International Statistical Review

International Statistical Review (2020), 88, S1, S135–S178 doi:10.1111/insr.12427

Stable Discovery of Interpretable Subgroups via Calibration in Causal Studies

Raaz Dwivedi[1], Yan Shuo Tan[2], Briton Park[2], Mian Wei[2], Kevin Horgan[6], David Madigan[5] and Bin Yu[1,2,3,4,7]

[1]*Department of EECS, University of California, Berkeley, Berkeley, CA, USA*
E-mail: raaz.rsk@berkeley.edu

[2]*Department of Statistics, University of California, Berkeley, Berkeley, CA, USA*

[3]*Division of Biostatistics, University of California, Berkeley, Berkeley, CA, USA*

[4]*Center for Computational Biology, University of California, Berkeley, Berkeley, CA, USA*

[5]*Khoury College of Computer Sciences, Northeastern University, Boston, MA, USA*

[6]*Protypia Inc,111 10th Avenue South, Suite 102, Nashville, TN, 37023, USA*

[7]*Chan Zuckerberg Biohub, San Francisco, CA, USA*

Summary

Building on Yu and Kumbier's predictability, computability and stability (PCS) framework and for randomised experiments, we introduce a novel methodology for Stable Discovery of Interpretable Subgroups via Calibration (StaDISC), with large heterogeneous treatment effects. StaDISC was developed during our re-analysis of the 1999–2000 VIGOR study, an 8076-patient randomised controlled trial that compared the risk of adverse events from a then newly approved drug, rofecoxib (Vioxx), with that from an older drug naproxen. Vioxx was found to, on average and in comparison with naproxen, reduce the risk of gastrointestinal events but increase the risk of thrombotic cardiovascular events. Applying StaDISC, we fit 18 popular conditional average treatment effect (CATE) estimators for both outcomes and use calibration to demonstrate their poor global performance. However, they are locally well-calibrated and stable, enabling the identification of patient groups with larger than (estimated) average treatment effects. In fact, StaDISC discovers three clinically interpretable subgroups each for the gastrointestinal outcome (totalling 29.4% of the study size) and the thrombotic cardiovascular outcome (totalling 11.0%). Complementary analyses of the found subgroups using the 2001–2004 APPROVe study, a separate independently conducted randomised controlled trial with 2587 patients, provide further supporting evidence for the promise of StaDISC.

Key words: Causal inference; randomised experiments; subgroup discovery; CATE modelling; calibration; stability; PCS framework; VIGOR study.

⋆Raaz Dwivedi and Yan Shuo Tan are joint first authors and contributed equally to this work.

1 Introduction

Since its inception, the field of statistics has aimed to produce tools to help scientists seek scientific truth. Scientific truth, however, is not of a singular quality. While some relations in physics like Hooke's law are made apparent using simple linear regression, questions dealing with complex, emergent phenomena such as the efficacy of drugs or job training programmes seem to have more contingent answers. It was the urge to formalise and investigate such questions that begot and nurtured the field of causal inference in statistics over the past century. One of the two most influential frameworks for causal inference, the Neyman–Rubin causal model (Holland, 1986), has its roots in Fisher and Neyman's (Fisher, 1936; Splawa-Neyman *et al*., 1990; Neyman & Iwaszkiewicz, 1935) work on randomised experiments for agriculture and was later codified by Rubin (1974), who was then interested in psychometrics.[1]

Historically, causal inference researchers have used traditional regression methods in their analyses, with econometricians in particular developing a comprehensive theory of drawing inference from linear models (Angrist & Pischke, 2008). This is rapidly changing, however, with recent works (Athey, 2018; Künzel *et al*., 2019; Chernozhukov *et al*., 2018; Molina & Garip, 2019) bringing in machine learning tools to tackle causal inference problems, one genre of which has been the investigation of heterogeneous treatment effects.

1.1 Heterogeneous Treatment Effects

In both randomised experiments as well as observational studies, apart from the treatment and response variables, additional pre-treatment information is often known about the study subjects. For instance, information on medical risk factors is collected in clinical trials, while demographic and socioeconomic data are collected in social science studies. Such side information has always been important because it allows us to adjust for confounding in observational studies and also to create more efficient estimators in randomised experiments (Lin, 2013; Imbens & Rubin, 2015).

In addition to these uses, researchers are also increasingly interested in drawing inference about how the effect of a treatment varies depending on an individual's observed covariates. The past decade in particular has witnessed a wave of innovation in the modelling and estimation of heterogeneous treatment effects. Underlying the hot topic of *precision medicine* (Collins & Varmus, 2015) is a realisation that how a patient responds to a particular drug or treatment depends on the patient's genetics, lifestyle and environment and that, consequently, accounting for these differences will allow doctors to deliver better and more targeted care. Moreover, this emphasis on understanding and exploiting heterogeneity is not unique to the biomedical sciences and has also arisen in economics (Imbens & Wooldridge, 2009), political sciences (Gerber *et al*., 2008; Feller & Holmes, 2009), online advertising (Michel *et al*., 2019) and many other fields (Feller & Holmes, 2009).

Broadly speaking, methodological research on heterogeneous treatment effects can be put into two categories: (i) conditional average treatment effect (CATE) function estimation (Imbens & Wooldridge, 2009; Gerber *et al*., 2008; Feller & Holmes, 2009; Cai *et al*., 2011; Foster *et al*., 2011; Tian *et al*., 2014; Bloniarz *et al*., 2016) and (ii) subgroup analysis (Wang *et al*., 2007; Peck, 2003; Athey & Imbens, 2016; Lipkovich *et al*., 2011), with the latter having a longer history. Here, we attempt a brief review of the existing literature and refer the readers to referenced papers for further background.

1.1.1 CATE estimation

For a binary treatment, the CATE is defined to be the expected difference between the potential outcome under treatment and that under no treatment, conditional on a subject's observed covariates (see Section 3 for formal definitions). While the average treatment effect (ATE) is a scalar quantity, the CATE is a function and thus far more challenging to estimate. Because one observes only one of the two potential outcomes for every individual—an issue referred to as the fundamental problem of missing data in causal inference (Holland, 1986)—one cannot directly solve this problem using the conventional supervised learning techniques.

