FACULTY OF PSYCHOLOGY AND
EDUCATIONAL SCIENCES

Research Group Quantitative Psychology and Individual Differences

Tree-based methodologies for the detection of treatment-subgroup interactions and the estimation of optimal treatment regimes in randomized controlled trials

Doctoral research project provided by
Lisa Doove
Supervisor: Prof. Dr. Iven Van Mechelen
Co-supervisors: Dr. Elise Dusseldorp
Dr. Katrijn Van Deun

April 2014

Abstract

For many medical and psychological problems, multiple treatment alternatives are available. An obvious question in such cases pertains to which treatment alternative to choose. In this project, we focus on contexts in which this question is to be dealt with in terms of static choices between different treatment alternatives as a whole, with the choices being personalized in terms of subgroups of clients that are unknown and that are to be learned from the data. We further focus on a setting with data from randomized controlled trials (RCTs) in which a large number of pre-treatment characteristics is available. Ultimately, we will address the problem at hand by looking for optimal treatment regimes (i.e., decision rules that assign to each client a treatment, out of the set of available treatment alternatives, based on his/her pre-treatment characteristics), with the optimal regime being the one leading to the best expected (potential) outcome in the population under study. The cornerstone for the development of such optimal treatment regimes is the detection of so-called qualitative treatment-subgroup interactions, that is, in the case of two treatment alternatives A and B, interactions which imply that for some subgroups of clients treatment A outperforms treatment B, whereas for other subgroups the reverse holds true.

The detection of qualitative treatment-subgroup interactions and the estimation of optimal treatment regimes involve quite some statistical challenges. Recently, a promising class of tree-based methods has been proposed to deal with some of these. Unfortunately, however, the methods in question also bear a number of conceptual and methodological limitations. In this project, we will address four of them.

First, the methods in question have been developed almost independently, and the relations between them are not yet fully understood. We will address this limitation by closely examining the relations between a selection of five tree-based methods and by applying them to data from one and the same RCT. Second, one tree-based method (called QUINT – Dusseldorp & Van Mechelen, 2014) is to be singled out because of its immediate relevance for optimal treatment assignment (due to its exclusive focus on qualitative treatment-subgroup interactions instead of treatment-subgroup interactions in general); however, the paper in which QUINT has been introduced is not easily accessible for non-methodologists. We will address this limitation by clearly and concisely explaining the conceptual basis of QUINT to non-methodologists and by illustrating its significance for psychological applications. Third, the criterion that is optimized by QUINT is an ad hoc criterion that is not at the level of inferences about a population of interest. To address this limitation we will develop a novel methodology that optimizes a population-based criterion while looking for a tree-based optimal treatment regime; we will further extensively evaluate the methodology's optimization performance and inferential qualities. Fourth, QUINT is limited to the case of two-arm randomized controlled trials, whereas quite a few RCTs include more than two treatment alternatives. To address this limitation, we will extend our novel methodology for constructing optimal tree-based treatment regimes to data from RCTs that involve more than two treatment alternatives or variants of the same treatment type, yet applied under different modalities.

Introduction

For many medical and psychological problems, multiple treatment alternatives are available. An obvious question in such cases pertains to which treatment alternative to choose. This question can take several specific forms depending on the context. Indeed, firstly, either a single static choice between different treatment alternatives as a whole could be at issue, or dynamic choices between different treatment steps in a succession of several treatment phases. Secondly, the research question may pertain to one universal choice for the full population of clients under study, or it may concern personalized choices, that is, custom-made choices for different subgroups of clients that can be characterized in terms of pre-treatment characteristics. Thirdly, in the case of personalized choices, these may be studied within a context of pre-specified subgroups of clients across which relative treatment effectiveness is hypothesized to vary, or within a context in which the subgroups are unknown and are to be learned from the data. In this PhD project, we focus on the case of a single static choice that is personalized in terms of subgroups of clients that are unknown and should be learned from the data.

The choice problem specified above is typically studied either in studies of randomized controlled trials (RCTs), in which clients are randomly assigned to the alternative treatment conditions, or in observational studies, in which the assignment of the clients to the alternative treatment conditions is outside the control of the investigator. In both types of studies, the clients are measured in terms of a set of pre-treatment characteristics, in addition to (at least) one outcome variable. In this project we will focus on data that are obtained from RCTs in which a large number of pre-treatment characteristics are available. However, for inferential purposes, at certain points a digression will be made to aspects of observational studies.

The main ingredients of the data we focus on in this project are: a treatment variable *T*, an outcome variable *Y*, and a set of pre-treatment characteristics $X_v \;\; (v = 1, \ldots, V),$ which may be continuous or categorical in nature. Given client $i = 1, \ldots, I,$ the observed data are then $(Y_i, T_i, \boldsymbol{X}_i).$ In addition to the observed outcome variable, we will also consider potential outcomes, making use of the so-called Neyman-Rubin Causal Model (Morgan & Winship, 2007; Rubin, 1974). The potential outcome of a treatment for a client pertains to the outcome that would be observed if the client in question were subject to that treatment, irrespective of whether this is actually the case. Potential outcomes are a key constituent of the concept of optimal treatment regimes. Formally speaking, a treatment regime is a decision rule that assigns to each client a treatment, out of the set of available treatment alternatives, based on his/her observed characteristics (Murphy, 2003). The optimal regime is the one leading to the greatest expected potential outcome in the population under study. The cornerstone for the development of optimal treatment regimes is the detection of meaningful so-called qualitative treatment-subgroup interactions (Rothwell, 2005), that is (in the case of two treatment alternatives A and B), interactions that imply that for some subgroups of clients treatment A outperforms treatment B, whereas for other subgroups the reverse holds true. The detection of such interactions, along with the development of optimal treatment regimes, can be considered crucial for personalized medicine (Tunis, Benner, & McClellan, 2010; Dehejia, 2005).

