

Journal of Biopharmaceutical Statistics

ISSN: 1054-3406 (Print) 1520-5711 (Online) Journal homepage: https://www.tandfonline.com/loi/lbps20

Studying treatment-effect heterogeneity in precision medicine through induced subgroups

Aniek Sies, Koen Demyttenaere & Iven Van Mechelen

To cite this article: Aniek Sies, Koen Demyttenaere & Iven Van Mechelen (2019) Studying treatment-effect heterogeneity in precision medicine through induced subgroups, Journal of Biopharmaceutical Statistics, 29:3, 491-507, DOI: 10.1080/10543406.2019.1579220

To link to this article: https://doi.org/10.1080/10543406.2019.1579220

View supplementary material

Published online: 22 Feb 2019.

Submit your article to this journal

Article views: 105

View related articles

View Crossmark data

Full Terms & Conditions of access and use can be found at
https://www.tandfonline.com/action/journalInformation?journalCode=lbps20

JOURNAL OF BIOPHARMACEUTICAL STATISTICS
2019, VOL. 29, NO. 3, 491–507
https://doi.org/10.1080/10543406.2019.1579220

Taylor & Francis
Taylor & Francis Group

Check for updates

Studying treatment-effect heterogeneity in precision medicine through induced subgroups

Aniek Sies [a], Koen Demyttenaere[b], and Iven Van Mechelen[a]

[a]Faculty of Psychology and Educational Sciences, KU Leuven, Leuven, Belgium; [b]Department of Neurosciences, KU Leuven, Leuven, Belgium

ABSTRACT

Precision medicine, in the sense of tailoring the choice of medical treatment to patients' pretreatment characteristics, is nowadays gaining a lot of attention. Preferably, this tailoring should be realized in an evidence-based way, with key evidence in this regard pertaining to subgroups of patients that respond differentially to treatment (i.e., to subgroups involved in treatment–subgroup interactions). Often a-priori hypotheses on subgroups involved in treatment–subgroup interactions are lacking or are incomplete at best. Therefore, methods are needed that can induce such subgroups from empirical data on treatment effectiveness in a *post hoc* manner. Recently, quite a few such methods have been developed. So far, however, there is little empirical experience in their usage. This may be problematic for medical statisticians and statistically minded medical researchers, as many (nontrivial) choices have to be made during the data-analytic process. The main purpose of this paper is to discuss the major concepts and considerations when using these methods. This discussion will be based on a systematic, conceptual, and technical analysis of the type of research questions at play, and of the type of data that the methods can handle along with the available software, and a review of available empirical evidence. We will illustrate all this with the analysis of a dataset comparing several anti-depressant treatments.

ARTICLE HISTORY
Received 20 December 2017
Accepted 1 December 2018

KEYWORDS
Precision medicine; treatment-effect heterogeneity; exploratory subgroup analysis; treatment–subgroup interactions; clinical trials

1. Introduction

During the past years, precision medicine has gained a great deal of attention. Moreover, it has been encouraged by politicians, regulatory instances, and research organizations (FDA 2013; Hamburg and Collins 2010; Horizon 2020 2020; The White House 2015). The objectives of precision medicine relate to the customization of health care, including medical decisions, practices, and products, to patients' pretreatment characteristics. In this paper, we will focus on customization of treatment choice specifically.

Preferably, tailoring treatment choice to patients' pretreatment characteristics should be realized in an evidence-based way. For this purpose, evidence regarding treatment-effect heterogeneity (Kravitz et al. 2004; Rothwell 1995; Tunis et al. 2010) is crucial, with the effect of a treatment being heterogeneous if patients vary in their response to it. This implies that when two treatment alternatives are being compared, the relative treatment effect (i.e., the difference in effect between the two alternatives) differs across patients. We can further draw a distinction between quantitative heterogeneity, referring to a difference in magnitude of the relative treatment effect, and qualitative heterogeneity, referring to a difference in sign of the relative treatment effect (Bailey 1994). In case of

CONTACT Aniek Sies aniek.sies@kuleuven.be Faculty of Psychology and Educational Sciences, KU Leuven, Tiensestraat 102, box 3713, Leuven 3000, Belgium

Supplemental data for this article can be accessed here.

© 2019 Taylor & Francis Group, LLC

two treatments, the latter means that for some patients Treatment 1 outperforms Treatment 2 (and, hence, that the relative treatment effect of Treatment 1 is positive), whereas for other patients, the reverse is true (implying that the relative treatment effect of Treatment 1 is negative).

Evidence about treatment-effect heterogeneity can be gathered by studying subgroups of patients that respond differentially to treatment. Formally speaking, this comes down to studying subgroups that are involved in treatment–subgroup interactions. No seldom this is done by defining subgroups a priori, and by including these in subsequent analyses (e.g., Cohen et al. (2013); Cronbach and Snow (1977); Shaffer (1991)). In quite a few cases, however, a-priori hypotheses on relevant subgroups are lacking or incomplete (e.g., in the sense that one may have a conjecture about which variables are involved in the subgroup definitions but not about the exact way in which this is the case) (Boonacker et al. 2011; Byar 1985). Under such circumstances, traditional approaches are insufficient, and new methods are needed that can induce subgroups involved in treatment–subgroup interactions from data on treatment effectiveness in a *post hoc* manner. Recently, quite a few such methods have been developed (e.g., Zhang et al. (2015); Laber and Zhao (2015); Zhao et al. (2015); Dusseldorp and Van Mechelen (2014); Zhang et al. (2012a); Lipkovich et al. (2011); Qian and Murphy (2011); Su et al. (2009); Zeileis et al. (2008)).

Lipkovich et al. (2017) provided a comprehensive overview of many of these methods and characterized the relations between them. Still, medical statisticians and statistically minded medical researchers may face quite a few challenges when applying the methods in question in practice, as there is little empirical experience in their usage, while there are many (nontrivial) choices to be made. Examples include choices with regard to selecting an appropriate method, setting tuning parameters, and (formally) evaluating and interpreting the output. The main purpose of this paper is to discuss the major concepts and considerations when using these methods. We will do so based on three different aspects: (1) a conceptual analysis of the type of research question that may be at play; (2) an analysis of the types of data that a method can handle and an overview of available software; and (3) a systematic review of specific comparative empirical evidence regarding the performance of the different methods under study.

In the remainder of this paper, we will first describe the choices the researcher is faced with, along with a few major considerations regarding them (Section 2). In Section 3, we will illustrate the choices and considerations with a dataset comparing anti-depressant treatments. The paper will end with some concluding remarks in Section 4.

2. Choices in applying methods to induce subgroups involved in treatment–subgroup interactions from the data

To be able to tailor treatment choice to patients' pretreatment characteristics using one of the methods to induce subgroups from the data, the researcher must make a large number of choices. Here, we distinguish between choices with regard to (a) which method to choose (Section 2.1), and (b) analyzing the data (Section 2.2).

2.1. *Choosing a method*

When a researcher is deciding on which method to use, four factors are important: The research question that has to be answered (Section 3.1.1), the data that are available (Section 3.1.2), the performance of possibly suitable method(s) (Section 3.1.3), and whether or not software or source code is publicly available (Section 3.1.4).

2.1.1. *The research question*

Carefully specifying the research question at the start of a study is of utmost importance to avoid committing a Type III error (i.e., the error of giving the right answer to the wrong problem (Kimball 1957)). To give the right answer to the right problem, the research question has to be well formulated so as

to allow the researcher to choose an appropriate method to answer it. At this point, extensive and iterative deliberation between substantial researchers and statisticians is of great importance (Kimball 1957).

The focus of the present paper is the study of treatment-effect heterogeneity by inducing subgroups involved in treatment–subgroup interactions from the data during the data-analytic process, with the subgroups in question being defined in terms of their associated pattern of pretreatment patient characteristics. Within this context, we may distinguish three types of research questions: (a) Can we identify one or more specific subgroup(s) of patients for whom some (new) treatment alternative works much better or much worse than the other (standard) treatment alternative(s)? (In statistical terms, this implies the search for subgroups of patients with an enhanced relative treatment effect.) (b) Can we capture possible treatment-effect heterogeneity by identifying subgroups with different relative treatment effects, with as research question: Can we partition the patients into different subgroups within each of which the difference in response to one treatment alternative as compared to the other treatment alternative(s) is as similar as possible for all patients (and dissimilar from the corresponding difference in the other subgroups)? (In statistical terms, this implies the search for subgroups involved in treatment–subgroup interactions.) (c) Can we identify subgroups underlying an optimal treatment regime (i.e., a rule that specifies which treatment alternative is better for which subgroup of patients), with optimality referring to maximizing the expected potential outcome if the regime would be used for treatment assignment of all patients in the population if interest? (In statistical terms, this implies the estimation of a subgroup-based decision rule or treatment regime, with a number of subgroups differing regarding which treatment alternative works better, possibly in addition to subgroups for which the two (or more) best alternatives work about equally well.) Note that, in the more general context of studying treatment-effect heterogeneity without necessarily relying on subgroups, other questions than the ones distinguished above may be of interest, such as "what is the proportion of patients that benefits from treatment" or "what is the amount of overall treatment effect heterogeneity" (for more examples of such questions, Henderson et al. (2017)).

