F; and if the first objective fails, to demonstrate the efficacy of the treatment in S. As there are multiple

tests, a multiple testing procedure can be used to control the family-wise error rate for the null hypotheses of no-treatment effect in S and in F [19, 20, 21].

4.2 Confirmatory Adaptive Subgroup Enrichment Designs

To improve the probability of success in drug development and to speed up the process, several authors have proposed confirmatory adaptive subgroup enrichment designs for phase II/III trials [22, 23, 24, 25, 26, 27]. In such designs, the trial is divided into several stages with possible subpopulation enrichment after each interim analysis, such as restricting future enrollment to only those subgroups that appeared to be benefitting from the experimental therapy. Based on the results of each interim analysis, the trial may continue as initially planned, be stopped early, or continue with adaptive modifications, for example revision of recruitment inclusion criteria, or sample size re-estimation. In the absence of enrichment, statistical inference is conducted to test whether there is a significant treatment effect in the whole population or in any of the predetermined subgroups. If there is subgroup enrichment during the trial, the objective becomes to demonstrate that the treatment effect is significant in the enriched subgroup. Because these designs are under a confirmatory setting, they are all based on frequentist methods with a small set of predefined subgroups. The family-wise error rate is controlled by closed testing principles in the strong sense.

Bayesian tools can be used in the go/no-go decision for the original population or specific subgroups. For example, [23] developed a decision tool based on predictive probabilities for rejecting certain null hypotheses to determine which of the two populations, full population or subpopulation, should be further investigated in the second stage. This approach has been applied to an oncology phase II/III trial [28].

4.3 Seamless Exploratory and Confirmatory Trials

One may also consider combining the exploratory and the confirmatory trial seamlessly. The trial is initiated with a broad patient population without prespecified subgroups, with the chance of restricting the inclusion criteria to a certain subgroup that is adaptively learned from the accumulating data in the trial. For example, the method proposed in [26] can also be applied to a confirmatory trial without prespecified subgroups. It can identify subgroups from the interim analysis after the first stage. The approach is statistically valid and flexible, however the first-stage data can not be used in the final analysis. The discussion in

[26] recommends that "for regulatory submissions […] the subgroups and decision rule be prespecified."

In [29] a class of adaptive subgroup enrichment designs is proposed that adaptively update the eligibility criteria without prespecified subgroups. The method is based on the construction of a clever frequentist test of the overall null hypothesis that no subpopulation benefits more from treatment than control. This test preserves the type I error regardless of the method used for making enrichment decisions and regardless of, possibly data dependent, time trends in the characteristics of the patients. If the overall null hypothesis is rejected, the subgroup estimation will be reported at the final analysis for regulatory labeling. [30] extended [29] by incorporating Bayesian decision tools into the interim decision making for enrichment and Bayesian inference for the treatment effect in the estimated subgroup. Though the hypothesis testing is statistically sound in [29] and [30], the treatment effect in the estimated subgroup may be subject to a resubstitution bias. To correct for the resubstitution bias in estimating treatment efficacy for the selected subgroup [31] propose cross-validation and bootstrap methods.

4.4 Exploratory Adaptive Enrichment Designs

Some recent literature has developed exploratory adaptive enrichment designs [32, 33] with response-adaptive randomization. The designs have the objective to assign patients in the trial to more desirable treatment arms as well as learning subgroups. These designs are a perfect fit for phase II umbrella trials or platform trials with multiple treatment arms. Xu et. al. [32] proposed SUBA, a Bayesian subgroup-based adaptive design using a random partition model for continuous biomarkers ($\boldsymbol{x}$) and binary outcome (y). The main contribution of SUBA is that it allows estimation of the cutoff values that define subgroups for continuous biomarkers. See the appendix for more details. Figure 3 illustrates the design. SCUBA [33] extends SUBA by allowing a more flexible partition of the continuous biomarkers. Both, SUBA and SCUBA can be used for subgroup identification after the trial is completed. For an example of SUBA applied to a real world trial see [34]. This is an ongoing trial at NorthShore University Health System, which adopts SUBA for adaptive allocation of patients to the best treatment arm based on posterior predictive probabilities.

Another adaptive enrichment design for a basket trial of targeted therapies across multiple cancers is discussed in [35]. The design is based on an approach similar to [14], with subgroups defined by binary covariates that include biomarkers and tumor types.