The search for an optimal personalized answer to the question of which treatment alternative to choose among a set of treatment alternatives within a context of subgroups of clients that are unknown and are to be learned from the data implies quite some statistical issues. A first critical issue pertains to the identification of the subgroups involved in treatment-subgroup interactions and the related building of optimal treatment regimes. Once a treatment regime has been built, a second issue pertains to estimating the performance of the treatment regime, that is, the population mean potential outcome under the regime. A final issue involves hypothesis testing, and in particular whether a personalized treatment assignment outperforms the strategy of assigning all clients to the marginally best treatment (which comes down to testing the presence of a qualitative treatment-subgroup interaction). These statistical issues are fairly challenging. A first challenge arises from the fact that, in many RCTs, typically a large number of pre-treatment characteristics is available in the data. Within the present context of unknown subgroups of clients that are to be learned from the data, this implies a vast search space for possible subgroups involved in qualitative treatment-subgroups interactions. A second challenge is situated on the level of treatment regime performance estimation: Estimators may behave poorly due to model misspecifications and may be biased due to capitalization on the data in the construction of the treatment regime. A third challenge concerns the risk of inferential errors. On the one hand, these include Type II errors, which reflect a possible lack of power to detect true treatment-subgroup interactions (e.g., Pocock, Assmann, Enos, & Kasten, 2002). Such a detection generally requires larger samples than the detection of main effects, and perhaps considerably larger than those enrolled in a number of traditional clinical trials. On the other hand, and even more importantly, one should also beware of Type I errors, that is, erroneous claims about the occurrence of apparent interactions that cannot be replicated in follow-up studies (Dixon & Simon, 1991; Pocock et al., 2002; Rothwell, 2005; Wang, Lagakos, Ware, Hunter, & Drazen, 2007). Otherwise, because of the presumed relatively high risk of Type I errors, a number of critics remembered subgroup analysis with the pet name of 'computerized data dredging' (Feinstein, 1998; Rothwell, 2005).

A broad range of methods is available for the study of treatment-subgroup interactions and the related development of optimal treatment regimes. Much work in this area primarily pertains to linear-model-based methods. Examples of models involved comprise factorial analysis of variance (ANOVA) with a first factor pertaining to treatment methods and a second one to subgroups (Shaffer, 1991), and (logistic) regression analyses with suitable interaction terms being included in the regression model (see, e.g., Brinkley, Tsiatis, & Anstrom, 2010; Dixon & Simon, 1991; Hayward, Kent, Vijan, & Hofer, 2006; Qian & Murphy, 2011), which optionally could be further used to define a class of treatment regimes within which an optimal regime could be looked for (Zhang, Tsiatis, Laber, & Davidian, 2012; Zhao, Zeng, Rush, & Kosorok, 2012). These linear-model-based methods are especially suitable in situations in which comprehensive a priori hypotheses exist about which subgroups of clients are involved in the interactions or in which a small number of characteristics are available to define the subgroups. However, they are not readily applicable within the context of unknown subgroups to be learned from data that include a (very) large number of pre-treatment characteristics. In such situations they should, at the very least, be supplemented with variable selection procedures (see, e.g., Gunter, Zhu, & Murphy, 2011).

A promising family of recently proposed tree-based methods does not require a priori hypotheses nor a limited number of potential moderator variables. Rather, these methods induce subgroups involved in treatment-subgroup interactions during the actual data analysis. In particular, they do so via recursive partitioning, which implies that the total group of clients is repeatedly split into child subgroups that vary in terms of relative treatment effectiveness. This family of methods constitutes the focus of the present project. First and foremost, it includes a group of fully tree-based methods. These are: Model-based recursive partitioning (MOB; Zeileis, Hothorn, & Hornik, 2008), Interaction Trees (Su, Tsa, Wang, Nickerson, & Li, 2009; Su, Zhou, Yan, Fan, & Yang, 2008), Simultaneous Threshold Interaction Modelling Algorithm (STIMA; Dusseldorp, Conversano, & Van Os, 2010), Subgroup Identification based on Differential Effect Search (SIDES; Lipkovich, Dmitrienko, Denne, & Enas, 2011), Virtual Twins (Foster, Taylor, & Ruberg, 2011), and QUalitative INteraction Trees (QUINT; Dusseldorp & Van Mechelen, 2014). Second, this family also includes a partially tree-based method to derive optimal treatment regimes. This method, which was originally developed within the context of observational studies to compare two treatment alternatives, starts by using a model-based estimation of the expected difference in treatment effectiveness at each of the observed data points; the model basis of the estimator in question includes a regression-type propensity model (possibly in addition to a regression-type outcome model). Subsequently, the estimation results are subjected to a classification tree method to build an optimal treatment regime (Zhang, Tsiatis, Davidian, Zhang, & Laber, 2012).

