

On Discovering Treatment-Effect Modifiers Using Virtual Twins and Causal Forest ML in the Presence of Prognostic Biomarkers

Erik Hermansson and David Svensson(✉)

AstraZeneca R&D Gothenburg, Gothenburg, Sweden
david.j.svensson@astrazeneca.com

Abstract. The recent years have seen a rapid development in the general Machine Learning area and a similar strong trend has taken place within the drug development domain, known as Precision Medicine. Traditionally, the main focus in clinical trials has been to estimate the overall treatment effect, but the reality of treatment effect heterogeneity has led to an interest in data-driven approaches for automatically identifying treatment effect modifiers and subgroups with enhanced treatment effect. These techniques are different from general purpose supervised learning due to its causal inference flavour and the notorious presence of prognostic effects in clinical trials. In this work, we focus on the popular method `Virtual Twins`; despite being among the earliest method, relatively little attention has been paid to the specifics of its implementation, in particular to the relative merits of arm-specific modelling versus a 'common-surface' approach. Since `Virtual Twins` is a generic algorithm (and not limited to any specific predictive base model such as the originally proposed `Random Forest`) we choose to base it on `XGboost`. We also study the forest-based `Casual Forest`, recently popular due to its (by-design) unbiased estimates of individual treatment effects. We compare the performance of these methods regarding their ability to disentangle prognostic and predictive effects. This is of critical importance in precision medicine, since only the latter are related to enhanced treatment effects; also a certain tendency to mistakenly declare prognostic biomarkers as predictive has been reported in the ML literature.

Keywords: Treatment effect heterogeneity · Precision medicine · Individual Treatment Effects (ITE) · Machine learning · Prognostic and predictive biomarkers

1 Introduction

A key step in the review of clinical trial data is to assess to potential presence of treatment effect heterogeneity. This arise in different ways in drug development, and one is the routine regulatory assessment of pivotal Phase III trials known as

© Springer Nature Switzerland AG 2021
O. Gervasi et al. (Eds.): ICCSA 2021, LNCS 12952, pp. 624–640, 2021.
https://doi.org/10.1007/978-3-030-86973-1_44

consistency assessment. This is conducted along the lines described in [1,10] and aims to assess a pre-defined set of subgroups, with the aim to identify if some patient groups lacks a proper response to the novel treatment and subsequently should be excluded when the drug is approved. The subgroups are agreed upon beforehand and are defined in terms of baseline characteristics (such as age, gender, medical history, key laboratory variables, etc.) and it is standard practice to dichotomize continuous variables (cuts decided prospectively).

In this work, we look at the more 'agnostic' approach to subgroups where machine learning (ML) is applied for discovering heterogeneous treatment effects and novel subgroups: while consistency assessment might be said to be a 'regulatory perspective', this is more the 'sponsor's perspective' [19]. The key difference is to go beyond inspection of pre-specified subgroups which normally excludes the possibility of higher order interactions.

The need for such analyses is motivated, e.g., by failed trials (where the overall result did not meet the expectations) where there is medical reasons to believe that some patients are likely to respond better than average and it was not captured by the simpler univariate, pre-specified, subgroup analysis. Different level of rigour of such ML approaches might exist, ranging from strictly aiming to define a biomarker signature (with an honest treatment effect estimate associated to the subgroup, as the basis for another trial investment) to merely short-listing the most likely effect modifying biomarkers (facilitating the scientific understanding of the drug's mechanism), see [27] for further details. We focus here on the latter aspect, which arguably is one core aspect in the practice; not all methods in the area automatically define a novel biomarker signature (e.g., Causal Forest and regression-based methods do not [19]) whereas most methods provide some quantification of importance of biomarkers.

2 Setting

We focus on parallel Randomized Clinical Trials (RCT), a standard design for Phase III drug development and implying that the 'ground truth' (individual response to each treatment) cannot be directly observed (e.g., [3]). This leads to a causal inference flavour and the concept of potential outcome; more formally, we observe data $\{(Y_i, \mathbf{x}_i, T_i)\}_i^n \sim \mathbb{P}$ where i is index for patient, $\mathbf{x}_i \in \mathbb{R}^p$ is the patient's baseline data, $T_i \in \{0, 1\}$ is the binary randomized treatment allocation (with $T_i = 0$ and $T_i = 1$ representing the control and active drug, respectively), and $Y_i \in \mathbb{R}$ is the outcome (i.e., assumed continuous). From now, we drop index i for patient. As standard in causal inference [24], let $Y^{(0)}$ and $Y^{(1)}$ denote the *potential* outcomes for a given patient on treatments 0 and 1, respectively; one will be factual and one counterfactual, due to the randomization. The target of interest [5] is then to estimate the Conditional Average Treatment Effect (**CATE**)):

$$\tau(\mathbf{x}) = \mathbf{E}[(Y^{(1)} - Y^{(0)})|X = \mathbf{x}], \tag{1}$$

which is the expected treatment effect for a patient with covariates $\mathbf{x}$, also known as the *individual treatment contrast*. The core aspect in precision medicine is to

characterize $\tau(\mathbf{x})$ across the covariate space. See [22] for the nuances of such analyses and their slightly different purposes. Regardless of exact purpose, it is of practical high importance to understand if only a subset of the p variables drive $\tau(\mathbf{x})$; these are the *predictive* variables, modifying the treatment effect.

2.1 Prognostic and Predictive Effects in Clinical Trials, and Potential Confounding Issues in the ML Analysis

A general form for a linear model is

$$Y_i = h(\mathbf{x}_i) + g(\mathbf{x}_i) \cdot (T_i - \pi) + \epsilon_i \tag{2}$$

where $\epsilon_i \sim N(0, \sigma^2)$ i.i.d, π treatment probability assignment to active arm ($= 0.5$ if 1:1 randomization), and with $g()$ and $h()$ unspecified [12]. For this class of models, the treatment effect for a patient with covariates $\mathbf{x}_i$ is $g(\mathbf{x}_i)$, i.e., independent of $h(\mathbf{x}_i)$; the latter only influences the outcome y_i (identically on both arms) and is said to be prognostic. The $g(\mathbf{x}_i)$ is the predictive part, and it is possible that not all p covariates contribute to it. Since this is of key importance in our further investigations, we illustrate the concepts using the following simple model: $Y = T + x_1 + x_3 + x_2 * T + x_3 * T + \epsilon$. Here x_1 is prognostic (predicting Y); x_2 is predictive since it modifies the treatment contrast; x_3 is both prognostic and predictive; if another variable x_4 was measured in the trial it is non-informative variable.