Table 1 provides an (non-exhaustive) overview of existing subgroup-based methods to study treatment-effect heterogeneity, sorted by the research question that each method primarily targets. Note that although many methods can also be used to answer other types of research questions than the one they primarily target, this will often lead to suboptimal results. For example, one could use a method that is primarily targeted at finding subgroups with similar relative treatment effects to estimate a treatment regime. Yet, this will often result in a treatment regime that does not maximize the expected outcome; moreover, the result may include subdivisions into subgroups for which the relative treatment effect is different in magnitude only, and which are therefore all associated with the same optimal treatment alternative.

Additionally, depending on the problem at hand and the researcher's goals, the research question may include the wish to obtain a merely pragmatic subdivision of the patients in subgroups that, for example, further constitute the basis of an effective treatment regime in absence of a simple characterization of the subgroups involved; alternatively, the research question may include the wish to obtain a clear characterization of the subgroups underlying treatment-effect heterogeneity in order to better understand this heterogeneity. For each of these goals, subgroup specifications with different forms may be optimal. In this regard, we may distinguish between subgroup specifications that can be translated into a rule (or a set of rules) in which (transformations of) the pretreatment characteristics are included versus subgroup specifications for which this is not possible. Furthermore, within the first category, we may distinguish rules that include all pretreatment characteristics from rules that include a limited subset of them only. The column "Form" of Table 1 shows for each method which of these forms the result takes.

2.1.2. *The data*

When choosing a method, the researcher should take the type of the data into account. This can differ in several respects. First, the study can be either a randomized clinical trial (RCT) or an observational study.[1] Second, the measurement scale of the outcome variable may be continuous

Table 1. Overview of methods that can be used to detect subgroups involved in treatment–subgroup interactions from the data in a *post hoc* manner, sorted by the research question that each method primarily targets.

Question method	Form	Software?	Ref
(a) Finding subgroups with enhanced treatment effect (EE)			
EE1 Simple Optimal Regime Approximation (SORA)	1A	No	Foster et al. (2014)
EE2 "Selecting a target population for a future comparative study"	1A	No	Zhao et al. (2013)
EE3 Subgroup Identification, Differential Effect Search (SIDES)	1A	Yes	Lipkovich et al. (2011)
EE4 Virtual Twins	1A	Yes	Foster et al. (2011)
EE5 Responder identification method	1A	No	Kehl and Ulm (2006)
(b) Partitioning based on similar relative treatment effects (SE)			
SE1 GUIDE "interaction" (Gi)	1A	Yes	Loh et al. (2015)
SE2 GUIDE "sum" (Gs)1	1A	Yes	Loh et al. (2015)
SE3 "Bayesian approach to subgroup identification"	1A	No	Berger et al. (2014)
SE4 Modified Covariate Approach	1A	No	Tian et al. (2014)
SE5 QUalitative INteraction Trees (QUINT)[b]	1A	Yes	Dusseldorp and Van Mechelen (2014)
SE6 "Analysis of RCT data for personalized treatment selections	1B	No	Cai et al. (2011)
SE7 Simultaneous Threshold Interaction Modelling (STIMA)	1A	Yes	Dusseldorp et al. (2010)
SE8 Interaction Trees	1A	Yes	Su et al. (2009)
SE9 Interaction Trees survival data	1A	No	Su et al. (2008)
SE10 Model-based Recursive Partitioning[a]	1A	Yes	Zeileis et al. (2008)
SE11 "Tree-structured subgroup analysis, censored survival data"	1A	No	Negassa et al. (2005)
SE12" Modelling interactions using fractional polynomials"	1A	No	Royston and Sauerbrei (2004)
SE13 "Bayesian subset analysis"	1B	Yes	Dixon and Simon (1991)
(c) Finding an optimal treatment regime (TR)			
TR1 Greedy Outcome Weighted Learning (GOWL)	1A	Yes	Zhu et al. (2016)
TR2 "Optimal treatment regimes via subgroup identification"	1A	Yes	Fu et al. (2016)
TR3 Minimum Impurity Decision Assignments (MIDAS)	1A	No	Laber and Zhao (2015)
TR4 Regularized Outcome Weighted Subgroup ident. (ROWSi)	1A	Yes	Xu et al. (2015)
TR5 "Treatment regimes represented as decision lists"	1A	Yes	Zhang et al. (2015)
TR6 Outcome Weighted Learning (OWL) with censored data	2(1B)[d]	No	Zhao et al. (2015)
TR7 Parametric Q-learning[c]	1B	Yes	Schulte et al. (2014)
TR8 "Optimal biomarker combinations for treatment selection"	2(1B)[4]	Yes	Huang and Fong (2014)
TR9 "Estimating treatment-effect heterogeneity using L2-SVM"[c]	1A	Yes	Imai and Ratkovic et al. (2013)
TR10 Outcome Weighted Learning (OWL)	2(1B)[4]	Yes	Zhao et al. (2012)
TR11 "Robust method for optimal treatment regimes"	1B	Yes	Zhang et al. (2012a)
TR12 "Classification-based treatment regimes"	1A	Yes	Zhang et al. (2012b)
TR13 "Variable selection for optimal treatment decision"[c]	1A	Yes	Lu et al. (2011)
TR14 l_1 penalized least squares method (l_1-PLS)[c]	1A	No	Qian and Murphy (2011)

[a]Although these methods may be used to partition patients into groups with similar relative treatment effects, they have been primarily developed for a more general purpose and may pick up main effects of pretreatment characteristics as well.
[b]This method specifically targets capturing qualitative treatment–subgroup interactions.
[c]These methods are primarily suited to estimate optimal treatment regimes, although they do not necessarily lead to a regime that is optimal in the sense that it optimizes the expected outcome (within a prespecified class) when used for treatment assignment in the full population.
[d]If an identity (linear) kernel is used, subgroup membership can be defined in terms of explicit rules that involve the pretreatment characteristics.
The third column indicates the form of the subgroup specifications: 1 = subgroups that can be defined in terms of rule(s) that involve (explicit transformations of) pretreatment characteristics [with 1A = rule(s) involving a small subset of the characteristics, and 1B = rule(s) involving all characteristics]; 2 = subgroups that cannot be defined in terms of such rules. The fourth column indicates whether software is publicly available or not.

(general or time-to-event with censoring) or categorical (binary or ordinal). Third, the number of treatment alternatives under comparison may be two or more than two. Finally, although most methods can, in principle, handle all types of independent variables (continuous, categorical with two or more categories), for some of them the use of certain types of variables may not yet be possible in practice. Table 2 shows for each method what type(s) of data it can handle (leaving aside the type of the independent variables, as this is mostly dependent on computational aspects of software packages or codes that may be under active development).

Table 2. Properties of the methods with regard to the data that each method can handle, including the type of study (RCT or observational (Obs)), the number of treatment alternatives (2 or > >2), and the type of outcome variable (continuous (Cont.), time-to-event, categorical, ordinal (Cat. Ord.), and categorical, binary (Cat. Bin.)), where + indicates that the method can handle the type of data indicated by the column in question, whereas − indicates that it cannot.