With regard to the fully tree-based methods, a major problem reads that these methods have been developed almost independently, and that the relations between them are not yet fully understood. Furthermore, within the group of fully tree-based methods, QUINT is to be singled out because of its relevance for optimal treatment assignment. In fact, QUINT has an exclusive focus on qualitative treatment-subgroup interactions while the other methods focus on treatment-subgroup interactions in general; yet, qualitative interactions are of particular relevance for optimal treatment assignment. Unfortunately, however, the QUINT method also has a number of disadvantages. A first, pragmatic disadvantage is that the paper in which the QUINT methodology has been introduced is not easily accessible for non-methodologists. A second disadvantage is that the criterion that is optimized by QUINT is an ad hoc criterion that is not at the level of inferences about a population. A third disadvantage is that QUINT, as most tree-based methods, is estimated on the basis of a greedy heuristic. A fourth disadvantage is that QUINT is limited to the case of two alternative treatments, whereas quite a few randomized controlled trials include more than two treatment alternatives.

As regards the partially tree-based method, its use primarily requires the specification of external regression models that are independent of the classification tree method to build an optimal treatment regime. This specification can be fairly challenging within the context of data that include a large number of pre-treatment characteristics. Anyhow, this method can serve as a useful benchmark for performance evaluation of fully tree-based methods.

In the present project, we will address the four disadvantages listed above. This results in the following four research objectives:

1) *Capturing the relation between existing recursive tree partitioning methods to study treatment-subgroup interactions.* We want to clarify the conceptual basis of the recursive tree partitioning methods, and capture the relation between them.

2) *Reviewing the conceptual basis of QUINT.* We want to clearly and concisely explain the conceptual basis of QUINT to non-methodologists. Moreover, we also want to illustrate the significance of QUINT for psychological applications.

3) *Construction of a globally optimal tree-based treatment regime with control of inferential quality.* We want to develop a novel fully tree-based methodology to construct globally optimal treatment assignment rules. The ultimate criterion that is to be optimized by this methodology should be the expected potential outcome of the resulting treatment regime. The methodology should rely on both greedy and global optimization approaches. Its performance should be compared with that of estimated optimal treatment regimes that may be derived from other fully and partially tree-based methods.

4) *Extension to multiple treatments.* We want to extend our methodology to data from randomized controlled trials that involve multiple treatments, or multiple variants of the same treatment type, yet applied under different modalities.

Each of these objectives will be dealt with in a separate work package. Below, we will successively present the resulting four work packages.

Work packages

Work package 1: Capturing the relation between existing recursive tree partitioning methods to study treatment-subgroup interactions

Five recently proposed recursive tree partitioning methods can be used to address the identification of subgroups involved in treatment-subgroup interactions. These are: MOB, Interaction Trees, STIMA, SIDES, and Virtual Twins. (Note that QUINT is not included in this list as the QUINT paper was still under review at the time of the activities for this work package.)

A major problem that may hamper the use of these methods for the detection of treatment-subgroup interactions is that the relations between the methods are not fully understood as they have been developed almost independently. This situation is further aggravated by the fact that some of the papers introducing these methods are not easily intelligible, even for methodologists.

Objectives

The objectives of this work package respond to the problem outlined above:

1) Clarify the conceptual basis of the different individual methods
2) Capture the relation between them

Methodology

We will start by outlining each one of the different methods. To then capture the relation between the methods, we will make a comparison between them on a theoretical/conceptual level. Next, we will provide an overview of the methods with respect to their goals, the type of data they can handle, their underlying model structure, their algorithmic process, and the availability of software. Based on this conceptual comparison, we will also examine questions that are important for end users in making a selection between the different methods.

In addition, we will apply the different methods to a randomized controlled clinical trial data set. Our focus will be on existing data from a randomized controlled trial in which clients have been randomly assigned to two alternative treatment conditions, with measurements of a broad range of pre-treatment characteristics of the clients, and of at least one outcome variable. Our primary interest will be in having a concrete instantiation of the type of output of each method. For this reason, we will use in the application the default settings of the tuning parameters of each method under study where possible. It will nevertheless also be useful to look for a synthesis of the different results. We will therefore consider recurrent elements that show up in the output of the different methods. Finally, we will examine the size of the possibly detected treatment-subgroup interactions. For this purpose we will rely on variance component analyses.

Output

This work package will result in a structured comparison of existing recursive tree partitioning methods to detect subgroups involved in treatment-subgroup interactions. In addition, it will contribute an illustrative application of the methods through analyses of data from a randomized controlled trial. Lastly, practical advice will be given to end users with regard to making a selection between the different methods. The output will comprise a publication in a peer-reviewed international journal, including supplementary materials to allow readers to replicate the results, with the data being publicly available.

Work package 2: Reviewing the conceptual basis of QUINT

The recently developed method QUINT allows the researcher to identify subgroups involved in treatment-subgroup interactions in situations where a large number of moderators are available in the data, without comprehensive a priori hypotheses on such subgroups, and with an exclusive focus on qualitative treatment-subgroup interactions (Dusseldorp & Van Mechelen, 2014). In fact, the goal of QUINT is to find the best partition of the total group of clients on the basis of background characteristics into two or three mutually exclusive subgroups that are characterized as follows: In the first subgroup, the clients assigned to treatment A show a clearly better outcome than the clients assigned to treatment B; in the second subgroup, the reverse is true; in the third (optional) subgroup, the clients assigned to A show more or less the same outcome as the clients assigned to B. The optimal partition then is such that the qualitative treatment-subgroup interaction related to the partition has the largest possible practical significance. To achieve this, the first two subgroups need to satisfy two conditions: (a) In both subgroups the difference in outcome between treatments A and B should be large, and (b) each of the subgroups should comprise many clients. QUINT uses a weighted compound criterion that implies that these two conditions are optimized simultaneously. The development of QUINT included an explicit account of the problem of inferential errors, in terms of an extensive simulation study that led to a number of strategies and recommendations to control for this problem.