Handling of prognostic effects in causal inference ML is a key aspect and a number of issues has been noted in the literature. It has been noted [25] that ML methods sometimes tend to confound prognostic variables with predictive ones, i.e., assigning high ranking to biomarkers which are not truly involved in any treatment interaction. In the practice of drug development, prognostic effects are often present (e.g., age, gender, disease severity are often prognostic for a given disease). These effects are typically of a stronger magnitude than treatment effects, and issues arise (revolving around prognostic effects overshadowing predictive effects) if a straightforward supervised regression approach is taken (see [19]). It is demonstrated in e.g., [21] that prognostic effects can be confounded with treatment effects in subgroups: while all prognostic baseline variables are expected to be balanced across arms due to the randomization, this does not necessarily holds in local areas of the covariate space. This is particularly an inherent problem with tree-based methods; assume x is a prognostic variable and positively correlated with the endpoint. It is expected to be evenly distributed across treatment arms in the full population, but in a (data-driven) tree-node x can now have different distributions. Since x is prognostic, it translates to a confounding with the treatment effect (e.g., [15,21]); in the node, the arm with higher x, the endpoint now (in distribution) is higher, and a naive analysis in the node would now render an artefact estimate. This is similar to inherent issues in observational data. In [25] a clear tendency of mistake prognostic effects for predictive was reported for methods (where the CRAN version of Virtual Twins was among the methods studied, [34]).

2.2 Methodology for Heterogeneous Treatment Effect ML

The literature has seen a multitude of proposed methods (e.g., [6,36,38]) and already in the 2016 paper [19] a long list of methods were described, ranging from penalized regression techniques to tree-ensemble methods. A common denominator is the current paradigm [8] to treat subgroup detection as a special case of model selection. As a background, we here only mention some well-known methods in this area before focusing in on two specific ones. The method `SIDES` [18] is a local method in the sense that it only models $\tau(\mathbf{x})$ in a sub-region of the covariate space, and has operator-choice complexity control regarding how complex a subgroup is allowed to be; `IT` [28] is an early method in this area, tree-based via interaction-test p-values; `GUIDE` [21] is also tree-based and fits local GLM models in nodes aiming to adjust for local imbalances of prognostic effects; it uses a split selection that is by design unbiased. Model-based recursive partitioning `MOB` is also tree-based and fits local models but uses a different splitting process than GUIDE's based on fluctuation tests [37], and has e.g., proposed for dose-finding trials [31]. QUINT [9] seeks qualitative interactions (i.e., with some sub-regions where the active drug might be harmful).

In this work, we compare two popular ensemble-based approaches; the 'notable' [11] **Virtual Twins** (VT) was originally proposed in 2011 [13] and became a popular choice in this area (e.g., [4,16,17,19,20,25,26,33,36,38,39]). It is a two-step approach where a supervised learning model learns to predict the outcome in the first step, then relying on factual and counterfactual predictions that renders prediction of individual treatment effects explored in a second modelling step. It is worth mentioning that this scheme is a general one in ITE modelling and not specific to VT (discussed in e.g., [3]), and we speculate that this fact has contributed to its popularity in the data science community, where the knowledge of fitting general purpose supervised learners is high. Recently, a generalized version of Random Forest has emerged, the **Causal Forest** (CF) [35] which uses a certain splitting of the training data such that the same data does not enter the tree building process and the estimation process. This is known to render unbiased estimation of individual treatment effects. Both VT and CF constructs variable importance of all predictors, offering a ranking of most important features, see next section for further details.

3 The ML Methods Compared

Before describing details of the methods we first make some general remarks: VT is a two-step procedure where the first step involves fitting a supervised learning model to the training data simply for learning to predict the outcome. This modelling then is applied for estimating the individual treatment contrasts $\tau(\mathbf{x})$. Since it appears reasonable to expect that the highest possible accuracy[1] in Step 1 would improve the overall performance of VT, we think it makes much sense to consider the XGboost rather than Random Forest. The XGboost

[1] In a cross-validation or test-data sense, but not in an overfitting sense.

[29] is arguably among the strongest general-purpose algorithms for supervised learning for tabular data (e.g., *'Undoubtedly, the reigning champions of learning from tabular currently are decision tree-based algorithms. Gradient boosted trees are often the tool of choice, evidenced from their popularity in Kaggle contests'*, [32]). It is also stressed in the recent book [11] that other base models can be used for VT. While our modernized VT is boosting-based, CF is bagging-based, i.e., based on Random Forest but with the twist that allows it to estimate individual treatment effects honestly due to separating the construction of trees from the estimation process. Another motivation to look at CF is its off-the-shelf nature; given almost non-existing practical risks of commit model fitting errors, one can consider the inclusion of CF as also offering a kind of validation of the results from the more tuning-heavy VT. For the above reasons we think it makes sense to contrast these methods. There is a practical consideration too - does it pay off to perform all the computationally optimization of the Boosting models (for VT) or will CF by default give comparable or even better results?

There is one further largely unexplored aspect that motivates our simulation study: what are the relative merits of using one- or two models in Step 1 in VT? This is elaborated on in the next section, and is related to a general modeling strategies in causal inference ML (e.g., S-learners, T-learners and hybrids [7]).

3.1 Virtual Twins Implementation and Inherent Modelling Choices; Separate Models or Shared Representation

VT is a two-step algorithm, schematically consisting of the following steps (but see the specifics further below):

- `STEP 1:` Fitting supervised regression ($\mathbf{F}$) for predicting potential outcomes $y^{(0)}$ and $y^{(1)}$; the difference of predictions $z = \hat{\mathbf{F}}(\mathbf{x}, T = 1) - \hat{\mathbf{F}}(\mathbf{x}, T = 0)$ represents predictions of individual treatment effects,
- `STEP 2:` Use another supervised regression ($\mathbf{G}$) modelling z as the response against the predictors $\mathbf{x}$; i.e., explore if baseline data $\mathbf{x}$ can predict the z. (Discussed below).