Type of study			# Treat.		Outcome			
Method[a]	RCT	Obs.	2	>2	Cont. General	Cont: Time-to-event	Cat.: Bin.	Cat.: Ord
EE1	+	−	+	−	+	−	+	+
EE2	+	−	+	−	+	+	+	+
EE3	+	+	+	−	+	+	+	+
EE4	+	−	+	−	+	−	+	+
EE5	+	−	+	−	−	+	−	−
SE1	+	−	+	+	+	+	−	+
SE2	+	−	+	+	+	+	−	+
SE3	+	−	+	−	+	−	−	−
SE4	+	−	+	−	+	+	+	+
SE5	+	−	+	−	+	−	−	+
SE6	+	−	+	−	+	+	+	+
SE7	+	−	+	−	+	−	+	+
SE8	+	−	+	−	+	−	−	−
SE9	+	−	+	−	−	+	−	−
SE10	+	−	+	+	+	−	+	+
SE11	+	−	+	−	−	+	−	−
SE12	+	−	+	+	+	+	+	+
SE13	+	−	+	−	+	−	−	−
TR1	+	+	+	+	+	+	+	+
TR2	+	+	+	+	+	+	+	+
TR3	+	+	+	+	+	−	+	+
TR4	+	+	+	−	+	+	+	+
TR5	+	+	+	+	+	+	+	+
TR6	+	−	+	−	−	+	−	−
TR7	+	−	+	+	+	−	+	+
TR8	+	−	+	−	+	−	+	+
TR9	+	−	+	−	−	−	+	−
TR10	+	−	+	−	+	−	+	+
TR11	+	+	+	−	+	−	+	+
TR12	+	+	+	−	+	−	+	+
TR13	+	+	+	−	+	−	+	+
TR14	+	−	+	+	+	−	+	+

[a]The abbreviations were introduced in Table 1

2.1.3. *Performance*

The method(s) that can be suitably used to answer the specific research question under study and that can handle the data at hand should be evaluated with regard to its/their performance. Three types of performance may be of interest at this point, namely, the value of the objective function that is to be optimized (e.g., the expected potential outcome in the population in the case of an optimal treatment regime or the model fit in a model-based approach), the recovery of the true model (on different levels, such as the level of the individual pretreatment characteristics involved in the true subgroup specifications and the level of the decision outcome), and the inferential error rate. Table 1 of the Supplementary Materials summarizes the available evidence with regard to these three performance criteria. For this purpose, it shows for each study the outcome variable, the scenario under which the result was found, and the actual result, in terms of the order of performance for the methods in the study.

If there is no universally best method with regard to all performance criteria, the research question may imply which performance aspect is most important. For example, if a researcher wants to interpret resulting subgroups, it may be important to evaluate the methods in terms of recovery of the true relevant pretreatment characteristics. Similarly, if a researcher wants to arrive at a merely pragmatically useful division of patients into subgroups and is not interested in a characterization of them, the value of the objective function that is to be optimized may primarily be of interest.

When using Table 1 of the Supplementary Materials to decide upon which method to use, it should be noted, however, that almost all evidence summarized in this table was gathered from non-neutral comparisons (i.e., comparisons of a newly proposed method by the authors of that method with existing methods). As such, publication bias may be at play (Boulesteix et al. 2015) and the studies may be overoptimistic with regard to the superiority of the newly proposed method for several reasons, as outlined by Boulesteix et al. (2013).

2.1.4. *Software*

Finally, a method can only be applied if software or source code is publicly available. If this is the case, the software can be documented or not. Table 1 shows for each method whether software or source code is publicly available. Table 2 of the Supplementary Materials shows for all methods for which this is the case, the name, form, and location of the software, whether it has been documented or not, and whether the authors that proposed the method were involved in developing it. However, one should keep in mind that many R-codes and software packages are under active development and that if software is currently lacking, it may be available at a later point in time. Similarly, source code that is undocumented today may become a documented software package in the near future.

Tables 1 and 2 of the paper, and Tables 1 and 2 of the Supplementary Materials may serve as a tool to arrive at a decision with regard to which method to use. Importantly, it is not necessary to choose a single method to analyze the data. On the contrary, when multiple suitable methods are available, it may be better to use several or even all of them. By doing so (which is sometimes referred to as methodological triangulation (Denzin 1978)), the weaknesses of one method can be overcome in part by other ones. Moreover, if the results of multiple methods converge, stronger conclusions may be drawn.

2.2. *Analyzing the data*

To analyze the data, often several choices regarding the setting of tuning parameters have to be made (Section 2.2.1). Subsequently, the result that has been obtained has to be formally evaluated and conclusions should be drawn carefully (Section 2.2.2).

2.2.1. *Setting tuning parameters*

Most methods for detecting subgroups involved in treatment–subgroup interactions from the data require the setting of a number of tuning parameters. Examples include the maximum number of leaves or maximum tree depth in case of recursive partitioning methods, the value of C (cost of misclassification of training data points) in case of support vector machines (SVM), and the value of the penalty parameter in case of penalized regression methods.

In general, there are three options for setting tuning parameters: The first one is to use the default settings, the second one is to manually set the tuning parameters, and the third one is to automatically set the tuning parameters using resampling methods. One may wish to refrain from using default settings without careful consideration of whether these are suitable or not. Rather, one may wish to base the setting of tuning parameters on statistical as well as substantive considerations. Statistical considerations may include the quality of the output, the stability of the result, the probability of replicability, and the probability of inferential errors. A key substantive consideration is the interpretability of the output. In general, for both types of considerations, parsimony may be an important principle as it may lead to more stability, a higher probability of replicability, a lower probability of Type I errors, and finally to better interpretable models. Finally, it may be advisable to perform some kind of sensitivity analysis to investigate to what extent the result may depend on the setting of specific tuning parameters.

2.2.2. *Formal evaluation and carefully drawing conclusions*

In this subsection, we will discuss the optional testing of a result (e.g., is the treatment effect in an enhanced subgroup significantly better than in the rest of the population?) and evaluation of a result in terms of its quality (e.g., the expected potential outcome of an optimal treatment regime if you would use it for treatment assignment of the whole population of interest).

It is well-known that in the context of inducing subgroups involved in treatment–subgroup interactions, one may face high risks of making inferential errors and of obtaining biased estimates of the quality of a result (see e.g., Rothwell (2005), Feinstein (1998), Lipkovich et al. (2017)). This may be caused by a number of factors, which can occur in different combinations, including (a) the omission of a formal test of the results, (b) subjecting the result to a (large) number of formal tests causing problems of multiple testing (see, e.g., Brookes et al. (2004); Sleight (2000); Lagakos (2006)), (c) performing underpowered studies (e.g., detecting interactions of the same size as a main effect with the same power requires a sample size that is four times as large (Brookes et al. 2004); interaction tests to tell quantitative and qualitative interactions apart, the power is even lower (Gail and Simon 1985; Lee et al. 2015; Pan and Wolfe 1997; Shaffer 1991)), and (d) using the same observations for both estimating and testing or evaluating a model (see, e.g., Friedman et al. (2009)). Below, we will discuss for each type of research question a number of possible tests and model evaluation measures of interest.

2.2.3. *Finding subgroups with an enhanced treatment effect*

When subgroups with enhanced effects have been induced, a possible test of interest could be based on the null hypothesis that the relative treatment effect in subgroup A (a subgroup with an enhanced effect) is similar to that in subgroup B (a subgroup without an enhanced effect) vs. the alternative hypothesis that the relative treatment effect in subgroup A is larger than that in subgroup B. In this case, a specific interaction between subgroup membership and treatment would be tested.

As to model assessment, one could estimate the magnitude of the difference of the relative treatment effect in the enhanced subgroup compared to that in the non-enhanced subgroup (see, e.g., Foster et al. (2011)). In this evaluation, one should choose which observations are to be used. A first option is to use all observations for both the data analysis and the assessment of the resulting model (i.e., resubstitution validation). However, this type of validation leads to a fairly extreme positive bias. Second, one could split the sample into a training set and a test set, where the training set is used to analyze the data, and the test set to assess the obtained model(s). This option, however, may only be appropriate in the case of a very large sample size. The third, and often only feasible option, is to rely on resampling techniques such as k-fold cross-validation or bootstrap (Efron 1983; Friedman et al. 2009). In k-fold cross-validation the data is split up into k equally sized folds. In each of k analyses, a different fold is left out for validation, while the remaining $k - 1$ folds are used to analyze the data. It is important in case of a multistep modeling procedure, that all steps of the analysis involving the outcome are included in the k cross-validation analyses as well (Friedman et al. 2009). The value of k can range from two (split half) to the number of observations in the sample (leave-one-out cross-validation). As to the bootstrap, B samples of the same size as the original sample are drawn with replacement from the original dataset; each bootstrap sample is analyzed and the resulting model(s) is/are assessed based on the original dataset. Different types of estimators of model performance can be used in the bootstrap analysis, including the ordinary bootstrap, the leave-one-out bootstrap Efron (1983), the .632 bootstrap Efron (1983), and the .632 + bootstrap Efron and Tibshirani (1997).