Over the past decade or so, researchers have made tremendous progress with CATE estimation and proposed numerous methods for it (Imbens & Wooldridge, 2009; Gerber *et al*., 2008; Feller & Holmes, 2009; Cai *et al*., 2011; Foster *et al*., 2011; Tian *et al*., 2014; Bloniarz *et al*., 2016). A large fraction of these (Imbens & Wooldridge, 2009; Feller & Holmes, 2009; Cai *et al*., 2011; Bloniarz *et al*., 2016) fall under the framework of meta-learners. These are 'meta-algorithms [that] decompose estimating the CATE into several regression sub-problems that can be solved with any regression or supervised learning method' (Künzel *et al*., 2019). Some of these meta-algorithms are fairly obvious. For instance, the *T*-learner strategy (Foster *et al*., 2011) comprises fitting models for the two response functions (the conditional expectation of each potential outcome) and then taking their difference. Others, such as the *X*-learner (Künzel *et al*., 2019) and *R*-learner (Nie & Wager, 2017) strategies, are more sophisticated and require more notation to explain (see Section 4.1 for further details). Not all proposed algorithms follow a meta-learner strategy, the popular causal tree and causal forest algorithms (Athey & Imbens, 2016; Wager & Athey, 2018) being prominent examples.

1.1.2 Concerns with model choice for CATE estimation

With such a diverse range of estimators, most of which come with hyperparameters, model choice becomes a primary concern. Some researchers have used asymptotic efficiency (Nie & Wager, 2017; Kennedy, 2020) to establish when certain estimators can be definitely favoured under (uncheckable) generative models. Such arguments, however, rely on smoothness assumptions and asymptotic data regimes that are typically hard to verify for the problems typically considered by causal inference researchers. Meanwhile, plug-in prediction accuracy on holdout test sets is frequently used to do model selection in supervised learning, but this is infeasible for CATE estimation owing to the data missingness we alluded to earlier. To circumvent this issue, researchers have formulated proxy loss functions (Schuler *et al*., 2018) for data-driven model choice, with ideas including using nearest neighbour matching (Rolling & Yang, 2014), kernel-based local linear squares fit (Cai *et al*., 2011) and influence functions (Alaa & Van Der Schaar, 2019). These model choice methods, however, have only been justified using simulations often in strong signal regime, a scenario that does not hold in many if not most real data problems (including the one considered in this work).

1.1.3 Concerns with model validation for CATE estimation

Before deciding which estimator to choose for a given task, we would first like to know whether there is even enough signal in the data to fit a generalisable model. Again, data missingness means that there is no clear answer to this problem. The proxy loss functions are not good substitutes for quantities like R^2 or area under the receiver operating characteristic curve scores because they can be noisy, and furthermore, they do not have an easily interpretable scale. This is especially concerning because randomised experiments often have low-signal strength.[3]

1.1.4 Subgroup analysis

An older approach to investigating heterogeneity is through 'subgroup analysis'. The goal here is to identify subgroups of subjects in the study over which the treatment effect is significantly larger or smaller than that the population average. Such a conception of heterogeneity has two advantages over CATE estimation: (a) It is less ambitious and thus promises to be more tractable given the low data regime in real settings, and (b) it is often more aligned with the downstream tasks involving decision making (e.g., identifying which subgroup of individuals to treat).

Traditionally, for subgroup analysis, researchers check the treatment effect over a predetermined list of subgroups that are suggested by prior domain knowledge. Doing this, however, ignores potential unforeseen heterogeneity in the data, and there has been much recent work on how to conduct a data-driven search for subgroups. Naive searching can quickly overfit,[3] so any search method has to balance aggressiveness of searching with the need to account for multiple testing. Proposed methods include using recursive partitioning (Su *et al*., 2009; Athey & Imbens, 2016), Cox modelling (Negassa *et al*., 2005), controlled partitioning with significance checks using data splits (Lipkovich *et al*., 2011) and several variants (Dusseldorp & Van Mechelen, 2014; Ballarini *et al*., 2018). Unfortunately, systematic analyses of these methods have usually provided unsatisfactory results in real data settings and in low-signal simulations (Ondra *et al*., 2016; Huber *et al*., 2019). We refer the readers to the book of Carini *et al*. (2014) (Chapter 8) and the review papers (Ondra *et al*., 2016; Huber *et al*., 2019) for further discussion on these methods.

Finally, we note that some researchers have proposed using CATE estimation as a stepping stone to finding subgroups. Such a strategy was proposed by Foster *et al*. (2011) with their Virtual Twins method, namely, the *T*-learner with random forests (RFs), while Chernozhukov *et al*. (2018) recapitulate this idea in the context of a broader call to perform inference on features of the CATE function rather than the function itself. In another line of work, Shahn & Madigan (2017) integrate (linear) CATE modelling with latent class mixture modelling in a Bayesian framework to allow for treatment effect heterogeneity in discrete levels. They then use the feature importance from the latent (logistic) model and the posteriors for the CATE, to estimate qualitatively, subgroups with large treatment effect.

1.2 The PCS Framework for Veridical Data Science

As argued in the previous section, obtaining reliable conclusions with respect to heterogeneous treatment effects is fraught with difficulty. On the one hand, poor signal and weak priors are prevalent, and on the other hand, missing potential outcomes means that test-set validation is not directly feasible. Methods validated on simulation studies may not work well for real data problems because their performance is often misleading. Furthermore, empirical evidence tells us that the relative and absolute performance of estimation algorithms is highly data and context dependent (Olson *et al*., 2018).[4] Given these problems, it is puzzling to see that much new methodology is being developed that is detached from solving real data problems.

In this paper, we re-analysed the 1999–2000 VIGOR study (an 8076-patient randomised clinical trial) and had to face precisely these challenges. To overcome them, we take advantage of the recent works on CATE estimation (Bloniarz *et al*., 2016; Künzel *et al*., 2019; Athey & Imbens, 2016; Nie & Wager, 2017; Wager & Athey, 2018) and build on the PCS framework for veridical data science recently introduced by Yu & Kumbier (2020). As a result, we develop a methodology called Stable Discovery of Interpretable Subgroups via Calibration (StaDISC),

which is generally applicable beyond this dataset. We now briefly review the PCS framework, before turning to the overview of our contributions and StaDISC in Section 1.3.