In `STEP 1`, the treatment variable T enters either via splitting the training data according to arm and fit separate models, or fitting a single model to the data where treatment assignment is present as a feature. This is a key aspect in this paper, further discussed in Sect. 3.2. In [13], the choice for $\mathbf{F}$ was Random Forest. Note that the treatment variable is not explicitly present as a feature in `STEP 2` since the focus there is to identify trends in z against the baseline predictors.

The modelling approach $\mathbf{G}$ in the `STEP 2` can be different from the one in `STEP 1`, e.g., $\mathbf{G}$ = CART was proposed in [13], when an explicit subgroup definition is sought; the union of tree nodes with better than average response then defines a subgroup with enhanced effect. If instead a ranking via Variable Importance (VIP) is the aim of the analysis (aiming to identify predictive x), any suitable supervised learner can be considered. Given the general inferiority of single trees versus tree ensembles, it makes sense to use more modern approaches, e.g., $\mathbf{G}$ = XGboost (see [11]), which is our choice.

3.2 A Key Modeling Choice: Arm-Specific or Single-Surface?

We think the following consideration is somewhat overlooked in the literature: originally, when VT emerged in 2011 [13] the recommendation was to not base the Step 1 on the training data *as given* (which is (Y, X, T)) but on an 'enriched' data set where certain pre-computed treatment-covariate interactions were manually entered as additional columns to the data set:

$$W = (Y, X, T, X \cdot I_{T=0}, X \cdot I_{T=1}) \tag{3}$$

where I_B denotes an indicator function (= 1 if B holds, otherwise zero). This may not be of obvious value, but the authors had noted that it improved numeric stability; they gave no further quantitative motivation. We now refer to this latter choice as 'expanded' training data. An intuition behind it (loosely speaking) may be that prognostic effects tend to be stronger than predictive ones in clinical trials, and by introducing new candidate treatment-related columns the ML is offered more chances to strike on the right path. (See e.g., [19] for illustrations of how prognostic effects tend to dominate the model fit if a standard tree-based ML approach is used, Sect. 5, and similar aspects are discussed in [7], Sect. 4). Note that in this formulation, a *single* Random Forest model is fitted to the expanded training data W. In the vignette of the `CRAN` package `aVirtualTwins` [34] it is explained that the user has modelling freedom: specifying base model of choice and, in particular, decide regarding the training data: arm-specific or expanded? The arm-specific option in Step 1 may be considered as the most natural starting point and we speculate that this sometimes (for that reason) is used in the practice.

The above considerations are generally known in causal inference ML (e.g., 'separate models', 'augment feature space', and 'shared representation' in [23]). In a recent paper [3], section 'Modeling choices and methods for individualized treatment effects' some principled pros and cons were discussed (although the explicit expansion (3) was not covered).

In this paper we look at the special case of these general considerations for the Virtual Twins method, and how this impacts the performance in terms of handling prognostic effects. We use notation **VT1** for the version based on training data W (a single surface approach), and **VT2** for arm-specific modelling.

3.3 Fitting Boosting Models

Regarding fitting Boosting models, efforts were made to avoid over-fitting and optimization was performed over a grid of hyper-parameters using cross validation (CV). (In the analysis of clinical trials, it is sometimes not feasible to split data into training, test and validation; the data is too expensive, the data sets are too small). The maximal number of trees considered was 1000 (preliminary experimental runs indicated that considerably fewer trees often were sufficient).

3.4 Causal Forest Implementation

We relied on the CRAN package `grf` [30] and treated CF as the 'off-the-shelf' option, using it based on `num_trees` = 10000 (Bagging-based methods allow a larger no. of trees without risking overfitting [35], and some preliminary runs confirmed that larger ensembles did not make a notable difference on the performance).

4 Simulation Experiments

We performed a simulation experiment were RCTs were simulated under two assumptions regarding total *trial size* ($N = 800$ and $N = 5000$), representing smaller and larger clinical trials. The number of *candidate biomarkers* were fixed to $p = 20$, a realistic and typical number (e.g., see [6,19,20,38]). (Smaller p is sometimes in the practice considered (e.g., [14]). Regarding larger, we stress that clinical trial data is not really 'Big Data', neither in n (trial size) or p (no. of covariates), and, hence, very large p (e.g., 1000) is outside the scope here). The x_j were i.i.d. standard normal ($\sim N(0,1)$), and the motivation for uncorrelated biomarkers is to enable to clearest possible interpretation of the results. However, we also included runs with a compound symmetry correlation structure (rendering $\text{corr}(x_j, x_k) = 0.3$) for all $j \neq k$. We generated *continuous endpoints* from models in Table 1.

Table 1. Simulation models. All models except model **M4** contain true predictive effects. Two models (**M7**, **M8**) deviate from having 5 prognostic variables.