2.2.4. *Capturing treatment-effect heterogeneity by inducing subgroups with similar relative treatment effects*

When subgroups with similar relative treatment effects have been induced, there may be several tests of interest. As a first possibility, one may test the null hypothesis that the relative treatment effect is the same for the whole population of patients (i.e., absence of treatment-effect heterogeneity) against

the alternative hypothesis that this is not the case. This is a global test of treatment–subgroup interaction. Note that often authors try to obtain evidence against this null hypothesis by testing the main effect of treatment within each subgroup specifically, which in many cases resulted in treatment–subgroup interactions having been claimed that could not have been replicated in subsequent research (see, e.g., Table 1 in Rothwell (2005)). A better option here would be to use global tests of interaction (Brookes et al. 2004), including permutation tests of this type (see, e.g., Potthoff et al. (2001); Foster et al. (2016)). A second possibility is to test the null hypothesis that no qualitative treatment–subgroup interaction is present against the alternative hypothesis that there is at least one such an interaction. Here, global tests for qualitative interactions can be used (e.g., Gail and Simon (1985); Pan and Wolfe (1997)). Finally, the null hypothesis of no treatment–subgroup interaction for a specific subgroup and its complement can be tested against the alternative hypothesis that there is such an interaction. This is a test of a subgroup-specific interaction.

As to model assessment, one could use a measure that evaluates the strength of the treatment–subgroup interaction (e.g., the difference in relative treatment effect in two subgroups, weighted by their pooled standard deviation). Again, one should think about which observations to use here, in order to avoid obtaining biased results.

2.2.5. *Inducing subgroups underlying an optimal treatment regime*

A first null hypothesis that could be tested regarding this research question is that the expected potential outcome under the optimal treatment regime is the same as when giving everyone the same (marginally best) treatment alternative. Recently, several methods have been proposed to test this global hypothesis (see e.g., Foster et al. (2016); Xu et al. (2015); Chakraborty et al. (2014); Loh et al. (2015); Luedtke and Van Der Laan et al. (2016); Laber et al. (2014)).

As to model assessment, the expected potential outcome if all patients would be assigned to treatment according to the estimated regime may be of interest. Again, it is important to avoid obtaining biased estimates by using a separate learning and test sample or by using resampling methods.

Finally, when drawing conclusions, it is important to report that the methods under study are exploratory in nature and that inferential errors are ubiquitous. This means that conclusions should be presented with caution, and that it is important to mention that the obtained results should be replicated. Indeed, the domain of subgroup analyses is notorious for a lack of replicability of results. This immediately relates to the fact that to reliably detect qualitative interactions (on which nontrivial treatment regimes are based), a large sample size is needed. Moreover, some critical subgroups involved in qualitative interactions may be very small. The above implies that to validate subgroup-based findings on HTE, it may be desirable to rely on samples in which certain subgroups are overrepresented (which can be accomplished by using a stratified sampling strategy). Additionally, an RCT may be preferable over an observational study to ensure unconfounded treatment assignment. That being said, although a stratified RCT may be the preferred approach, observational data (such as EMR data) can be useful as well for this purpose. However, when using such data, more caution is required to avoid under-represented critical subgroups and to account for a possible confounding in treatment assignment (e.g., by using a propensity model).

Apart from recognizing the need for replication, the degree of caution may vary depending on whether efforts have been made to substantiate the results, for example, by performing statistical tests, sensitivity analyses, or methodological triangulation. Another way of substantiating conclusions is by justifying the obtained result based on theoretical or substantive insights regarding when or for which types of patients (e.g., in terms of pathology) ingredients of a treatment alternative may be especially active.

3. Illustrative application

In this section, we will illustrate a number of choices a researcher may be faced with when applying methods for detecting subgroups involved in treatment–subgroup interactions, using the Combining

Medication to Enhance Depression Outcomes (CO-MED) dataset (Rush et al. 2011). In this RCT dataset (including 665 participants with moderately severe nonpsychotic chronic and/or recurrent major depressive disorder), the efficacy of two combination treatments (sustained-release bupropion + escitalopram and extended-release venlafaxine + mirtazapine) was compared to the efficacy of a mono-treatment (escitalopram + placebo). In the original study, Rush et al. (2011) found no differences between the three treatment arms in terms of remission rate, response rate, and the percentage of change in the QIDS-SR score (a self-report measure of depressive symptom severity derived from the Inventory of Depressive Symptomatology (IDS) (Rush et al. 2003)) at 12 weeks; yet the combination of venlafaxine + mirtazapine was found to have a greater risk of adverse effects. As such, we will leave the latter alternative out of account in the primary analysis of this application; however, we will use it in a secondary (*post hoc*) analysis.

Although Rush et al. (2011) discretized the outcome variable (QIDS-SR score) into a binary remission variable, we chose not to do so because of the associated loss of information. Instead, to account for the wide range of observed QIDS-SR scores at baseline (min = 5, max = 27), we will use the reduction of pre-existing levels of depression symptoms as our outcome variable, by subtracting the QIDS-SR score in week 12 from the QIDS-SR score at baseline (note that computed in this way, higher values imply a better result). Of course, although statistically less efficient, remission rates may still be clinically relevant, and the same holds for response rates (Rush et al. 2006, 2011). Therefore, we will calculate these measures *post hoc* in the subgroups (with remission being defined as a QIDS-SR score of less than 6 in week 12, and response as a reduction of at least 50% in QIDS-SR score between baseline and week 12).

We included all 54 independent variables in our analysis, except for the ones with more than 10 missing responses at baseline. Furthermore, about 167 patients dropped out from the study before Week 12. We discretized variables related to the number of symptoms experienced with regard to certain psychiatric disorders on the basis of established theoretical cut-offs (Zimmerman and Chelminski 2006; Zimmerman and Mattia 2001), such that the resulting discretized variables represent whether a patient was diagnosed with the disorder or not.

Furthermore, there were quite some drop-outs. As we did not find any pronounced patterns in the data in relation to this drop-out, and as most methods under study cannot handle missing data and cannot easily be extended with some multiple imputation procedure, we decided to do a complete case analysis. As a sensitivity check, we will calculate *post hoc* in each subgroup the mean outcomes for the treatment alternatives based on a completed dataset for which we imputed the missing responses by carrying the last observation forward and a completed dataset for which we imputed missing responses by the global mean outcome.

3.1. *Choosing a method*

3.1.1. *The research question*

Although Rush et al. (2011) found no evidence that treating patients with a combination of bupropion and escitalopram was significantly better on average than treating them with escitalopram alone, it is known that certain moderators (e.g., gender, age, depression severity, depression subtype) can cause a differential response to anti-depressant treatment (Shelton and Trivedi 2011). Consequently, the question arises whether it could be that for some patients, treatment with a combination of the two antidepressants is beneficial, whereas for other patients treatment with escitalopram alone is more effective. An identification of these subgroups would help medical researchers to select effective treatment for individual patients (Shelton and Trivedi 2011). This leads to the following research question: Can we induce subgroups at the basis of a decision rule that specifies what type of patients should preferably receive escitalopram only and what type should preferably receive escitalopram + bupropion in order to obtain the maximal benefit in the population of patients under study? Moreover, an insightful characterization of the subgroups involved in terms of a limited subset of pretreatment characteristics would corroborate the obtained result both

from a clinical and a statistical viewpoint. This implies that we will be looking for an optimal treatment regime based on subgroups that can be represented by a rule in which a limited subset of pretreatment characteristics is included. Suitable methods for this purpose according to Table 1 are TR1, TR2, TR3, TR5, and TR12.

3.1.2. *The data*

The data stem from an RCT. The primary outcome variable is the continuous QIDS-SR score (Rush et al. 2003). As indicated before, we will focus on the escitalopram mono-treatment and the escitalopram + bupropion combination treatment arms. According to Table 2, all methods that are suitable to answer our research question can handle this kind of data.

3.1.3. *Performance*

The objective function that is to be optimized is the expected outcome if the estimated regime would be used to assign treatment to everyone in the population. Ideally, all possibly suitable methods should have been directly compared to each other in this regard. However, according to Table 1 of the Supplementary Materials, this is only the case for TR1 and TR3, where TR1 outperformed TR3. Therefore, we will not consider TR3 as a possibly suitable method. According to Table 1 of the Supplementary Materials, TR5 and TR12 have further been compared to other methods than the ones that we are interested in. In this comparison, their performance appeared to be satisfactory. Therefore, we will maintain them as possibly suitable methods. From its part, TR2 has not been evaluated at all with regard to expected outcome; therefore, we will not consider it as a possibly suitable method. As we want to characterize the possibly resulting subgroups, besides performance in terms of the expected outcome of the estimated regime, we are also interested in the recovery rates of the regime's true underlying structure. Additionally, the inferential error rates of the methods under study are of interest as well. Unfortunately, none of the methods have been evaluated in these respects. Consequently, we end up with TR1, TR5, and TR12 as suitable methods to analyze our data.