5 Conclusion

We have reviewed several approaches to Bayesian subgroup analysis. An important consideration in choosing among different approaches is the intended aim of the subgroup analysis. As the simulation showed, some approaches are better at finding parsimonious groups, such as those presented as model selection and those combined with a utility function. This might be important when the subgroup analysis is carried out to plan the eligible population for a future trial. Utility functions also have the added benefit that they can account for other considerations that are not captured by enhanced treatment effects alone. Alternatively, an investigator might be mainly interested in an informed individualized treatment choice, in which case an approach that allows for more complex subgroups, such as the potential outcomes approach, might be preferred. Focusing on the nature of the subgroup as a random quantity enables investigators to explicitly describe the uncertainty of the subgroup report, and make an informed choice to be more or less conservative.

A common context for subgroup analysis is the planning of future trials, or the closely related problem of adaptation in a current trial. We discussed some examples. An interesting aspect of this application is that one can use Bayesian inference to find interesting subpopulation, but could then proceed with a purely frequentist design for the trial.

Planning for future trials is not the only application. Subgroup analysis might also be carried out as data analysis for previously completed trials, particularly where different trials demonstrate conflicting conclusions. Subgroup analysis and enrichment designs should be used for exploratory purposes, unless subgroups are predefined and a rigorous multiplicity adjustment plan is in place. One of the most difficult problems in statistics is adjusting for multiplicity, and subgroup analysis is vulnerable to explicit as well as implicit multiple comparisons. For example, when estimating the biomarker cutoff values, implicit multiplicity could affect final analysis results. We recommend conducting careful simulations to check the operating characteristics of the subgroup analysis methods or enrichment designs.

References

[1] Robert C. *The Bayesian Choice*. Springer-Verlag: New York 2007.

[2] Jones HE, Ohlssen DI, Neuenschwander B, Racine A, Branson M. Bayesian models for subgroup analysis in clinical trials. *Clinical Trials.* 2011;8:129-143.

[3] Dixon DO, Simon R. Bayesian subset analysis. *Biometrics.* 1991;47:871-881.
[4] Mitchell TJ, Beauchamp JJ. Bayesian variable selection in linear regression. *Journal of the American Statistical Association.* 1988;83:1023–1032.
[5] Madigan D, Raferty AE. Model selection and accounting for model uncertainty in graphical models using Occam's window. *Journal of the American Statistical Association.* 1994;89:1535–1546.
[6] George EI, McCulloch RE. Approaches for Bayesian variable selection. *Statistica Sinica.* 1997;7:339–373.
[7] Ishwaran H, Rao JS. Spike and slab variable selection: frequentist and Bayesian strategies. *The Annals of Applied Statistics.* 2005;33:730–773.
[8] Johnson VE, Rossell D. On the use of non-local prior densities in Bayesian hypothesis tests. *Journal of the Royal Statistical Society. Series B (Statistical Methodology).* 2010;72:143–170.
[9] Jefferys WH, Berger JO. Ockham's razor and Bayesian analysis. *American Scientist.* 1992;80:64–72.
[10] Scott JG, Berger JO. Bayes and empirical-Bayes multiplicity adjustment in the variable-selection problem. *The Annals of Statistics.* 2010;38:2587–2619.
[11] Berger JO, Wang X, Shen L. A Bayesian approach to subgroup identification. *Journal of Biopharmaceutical Statistics.* 2014;24:110-129.
[12] Sivaganesan S, Laud PW, Mueller P. A Bayesian subgroup analysis with a zero-enriched Polya Urn scheme. *Statistics in Medicine.* 2011;30:312-323.
[13] Foster JC, Taylor JM, Ruberg SJ. Subgroup identification from randomized clinical trial data. *Statistics in Medicine.* 2011;30:2867-2880.
[14] Morita S, Müller P. Bayesian population finding with biomarkers in a randomized clinical trial. *Biometrics.* 2017;73:1355-1365.
[15] Schnell PM, Tang Q, Offen WW, Carlin BP. A Bayesian credible subgroups approach to identifying patient subgroups with positive treatment effects. *Biometrics.* 2016;72:1026-1036.
[16] Schnell PM, Müller P, Tang Q, Carlin BP. Multiplicity-adjusted semiparametric benefiting subgroup identification in clinical trials. *Clinical Trials.* 2018;15:75-86.
[17] Schnell P, Tang Q, Müller P, Carlin BP. Subgroup inference for multiple treatments and multiple endpoints in an Alzheimerâ€™s disease treatment trial. *The Annals of Applied Statistics.* 2017;11:949–966.
[18] Ondra T, Dmitrienko A, Friede T, et al. Methods for identification and confirmation of targeted subgroups in clinical trials: A systematic review. *Journal of Biopharmaceutical Statistics.* 2016;26:99-119.
[19] Bauer P, Kohne K. Evaluation of experiments with adaptive interim analyses. *Biometrics.* 1994;50:1029–1041.
[20] Bauer P, Kieser M. Combining different phases in the development of medical treatments within a single trial. *Statistics in Medicine.* 1999;18:1833–1848.
[21] Hommel G. Adaptive modifications of hypotheses after an interim analysis. *Biometrical Journal: Journal of Mathematical Methods in Biosciences.* 2001;43:581–589.
[22] Wang SJ, James Hung H, O'Neill RT. Adaptive patient enrichment designs in therapeutic trials. *Biometrical Journal: Journal of Mathematical Methods in Biosciences.* 2009;51:358–374.