Case:	Model:
M1	$y = -1 + 3 \cdot (x_1 + x_2 + x_3 + x_4 + x_5) + T \cdot I_{(X_7>0, X_8>0)} + \epsilon$
M2	$y = -1 + 3 \cdot (x_1 + x_2 + x_3 + x_4 + x_5) + T \cdot I_{(X_1>0)} + \epsilon$
M3	$y = -1 + 3 \cdot (x_1 + x_2 + x_3 + x_4 + x_5) + T \cdot I_{(X_{20}>0)} + \epsilon$
M4	$y = -1 + 3 \cdot (x_1 + x_2 + x_3 + x_4 + x_5) + \epsilon$
M5	$y = -1 + 3 \cdot (x_1 + x_2 + x_3 + x_4 + x_5) + T \cdot (x_{10} + x_{11} + x_{12} + x_{13} + x_{14} + x_{15}) + \epsilon$
M6	$y = -1 + 3(I_{(X_1>0)} + I_{(X_2>0)} + I_{(X_3>0)} + I_{(X_4>0)} + I_{(X_5>0)}) + T \cdot I_{(X_5>0)} + \epsilon$
M7	$y = -1 + 3 \cdot (x_1 + x_2 + x_3 + x_4 + x_5 + x_6 + x_7 + x_8 + x_9 + x_{10}) + I_{(X_{20}>0)} \cdot T + \epsilon$
M8	$y = -1 + 3 \cdot (x_1 + x_2 + x_3 + x_4 + x_5 + x_6 + x_7 + x_8 + x_9 + x_{10}) + I_{(X_1>0)} \cdot T + \epsilon$
M9	$y = -1 + 3 \cdot (x_1 + x_2 + x_3 + x_4 + x_5) + \sin(X_1) \cdot T + \epsilon$
M10	$y = -1 + 10 \cdot (x_1 + x_2 + x_3 + x_4 + x_5) + T \cdot I_{(X_1>0)} + \epsilon$
M11	$y = -1 + (x_1 + x_2 + x_3 + x_4 + x_5)^3 + X_1^3 \cdot T + \epsilon$

Other types of endpoints commonly seen in clinical trials were excluded due to the current implementation of Causal Forest is restricted to numeric ones. Another reason is that the continuous case is an informative start case and

if methods have difficulties here, the problems are not likely to diminish for other endpoints (e.g., due to [2]). It is furthermore not uncommon to make comparisons in the continuous case (e.g., [6,18,36]). All simulations contained ≥ 5 *prognostic variables* with varying prognostic strength. Including many prognostic reflects reality in clinical trials, where (by design) one includes variables known to influence at least the outcome; due to limited scientific knowledge in the drug development some variables will by necessity be non-informative, and such were included too.

We vary from model to model how many variables were truly predictive (effect modifying); some scenarios were simple (making sure such are easily detected by the machines) and some more complex. One model lacked true predictive effects (**M4**). Some models have variables with both predictive and prognostic properties (e.g., (**M2**). The predictive effects were designed to be of lesser magnitude than the prognostic effects, again to reflect reality. Most models had a linear additive structure, representing simpler cases and if these already present difficulties for the ML, more complex cases will be at least as hard. The *predictive effects* were either introduced as the presence of a true subgroup via dichotomization (a true planted subgroup, e.g., (**M2**)) or as gradual covariate-treatment interactions (no clear splits involved, e.g., (**M5**)). The focus here is simply to identify these variables rather than to define novel subgroups; this is not atypical for practical applications where an understanding of drivers often is a key consideration.

Each model scenario was iterated $k = 100$ times, and for each replicate the resulting Variable Importance was tracked (from each of the VT, and CF). For clarity, the VIP for Virtual Twins is derived from the 2nd step in the algorithm. The VIPs are not directly comparable across models, but within each it is informative to compare the relative ranking of the $x_1, ..., x_{20}$. We define `Power` as the proportion simulation iterations when a truly predictive (treatment-effect modifying) variable was top ranked, and `Type-1-Error` as the number of iterations when a none-predictive variable was top-ranked. This is non-standard terminology in ML, e.g., no formal statistical test is actually conducted on some nominal significance level, but the terminology captures the risk of getting a false positive result. We furthermore look at to what extent this Type-1-Error-rate is driven by prognostic variables.

Another key aspect is the individual treatment effects themselves and how the ability to top-rank predictive variables correlates with estimation accuracy. We hence tracked (for each of VT1, VT2 and CF) the Root Mean Square Error (`RMSE`) of z and $\hat{z}$ in each simulation iteration, where the former here denotes is the theoretical true ITE and the latter the estimated ITE. The z are known in simulations (e.g., for model **M2** it is $z = I(x_1 > 0)$). We then average over the 100 iterations. We are focusing on the *relative performance* of VT1, VT2 and CF *within* each simulation scenario (i.e., RMSE and VIP for a fixed ML method is not directly comparable across simulation models).

5 Results

Before summarizing results across the landscape of models and iterations, we first discuss some specific results. In Fig. 1 we see how the VIP distribute across the $x_1, ..., x_{20}$ for simulation model **M1**. In this specific case, it is clear that only the single-surface VT (VT1) had some success in identifying the true predictive variables x_7 and x_8; arm-specific VT (VT2) was much worse and consistently claimed prognostic variables as predictive (i.e., the box-plots for x_j, $j = 1, ..., 5$ are clearly located far above all other boxplots). VT1 had some tendency to top rank the prognostic $x_1, x_2, ..., x_5$. (How often all this happened is summarized in Fig. 2, and described further below). Causal Forest was between VT1 and VT2 in terms of performance in this example.

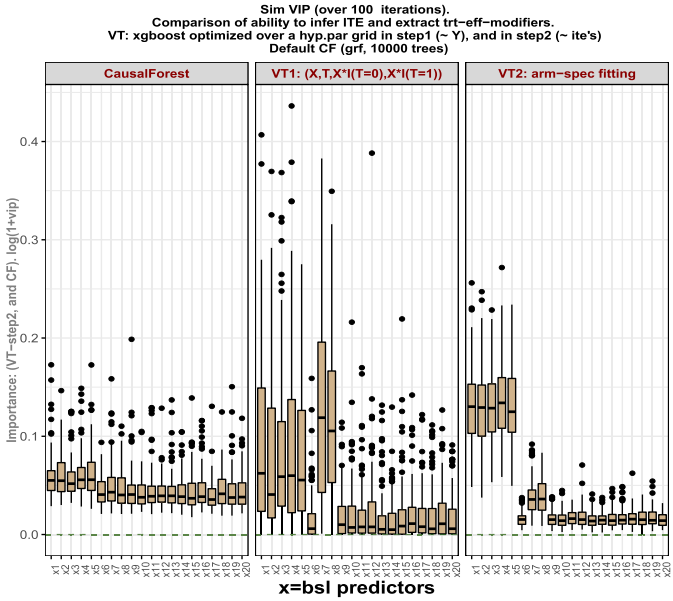

Fig. 1. Variable Importance (VIP) results for simulation model **M1**, 100 iterations (trial size 800).