The QUINT method has been extensively described by Dusseldorp and Van Mechelen (2014). However, their paper is rather technical and the objective function that is maximized by QUINT is built up over several subsections. The methodology is therefore not easily accessible and surveyable for non-methodologists. There is a need to communicate the methodology to a wider audience, and to show its significance for psychological applications.

Objectives

The objectives of this work package are twofold:

1) Clearly and concisely explain the conceptual basis of QUINT, with a focus on the big picture and underlying concerns
2) Show the significance of QUINT for psychological applications

Methodology

We will review the conceptual basis of QUINT, where equations will be avoided whenever possible. We will start from a description of the context in which QUINT operates, and the goal of the method. A second step will be to explain the QUINT criterion and its underlying stepwise tree building algorithm. Taking into account the risk of inferential errors, subsequently, the importance and working of the stopping criteria and pruning procedure will be reviewed. In addition, results of the simulation study by Dusseldorp and Van Mechelen (2014) will be summarized, including several resulting guidelines to assess the risk of inferential errors.

To show the relevance of QUINT for psychological applications, we will subject existing data from a randomized controlled trial to a reanalysis with QUINT. We will emphasize the added value of this reanalysis, and discuss the results with regard to the interpretation of the tree, which might yield hypotheses on the mechanisms underlying relative treatment effectiveness in the randomized controlled trial under study.

Output

The output will comprise a conceptual article on QUINT, including an application showing how QUINT can advance the field of psychology.

Work package 3: Construction of a globally optimal tree-based treatment regime with control of inferential quality.

The criterion that is maximized by QUINT is an ad hoc criterion. Unfortunately, it does not allow us to make inferences about a population of interest. In addition, QUINT uses a stepwise optimization approach relying on a greedy heuristic, where locally-optimal decisions are made at each partition step. QUINT is therefore not guaranteed to give a globally optimal tree solution.

Objectives

In this work package we will address two objectives:

1) Develop a methodology that optimizes a population-based criterion while looking for a tree-based optimal treatment regime
2) Control the quality of the methodology

Methodology

Concepts. We will focus on a setting in which data are collected from a two-arm randomized controlled trial. This implies that the actual treatment assignment, denoted by T = 0 or 1, is independent of any pre-treatment characteristics, which are denoted by $X_v \ (v = 1, \ldots, V)$. Let Y be the observed outcome and, without loss of generality, assume larger values of Y to be more desirable. In addition, let $Y^*(t)$ denote the potential outcome, that is, the outcome that would be observed were a client subject to treatment t. Given client $i = 1, \ldots, I,$ the observed data are then $(Y_i, T_i, \boldsymbol{X}_i)$. We further assume that the observed outcome is the potential outcome that would be observed under the treatment actually received, so that $Y = Y^*(1)T + Y^*(0)(1-T);$ this is also known as the consistency assumption. Lastly, we assume that there is no interference among clients, also known as the Stable Unit Treatment Value Assumption (SUTVA). The overall population mean potential outcome were all clients in the population subject to treatment *t* is represented by $E[Y^*(t)]$; under the assumptions made above, we can deduce that $E[Y^*(t)]$ is equal to $E_{\boldsymbol{X}}\{E[Y^*(t) \mid \boldsymbol{X}]\} = E_{\boldsymbol{X}}\{E[Y^*(t) \mid \boldsymbol{X}, T = t]\} = E_{\boldsymbol{X}}\{E[Y \mid \boldsymbol{X}, T = t]\}$, where $E_{\boldsymbol{X}}\{\cdot\}$ denotes the expectation with respect to the marginal distribution of $\boldsymbol{X}$.

In this setting, a treatment regime is a function *g* that maps values of $\boldsymbol{X}$ to {0,1}, as a formalization of the rule that a client with characteristic pattern $\boldsymbol{X} = \boldsymbol{x}$ is to receive Treatment 0 if $g(\boldsymbol{x}) = 0$ and Treatment 1 if $g(\boldsymbol{x}) = 1$. Under the above assumptions, the potential outcome that would result from assigning treatment to a randomly chosen client according to *g* is given by $Y^*(g) = Y^*(1)g(\boldsymbol{X}) + Y^*(0)[1 - g(\boldsymbol{X})]$. The optimal treatment regime then is the one leading to the greatest expected potential outcome $E[Y^*(g)]$ in the population under study. Limiting ourselves to some class of treatment regimes denoted by *G*, we may define the optimal treatment regime, g^{opt}, as the one leading to the largest value of $E[Y^*(g)]$ among all $g \in G$, that is to say, $g^{\mathrm{opt}}(\boldsymbol{X}) = \arg\max_{g \in G} E[Y^*(g(\boldsymbol{X}))]$.