[23] Brannath W, Zuber E, Branson M, et al. Confirmatory adaptive designs with Bayesian decision tools for a targeted therapy in oncology. *Statistics in Medicine.* 2009;28:1445–1463.
[24] Jenkins M, Stone A, Jennison C. An adaptive seamless phase II/III design for oncology trials with subpopulation selection using correlated survival endpoints. *Pharmaceutical statistics.* 2011;10:347–356.
[25] Magnusson BP, Turnbull BW. Group sequential enrichment design incorporating subgroup selection. *Statistics in Medicine.* 2013;32:2695–2714.
[26] Mehta CR, Gao P. Population enrichment designs: case study of a large multinational trial. *Journal of Biopharmaceutical Statistics.* 2011;21:831-845.
[27] Friede T, Parsons N, Stallard N. A conditional error function approach for subgroup selection in adaptive clinical trials. *Statistics in Medicine.* 2012;31:4309–4320.
[28] Martín M, Chan A, Dirix L, et al. A randomized adaptive phase II/III study of buparlisib, a pan-class I PI3K inhibitor, combined with paclitaxel for the treatment of HER2–advanced breast cancer (BELLE-4). *Annals of Oncology.* 2016;28:313–320.
[29] Simon N, Simon R. Adaptive enrichment designs for clinical trials. *Biostatistics.* 2013;14:613-625.
[30] Simon N, Simon R. Using Bayesian modeling in frequentist adaptive enrichment designs. *Biostatistics.* 2018;19:27-41.
[31] Zhang Z, Chen R, Soon G, Zhang H. Treatment evaluation for a data-driven subgroup in adaptive enrichment designs of clinical trials. *Statistics in Medicine.* 2018;37:1–11.
[32] Xu Y, Trippa L, Müller P, Ji Y. Subgroup-based adaptive (SUBA) designs for multi-arm biomarker trials. *Statistics in Biosciences.* 2016;8:159–180.
[33] Guo W, Ji Y, Catenacci DV. A subgroup cluster-based Bayesian adaptive design for precision medicine. *Biometrics.* 2017;73:367.
[34] Simon KC, Tideman S, Hillman L, et al. Design and implementation of pragmatic clinical trials using the electronic medical record and an adaptive design. *JAMIA Open.* 2018;1:99–106.
[35] Xu Y, Müller P, Tsimberidou AM, Berry D. A nonparametric Bayesian basket trial design. *Biometrical Journal.* 2018;0.

Table 1: Comparison of different approaches to Bayesian subgroup analysis by the implied shrinkage, the inference target, and how they consider parsimony.

Shrinkage	Inference Target	Parsimony
2.1. Regression:		
priors on the treatment and treatment × covariate interaction coefficients	treatment × covariate interaction coefficients	no explicit criterion, though could be modified through specifying alternate priors
2.2. Model selection:		
shrinkage prior on model parameters	competing models	restricting the model space
2.3. Potential outcome framework:		
priors for the mean outcomes in the leaves of the tree	enhanced treatment effect	none
2.4. Decision problem:		
implicit in the underlying probability model	optimal subgroup report (action)	utility function and restricting the model space
2.5. Random Quantity:		
implicit in the underlying probability model	a random subset (B) in the covariate space	none