Another case is displayed in Fig. 6, simulation model **M7** with rather many (ten) prognostic effects present. Then, both CF and VT2 often mistakenly top-ranked prognostic effects; this is in contrast to the single-surface VT. Furthermore the case (**M11**, $N = 5000$) is displayed in Fig. 5 where x_1 was both prognostic and predictive; here CF and VT1 performed well, whereas VT2 did (to a substantial degree) top-rank purely prognostic variables. However, the performance was very similar between VT1 and VT2 when $N = 800$ (and the same for a few other cases, e.g. **M5**, **M6**, **M9**).

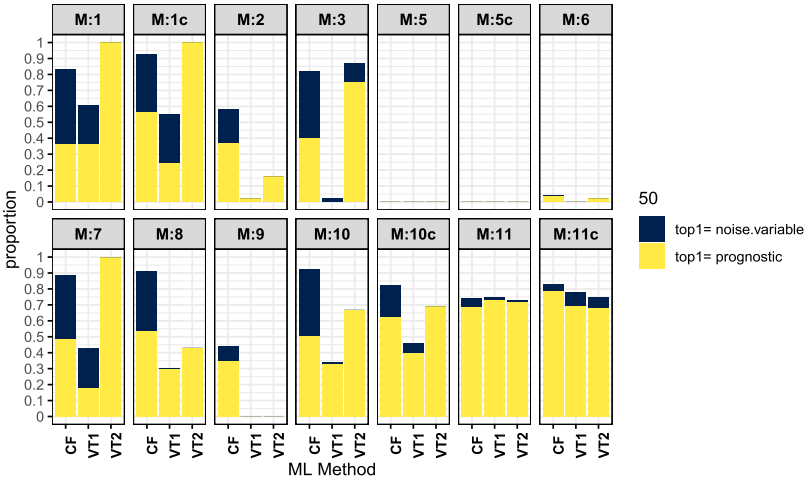

Fig. 2. Type-1-Error rates: Proportions incorrect top-ranking over 100 simulation iterations (trial size 800); total bar height displays the proportion of runs when a non-predictive variable was ranked as the most important one, with light colour scheme displaying how often this is due to a prognostic variable. (**CF** = Causal Forest, **VT1** = single-surface Virtual Twins, **VT2** = arm-specific Virtual Twins. 'c' = cases with correlation 0.3 between predictors).

The above mentioned results were typical in the sense that (1) non-informative variables were seldom top-ranked, and (2) the prognostic variables tended to drive the Type-1-Error rate (see also the Appendix for further examples). All this is summarized in what now follows.

The results regarding Type-1-Error and, in particular, how prognostic variables drive this error rate, is summarized in Fig. 2 (for trial size = 800). In that graph, the total height of a bar displays the Type-1-Error (e.g., for CF and simulation model **M1** the Type-1-Error was 0.83). The light colour scheme displays

Table 2. Simulation Result regarding Power: Proportion of simulations over 100 runs when a truly predictive variable had highest VIP. (Abbreviations: **CF** = Causal Forest, **VT1** = single-surface Virtual Twins, **VT2** = arm-specific Virtual Twins). The results for the simulation cases with correlation gave similar results to their non-correlation counterpart. Our definition of Power lacks meaning if no true predictive variables are present, hence the NA for model **M4** .

ML approach	N	**M:1**	**M:2**	**M:3**	**M:4**	**M:5**	**M:6**	**M:7**	**M:8**	**M:9**	**M:10**	**M:11**
CF	800	0.17	0.42	0.18	NA	1	0.96	0.11	0.09	0.56	0.08	0.26
VT1	800	0.39	0.98	0.98	NA	1	1	0.57	0.70	1	0.66	0.25
VT2	800	0.00	0.84	0.13	NA	1	0.98	0.00	0.57	1	0.33	0.27
CF	5000	0.74	1	1	NA	1	1	0.55	0.43	1	0.46	0.66
VT1	5000	1	1	1	NA	1	1	1	1	1	1	0.81
VT2	5000	0.00	1	1	NA	1	1	0.8	1	1	0.83	0.37

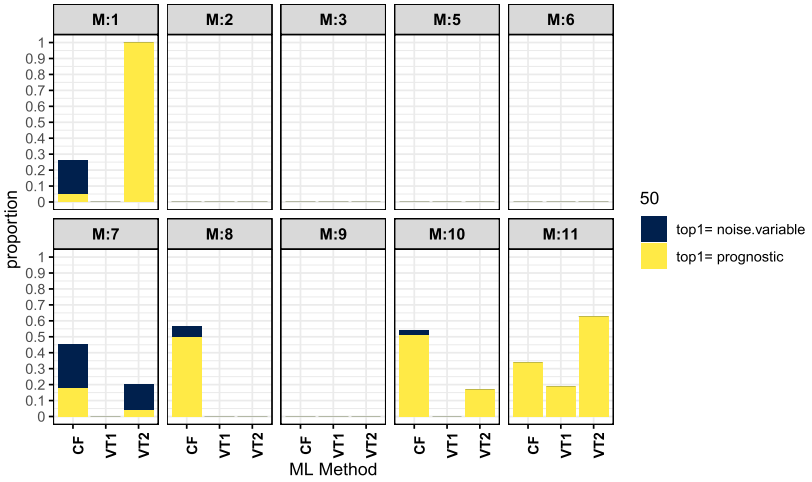

Fig. 3. Error rates: Proportions incorrect top-ranking over 100 simulation iterations for the 'ANTI-NULL' simulation models (with a true predictive effect present). Trial size was $n = 5000$. (Cases with correlation not displayed due to identical with their corresponding non-correlation results.

how the error was due to an incorrect top-1-ranking of a prognostic variable (e.g., for (CF, **M1**) it was 44% indicated by yellow. This is not directly matching the scale of the y axis, since it is a proportion of a proportion). A clear trend across the landscape of simulations is how the prognostic variables are the dominating driving factor behind the error rate (e.g., case **M11**), in particular, prominently dominating over contributions from truly non-informative variables. The corresponding results for trial size 5000 is found in Fig. 3.