In this project, we will focus on treatment regimes that rely on a tree structure, which implies that decisions with regard to treatment assignment are based on leaf membership. Let S_ℓ $(\ell = 1, ..., L)$ be the leaves associated with such a tree (with the leaves being defined in terms of conjunctions of thresholds on pre-treatment characteristics, such as, e.g., $[(X_1 > z_1)$ and $(X_2 \le z_2)])$. A tree-based treatment regime can then be formalized as a function $\tilde{g} : \{S_1, \ldots, S_\ell, \ldots, S_L\} \to \{0,1\}$. Let us denote by $i \in S_\ell$ that client *i* belongs to leaf S_ℓ; by abuse of notation, let us further denote by $(X_1, \ldots, X_V) \in S_\ell$ that the pattern of pre-treatment characteristics meets the definition of leaf S_ℓ. Then, the expected overall population potential outcome were all clients in the population to receive treatment according to treatment regime $\tilde{g}$ is given by:

$$E\left[Y^*\left(\tilde{g}\right)\right] = \sum_{\ell=1}^{L} E\left[Y^*\left(\tilde{g}\left(S_\ell\right)\right) \mid \left(X_1, \ldots, X_V\right) \in S_\ell\right] \cdot \mathrm{pr}\left[\left(X_1, \ldots, X_V\right) \in S_\ell\right], \qquad (1)$$

where $\mathrm{pr}[(X_1, \ldots, X_V) \in S_\ell]$ denotes the probability of a client to belong to leaf S_ℓ.

Methods. In order to estimate an optimal tree-based treatment regime we will first identify an estimator of $E[Y^*(\tilde{g})]$ as specified in Equation (1). Subsequently, we will optimize it to obtain an estimator $\hat{\tilde{g}}^{\text{opt}}(\boldsymbol{X})$ for $\tilde{g}^{\text{opt}}(\boldsymbol{X})$. In order to define an estimator of $E[Y^*(\tilde{g})]$, let $\overline{Y}_\ell$ denote the mean outcome in leaf S_ℓ, and $\overline{Y}_{\ell|T=0}$ and $\overline{Y}_{\ell|T=1}$ the conditional mean outcomes for the two treatment groups in leaf S_ℓ. Our proposed estimator for $E[Y^*(\tilde{g})]$ is then given by:

$$\sum_{\ell=1}^{L} \overline{Y}_{\ell|T=\tilde{g}(S_\ell)} \times \frac{\# S_\ell}{I}, \tag{2}$$

where $\# S_\ell$ denotes the cardinality of leaf S_ℓ. If we denote by $I_{[T_i=\tilde{g}(S_\ell)]}$ an indicator function, which is such that, when $I_{[T_i=\tilde{g}(S_\ell)]}=1$, $Y^*(\tilde{g})$ is observed, and when $I_{[T_i=\tilde{g}(S_\ell)]}=0$, $Y^*(\tilde{g})$ is missing, the proposed estimator can be rewritten as

$$\frac{1}{I}\sum_{\ell=1}^{L} \frac{\sum_{i \in S_\ell} I_{[T_i=\tilde{g}(S_\ell)]} Y_i}{\text{p}_\ell^{\tilde{g}}}, \tag{3}$$

with $\text{p}_\ell^{\tilde{g}} = \frac{1}{\# S_\ell} \sum_{i \in S_\ell} I_{[T_i=\tilde{g}(S_\ell)]}$, that is, the proportion of clients in subgroup S_ℓ who have been actually assigned to the treatment that according to regime $\tilde{g}$ should be associated with subgroup S_ℓ. The estimator weights the observed outcomes in subgroup S_ℓ that line up with regime $\tilde{g}$ by the inverse of the observed proportion $\text{p}_\ell^{\tilde{g}}$. This criterion bears similarity to the inverse probability weighted estimator of the mean potential outcome (see, e.g., Zhang, Tsiatis, Laber, & Davidian, 2012). Estimators of an inverse probability weighting type may, under certain conditions, allow for a consistent estimation of mean potential outcomes. Based on a missing data analogy, they rely on a weighting by the inverse of the estimated probability of nonmissingness (conditional on the pre-treatment characteristics), which is a direct function of the estimated propensity score (i.e., the probability of actually receiving Treatment 1). Propensity scores were originally introduced as a tool to estimate treatment effects from observational data (Rosenbaum & Rubin, 1983). In the setting of randomized controlled trials, the value of the propensity score is known exactly (i.e., the probabilitites are under control of the investigator);

however, even in that setting it could still be beneficial to rely on estimated propensity scores as an adjustment for chance imbalance of prognostic variables (Williamson, Forbes, & White, 2014). Our Estimator (3) relies on observed proportions rather than on estimates based on a propensity score model. By analogy with Williamson et al. (2014), inverse weighting using observed proportions could be considered as an adjustment for chance imbalance of treatment assignment conditional on the pre-treatment characteristics.

An estimator for $\tilde{g}^{\text{opt}}$ may be obtained by maximizing the estimator for $E[Y^*(\tilde{g})]$, as specified in (3), across a class of tree-based regimes G to obtain $\hat{\tilde{g}}^{\text{opt}}(\boldsymbol{X})$. For this optimization we will develop two tree partitioning methods, based on a stepwise and a global optimization approach, respectively, that both aim at maximizing (3).

The stepwise approach will be based on a greedy heuristic that looks in each step for the triplet consisting of a parent node, a variable X_V, and a split point, that implies the largest increase in (3). The main advantage of this approach will be that it will allow for building trees in presence of a large number (*V*) of pre-treatment characteristics within a feasible amount of time. As already mentioned, however, the stepwise approach is not guaranteed to give an optimal solution for the maximization of (3), as locally-optimal decisions are made at each partition step.