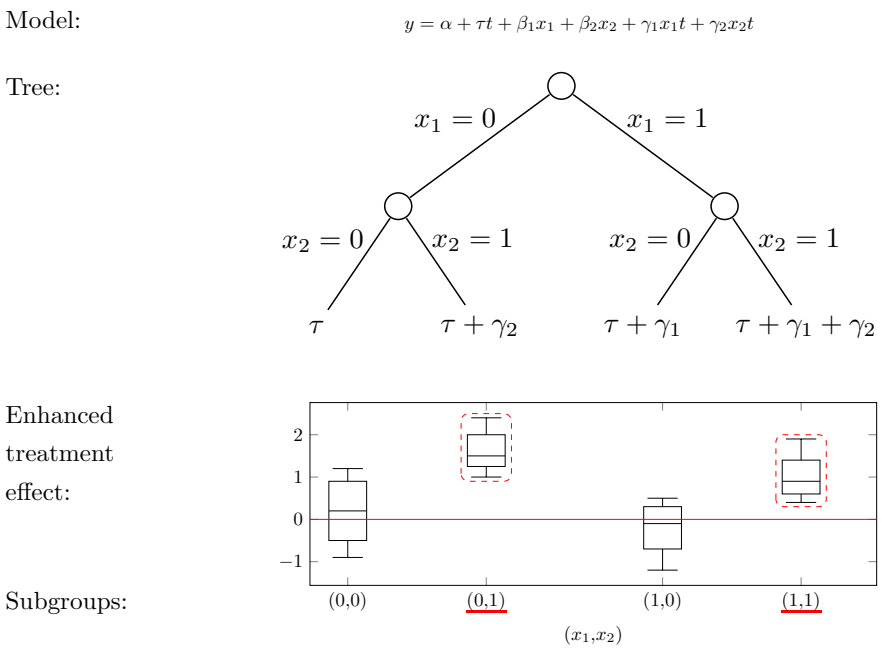

Figure 1: Dixon and Simon approach for two binary covariates. The model is written at the top, the patient index (i) is omitted for clarity. Below, there is an illustration of the inherent tree structure. The terminal nodes show the estimated treatment effect for a patient in each of the four subgroups. For each combination of covariates $\boldsymbol{x_i} = (x_{i1}, x_{i2})$ the boxplot shows the distribution of the treatment effect $E(y_i|\boldsymbol{x_i}, t_i = 1, data) - E(y\_i|\boldsymbol{x_i}, t_i = 1, data)$, estimated under the posterior distribution (that is, conditioning on the observed data). In this figure the enhanced treatment effects for the winning subgroups are circled in red, and the subgroups are underlined.

Table 2: Comparison of Methods. The table shows the probability (under repeat simulation) of $\boldsymbol{x} \in \hat{B}$ for scenario 1, that is a future patient with the indicated baseline covariates $\boldsymbol{x}$ being included in the reported subgroup. For VT, the first number corresponds to a cutoff of 0.1, while the number in parentheses corresponds to a cutoff of 0.05. For BW the first number refers to all patients who would be recommended for treatment, and the number in parentheses refers to the patients who would be designated as benefiting most. The column $E(n)$ shows the expected number of patients with the respective covariates according to the simulation truth. Covariate combinations that correspond to an enhanced treatment effect under the simulation truth are marked in bold face.

x	E(n)	DS	VT	BW	PU	BaPoFi	CS-I	CS-E
Overall Treatment Effect		0.01	0.00 (0.00)	0.00 (0.00)	0.00	0.00	0.00	0.00
(0,0,0)	75	0.01	0.00 (0.02)	0.24 (0.00)	0.00	0.00	0.00	0.03
(1,0,0)	25	0.03	0.06 (0.13)	0.24 (0.00)	0.03	0.00	0.00	0.05
(0,1,0)	75	0.02	0.02 (0.09)	0.24 (0.00)	0.06	0.00	0.00	0.06
(1,1,0)	25	0.03	0.05 (0.18)	0.24 (0.00)	0.06	0.00	0.00	0.05
(0,0,1)	37.5	1.00	1.00 (1.00)	1.00 (1.00)	1.00	1.00	0.99	1.00
(1,0,1)	12.5	1.00	1.00 (1.00)	1.00 (1.00)	1.00	1.00	0.92	1.00
(0,1,1)	37.5	1.00	1.00 (1.00)	1.00 (1.00)	1.00	1.00	0.99	1.00
(1,1,1)	12.5	1.00	1.00 (1.00)	1.00 (1.00)	1.00	1.00	0.95	1.00