Results regarding *power* are found in Table 2. The RMSE estimation accuracy is summarized in Table 3, and the relative performance of VT2 versus VT1 (in RMSE) is summarized in Fig. 4.

Table 3. Simulation results for accuracy of estimating individual treatment effects: `RMSE`(truth, estimated) averaged over over 100 iterations (case $N = 800$ and $N = 5000$), for each method. (Abbreviations: **CF** = Causal Forest, **VT1** = single-surface Virtual Twins, **VT2** = arm-specific Virtual Twins). The results for the simulation cases with correlation gave identical results to their non-correlation counterpart.

ML approach	N	**M1**	**M2**	**M3**	**M4**	**M5**	**M6**	**M7**	**M8**	**M9**	**M10**	**M11**
CF	800	0.44	0.76	0.75	0.29	1.47	0.47	0.71	0.68	0.48	0.99	4.44
VT1	800	0.43	0.82	0.80	0.30	1.45	0.48	0.62	0.63	0.49	0.73	12.12
VT2	800	1.13	1.38	1.37	1.09	1.82	0.66	1.88	1.87	1.16	2.92	21.43
CF	5000	0.33	0.79	0.81	0.12	1.28	0.49	0.52	0.51	0.40	0.64	4.83
VT1	5000	0.34	0.83	0.82	0.12	1.53	0.49	0.43	0.44	0.44	0.47	7.52
VT2	5000	0.71	1.04	1.05	0.62	1.69	0.53	1.13	1.13	0.75	1.68	14.38

6 Discussion

The literature has seen some simulation benchmarking of competing methods, such as [20,36,38], but their focus tended to be on other metrics such as accuracy of estimation of treatment effect in the identified subgroup, probability of first tree split being on a predictive variable, and selection bias regarding predictors. Albeit clearly important, there are also other important aspects that have received relatively little attention in the literature: the ability to disentangle prognostic and predictive variables, and the relative efficiency of ways of implementing VT (arm-specific versus one-surface approaches). Some of this was the topic of [25] where the default VT (single model, Random Forest-based) was among the evaluated methods, but the question from Sect. 3.2 was not explicitly addressed.

While we looked at two ways of implementing VT using the modern XGboost, we also compared to the recently proposed Causal Forsest which acted as the off-the-shelf choice (lacking the considerable tuning steps required for VT). Given that CF is known to provide unbiased ITE estimates by design, it is of interest to compare if VT would be far behind CF in accuracy.

We gave partial empirical answers regarding the implementation aspect of VT and saw that the expansion of training data as (3) indeed have clear benefits. This latter approach (combined with a 'common-surface' version of VT) strongly outperformed the arm-specific version across most simulation models regarding VIP. The latter was highly sensitive to prognostic variables and struggled with detecting the truly predictive ones. Generally, across all scenarios, the runs confirm a clear tendency of prognostic effects driving the Type-1-Error (see Figs. 2 and 3).

In terms of RMSE, VT1 outperformed the arm-specific VT considerably, see Fig. 4.

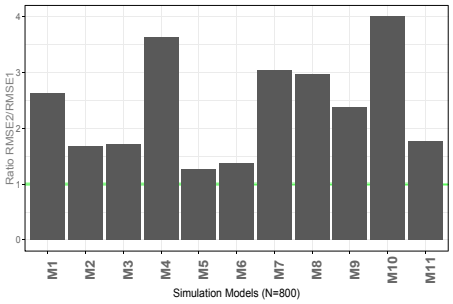

Fig. 4. *Relative performance of Arm-specific Virtual Twins (VT2) and one-surface VT (VT1). The y-axis shows* $RMSE_{\mathbf{VT2}}/RMSE_{\mathbf{VT1}}$*, where RMSE were averaged across the* 100 *iterations. Bars above* 1 *favours the version of Virtual Twins based on* $(Y, X, T, X \cdot I_{T=0}, X \cdot I_{T=1})$ with a single Boosting model fitted in Step 1. In no scenario did VT2 perform equal or better than VT1.

It was also striking that CF was relatively non-robust to prognostic variables, in the sense that it often did not rank the predictive over the prognostic ones (e.g., **M1**, **M3**, **M7**, **M8**, **M10**, **M11**). To our knowledge, this has not been investigated in any papers. Causal Forest has been proposed for Real World Evidence data, where one might assume rather large number of observations, and it is possible that the strength of CF lies more in such settings rather than in more limited RCTs; this is also in line with the $N = 5000$ results in Fig. 3. Having said that, VT1 might have a practical advantage over CF for of allowing for other types of endpoints in contrast to the current implementation of CF (restricted to continuous endpoints)

One notable conclusion from the simulations is that VT1 (extended training data, one model in Step 1) outperformed CF across the landscape of simulations in terms of both power and VIP: **M1**, **M2**, **M3**, **M7**, **M10** (with a stronger advantage to VT1 when the trial size was smaller).

It is interesting to contrast the results from models **M2** and **M10**; these are structurally identical but with higher prognostic strength in the latter case. When the RCT was smaller ($N = 800$), Causal Forest performed surprisingly bad in terms of ranking, and its power did deteriorate when the prognostic strength increased. However when the data sets were larger ($N = 5000$) CF performed perfect for the moderate prognostic strength (**M2**) but with a 54.0% Type-1-Error for the higher prognostic strength (**M10**).

Scenario (**M5**) is notable: when many (six) predictive effects were present, all models performed very well and the VIP for the predictive x did stand out strongly. This might be considered as one of the simpler simulation cases due to mutually exclusive prognostic and predictive effects, and it is relieving to see that all methods performed well here. (Still, VT1 outperformed VT2 in terms of estimating ITEs, see Table 3).

In our empirical comparisons, we have here excluded another possible implementation of VT, where a single Boosting model is fitted to the training data 'as is', i.e., (Y, T, X), without any treatment-covariate column extensions. This is sometimes referred to as the 'Augmented Feature Space' approach [23]. We noted model fitting instabilities in the 2nd step of VT, where often the individual differences were estimated to identical zero (and rendering VIP meaningless). We speculate that the prognostic effects totally dominated the model fit, suppressing the impact of treatment allocation, i.e., running into the issues discussed in [19]. Hence, our conclusion is that the extended single-surface VT is superior to arm-specific VT, which in turn is superior to single-surface VT without extensions.