To obtain a tree-based regime that globally optimizes (3) we will rely on dynamic programming (Hubert, Arabie, & Meulman, 2001; Van Os, 2001). To clarify this, explicit enumeration methods for combinatorial problems such as global estimation of partitioning trees suffer from a combinatorial explosion. To meet this limitation, implicit enumeration techniques try to reduce the size of the enumeration problem while implicitly ensuring the complete evaluation. Dynamic programming is a general principle for reducing the size of the enumeration problem, based on decomposing the problem into subproblems, and storing the optimal solution for each subproblem. It should be noted that quite some issues are implied by an efficient implementation of dynamic programming to identify globally optimal tree-based treatment regimes. A first major issue is to avoid redundant calculations. A second issue concerns the efficient storage of optimal solutions of previously searched nodes. At this point we will rely on the principles dicussed by Van Os (2001), who reviews so-called retrace and swap strategies to reduce storage requirements. A final issue concerns possibly excessive running times when a very large number of pre-treatment characteristics are available in the data. To cope with this we will rely on different strategies to reduce the search space of characteristics, including: (1) using the stepwise partitioning tree method to do an intitial search for relevant characteristics and possible tree structures, (2) pre-specifying a minimum required sample size in the subgroups, which prevents the algorithm from considering extreme partitions, and (3) limiting the number of possible splits for characteristics that are continous. The latter family of strategies will not only reduce the search space, but may also lead to an increased stability of the partitioning trees found, and possibly reduce variable selection bias (see, e.g., Loh, 2002).

A major challenge in the development of a methodology to construct optimal tree-based treatment regimes pertains to the development of suitable procedures for model selection. An important target in this context is the complexity of the partitioning tree in terms of the number of splits and the number of variable-split combinations involved. To address this challenge, we will start from standard techniques used in tree-based analysis for pruning (see, e.g., LeBlanc & Crowley, 1993). An important bottleneck in our proposed methodology, that can also be expected to hamper the model selection procedure, is the fact that the estimated criterion values (3) will be strongly positively biased as they will have been determined on the basis of the same data as the ones that were used to build the tree. To deal with this, we will develop tailor-made bootstrap-based bias correction procedures (Efron, 1983; LeBlanc & Crowley, 1993).

Yet another major challenge involves hypothesis testing, in particular to check whether, given the induced set of subgroups, the optimal treatment regime resulting from the tree partitioning methods significantly outperforms the strategy of assigning all clients to the marginally best treatment (in terms of expected potential outcome). This is a critical test that is equivalent to an overall test of the presence of a qualitative treatment-subgroup interaction for the induced set of subgroups. To develop such a test we will start from previous proposals of statistics for the testing of qualitative treatment-subgroup interactions (see, e.g., Bayman, Chaloner, & Cowles, 2010; Gail & Simon, 1985; Shaffer, 1991). However, in this project these will figure in pseudo tests as the set of subgroups is induced by maximizing the performance of the optimal treatment regime from the same data as the ones used for the hypothesis testing.

Evaluation. We want to evaluate the performance of the newly proposed stepwise and global procedures for building tree-based optimal treatment regimes. Moreover, we want to do so in a broad range of data settings that vary in terms of several kinds of data characteristics. At this point, we will consider five important evaluation aspects. A first evaluation aspect concerns the optimization performance of the stepwise approach. This aspect pertains to whether the stepwise algorithm is successful in identifying a treatment regime with optimal estimated mean potential outcome. This question is linked to the greedy nature of this algorithm: The stepwise approach guarantees that within each split, an optimal partitioning is found but not that the solution after several splits is still globally optimal. We would like to get an idea to what extent this is the case. A second evaluation aspect concerns the estimated criterion values (3). At this point, two sub questions can be distinguished. Given an estimated treatment regime, the first question pertains to the quality of its estimated criterion value. This relates to the possible bias (i.e., optimism) in the estimation of the criterion as already discussed above, along with the quality of the to be developed bias correction procedures. Regarding the second question, one may wish to know the quality of the estimated treatment regime compared to the true optimal treatment regime, as well as to the treatment regimes that may be derived from other tree-based methods, including, in particular, Virtual Twins (Foster, Taylor, & Ruberg, 2011), and the partially tree-based methods developed by Zhang, Tsiatis, Davidian et al. (2012). A third evaluation aspect concerns recovery performance, which refers to the extent to which the stepwise and global algorithms are successful in retrieving the true tree structure underlying the data (if such a structure underlies the data, indeed). This aspect comprises recovery of the complexity of the true tree (i.e., the quality of the model selection procedure), recovery of the structure of the true splitting variables and split points, and recovery of the assignment of the leaves to the two treatment alternatives. At this point, too, we would like to make a comparison with Virtual Twins and the partially tree-based methods proposed by Zhang, Tsiatis, Davidian et al. (2012). A fourth evaluation aspect pertains to the performance of the to be developed hypothesis testing procedure. That is, we want to know the probability that the methodology decides wrongly that (a) a qualitative interaction is present in data generated from a true tree structure without a qualitative treatment–subgroup interaction (Type I error), and (b) a qualitative interaction is not present in data generated from a true tree structure with a qualitative treatment–subgroup interaction (Type II error).

We will set up an extensive series of simulation studies to evaluate all four performance aspects listed above. In these studies we will make use of data simulations based on formal models that involve different optimal treatment regimes. The optimal treatment regimes will differ in complexity in terms of the number of moderator-threshold combinations involved. Moreover, we will include one scenario in which the optimal treatment regime will be to assign Treatment 1 to the entire population. In the data generation, a broad range of data characteristics will be systematically manipulated (e.g., sample size, number of possible moderators), and we will simulate a number of replications for each scenario.