The PCS framework bridges, unifies and expands on ideas from machine learning and statistics for the entire data science life cycle. The letters in PCS stand for the three core principles of data science, namely, predictability, computability and stability. In a nutshell, the PCS framework advocates using both predictability and stability analysis, argued and documented in a PCS documentation, for reliable and reproducible scientific investigations, thereby providing a way for bridging Breiman's Two Cultures (Breiman, 2001). More specifically, predictability emphasises reality checks for the modelling stage, by integrating the use of data-driven validation such as out-of-sample testing favoured by machine learning and that of goodness-of-fit measures that have a rich history in traditional statistics. Stability, besides encompassing sampling variability, expands to other stability or robustness concerns of the contingency of modelling conclusions to researcher 'judgement calls'. These calls include the choices made by the researcher at various stages of the data science life cycle, including data cleaning in addition to the modelling decisions such as model choices and data perturbations. Computability reflects the need to keep computational feasibility and efficiency in mind when constructing any modern data analysis pipeline, especially those that subscribe to the first two principles, which are usually more demanding computationally.

The PCS framework addresses to a certain extent Professor Efron's concern (Efron, 2020) that machine learning methods (or pure prediction algorithms) are not ready to be used on scientific problems.[5] The PCS framework adds a paramount consideration of stability to predictability and computability that are hallmarks of machine learning. It guides researchers in validating machine learning and statistical methods with respect to the specific task they are to be applied and extracting data conclusions that can be relied upon. As one of us has previously discussed (Yu & Barter, 2020), even though 100% truth is beyond reach, a useful goal is an 'accurate approximation for a particular domain, and relative to a particular performance metric', which is a more precise articulation of George Box's belief that 'all models are wrong, but some are useful'.

1.3 Our Contributions

This paper makes three main contributions. First, we seek subgroups with demonstrable heterogeneous treatment effects in the dataset from the 1999–2000 VIGOR study. Complementary analyses with the 2001–2004 APPPROVe study provides additional evidence for the heterogeneity in treatment effect for the found subgroups. Enroute, building on the recent CATE literature and the PCS framework, we develop a new methodology, which we call StaDISC. We provide an overview of this methodology toward the end of this section. Finally, this paper also serves as the first articulation of the PCS framework in the context of causal inference, with StaDISC providing a template for more informative understanding of heterogeneous outcomes.

1.3.1 Organisation

The rest of the paper is organised as follows. In Section 2, we start with a brief history of the VIGOR study, and then we describe the dataset and data engineering and splitting done by us. Section 3 reviews the Neyman–Rubin model briefly with basic notations introduced. The development of the StaDISC methodology (overviewed below) is carried out in Sections 4 to 6 with the final subgroups reported in Section 6.3. Results for the complementary analyses of the found subgroups with the APPROVe study are presented in Section 7. We conclude in Section 8 with a recap of our results and a discussion of the relevance of our discoveries in medicine, and we discuss several directions for future work with StaDISC. Most of the figures and tables are

deferred to the appendix. Moreover, in accordance with the PCS framework's requirement for clear and careful documentation, we provide our code, data cleaning and statistical analyses in the form of Jupyter notebooks on GitHub (https://github.com/Yu-Group/stadisc).

1.3.2 Overview of StaDISC

First of all, a given dataset (deemed approximately i.i.d.) is divided into a holdout test set $\mathbf{S}_{\text{TEST}}$ and a training set $\mathbf{S}_{\text{TRAIN}}$ (per outcome). For hyperparameter tuning, we use four-fold cross-validation with the training data $\mathbf{S}_{\text{TRAIN}}$.[6] For any set of training folds, we refer to the leftout fold as the corresponding validation fold. The test set is used only once at the final step of checking the significance of the interpretable subgroups found by our methodology. See Section 2.3 for more details on data splitting and Section 4.1 for the fitting of CATE estimators. With this set-up at hand, StaDISC can be summarised in three steps: a predictive reality check in Section 4 based on calibration, stability-driven ranking and aggregation of CATE estimators in Section 5 and finally the `CellSearch` procedure for finding interpretable subgroups in Section 6. In Section 4, we introduce a novel calibration-based pseudo-R^2 score for CATE estimators denoted by $\mathcal{R}^2_{\text{C}}$, which involves placing individuals (in both training and validation folds) into equally sized bins based on their predicted CATE value, with quantiles of the predicted CATE distribution on the training folds as thresholds for the CATE estimators. Using such a binning and the $\mathcal{R}^2_{\text{C}}$-scores, we show that 18 popular CATE estimators generalise poorly for the VIGOR data on the validation folds of the training data. However, we find that certain quantile-based bins (referred to as quantile-based top subgroups) do generalise well in the sense of having significantly stronger subgroup CATE on both training and validation folds. This provides the starting point of the next step. In Section 5, we use the *t*-statistics of the treatment effect over the quantile-based top subgroups and its stability over seven different appropriate data perturbations to rank, screen and finally average the screened CATE estimators (the ensemble CATE estimator). Section 6 details the last step of StaDISC, where we introduce the `CellSearch` procedure to find a stable and interpretable representation of the quantile-based top subgroup of the ensemble from the previous step, and then we check its performance on the holdout test set (which was used only for final testing).

As a final overview remark, we note that we use poor performance and good/bad generalisation in a slightly loose sense throughout the paper. We only use the holdout test set at the final stage, for verifying the CATE estimates of discovered subgroups. Nonetheless, we use the phrase *poor generalisation* to refer to worse-than-expected-performance, where the performance metric varies across results, on the validation folds.

2 Dataset from the VIGOR Study

In this paper, we are interested in finding subgroups of patients who benefit from the treatment in the dataset from the Vioxx Gastrointestinal Outcomes Research (VIGOR) study (Bombardier *et al*., 2000). In the process of seeking such subgroups, we develop the new StaDISC methodology. In this section, we provide an overview of this study and the dataset, and we also explain our data pre-processing and feature engineering.