3.1.4. *Software*

Whereas for all three methods selected above, source code or software is publicly available, only for TR12 a documented R-package is available. Therefore, we will use TR12 as the primary method to analyze our data, whereas we will use the other two methods for secondary analyses and compare the resulting treatment regimes with TR12. For TR1, we will use code provided by the author in the supplementary materials of the paper, supplemented by some R-code that we wrote ourselves to implement a cross-validation approach to obtain the optimal tuning parameter in the tree pruning. For TR5, we will use the R-package dtrlist, which is based on the algorithm proposed in Zhang et al. (2016). This algorithm was developed for list-based dynamic treatment regimes, yet can be used for the one decision point case as well. The splitting criteria used in this algorithm are modified compared to the criteria proposed by TR5 to avoid (asymptotically) becoming stuck in a local mode.

3.2. *Analyzing the data*

3.2.1. *Specifying the method and setting tuning parameters*

Within the framework of TR12, the method used to maximize the objective function has to be chosen. We chose to use Classification and Regression Trees (CART) for this purpose, as it leads to a most insightful result and as documented R-code for it is available within the rpart package (Therneau et al. 2015). Also, an estimator for the expected potential outcome has to be chosen. Possibilities include an estimator based on an outcome model only (Zhang et al. 2012b), an Inverse Probability Weighted Estimator (IPWE, in which observations are weighted by a propensity model; see e.g., Bang and Robins (2005)), or an Augmented Inverse Probability Weighted Estimator (AIPWE, in which observations are weighted by a propensity model and augmented by an outcome model; see e.g., Bang and Robins (2005)). We chose to use the AIPWE estimator, as it is more robust

and efficient than the other two estimators, and as it performed best according to simulations performed by Zhang et al. (2012a). Finally, various tuning parameters must be set in order to use the rpart package. We set them as follows such that a parsimonious and stable result will be favored:

- The minimum number of observations in a leaf: 50 (which suffices to calculate reliable statistics in the leafs)
- The minimum number of observations in a node: 100 (twice the minimum number of observations in a leaf)
- The maximum depth of the tree: 3 (as a subgroup characterization based on three variables seems to be of the maximum complexity that is still interpretable)
- The minimum increase in complexity parameter: .001 (taking into account that splits that are not worthwhile will be pruned off afterwards)
- The number of folds used in cross-validation: 10 (as 10-fold cross-validation is generally recommended in the literature for its good performance regarding model selection)
- The pruning complexity parameter: The value for which the cross validated error is within one standard error above the value with the lowest cross-validated error (such that the most parsimonious result within one standard error from the best result will be chosen)

For TR1, we used 10-fold cross-validation to determine the best tuning parameter for the tree-pruning, with the same standard-error rule used in the pruning of TR12. For TR5, we used as maximum depth of the decision list 3, comparable to the maximum depth of the tree-based treatment regime resulting from TR12.

3.2.2. *Formal evaluation*

We decided to formally evaluate and interpret the most parsimonious treatment regime that we would obtain from TR12 when running rpart multiple times (repeating rpart multiple times on the same dataset may lead to different results because of the cross-validation procedure used for the pruning that is based on a random split of the data). The regime that we obtained in this way is the one shown in Figure 1. According to this regime, patients without panic disorder should preferably receive the mono-treatment (escitalopram + placebo), whereas patients with panic disorder should receive the combination treatment (escitalopram + bupropion). The mean outcomes for the other endpoints (calculated *post hoc*) support the obtained regime (see Figure 1); the same holds for the mean outcomes (calculated *post hoc*) for the last observation carried forward and single mean imputation datasets.

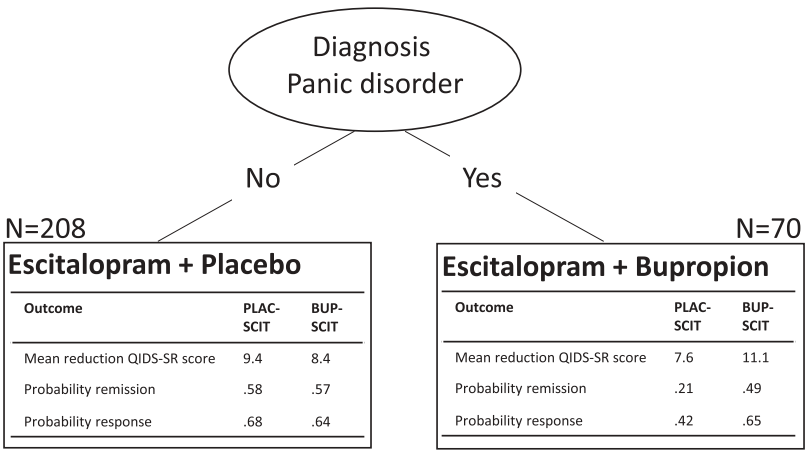

Figure 1. Estimated optimal treatment regime for the CO-MED data using TR12, with for each subgroup (leaf) the optimal treatment alternative (in bold), followed by the mean reduction in QIDS-SR score (primary outcome variable), the probability of remission (calculated *post hoc*), and the probability of response (calculated *post hoc*) under each treatment alternative.

We subsequently performed the interaction test proposed by Gail and Simon (1985) to check whether the interaction between treatment and panic disorder is significant indeed. As suggested by Gail and Simon (1985), we first tested for quantitative heterogeneity. This test was significant with α=.05 (H = 9.98, df = 1, p< .01). Second, we tested for qualitative heterogeneity. At this point, we could not reject the null hypothesis of no qualitative interaction between treatment and panic disorder at α=.05 (min(Q^+, Q^-)=2.66), but we could do so at α= .10.

Finally, we used a .632+ bootstrap correction (Efron and Tibshirani 1997) to obtain an honest estimate of the expected outcome if the obtained regime would be used to assign everyone in the population to treatment, which turned out to be 9.27. The mean outcome for the trivial regime of assigning all patients to escitalopram treatment was 8.95, and for the trivial regime of assigning all patients to escitalopram + bupropion treatment 9.04.

The result of TR1 was a trivial tree, based on the root node only. According to this regime, all patients should receive mono-treatment. The result of TR5 is shown in Figure 2. According to this regime, females with an IDS score lower than 48.5 should receive the combination treatment. Otherwise, persons without a panic disorder who do not abuse drugs should receive mono-treatment. All others should receive combination treatment.

3.2.3. *Conclusions*

We found somewhat mixed results: According to the regime that we obtained with TR12, patients with panic disorder should preferably receive the combination treatment (escitalopram+ bupropion), while patients without panic disorder should receive the mono-treatment (escitalopram + placebo). Yet the difference in mean outcomes between the two treatments is rather convincing for the patients with panic disorder, it is rather small for patients without. This implies that, instead of a qualitative interaction, predominantly a quantitative interaction seems to be at play here. These mixed results were echoed by the Gail and Simon interaction tests (in that the quantitative interaction test was clearly significant, whereas the qualitative interaction test was marginally significant only), and by the estimated expected outcome of the regime (which was only slightly larger than the mean outcome of the trivial regime of assigning all patients to the marginally best treatment alternative, viz., the escitalopram + bupropion combination treatment). The secondary analyses with TR1 and TR5, confirmed this mixed picture as well: The result of TR1 was a trivial tree, indicating that there is no qualitative treatment-effect

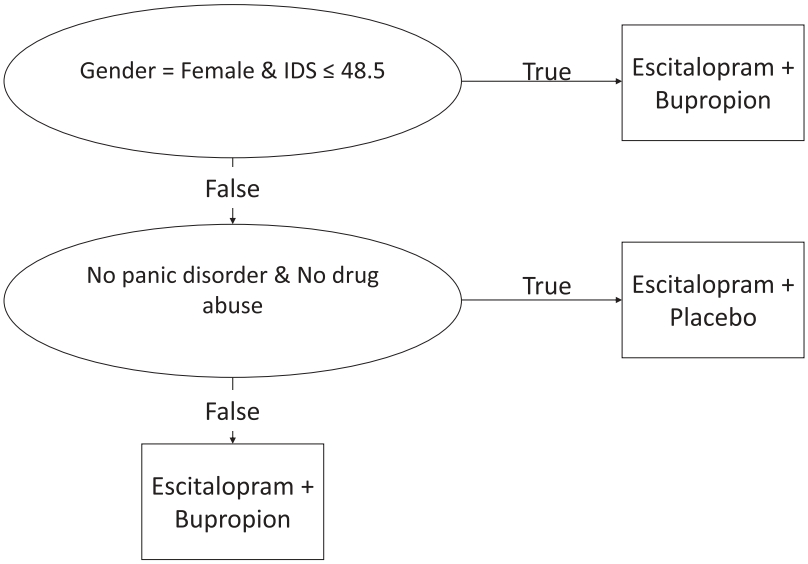

Figure 2. Estimated list-based optimal treatment regime for the CO-MED data using TR5.

heterogeneity at play. The result of TR5 was more complex, pointing at qualitative treatment-effect heterogeneity, not only between persons without and with panic disorder but also between males and females, between different scores on the IDS, and between persons abusing drugs or not. However, we should take into account that Sies and Van Mechelen (2017) reported on the basis of a small simulation study that TR5 virtually never results in a trivial decision list (i.e., assigning all patients to the same treatment), indicating that the "Type I error risk" for this method is probably very high.