We will apply to each data set the newly proposed stepwise and global optimization approach for building tree-based optimal treatment regimes, as well as Virtual Twins and the partially tree-based methods. For the partially tree-based method, we will specify logistic regression models for the propensity scores and ordinary regression models for the outcome.

On the basis of all this we will address the four evaluation aspects listed above as follows (while investigating the influence of the manipulated data characteristics through suitable analyses of variance). Firstly, regarding the optimization performance of the stepwise approach, we will compare the criterion values of the solutions provided by the stepwise algorithm with those of the global optimization approach if feasible, and, if not, with the estimated mean potential outcome of the true structure that generated the data. Secondly, to assess the quality of the bias-corrected estimated criterion value of the result of the tree partitioning methods $\hat{g}^{\text{opt}}$, we will compare, along the lines of Zhang, Tsiatis, Davidian, et al. (2012), this value with a Monte Carlo estimation of $E[Y*(\hat{g}^{\text{opt}})]$ obtained by averaging the value of (3) for a large number of simulated replications of each data set under study. Moreover, to assess the optimality of the estimated treatment regimes $\hat{g}^{\text{opt}}$ identified by our newly proposed methodology, we will compare the Monte Carlo estimation of $E[Y*(\hat{g}^{\text{opt}})]$ with the mean potential outcome of the true structure that was used to generate the data and with a Monte Carlo estimation of the mean potential outcome of the treatment regimes derived from Virtual Twins and the partially tree-based methods. Thirdly, we will compare the complexity, splitting variables, and split points of the tree structures of the treatment regimes estimated by the newly proposed, Virtual Twins and partially tree-based approaches with their true counterparts. In addition, we will compute the agreement between the assignment to the treatment alternatives for the estimated treatment regimes and the true assignment. Fourthly, regarding the hypothesis testing, making use of suitable simulated data sets we will investigate the Type I and Type II error rates.

Ouput

The output will comprise a paper in which the new methodology is extensively described and with extensive data on its performance. The work package will also contribute publicly available software, with clear guidelines on how to use it properly, and with clear information on the minimum requirements for arriving ar correct inferences.

Work package 4: Extension to multiple treatments.

Quite a few randomized controlled trials include more than two treatment alternatives, or multiple variants of the same treatment type, yet applied under different modalities. It would therefore be most useful to extend our methodology to deal with such cases. This first requires two small extensions on a conceptual level.

A first conceptual extension pertains to the concept of a potential outcome. This can be readily extended to a many-valued treatment variable (see, e.g., Morgan & Winship, 2007, pp. 53-57), the key difference with the two-valued situation being that now many more potential outcomes are unobservable rather than observable.

A second extension pertains to the concept of treatment regimes. This concept could again be extended immediately to the many-valued case, the only difference being that treatment variable T may now take values $\{0,1,\ldots,A\}$ so that a treatment regime is denoted by a function $\tilde{g}:\{S_1,\ldots,S_\ell,\ldots,S_L\}\rightarrow\{0,1,...,A\}$. As in the two-valued case, the criterion we will optimize is the expected potential outcome of the treatment regime.

Objectives

The objectives of this work package are twofold:

1) Extension of the methodology to optimize a population-based criterion while looking for a tree-based optimal treatment regime

2) Evaluation of the performance of the extended methodology

Methodology

On a methodological level, the extension to the many-valued treatment case will involve limited extensions only regarding the criterion that is to be optimized and the leaf assignment to the different treatment alternatives. Yet, a much more significant bottleneck can be expected on the level of the performance of the extended methodology. Reasons for this include the fact that now many more outcomes are unobservable rather than observable, and that the solution space of tree-based treatment regimes will become even larger (which will probably result in a stronger bias for the estimator of the criterion values). As a result of all this, it can be expected that considerably larger sample sizes will be needed for arriving at correct inferences in the many-valued setting. Anyhow, in the evaluation of the performance of the extended methodology, all evaluation aspects and questions listed in Work package 3 will be revisited. Otherwise, one may note that, in order to compare the performance of our extended methodology with the partially tree-based optimal treatment regime methods, an extension of the Virtual Twins and partially tree-based methods will be required as well. This extension will go with propensity score models based on multinomial logistic regression, and estimators of the expected difference in treatment effectiveness for all pairs of treatment alternatives. Subsequently, the classification tree method to build an optimal treatment regime can be readily extended to the many-valued situation (Breiman, Friedman, Stone, & Olshen, 1984).

Output

The output will comprise a paper in which the conceptual basis of the extended methodology is concisely explained and which includes data on its performance. The work package will also contribute an extension of the to be developed software in Work Package 3.

Time-line

In the schedule below, it is indicated when we plan to work on the different work packages of this PhD project. During the past year and a half, Work package 1 has been finished, which resulted in a paper published in *Advances in Data Analysis and Classification*. In addition, Work package 2 has resulted in a paper that has been submitted for publication. The output of Work package 3 will comprise two papers, one of which is a software paper, and the output of Work package 4 will comprise one paper.