2.1 VIGOR Study History and Description

The VIGOR study was a randomised head-to-head trial comparing two drugs used to alleviate pain and inflammation for patients with rheumatoid arthritis: a 'new' cyclooxygenase-2 (COX-2) inhibitor drug rofecoxib (Vioxx) recently approved and developed by Merck, and

naproxen, a standard nonsteroidal anti-inflammatory drug (NSAID) already in routine clinical use for many years. NSAIDs, although effective for treating pain and inflammation, cause serious gastrointestinal (GI) side effects in a small proportion of patients with frequent use. The rationale for the development of COX-2 inhibitors, such as Vioxx, was reduced GI toxicity as compared with traditional NSAIDs. Previously conducted short-term clinical studies were supportive of this hypothesis, although concerns about potential cardiovascular toxicity associated with Vioxx had also been raised.

2.1.1 Aim of the study

The VIGOR study was designed to provide more conclusive evidence of the superior GI safety of Vioxx. The study was conducted in the years 1999–2000 by Merck with the primary hypothesis that its drug Vioxx would have fewer GI side effects than naproxen for the treatment of rheumatoid arthritis. The study population comprised 8076 patients 'with rheumatoid arthritis who were at least 50 years old (or at least 40 years old and receiving long-term glucocorticoid therapy) and who were expected to require NSAIDs for at least one year'. This population was known to be at relatively high risk of GI side effects with NSAIDs.[7] The patients in the control arm were assigned the drug naproxen, while the patients in the active treatment arm were assigned Vioxx.

2.1.2 Details and findings of the study

Patients were followed up for a median time of 9 months, and the primary end point was time to first occurrence of a confirmed clinical upper GI event defined as 'gastroduodenal perforation or obstruction, upper gastrointestinal bleeding, and symptomatic gastroduodenal ulcers'. The original study report (Bombardier *et al*., 2000) performed a survival analysis using a Cox proportional hazard model and estimated the relative risk for patients in the treatment arm compared with those in the control arm to be 0.5, with a confidence interval of 0.3 to 0.6.[8]

The study authors also conducted a subgroup analysis for the GI events, analysing subgroups defined by gender, age, nationality, steroids, PUB (perforations, ulcers and bleeding) history (prior history of GI events) and presence of *Helicobacter pylori* antibodies. The rationale was that certain patients were known to be at increased risk of GI events, and they wanted to see if the benefit of Vioxx extended to these high-risk patients. The conclusion from the subgroup analysis was that the risk ratio for every subgroup remained significant, while differences of the ratios between subgroups were not significant.

However, VIGOR demonstrated that Vioxx was associated with an increased risk of thrombotic cardiovascular events (henceforth referred to as CVT events), an aspect that was not emphasised in the original report of the study (Bombardier *et al*., 2000). The study authors suggested that apparent association of Vioxx with CVT events was actually the result of naproxen preventing CVT events. However, placebo-controlled studies confirmed that Vioxx did indeed cause CVT events, and this ultimately led to the withdrawal of Vioxx from the market. We refer the reader to the articles (Krumholz *et al*., 2007; Ross *et al*., 2009) for more context on the VIGOR study and its consequences thereafter.

2.1.3 Goal of our investigation into the VIGOR study

In this work, we perform analysis for both the GI and CVT events. While the GI event was an infrequent event (experienced by around 2% patients) in the study, the less common CVT event (around 0.6% were reported to have a confirmed CVT event) was considered to be more significant medically. As the earlier works already established that Vioxx led to an overall decrease

in the GI risk but an increase in the cardio risk on the overall population of the study, an important by-product of this work is finding clinically relevant and interpretable subgroups of interest for which Vioxx provided a significant decrease in the risk for the GI event but did not increase the risk for the CVT event. Interpretability of the subgroup, as well as the transparency of the search procedure, is important from a clinical view point, as the doctors can then better justify their choice to favour prescribing the drug for patients in the discovered subgroup.

We present detailed results for both the GI and CVT events throughout this paper, while occasionally deferring some details to the Appendix. To perform our analysis, we created a dataset with the two outcomes—GI and CVT events—as discussed above, a treatment indicator and 16 binary features. The data processing necessary to create this dataset is the topic of the next section.

2.2 Feature Selection and Engineering

The VIGOR study collected an extensive range of patient data, including demographic details, prior medical history, and the timing and details of adverse events during the clinical experiment. From this, we extracted 16 clinically relevant binary features, which we report in Table 1 together with covariate balance details. Also, see Figure 1 for a visual comparison. We now describe some of the decisions we took with respect to feature engineering, as well as the meaning of the selected features.

The medical history risk factors and drug use information were all already binary and were selected by the VIGOR study designers as being medically relevant. For instance, it is known

Table 1. *Overview of the baseline covariates in the control and treatment arm of the VIGOR study.*

Covariate (ABBRV)	**Control no. (%)**	**Treatment no. (%)**
Overall population	4029 (49.9)	4047 (50.1)
Demographics		
Whether *gender* is male (MALE = 1)	814 (20.2)	824 (20.4)
Whether *race* is white (WHITE = 1)	2752 (68.3)	2764 (68.3)
Whether *country* is US (US = 1)	1750 (43.4)	1748 (43.2)
Whether *adjusted age*† >65 (ELDERLY = 1)	1172 (29.1)	1136 (28.1)
Whether *body mass index* >30 (OBESE = 1)	1060 (26.3)	1106 (27.3)
Lifestyle		
Whether patient *smokes* ≥1 cig./day (SMOKE = 1)	1879 (46.6)	1919 (47.4)
Whether patient has ≥1 *alcoholic drinks*/week (DRINK = 1)	1045 (25.9)	1053 (26.0)
Prior medical history		
Of *GI PUB events* (PPH = 1)	317 (7.9)	313 (7.7)
Of *hypertension* (HYPGRP = 1)	1168 (29.0)	1217 (30.1)
Of *hypercholesterolemia* (CHLGRP = 1)	293 (7.3)	343 (8.5)
Of *diabetes* (DBTGRP = 1)	254 (6.3)	240 (5.9)
Of *atherosclerotic cardiovascular disease* (ASCGRP = 1)	216 (5.4)	238 (5.9)
Indicating use of *aspirin* under FDA guidelines (ASPFDA = 1)	151 (3.7)	170 (4.2)
Prior usage of drugs		
Whether used *glucocorticoids/steroids* (PSTRDS = 1)	2253 (55.9)	2244 (55.4)
Whether used *naproxen* (PNAPRXN = 1)	747 (18.5)	759 (18.8)
Whether used *NSAIDs* (PNASIDS = 1)	3341 (82.9)	3344 (82.6)
Outcomes		
Whether *GI event* occurred (GI = 1)	121 (3.0)	56 (1.4)
Whether *CVT event* occurred (CVT = 1)	18 (0.4)	41 (1.0)