The result that adding bupropion to an escitalopram treatment may be beneficial for depressed patients with a comorbid panic disorder may at first sight look inconsistent with the existing literature on panic disorder. Indeed, except from a few small open-label studies and case reports (e.g., Simon et al. (2002); Serafini et al. (2011)), available evidence mainly suggests that bupropion is not effective for treating panic disorder (e.g., Young (1996); Sheehan et al. (1983)). Also, international guidelines suggest that Selective Serotonin Reuptake Inhibitors (SSRIs) (such as escitalopram) should be prescribed as the treatment of choice for panic disorder (e.g., Royal Australian and New Zealand College Of Psychiatrists Clinical Practice Guidelines Team For Panic Disorder And Agoraphobia (2003); National Institute for Health and Clinical Excellence (2011)). Yet, all results in question pertain to patients with a panic disorder only, whereas the result we found pertains to patients with depression and a comorbid panic disorder. Besides, the target outcome variable in our study is the reduction of depression symptoms instead of the reduction of anxiety-related symptoms. Almost no data are available on the effectiveness of bupropion for patients with depression and comorbid panic disorder in terms of a reduction of depression symptoms. Yet, it is known that, in general, depression with a comorbid anxiety disorder (such as panic disorder) is associated with a worse response to treatment (e.g., Souery et al. (2007); Chekroud et al. (2016)). Hence, the mechanism underlying our estimated treatment regime could read that, for a depression with a comorbid anxiety disorder, a combination treatment (which involves multiple working mechanisms) might be superior to a mono-treatment with SSRIs (which involves a serotonergic working mechanism only).

To further examine this hypothesis, we made an appeal to the data of the venlafaxine + mirtazapine treatment arm (which we left out from the primary analysis because of the greater risk of adverse effects that goes with this combination treatment). Specifically, we calculated *post hoc* the mean reduction in QIDS-SR score for the patients that received venlafaxine + mirtazapine within the two subgroups at the basis of our estimated treatment regime. It turned out that for patients without panic disorder this was 8.2 (which is again slightly worse than the corresponding mean score for the escitalopram mono-treatment), and for patients with a comorbid panic disorder it was 10.7 (which is again clearly better than the corresponding mean score for the escitalopram mono-treatment). This implies clear support for the hypothesis that patients with depression and a comorbid anxiety disorder benefit most from a combination treatment.

Obviously, to consolidate this result, follow-up research is needed. For this purpose, new data should be collected using a stratified design (with stratification based on the presence/absence of panic disorder) and equal, relatively large numbers of patients with and without panic disorder. On the basis of such data, one may primarily try to replicate the superiority of a combination treatment for patients with depression and comorbid panic disorder. As a secondary goal, one could also examine whether for patients without a comorbid panic disorder, a mono-treatment is approximately equally effective as a combination treatment, vs. whether for such patients a mono-treatment is superior (which, statistically speaking, comes down to testing whether the interaction under study is quantitative versus qualitative in nature).

4. Concluding remarks

Nowadays, studying treatment-effect heterogeneity to tailor treatment choice to patients' pretreatment characteristics is one of the most prominent challenges for medical statisticians and statistically minded medical researchers. In this paper, we discussed a number of considerations that may be useful for medical researchers to conduct such a study. These considerations may prove guidance for a number of nontrivial choices that have to be made during the data analysis.

Key contributions of the paper include a classification scheme for methods to study treatment-effect heterogeneity, an overview of available software, and a systematic review of comparative empirical evidence regarding the performance of the different methods under study. One main advantage of the first of these, that is to say, the proposed classification scheme, is that it is based on classification principles such as the type of research question that a method primarily addresses and the type of data that it can handle, which connect directly to basic aspects of each data-analytic process (and which are therefore of obvious importance when choosing a method); on top of that, they are also well accessible for medical researchers. That being said, it should be mentioned that other classification schemes have been developed in the literature as well. Perhaps the most prominent of these is the scheme proposed by Lipkovich and Dmitrienko (2014), which was also used by Lipkovich et al. (2017) and Ondra et al. (2016), with three classes that have been labeled in terms of what the authors call different forms of modelling. Otherwise, the classification scheme proposed in the present paper cannot be directly mapped onto the classification scheme of Lipkovich and Dmitrienko (2014).

One may finally note that good choices in a data-analytic process when studying treatment-effect heterogeneity to tailor treatment choice are choices that lead to answers that are fair and sound, both from a substantive point of view (e.g., choices that lead to interpretable results), and from a methodological point of view (e.g., choices that lead to acceptable inferential error risks or stabilities of the results). To obtain such fair and sound answers, input from medical researchers as well as medical statisticians is needed, and subsequently a careful reflection on this input. Finally, in reporting the results, a golden rule is to be transparent, explicit, and honest.

Note

1. Generally, methods that can handle observational data are based on inversely weighting each observation by the probability of receiving the treatment that it actually received using a propensity model (which models the probability of receiving treatment T, given a covariate pattern X). Often, whether or not these methods are consistent depends on whether or not the propensity model is correctly specified.

Acknowledgments

Data used in the preparation of this manuscript were obtained and analyzed from the controlled access datasets distributed from the NIMH-supported National Database for Clinical Trials (NDCT). NDCT is a collaborative informatics system created by the National Institute of Mental Health to provide a national resource to support and accelerate discovery related to clinical trial research in mental health. Dataset identifiers: Study ID: N01 MH090003-02, clinical trial ID: NCT00590863. This manuscript reflects the views of the authors and may not reflect the opinions or views of the NIMH or of the Submitters submitting original data to NDCT.

Funding

The research reported in this paper was supported in part by the Research Foundation – Flanders(G080219N), the Research Fund of KU Leuven (GOA/15/003) and by the Interuniversity Attraction Poles program financed by the Belgian government (IAP/P7/06).

ORCID

Aniek Sies http://orcid.org/0000-0002-3166-1413

References

Bailey, K. 1994. Generalizing the results of randomized clinical trials. *Controlled Clinical Trials* 15 (1):15–23. doi:10.1016/0197-2456(94)90024-8.

Bang, H., and J. M. Robins. 2005. Doubly robust estimation in missing data and causal inference models. *Biometrics* 61 (4):962–973. doi:10.1111/j.1541-0420.2005.00377.x.

Berger, J., X. Wang, and L. Shen. 2014. A Bayesian approach to subgroup identification. *Journal of Biopharmaceutical Statistics* 24 (1):110–129. doi:10.1080/10543406.2013.856026.

Boonacker, C., A. Hoes, K. van Liere-Visser, A. Schilder, and M. Rovers. 2011. A comparison of subgroup analyses in grant applications and publications. *American Journal of Epidemiology* 174 (2):219–225. doi:10.1093/aje/kwr160.

Boulesteix, A.-L., S. Lauer, and M. J. Eugster. 2013. A plea for neutral comparison studies in computational sciences. *PloS one* 8 (4):e61562. doi:10.1371/journal.pone.0061562.

Boulesteix, A.-L., V. Stierle, and A. Hapfelmeier. 2015. Publication bias in methodological computational research. *Cancer Informatics* 14 (Suppl 5):11–19.

Brookes, S. T., E. Whitely, M. Egger, G. D. Smith, P. A. Mulheran, and T. J. Peters. 2004. Subgroup analyses in randomized trials: Risks of subgroup-specific analyses; power and sample size for the interaction test. *Journal of Clinical Epidemiology* 57 (3):229–236. doi:10.1016/j.jclinepi.2003.08.009.

Byar, D. 1985. Assessing apparent treatment—Covariate interactions in randomized clinical trials. *Statistics in Medicine* 4 (3):255–263. doi:10.1002/(ISSN)1097-0258.

Cai, T., L. Tian, P. Wong, and L. Wei. 2011. Analysis of randomized comparative clinical trial data for personalized treatment selections. *Biostatistics* 12 (2):270–282. doi:10.1093/biostatistics/kxq060.

Chakraborty, B., E. B. Laber, and Y.-Q. Zhao. 2014. Inference about the expected performance of a data-driven dynamic treatment regime. *Clinical Trials* 11 (4):408–417. doi:10.1177/1740774514537727.