A future direction of this research involves simulation comparisons regarding other types of endpoints commonly seen in clinical trials (at least regarding the implementations of VT, since Boosting models are not restricted to continuous response variables); furthermore, a larger set of simulation models might be considered, and the inclusion of other types of performance metrics (e.g., accuracy of estimating effect in planted subgroups). Also, other types of ML/ITE methods might be considered, such as Penalized Regression techniques, Model-Based Recursive Partitioning and GUIDE. This is in particular interesting in settings with more than one active treatment arm, where VT is currently only applicable to pairwise comparisons whereas some other techniques allow a single model fit to the entire data set.

7 Appendix

Here we add some additional plots (referred to in the result section). These cases are qualitatively different in the sense that one model has mutually exclusive prognostic and predictive effects whereas the other one has not. Furthermore one model has non-linear effects.

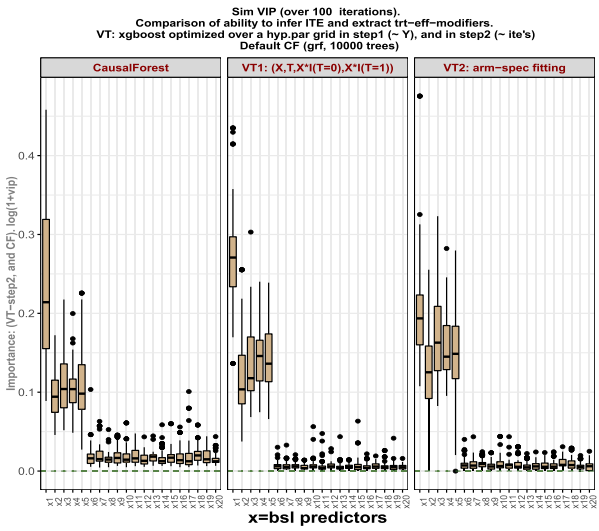

Fig. 5. Variable Importance (VIP) results for simulation model **M11**, 100 iterations with trial size $N = 5000$. The x_1 was the only predictive variable, and $x_1, ..., x_5$ were prognostic.

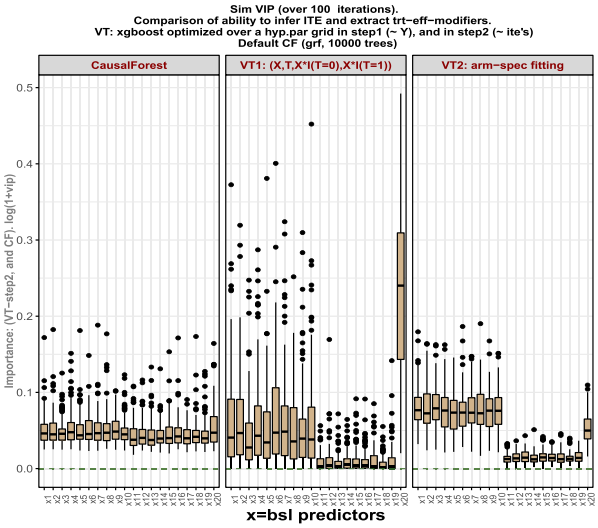

Fig. 6. Variable Importance (VIP) results for **M7**, 100 iterations with trial size $N = 800$, x_{20} predictive, $x_1, ..., x_{10}$ prognostic. It is notable that both VT2 and CF performed poorly.

References

1. Alosh, M.: Statistical Considerations on Subgroup Analysis: Interpretation of clinical trial findings and study design for targeted subgroup. Conference paper, FDA/DIA Statistics Forum, At North Bethesda, Maryland, US, (2014)
2. Altman, D., Royston, P.: The cost of dichotomising continuous variables. BMJ 332. https://doi.org/10.1136/bmj.332.7549.1080(2006)
3. Bica, I., Alaa, A., Lambert, C., Van Der Schaar M.: Clinical Pharmacology & Therapeutics published by Wiley Periodicals LLC on behalf of American Society for Clinical Pharmacology and Therapeutics (2020)
4. Bul, K., Doove, L., Franken, H., Van der Oord, S., Kato, P., Maras, A.: Results of the application of Virtual Twins with as outcome Social Skills Rating System (SSRS) (2018). https://doi.org/10.1371/journal.pone.0193681.g004
5. Chen, S., Tian, L., Cai, T., Yu, M.: A general statistical framework for subgroup identification and comparative treatment scoring. Biometrics **73**(4), 1199–1209 (2017). https://doi.org/10.1111/biom.12676
6. Chen, Y., Markatou, M.: A comparative study of subgroup identification methods for differential treatment effect: Performance metrics and recommendations. Statistical Methods in Medical Research (2017)
7. Curth, A., Van Der Schaar, M.: Nonparametric Estimation of Heterogeneous Treatment Effects: From Theory to Learning Algorithms. arXiv:2101.10943. [stat.ML] (2021)
8. Dmitrienko, A., Muysers, C., Fritsch, A., Lipkovich, I.: General guidance on exploratory and confirmatory subgroup analysis in late-stage clinical trials. J. Biopharmaceutical Stat. **26**, 71–98 (2016)