†Adjusted age denotes age multiplied by the ratio of the life expectancy in the USA to that in the individual's country of residence.

PUB stands for perforations, ulcers and bleeding.

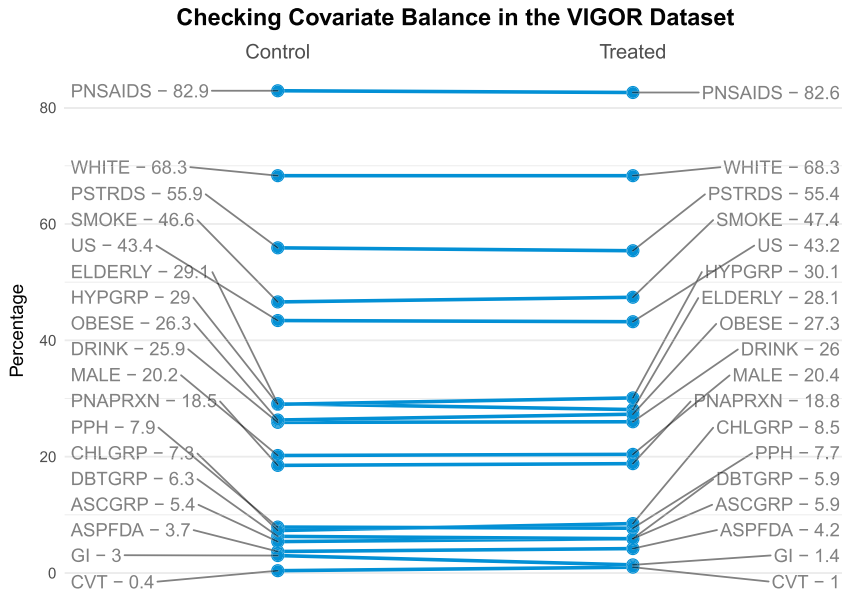


Figure 1. *A visual illustration showing the covariate balance and the outcome imbalance (GI and CVT) between the control and treatment population for the VIGOR study. The abbreviations are detailed in Table 1, and the number next to the abbreviation (ABBRV) denotes the % of the study size taking value 1 for that ABBRV in the respective arm. Note that the study size was 8076 total patients, and treatment and control arms comprise 4029 (49.9%) and 4047 (50.1%) individuals, respectively. [Colour figure can be viewed at* wileyonlinelibrary.com*]*

that use of glucocorticoids predisposes patients to GI events in the context of concomitant NSAID administration (Hernández-Díaz & Rodríguez, 2001). One feature that deserves special interest is ASPFDA. This was an indicator for patients in the study who 'met the criteria of the Food and Drug Administration (FDA) for the use of aspirin for secondary cardiovascular prophylaxis but were not taking low-dose aspirin therapy' (Bombardier *et al*., 2000) and was thought to be an especially strong risk factor for cardiovascular events. Patients who were actually undergoing aspirin therapy were excluded from the study.

On the other hand, some of the demographic and lifestyle risk factors required some engineering. The goal of the feature engineering was to simplify the data using prior information, so as to avoid overfitting and to simplify downstream data analysis. While the study collected more precise data on the patient's country of residence and their race, in both cases, a single level ('US' and 'white' respectively) contained a large fraction of the data, and we used these to binarise the two features. We also applied a similar logic to the smoking and alcohol lifestyle risk factors. We used height and weight information to calculate the body mass index for every patient, and then we used a threshold value of 30 to obtain an indicator for obesity.[9] Finally, we calculated the adjusted age for every patient (by multiplying their numerical age by the ratio of the life expectancy in the USA to that in their country of residence), and then we used a threshold value of 65 to define an indicator for being elderly. Finally, there was no direct indicator for patients with a prior history of GI event, so we made use of the medical history files to impute this. See Appendix C1 for more details.

The dataset was fairly complete [as is the case for most randomised controlled trials (RCTs)], with only a single patient missing an entry for each lifestyle risk factor (we filled in this with a 1), while 35 patients were missing entries for either height or weight, leading to a missing entry for the obesity indicator (we filled this in with a 0). Furthermore, the features also

have weak pairwise correlations except for the fact that the subgroup with ASPFDA = 1 (321 patients) is a subset of that with ASCGRP = 1 (454 patients).

2.3 Data Splitting

As a known best practice included in the PCS framework, for each outcome, we created a holdout test set comprising 20% of the individuals, which we did not touch in our further investigations until the very last stage of our analysis, that is, when we wanted to verify our results. Because of the rarity of events for both outcomes, we stratified the split by both the treatment and the outcome simultaneously; such a stratification ensures that the outcome remains balanced across the test-train splits. Let Y denote the binary outcome of interest (GI or CVT event) and T denote the treatment indicator. Then such a stratification (implemented as `model_selection.train_test_split` function in the sklearn library (Pedregosa *et al*., 2011)) is done by first categorising the study subjects in four categories $\{\{T = 0, Y = 0\}, \{T = 1, Y = 0\}, \{T = 0, Y = 1\}$ and $\{T = 1, Y = 1\}\}$—once with Y denoting the GI event and once with Y denoting the CVT event. Then we select a randomly sampled (without replacement) 20% of the subjects from each category together as the test set $\mathbf{S}_{\text{TEST}}$, with the remaining subjects form the training set $\mathbf{S}_{\text{TRAIN}}$.