Chekroud, A., R. Zotti, Z. Shehzad, R. Gueorguieva, M. Johnson, M. Trivedi, T. Cannon, J. Krystal, and P. Corlett. 2016. Cross-trial prediction of treatment outcome in depression: A machine learning approach. *The Lancet Psychiatry* 3 (3):243–250. doi:10.1016/S2215-0366(15)00471-X.

Cohen, J., P. Cohen, S. West, and L. Aiken. 2013. *Applied multiple regression/correlation analysis for the behavioral sciences*. Mahwah: Routledge.

Cronbach, L., and R. Snow. 1977. *Aptitudes and instructional methods: a handbook for research on interactions*. . New York: Irvington Publishers.

Denzin, N. 1978. *The research act: a theoretical introduction to sociological methods*. New York, NY: McGraw-Hill.

Dixon, D. O., and R. Simon. 1991. Bayesian subset analysis. *Biometrics* 47 (3):871–881. doi:10.2307/2532645.

Dusseldorp, E., C. Conversano, and B. Van Os. 2010. Combining an additive and tree-based regression model simultaneously: Stima. *Journal of Computational and Graphical Statistics* 19 (3):514–530. doi:10.1198/jcgs.2010.06089.

Dusseldorp, E., and I. Van Mechelen. 2014. Qualitative interaction trees: A tool to identify qualitative treatment–Subgroup interactions. *Statistics in Medicine* 33 (2):219–237. doi:10.1002/sim.5933.

Efron, B. 1983. Estimating the error rate of a prediction rule: Improvement on cross-validation. *Journal of the American Statistical Association* 78 (382):316–331. doi:10.1080/01621459.1983.10477973.

Efron, B., and R. Tibshirani. 1997. Improvements on cross-validation: The 632+ bootstrap method. *Journal of the American Statistical Association* 92 (438):548–560.

FDA. 2013. Paving the way for personalized medicine: fda's role in a new era of medical product development. https://www.fdanews.com/ext/resources/files/10/10-28-13-Personalized-Medicine.pdf

Feinstein, A. 1998. The problem of cogent subgroups: A clinicostatistical tragedy. *Journal of Clinical Epidemiology* 51 (4):297–299. doi:10.1016/S0895-4356(98)00004-3.

Foster, J., J. Taylor, N. Kaciroti, and B. Nan. 2014. Simple subgroup approximations to optimal treatment regimes from randomized clinical trial data. *Biostatistics* 16 (2):368–382. doi:10.1093/biostatistics/kxu049.

Foster, J., J. Taylor, and S. Ruberg. 2011. Subgroup identification from randomized clinical trial data. *Statistics in Medicine* 30 (24):2867–2880. doi:10.1002/sim.v30.24.

Foster, J. C., B. Nan, L. Shen, N. Kaciroti, and J. M. Taylor. 2016. Permutation testing for treatment–Covariate interactions and subgroup identification. *Statistics in Biosciences* 8 (1):77–98. doi:10.1007/s12561-015-9125-9.

Friedman, J., T. Hastie, and R. Tibshirani. 2009. *The elements of statistical learning, volume 1*. Berlin: Springer series in statistics Springer.

Fu, H., J. Zhou, and D. Faries. 2016. Estimating optimal treatment regimes via subgroup identification in randomized control trials and observational studies. *Statistics in Medicine* 35 (19):3285–3302. doi:10.1002/sim.6920.

Gail, M., and R. Simon. 1985. Testing for qualitative interactions between treatment effects and patient subsets. *Biometrics* 21 (2):361–372. doi:10.2307/2530862.

Hamburg, M., and F. Collins. 2010. The path to personalized medicine. *New England Journal of Medicine* 363 (4):301–304. doi:10.1056/NEJMoa1011205.

Henderson, N. C., T. A. Louis, G. L. Rosner, and R. Varadhan. 2017. Individualized treatment effects with censored data via fully nonparametric Bayesian accelerated failure time models. *arXiv preprint arXiv:1706.06611*.

Horizon 2020. 2015. Health, demographic change and wellbeing. https://ec.europa.eu/programmes/horizon2020/en/h2020-section/health-demographic-change-and-wellbeing

Huang, Y., and Y. Fong. 2014. Identifying optimal biomarker combinations for treatment selection via a robust kernel method. *Biometrics* 70 (4):891–901. doi:10.1111/biom.12204.

Imai, K., Ratkovic, M. 2013. Estimating treatment effect heterogeneity in randomized program evaluation. *The Annals of Applied Statistics* 7(1):443–470. doi:10.1214/12-AOAS593.

Kehl, V., and K. Ulm. 2006. Responder identification in clinical trials with censored data. *Computational Statistics & Data Analysis* 50 (5):1338–1355. doi:10.1016/j.csda.2004.11.015.
Kimball, A. 1957. Errors of the third kind in statistical consulting. *Journal of the American Statistical Association* 52 (278):133–142. doi:10.1080/01621459.1957.10501374.
Kravitz, R., N. Duan, and J. Braslow. 2004. Evidence-based medicine, heterogeneity of treatment effects, and the trouble with averages. *Milbank Quarterly* 82 (4):661–687. doi:10.1111/milq.2004.82.issue-4.
Laber, E., and Y. Zhao. 2015. Tree-based methods for individualized treatment regimes. *Biometrika* 102 (3):501–514. doi:10.1093/biomet/asv028.
Laber, E. B., D. J. Lizotte, M. Qian, W. E. Pelham, and S. A. Murphy. 2014. Dynamic treatment regimes: Technical challenges and applications. *Electronic Journal of Statistics* 8 (1):1225. doi:10.1214/14-EJS920.
Lagakos, S. 2006. The challenge of subgroup analyses-reporting without distorting. *New England Journal of Medicine* 354 (16):1667–1669. doi:10.1056/NEJMp068070.
Lee, S., M. Lei, and G. Brody. 2015. Confidence intervals for distinguishing ordinal and disordinal interactions in multiple regression. *Psychological Methods* 20 (2):245. doi:10.1037/met0000033.
Lipkovich, I., and A. Dmitrienko. 2014. *Biomarker identification in clinical trials*. New York, NY: Chapman and Hall/CRC Press.
Lipkovich, I., A. Dmitrienko, J. Denne, and G. Enas. 2011. Subgroup identification based on differential effect search – A recursive partitioning method for establishing response to treatment in patient subpopulations. *Statistics in Medicine* 30 (21):2601–2621.
Lipkovich, I., A. Dmitrienko, and S. R. D'Agostino. 2017. Tutorial in biostatistics: Data-driven subgroup identification and analysis in clinical trials. *Statistics in Medicine* 36 (1):136–196. doi:10.1002/sim.7064.
Loh, W., X. He, and M. Man. 2015. A regression tree approach to identifying subgroups with differential treatment effects. *Statistics in Medicine* 34 (11):1818–1833. doi:10.1002/sim.v34.11.
Lu, W., H. Zhang, and D. Zeng. 2011. Variable selection for optimal treatment decision. *Statistical Methods in Medical Research* 22 (5):493–504. doi:10.1177/0962280211428383.
Luedtke, A. R., Van Der Laan, M. J. 2016. Statistical inference for the mean outcome under a possibly non-unique optimal treatment strategy. *The Annals of Statistics* 44(2):713–742. doi:10.1214/15-AOS1384.
National Institute for Health and Clinical Excellence. 2011. Generalised anxiety disorder and panic disorder in adults: Management. NICE guideline (CG113).
Negassa, A., A. Ciampi, M. Abrahamowicz, S. Shapiro, and J. Boivin. 2005. Tree-structured subgroup analysis for censored survival data: Validation of computationally inexpensive model selection criteria. *Statistics and Computing* 15 (3):231–239. doi:10.1007/s11222-005-1311-z.
Ondra, T., A. Dmitrienko, T. Friede, A. Graf, F. Miller, N. Stallard, and M. Posch. 2016. Methods for identification and confirmation of targeted subgroups in clinical trials: A systematic review. *Journal of Biopharmaceutical Statistics* 26 (1):99–119. doi:10.1080/10543406.2015.1092034.
Pan, G., and D. Wolfe. 1997. Test for qualitative interaction of clinical significance. *Statistics in Medicine* 16 (14):1645–1652. doi:10.1002/(ISSN)1097-0258.
Potthoff, R. F., B. L. Peterson, and S. L. George. 2001. Detecting treatment-by-centre interaction in multi-centre clinical trials. *Statistics in Medicine* 20 (2):193–213.
Qian, M., and S. Murphy. 2011. Performance guarantees for individualized treatment rules. *Annals of Statistics* 39 (2):1180. doi:10.1214/10-AOS864.
Rothwell, P. 1995. Can overall results of clinical trials be applied to all patients? *The Lancet* 345 (8965):1616–1619. doi:10.1016/S0140-6736(95)90120-5.
Rothwell, P. 2005. Subgroup analysis in randomised controlled trials: Importance, indications, and interpretation. *The Lancet* 365 (9454):176–186. doi:10.1016/S0140-6736(05)17709-5.
Royal Australian and New Zealand College of Psychiatrists Clinical Practice Guidelines Team for Panic Disorder and Agoraphobia. 2003. Australian and New Zealand clinical practice guidelines for the treatment of panic disorder and agoraphobia. *Australian & New Zealand Journal of Psychiatry* 37(3):641–656. doi:10.1080/j.1440-1614.2003.01254.x.
Royston, P., and W. Sauerbrei. 2004. A new approach to modelling interactions between treatment and continuous covariates in clinical trials by using fractional polynomials. *Statistics in Medicine* 23 (16):2509–2525. doi:10.1002/(ISSN)1097-0258.
Rush, A., H. Kraemer, H. Sackeim, M. Fava, M. Trivedi, E. Frank, P. Ninan, M. Thase, A. Gelenberg, D. Kupfer, et al. 2006. Report by the ACNP task force on response and remission in major depressive disorder. *Neuropsychopharmacology* 31 (9):1841–1853.
Rush, A., M. Trivedi, H. Ibrahim, T. Carmody, B. Arnow, D. Klein, J. Markowitz, P. Ninan, S. Kornstein, R. Manber et al. 2003. The 16-item quick inventory of depressive symptomatology (QIDS), clinician rating (QIDS-C), and self-report (QIDS-SR): A psychometric evaluation in patients with chronic major depression. *Biological Psychiatry* 54(5):573–583. doi:10.1016/S0006-3223(02)01866-8.
Rush, A., M. Trivedi, J. Stewart, A. Nierenberg, M. Fava, B. Kurian, D. Warden, D. Morris, J. Luther, M. Husain et al. 2011. Combining medications to enhance depression outcomes (co-med): Acute and long-term outcomes of a