9. Dusseldorp, E., Doove, L., Mechelen, I.: Quint: An R package for the identification of subgroups of clients who differ in which treatment alternative is best for them. Behav. Res. Methods **48**(2), 650–663 (2015). https://doi.org/10.3758/s13428-015-0594-z
10. European Medicines Agency. Guideline on the investigation of subgroups in confirmatory clinical trials (2014)
11. Fang, L., Su, C.: Statistical Methods in Biomarker and Early Clinical Development. Springer. https://doi.org/10.1007/978-3-030-31503-0 (2020)
12. Foster, J.C., Nan, B., Shen, L., Kaciroti, N., Taylor, J.M.G.: Permutation testing for treatment–covariate interactions and subgroup identification. Stat. Biosci. **8**(1), 77–98 (2015). https://doi.org/10.1007/s12561-015-9125-9
13. Foster, J.C., Taylor, J.M.C., Ruberg, S.J.: Subgroup identification from randomized clinical trial data. Stat. Med. **30**, 2867–2880 (2011)
14. Gottlow, M., et al.: Application of structured statistical analyses to identify a biomarker predictive of enhanced tralokinumab efficacy in Phase III clinical trials for severe, uncontrolled asthma. BMC Pulmonary Medicine 19–129 (2019)
15. Hemmings, R., Koch, A.: Commentary on: Subgroup analysis and interpretation for phase 3 confirmatory trials: White Paper of the EFSPI/PSI working group on subgroup analysis by Dane, Spencer, Rosenkranz, Lipkovich, and Parke. Wiley (2019). https://doi.org/10.1002/pst.1935
16. Hou, J., et al.: Subgroup identification in personalized treatment of alcohol dependence. Alcohol Clin. Exp. Res. **39**(7), 1253–1259 (2015)
17. Jia, J., Tang, Q., Xie, W., Rode, R.: A Novel Method of Subgroup Identification by Combining Virtual Twins with GUIDE (VG) for Development of Precision Medicines. arXiv: Applications 167–180 (2020)
18. Lipkovich, I., Dmitrienko, A.: Strategies for identifying predictive biomarkers and subgroups with enhanced treatment effect in clinical trials using SIDES. J. Biopharmaceutical Stat. **24**, 130–153 (2014)
19. Lipkovich, I., Dmitrienko, A., D'Agostino, R., Sr.: Tutorial in biostatistics: data-driven subgroup identification and analysis in clinical trials. Stat, In Medicine (2016)
20. Loh W., Cao L., Zhou P. Subgroup identification for precision medicine: A comparative review of 13 methods. Data Mining Knowl Discov. (2019)
21. Loh, W., Man, M., Wang, S.: Subgroups from Regression Trees with adjustment for Prognostic effects and post-selection inference. Statistics in Medicine (2018)
22. Marchenko, O.V., Katenka, N.V. (eds.) Quantitative Methods in Pharmaceutical Research and Development, Chap. 6. Springer Nature Switzerland AG (2020). https://doi.org/10.1007/978-3-030-48555-9_6
23. Research Pillar: Individualized treatment effect inference. Webpage. https://www.vanderschaar-lab.com/individualized-treatment-effect-inference/
24. Rubin, D.B.: Causal inference using potential outcomes: design, modeling, decisions. J. Am. Stat. Assoc. **100**(469), 322–331 (2005)
25. Sechidis, K., Papangelou, K., Metcalfe, P.D., Svensson, D., Weatherall, J., Brown, G.: Distinguishing prognostic and predictive biomarkers: an information theoretic approach. Bioinformatics 1 **34**(23), 4139 (2018) https://doi.org/10.1093/bioinformatics/bty515
26. Sharmin Akter, S.: Subgroup identification with Virtual Twins and Guide algorithms - an application to adult fitness data. Thesis, Ball State University (2020). https://cardinalscholar.bsu.edu/handle/123456789/202449

27. Ting, N., et al. (eds.): Design and Analysis of Subgroups with Biopharmaceutical Applications, Emerging Topics in Statistics and Biostatistics. Springer Nature Switzerland AG 2020 (2020). https://doi.org/10.1007/978-3-030-40105-4_3
28. Su, X., Zhou, T., Yan, X., Fan, J., Yang, S.: Interaction trees with censured survival data. Int. J. Biostat. **4**, 1–26 (2008)
29. Chen, T., Guestrin, C.: XGboost: A Scalable Tree Boosting System. arXiv:1603.02754v3 [cs.LG]
30. Tibshirani, J.: Generalized Random Forest, CRAN package (2018). https://cran.r-project.org/web/packages/grf/index.html
31. Thomas, M., Bornkamp, B., Seibold, H.: Subgroup identification in dose-finding trials via model-based recursive partitioning. Stat. Med. **37**, 1608–1624 (2018)
32. Tune, P.: Towards Data Science Blogpost: The unreasonable in effectivness of Deep Learning on Tabular Data (2020). https://towardsdatascience.com/the-unreasonable-ineffectiveness-of-deep-learning-on-tabular-data-fd784ea29c33
33. Turner, E.: Predictive Variable Selection For Subgroup Identification. Thesis, University of Manchester, Faculty of Science and Enginering, School of Computer Science (2017). https://www.research.manchester.ac.uk/portal/files/63667893/FULL_TEXT.PDF
34. Vieille, F., Foster, J.: aVirtualTwins: Adaptation of Virtual Twins Method from Jared Foster. R package version 1.0.1 (2018). https://CRAN.R-project.org/package=aVirtualTwins
35. Wager, S., Athey, A.: Estimation and inference of heterogeneous treatment effects using random forests. J. Am. Stat. Assoc. **113**(523), 1228–1242 (2018). https://doi.org/10.1080/01621459.2017.1319839
36. Yang, et al.: Look before you leap: systematic evaluation of tree-based statistical methods in subgroup identification. J. Biopharm. Statistics (2019)
37. OI Zeileis, A., Hornik, K.: Generalized M-Fluctuation Tests for Parameter Instability. Stat. Neerland. 61. https://doi.org/10.1111/j.1467-9574.2007.00371.x (2003)
38. Zhang, Z., Seibold, H., Vettore, M., Song, W., François, V.: Subgroup identification in clinical trials: an overview of available methods and their implementations with R. Annal. Translational Med. **6**(7) (2018). https://doi.org/10.21037/atm.2018.03.07
39. Zink, R.C., Shen, L., Wolfinger, R.D., Showalter, H.D.H.: Assessment of methods to identify patient subgroups with enhanced treatment response in randomized clinical trials. In: Chen, Z., Liu, A., Qu, Y., Tang, L., Ting, N., Tsong, Y. (eds.) Applied Statistics in Biomedicine and Clinical Trials Design. IBSS, pp. 395–410. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-12694-4_24