Also, keeping in mind the rarity of the signals, we do not create an additional validation set, and instead we use the training data via a stratified four-fold cross-validation, where the folds are split uniformly at random, again stratified jointly according to T and YY. For such a split, each fold has around 35 GI events and 11 CVT events among the 1615 patients. We note that for a given outcome (say GI event), we use the same four-fold CV split—referred to as the *original split* and denoted as `cv_orig`—for tuning the hyperparameters for all the CATE estimators via cross-validation. We also use two *additional* stratified four-fold cross-validation (random) splits in several results throughout the paper, and we denote them by {`cv_0`, `cv_1`}. No hyperparameter tuning is done on these additional splits, and we simply use the tuned parameters from the `cv_orig` split for fitting the estimators on different sets of training folds of these additional splits. Note that for any four-fold CV split, there are four possible pairs of training-validation folds, denoted generically by $\mathbf{S}_{\text{TF}}$ and $\mathbf{S}_{\text{VF}}$, respectively. Mathematically, given disjoint folds from one four-fold CV split, namely, $\{\mathbf{S}_{\mathfrak{f}}\}_{\mathfrak{f}=1}^{4}$ of the training data $\mathbf{S}_{\text{TRAIN}}$ such that $\mathbf{S}_{\text{TRAIN}} = \cup_{\mathfrak{f}=1}^{4}\mathbf{S}_{\mathfrak{f}}$, the four pairs of training-validation folds are be denoted by $\left\{(\mathbf{S}_{\text{TF}} = \mathbf{S}_{\text{TRAIN}} \backslash \mathbf{S}_{\mathfrak{f}}, \mathbf{S}_{\text{VF}} = \mathbf{S}_{\mathfrak{f}}), \mathfrak{f} = 1, 2, 3, 4\right\}$.

3 Review on Neyman–Rubin Model and Notation

Throughout this paper, we assume the standard set-up for a completely randomised experiment under the Neyman–Rubin counterfactual framework. We assume that we observe a population of size N, in which the treatment variable T is completely randomised. For each individual i, there are two *potential outcomes*: $Y_i(0)$ when the individual i is assigned to the control arm $T_i = 0$, and $Y_i(1)$ when they are assigned to the treatment arm, $T_i = 1$. The individual treatment effect for individual i is defined as the difference of the two potential outcomes $\tau_i = Y_i(1) - Y_i(0)$. But this quantity is unobservable because for each individual we only observe one outcome corresponding to the arm that they are assigned to, that is, $Y_{i,\text{obs}} = Y_i(T_i)$, which we denote by Y_i for brevity. For each individual i, we also observe a vector of covariates $X_i \in \mathcal{X}$. As is conventional with other research into heterogeneous treatment effects, we perform inference by assuming that the samples are drawn i.i.d. from an infinite population.[10]

We now define the various quantities of interest studied throughout this paper. Let $\mathbf{G}$ be a measurable subset of the feature space $\mathcal{X}$. The average treatment effect (ATE), CATE, and the subgroup CATE are respectively defined as follows:

$$\text{ATE} : \tau_{\text{ATE}} := \mathbb{E}\left[Y(1)\right] - \mathbb{E}\left[Y(0)\right], \tag{1a}$$

$$\text{CATE} : \tau(x) := \mathbb{E}\left[Y(1)\ X = x\right] - \mathbb{E}\left[Y(0)\ X = x\right], \text{ for any } x \in \mathcal{X}, \tag{1b}$$

$$\text{subgroup CATE} : \tau_{\mathbf{G}} := \mathbb{E}\left[\tau(X)\ X \in \mathbf{G}\right], \text{ for measurable subset } \mathbf{G} \subset \mathcal{X}, \tag{1c}$$

where the expectation is taken with respect to the i.i.d. drawn from the infinite population.

At a high level, the goal of this work is to provide a systematic framework to find subgroups $\mathbf{G} \subset \mathcal{X}$, which (i) include non-trivial fraction of the observed data, (ii) are relevant and interpretable relevant for the domain problem at hand and (iii) most importantly have significant subgroup CATE; that is, $\tau_{\mathbf{G}}$ has significantly larger magnitude than τ_{ATE}.

3.1 Neyman Difference-in-Means Estimates for Finite Samples

We will often use the classical Neyman difference-in-means estimator to provide plug-in estimates for the ATE and subgroup CATE values. Formally, we denote the two study arms by the following:

$$\text{Treatment arm}\mathbf{T} := \{i \in [n] : T_i = 1\} \text{ and Control arm } \mathbf{C} := \{i \in [n] : T_i = 0\}. \tag{2a}$$

Throughout this paper, we will abuse notation: for any group $\mathbf{G} \subset \mathcal{X}$, we will use the same symbol to refer the subpopulation of individuals that belong to it. This allows us to denote the restriction of the two arms of the study to the subgroup, as follows:

$$\mathbf{T} \cap \mathbf{G} := \mathbf{T} \cap \{i \in [n] : X_i \in \mathbf{G}\} \text{ and } \mathbf{C} \cap \mathbf{G} := \mathbf{C} \cap \{i \in [n] : X_i \in \mathbf{G}\}. \tag{2b}$$

For a finite set $\mathcal{A}$, let $|\mathcal{A}|$ denote the number of elements in the set. With this notation at hand, the plug-in estimators for the average treatment effect τ_{ATE} and the subgroup average treatment effect $\tau_{\mathbf{G}}$ are given by

$$\hat{\tau}_{\text{ATE}} = \frac{1}{|\mathbf{T}|} \sum_{i \in \mathbf{T}} Y_i(1) - \frac{1}{|\mathbf{C}|} \sum_{i \in \mathbf{C}} Y_i(0), \qquad \text{and} \tag{3a}$$

$$\hat{\tau}_{\mathbf{G}} = \frac{1}{|\mathbf{T} \cap \mathbf{G}|} \sum_{i \in \mathbf{T} \cap \mathbf{G}} Y_i(1) - \frac{1}{|\mathbf{C} \cap \mathbf{G}|} \sum_{i \in \mathbf{C} \cap \mathbf{G}} Y_i(0). \tag{3b}$$

For randomised experiments, both estimates $\hat{\tau}_{\text{ATE}}$ and $\hat{\tau}_{\mathbf{G}}$ are unbiased (Splawa-Neyman *et al*., 1990), and standard error estimates are available for it (Imbens & Rubin, 2015). On the other hand, the precision of $\hat{\tau}_{\mathbf{G}}$ degrades as the size of the subgroup shrinks. For the same reason, a direct difference-in-means estimator for CATE (1b) is almost never feasible, as for most values of $x \in \mathcal{X}$ (e.g., when $\mathcal{X}$ is continuous, or combinatorially very large), there might not exist any sample with covariate equal to x.