single-blind randomized study. *American Journal of Psychiatry* 168(7):689–701. doi:10.1176/appi.ajp.2011.11020284.
Schulte, P., A. Tsiatis, E. Laber, and M. Davidian. 2014. Q-and a-learning methods for estimating optimal dynamic treatment regimes. *Statistical Science: A Review Journal of the Institute of Mathematical Statistics* 29 (4):640.
Serafini, G., M. Pompili, P. Fusar-Poli, G. Porfiri, G. Giordano, S. Ferracuti, P. Girardi, and R. Tatarelli. 2011. Bupropion and panic disorder: Case report and review of the literature. *The Journal of Neuropsychiatry and Clinical Neurosciences* 23 (2):E47–E50. doi:10.1176/jnp.23.2.jnpe47.
Shaffer, J. 1991. Probability of directional errors with disordinal (qualitative) interaction. *Psychometrika* 56 (1):29–38.
Sheehan, D., J. Davidson, T. Manschreck, and J. Van Wyck Fleet. 1983. Lack of efficacy of a new antidepressant (bupropion) in the treatment of panic disorder with phobias. *Journal of Clinical Psychopharmacology* 3 (1):28–31. doi:10.1097/00004714-198302000-00006.
Shelton, R., and M. Trivedi. 2011. Moderators of antidepressant response in major depression. *The Journal of Clinical Psychiatry* 72 (10):32. doi:10.4088/JCP.09m05440blu.
Sies, A., and I. Van Mechelen. 2017. Comparing four methods for estimating tree-based treatment regimes. *The International Journal of Biostatistics* 13 (1). doi:10.1515/ijb-2016-0068.
Simon, N., N. Emmanuel, J. Ballenger, J. Worthington, G. Kinrys, N. Korbly, F. Farach, and M. Pollack. 2002. Bupropion sustained release for panic disorder. *Psychopharmacology Bulletin* 37 (4):66–72.
Sleight, P. 2000. Debate: Subgroup analyses in clinical trials: Fun to look at-but don't believe them. *Current Controlled Trials Cardiovascular Medicine* 1 (1):25–27. doi:10.1186/CVM-1-1-025.
Souery, D., P. Oswald, I. Massat, U. Bailer, J. Bollen, K. Demyttenaere, S. Kasper, Y. Lecrubier, S. Montgomery, A. Serretti et al. 2007. Clinical factors associated with treatment resistance in major depressive disorder: Results from a European multicenter study. *Journal of Clinical Psychiatry* 68(7):1062–1070. doi:10.4088/JCP.v68n0713.
Su, X., C. Tsai, H. Wang, D. Nickerson, and B. Li. 2009. Subgroup analysis via recursive partitioning. *The Journal of Machine Learning Research* 10:141–158.
Su, X., T. Zhou, X. Yan, J. Fan, and S. Yang. 2008. Interaction trees with censored survival data. *The International Journal of Biostatistics* 4 (1). doi:10.2202/1557-4679.1071.
The White House. 2015. FACT SHEET: President Obama's Precision Medicine Initiative.
Therneau, T., B. Atkinson, and B. Ripley. 2015. Package 'rpart'.
Tian, L., A. Alizadeh, A. Gentles, and R. Tibshirani. 2014. A simple method for estimating interactions between a treatment and a large number of covariates. *Journal of the American Statistical Association* 109 (508):1517–1532. doi:10.1080/01621459.2014.951443.
Tunis, S., J. Benner, and M. McClellan. 2010. Comparative effectiveness research: Policy context, methods development and research infrastructure. *Statistics in Medicine* 29 (19):1963–1976. doi:10.1002/sim.v29:19.
Xu, Y., M. Yu, Y. Zhao, Q. Li, S. Wang, and J. Shao. 2015. Regularized outcome weighted subgroup identification for differential treatment effects. *Biometrics* 71 (3):645–653. doi:10.1111/biom.12322.
Young, S. J. 1996. Panic associated with combining fluoxetine and bupropion. *The Journal of Clinical Psychiatry* 57 (4):177.
Zeileis, A., T. Hothorn, and K. Hornik. 2008. Model-based recursive partitioning. *Journal of Computational and Graphical Statistics* 17 (2):492–514. doi:10.1198/106186008X319331.
Zhang, B., A. Tsiatis, E. Laber, and M. Davidian. 2012b. A robust method for estimating optimal treatment regimes. *Biometrics* 68 (4):1010–1018. doi:10.1111/j.1541-0420.2012.01763.x.
Zhang, B., A. Tsiatis, M. Davidian, M. Zhang, and E. Laber. 2012a. Estimating optimal treatment regimes from a classification perspective. *Stat* 1 (1):103–114. doi:10.1002/sta.411.
Zhang, Y., E. Laber, A. Tsiatis, and M. Davidian. 2015. Using decision lists to construct interpretable and parsimonious treatment regimes. *Biometrics* 71 (4):895–904. doi:10.1111/biom.12354.
Zhang, Y., E. B. Laber, A. Tsiatis, and M. Davidian. 2016. Interpretable dynamic treatment regimes. *arXiv preprint arXiv:1606.01472*.
Zhao, L., L. Tian, T. Cai, B. Claggett, and L. Wei. 2013. Effectively selecting a target population for a future comparative study. *Journal of the American Statistical Association* 108 (502):527–539. doi:10.1080/01621459.2013.763566.
Zhao, Y., D. Zeng, A. Rush, and M. Kosorok. 2012. Estimating individualized treatment rules using outcome weighted learning. *Journal of the American Statistical Association* 107 (499):1106–1118. doi:10.1080/01621459.2012.682532.
Zhao, Y., D. Zeng, E. Laber, R. Song, M. Yuan, and M. Kosorok. 2015. Doubly robust learning for estimating individualized treatment with censored data. *Biometrika* 102 (1):151–168. doi:10.1093/biomet/asu050.
Zhu, R., Y. Zhao, G. Chen, S. Ma, and H. Zhao. 2016. Greedy outcome weighted tree learning of optimal personalized treatment rules. *Biometrics 73 (2):391-400.*
Zimmerman, M., and I. Chelminski. 2006. A scale to screen for DSM-IV axis I disorders in psychiatric out-patients: Performance of the psychiatric diagnostic screening questionnaire. *Psychological Medicine* 36 (11):1601–1611. doi:10.1017/S0033291706008257.
Zimmerman, M., and J. Mattia. 2001. A self-report scale to help make psychiatric diagnoses: The psychiatric diagnostic screening questionnaire. *Archives of General Psychiatry* 58 (8):787–794. doi:10.1001/archpsyc.58.8.787.