	January-March	**April-June**	**July-September**	**October-December**
2012				Acquire theoretical background, start WP 2
2013	WP 1	WP 1	WP 1 (revision of paper), WP 2	WP2
2014	PhD project	PhD project, WP 3	WP 3	WP3
2015	WP 3	WP 3, WP 4	WP 4	WP 4
2016	WP 4	finishing PhD, defense	revisions of papers	

References

Bayman, E. O., Chaloner, K., & Cowles, M. K. (2010). Detecting qualitative interaction: A Bayesian approach. *Statistics in Medicine, 29*, 455-463.

Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A. (1984). *Classification and Regression Trees* (1st ed.). New York: Chapman and Hall/CRC.

Brinkley, J., Tsiatis, A., & Anstrom, K. J. (2010). A generalized estimator of the attributable benefit of an optimal treatment regime. *Biometrics, 66*, 512-522.

Dehejia, R. H. (2005). Program evaluation as a decision problem. *Journal of Economics, 125*(1-2), 141-173.

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

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

Dusseldorp, E., Conversano, C., & Van Os, B. J. (2010). Combining an additive and tree-based regression model. *Journal of Computational and Graphical Statistics, 19*(3), 514-530.

Efron, B. (1983). Estimating the error rate of a prediction rule: improvements on cross-validation. *Journal of the American Statistical Association, 78*, 316-331.

Feinstein, A. R. (1998). The problem of cogent subgroups: A clinicostatistical tragedy. *Journal of Clinical Epidemiology, 51*, 297-299.

Foster, J., Taylor, J., & Ruberg, S. (2011). Subgroup identification from randomized clinical trial data. *Statistics in Medicine, 30*(24), 2867-2880.

Gail, M., & Simon, R. (1985). Testing for qualitative interactions between treatment effects and patient subsets. *Biometrics, 41*, 361-372.

Gunter, L., Zhu, J., & Murphy, S. A. (2011). Variable selection for qualitative interactions in personalized medicine while controlling the family-wise error rate. *Journal of Biopharmaceutical Statistics, 21*, 1063-1078.

Hayward, R. A., Kent, D. M., Vijan, S., & Hofer, T. P. (2006). Multivariable risk prediction can greatly enhance the statistical power of clinical trial subgroup analysis. *BMC Medical Research Methodology, 6*, 18.

Hubert, L., Arabie, P., & Meulman, J. (2001). *Combinatorial data analysis: Optimization by dynamic programming.* Philadelphia: SIAM.

LeBlanc, M., & Crowley, J. (1993). Survival trees by goodness of split. *Journal of the American Statistical Association, 88*, 457-467.

Lipkovich, I., Dmitrienko, A., Denne, J., & Enas, G. (2011). Subgroup identification based on differential effect search-a recursive partitioning method for establishing response to treatment in patient subpopulations. *Statistics in Medicine, 30*(21), 2601-2621.

Loh, W.-Y. (2002). Regression trees with unbiased variable selection and interaction detection. *Statistica Sinica, 12*, 361-386.

Morgan, S. L., & Winship, C. (2007). Counterfactuals and causal inference: Methods and principles for social research. New York: Cambridge University Press.

Murphy, S. A. (2003). Optimal dynamic treatment regimes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 65*, 331-355.

Pocock, S. J., Assmann, S. E., Enos, L. E., & Kasten, L. E. (2002). Subgroup analysis, covariate adjustment and baseline comparisons in clinical trial reporting: Current practice and problems. *Statistics in Medicine, 21*, 2917-2930.

Qian, B. M., & Murphy, S. A. (2011). Performance guarantees for individualized treatment rules. *The Annals of Statistics, 39*, 1180-1210.

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika, 70*, 41-55.

Rothwell, P. M. (2005). Subgroup analysis in randomized controlled trials: Importance, indications, and interpretation. *The Lancet, 365*, 176-186.

Rubin, D. B. (1974). Estimating causal effects of treatment in randomized and nonrandomized studies. *Journal of Educational Psychology, 66*, 688-701.

Shaffer, J. (1991). Probability of directional errors with disordinal (qualitative) interaction. *Psychometrika, 56*(1), 29-38.

Su, X., Tsa, i. C., Wang, H., Nickerson, D. M., & Li, B. (2009). Subgroup analysis via recursive partitioning. *The Journal of Machine Learning Research, 10*, 141-158.

Su, X., Zhou, T., Yan, X., Fan, J., & Yang, S. (2008). Interaction trees with censored survival data. *The International Journal of Biostatistics, 4*(1), 2.

Tunis, S. R., Benner, J., & McClellan, M. (2010). Comparative effectiveness research: policy context, methods. *Statistics in Medicine, 29*(19), 1963-1976.

Van Os, B. J. (2001). *Dynamic programming for partitioning in multivariate data analysis (Doctoral dissertation).* Leiden: Leiden University.

Wang, R., Lagakos, S. W., Ware, J. H., Hunter, D. J., & Drazen, J. M. (2007). Statistics in medicine: Reporting of subgroup analyses in clinical trials. *New England Journal of Medicine, 357*, 2189-2194.

Williamson, E. J., Forbes, A., & White, I. R. (2014). Variance reduction in randomised trials by inverse probability weighting using the propensity score. *Statistics in Medicine, 33*, 721-737.

Zeileis, A., Hothorn, T., & Hornik, K. (2008). Model-based recursive partitioning. *Journal of Computational and Graphical Statistics, 17*(2), 492-514.

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

Zhang, B., Tsiatis, A. A., Laber, E. B., & Davidian, M. (2012). A robust method for estimating optimal treatment regimes. *Biometrics, 68*, 1010-1018.

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