4 Calibration as a Prediction (Reality) Check for CATE Estimators

Following the predictability principle of the PCS framework, any statistical model must pass a test of out-of-sample prediction accuracy before we should have any trust in it. This principle is in line with the ethos of the scientific method, which correlates the strength of a hypothesis with the rigour of prior attempts to falsify it (Popper, 1959). As discussed in Section 1.1, however, no such test currently exists for CATE models. The missing potential outcomes mean that we do not have a plug-in estimate for any risk function that measures the discrepany between the true CATE and the estimate CATE functions. Furthermore, unlike R^2 and area under the receiver operating characteristic curve scores, the proxy loss functions proposed for model choice (see Section 1.1 and the references therein) do not have interpretable scales.

To mitigate this problem, we develop a prediction accuracy check that can be applied to any CATE estimator. This check makes use of the ideas from the calibration literature (Dawid, 1982; DeGroot & Fienberg, 1983; Guo *et al*., 2017), and while passing the check is not a sufficient condition for a CATE estimator to have good performance, it is at least a necessary one. Even though our StaDISC approach is motivated by and grounded in the analysis of CATE estimators fitted to the VIGOR study data, we believe it is a general methodology useful for other causal inference problems.

The rest of this section is organised as follows. We discuss the 18 CATE estimators used in our analysis of the VIGOR data in Section 4.1. We then introduce the calibration-based scores for prediction checks in Section 4.2, and we apply it to the CATE estimators trained with VIGOR data in Section 4.3. Finally, in Section 4.4, we show how despite the poor performance on the overall data the CATE estimators have good generalisation locally, thereby setting the stage for identifying subgroups with subgroup CATE significantly larger than ATE in Section 5.

4.1 CATE Estimators Applied on the VIGOR Dataset

We now describe the 18 CATE estimators used in this work, 14 of which follow meta-learner strategies. Descriptions of the meta-learner strategies can be found in Künzel *et al*. (2019) and Nie & Wager (2017). Here, we simply list our choices of base learners for each meta-learner. The base learners are all drawn from a pool comprising lasso, logistic regression, RF, and gradient-boosted trees (GB). In our statistical analyses, we used implementations of the former three algorithms from the `scikit-learn` package (Pedregosa *et al*., 2011) and the `XGBoost` implementation of the latter (Tian *et al*., 2014). Furthermore, for code cleanliness, we made use of the meta-learner interface provided by the `causalml` package (Chen *et al*., 2020). In additional to estimators based on meta-learners, we also considered two versions each of causal tree (Athey & Imbens, 2016) and causal forest (Wager & Athey, 2018). The versions differ in terms of their hyperparameter choices. We used `causalml`'s implementation of the former. For the latter, we were not able to find a well-documented python implementation of the algorithm, so we built one around `causalml`'s causal tree implementation.

1. *S-learners (two* estimators): We used RF and GB as the base learners, denoted by `s_rf` and `s_xgb`, respectively.
2. *T-learners* (four estimators): We used lasso, logistic regression, RF and GB as base learners. These are denoted as `t_lasso`, `t_logistic`, `t_rf` and `t_xgb`.
3. *X-learners* (four estimators): We used lasso, logistic regression, RF and GB as base learners for the first stage and lasso as the only base learner for the second stage. These are denoted as `x_lasso`, `x_logistic`, `x_rf` and `x_xgb`.

4. *R-learners* (four estimators): In the case of randomised experiments, the R-learner requires a choice of base learner for the conditional expectation of the response with the treatment variable partialed out, and a choice of base learner for the treatment effect. We use four such pairs, each member of which was chosen uniformly at random from the base learners (with logistic regression excluded due to its similarity to lasso). Doing this, we got {lasso, lasso}, {lasso, GB}, {RF, lassso} and {RF, RF}. These are denoted as `r_lassolasso`, `r_lassoxgb`, `r_rflasso` and `r_rfrf`.
5. *Causal tree and causal forest* (four estimators): We used two versions each of the causal tree and causal forest algorithms, which we have denoted as `causal_tree_1`, `causal_tree_2`, `causal_forest_1` and `causal_forest_2`. Each pair of estimators differs in their hyperparameter choices. Specifically, `causal_tree_1` and `causal_forest_1` both use a minimum of 50 samples per leaf node, whereas `causal_tree_2` and `causal_forest_2` both use a minimum of 200 samples per leaf node. All other hyperparameter choices are standard and can be found in our documentation on GitHub.

Here, we briefly justify our choice of the 18 CATE estimators listed above. First, we chose our pool of base learners because they are representative of the most popular supervised learning algorithms in use today, with neural networks omitted because of the poor signal and small size of the dataset. The *T*-learner framework is perhaps the simplest way of fitting a CATE model and has been used and studied by many different authors. Using lasso as the base learners was proposed and analysed by Bloniarz *et al*. (2016) and Imai & Ratkovic (2013). Meanwhile, Foster *et al*. (2011) proposed using RF as the base learner. The X-learner (Künzel *et al*., 2019) and R-learner (Nie & Wager, 2017) frameworks have both been used by many recent works. The former has demonstrated favourable performance over other estimators in data challenges organised by the Atlantic Causal Inference Conference, while the latter has optimality guarantees under some assumptions and has been further supported by some follow-up work (Schuler *et al*., 2018). We included two S-learner estimators for completion, because all four meta-learner frameworks are supported by the `causalml` package. The causal tree (Athey & Imbens, 2016) and causal forest (Wager & Athey, 2018) estimators have similarly been used in much recent work, with the latter attaining the status of being a benchmark of sorts for CATE estimation methods in many simulations.