Table 3: Comparison of Methods. The table shows the probability (under repeat simulation) of $\boldsymbol{x} \in \hat{B}$ for scenario 2, that is a future patient with the indicated baseline covariates $\boldsymbol{x}$ being included in the reported subgroup. For VT, the first number corresponds to a cutoff of 0.1, while the number in parentheses corresponds to a cutoff of 0.05. For BW the first number refers to all patients who would be recommended for treatment, and the number in parentheses refers to the patients who would be designated as benefiting most. The column $E(n)$ shows the expected number of patients with the respective covariates according to the simulation truth. Covariate combinations that correspond to an enhanced treatment effect under the simulation truth are marked in bold face.

x	E(n)	DS	VT	BW	PU	BaPoFi	CS-I	CS-E
Overall Treatment Effect		0.01	0.00 (0.00)	0.00 (0.00)	0.00	0.00	0.00	0.00
(0,0,0)	75	0.00	0.00 (0.02)	0.32 (0.00)	0.00	0.00	0.00	0.03
(1,0,0)	25	0.06	0.12 (0.20)	0.33 (0.02)	0.08	0.00	0.00	0.09
(0,1,0)	75	0.00	0.03 (0.09)	0.32 (0.00)	0.10	0.00	0.00	0.04
(1,1,0)	25	0.08	0.12 (0.22)	0.33 (0.02)	0.10	0.00	0.00	0.07
(0,0,1)	37.5	1.00	0.99 (1.00)	0.99 (0.98)	1.00	1.00	0.99	1.00
(1,0,1)	12.5	1.00	1.00 (1.00)	1.00 (1.00)	1.00	1.00	0.96	1.00
(0,1,1)	37.5	1.00	1.00 (1.00)	0.99 (0.98)	1.00	1.00	0.98	1.00
(1,1,1)	12.5	1.00	1.00 (1.00)	1.00 (1.00)	1.00	1.00	0.99	1.00

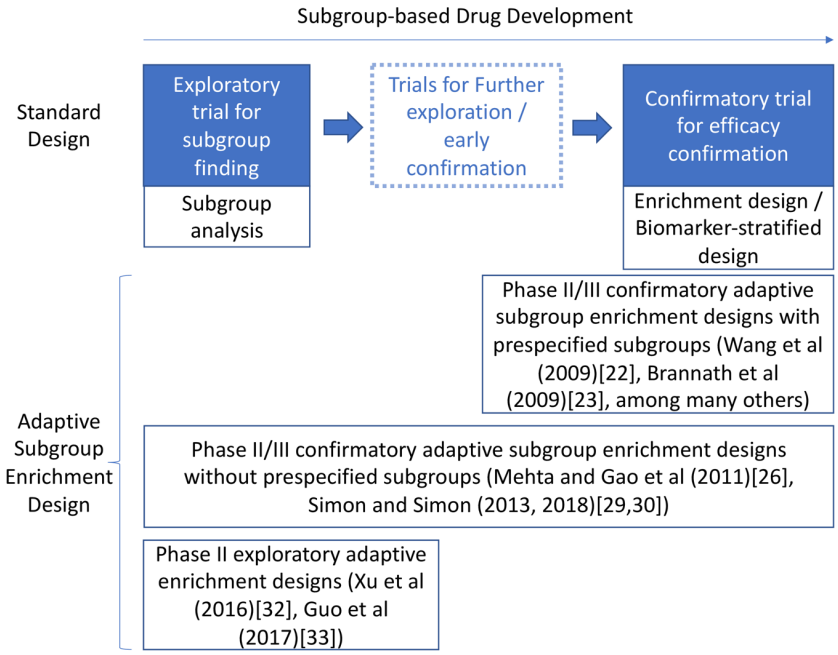

Figure 2: Overview of different subgroup-based designs.

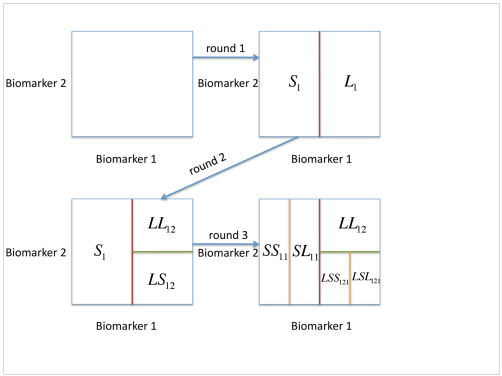

Figure 3: SUBA: subsets of the covariate space defined by recursive tree-like splits.