All CATE estimators based on meta-learners had the hyperparameters of their component base learners tuned via four-fold CV using `cv_orig`. A common hyperparameter grid was used for each base learner type, with details deferred to our documentation on GitHub.

4.2 A calibration-based score for CATE estimators

To develop a reality check scheme for CATE estimators, we now build on the literature of calibration of probability scores.

A binary classifier is said to be well-calibrated if the class probabilities that it predicts for each sample point are close to the true class probabilities. This property is desirable in many situations, such as weather forecasting, where we would like it to rain on close to 40% of the days on which a 40% chance of rain is forecast. Unfortunately, machine learning models are often not naturally calibrated, with neural networks in particular being overconfident in their estimated class probabilities (Guo *et al*., 2017). Furthermore, because class probabilities are unobserved, we cannot directly train a model to predict these values using supervised learning. While researchers have proposed various solutions to this problem, the common theme is to

bin the observations by their *predicted class probabilities* and then to use the observed class distribution over the bin to obtain plug-in estimates of the true class probabilities.

The concept of calibration has a long history (Dawid, 1982; DeGroot & Fienberg, 1983), and it has also been referred to as validity (Miller, 1962) or reliability (Murphy, 1973). Starting for evaluation of weather forecasts in the 1950s (Brier, 1950), calibration has been widely used as a generic scheme to compare several forecasters (DeGroot & Fienberg, 1983). Related ideas have been used to calibrate a wide range of methods, including Bayesian models (Dawid, 1982), support vector machines, boosted trees, RFs (Niculescu-Mizil & Caruana, 2005; Naeini *et al*., 2015) and more recently deep neural networks (Guo *et al*., 2017).

4.2.1 Binning via estimated CATE values

We now begin to define our calibration-based prediction accuracy measure for CATE estimators. While our scores—to be defined below—are easy to interpret, defining them formally requires a bit of notation, which we now describe.

Consider the training set $\mathbf{S}_{\text{TRAIN}}$ and let $\mathbf{S}_{\mathfrak{f}}, \mathfrak{f} = 1, 2, 3, 4$ denote its four-fold (random) CV split. Fix a fold $\mathfrak{f}$ and let $\mathbf{S}_{\text{TF}} = \mathbf{S}_{\text{TRAIN}} \backslash \mathbf{S}_{\mathfrak{f}}$ denote the training folds used to fit the CATE estimator $\mathbf{M} : \mathcal{X} \to \mathbb{R}$, and let $\mathbf{S}_{\text{VF}} = \mathbf{S}_{\mathfrak{f}}$ denote the left-out fold, which we also call as validation fold, for the estimator $\mathbf{M}$. Let $\mathfrak{m}_{\mathfrak{q}}$ denote the $\mathfrak{q}$-th quantiles of the CATE estimator $\mathbf{M}$ on the training folds of the data:

$$\mathfrak{m}_{\mathfrak{q}} = \min \left\{ c \;\middle|\; \frac{\#\{i \in \mathbf{S}_{\text{TF}} : \mathbf{M}(x_i) \leq c\}}{|\mathbf{S}_{\text{TF}}|} \geq \mathfrak{q} \right\}, \quad \text{for any } \mathfrak{q} \in (0, 1), \tag{4}$$

where by convention we set $\mathfrak{m}_0 = -\infty$ and $\mathfrak{m}_1 = \infty$. Then given a grid of q-values denoted by $\{\mathfrak{q}_1 \leq \mathfrak{q}_2 \leq \ldots \leq \mathfrak{q}_{K-1}\}$ in the interval $(0, 1)$, we split the real line into K bins, as follows:

$$\mathfrak{m}_0 < \mathfrak{m}_{\mathfrak{q}_1} \leq \mathfrak{m}_{\mathfrak{q}_2} \leq \ldots \leq \mathfrak{m}_{\mathfrak{q}_{K-1}} < \mathfrak{m}_1.$$

We use this binning to induce a partition of $\mathcal{X}$ into K *quantile-based subgroups* given by

$$\mathbf{G}_j := \mathbf{G}_j(\mathbf{M}) = \left\{ x \in \mathcal{X} | \mathbf{M}(x) \in [\mathfrak{m}_{\mathfrak{q}_j}, \mathfrak{m}_{\mathfrak{q}_{j+1}}] \right\} \text{ for } j = 0, 1, \ldots K-1. \tag{5a}$$

Given a set of individuals $\mathbf{S}$ (say, training folds $\mathbf{S}_{\text{TF}}$ or validation fold $\mathbf{S}_{\text{VF}}$), let $\overline{\mathbf{M}}_{\mathbf{G}_j \cap \mathbf{S}}$ denote the mean of the predicted CATE from the estimator $\mathbf{M}$ on the subgroups $\mathbf{G}_j \cap \mathbf{S}$:

$$\overline{\mathbf{M}}_{\mathbf{G}_j \cap \mathbf{S}} := \frac{1}{\left|\mathbf{G}_j \cap \mathbf{S}\right|} \sum_{i \in \mathbf{G}_j \cap \mathbf{S}} \mathbf{M}(X_i), \text{ where } \mathbf{G}_j \cap \mathbf{S} = \left\{ i \in \mathbf{S} | X_i \in \mathbf{G}_j \right\}. \tag{5b}$$

Similarly, recall that $\hat{\tau}_{\mathbf{G}_j \cap \mathbf{S}}$ denotes the plug-in estimate for the subgroup CATE for the subgroup $\mathbf{G}_j$

$$\hat{\tau}_{\mathbf{G}_j \cap \mathbf{S}} := \frac{1}{\left|\mathbf{T} \cap \mathbf{G}_j \cap \mathbf{S}\right|} \sum_{i \in \mathbf{T} \cap \mathbf{G}_j \cap \mathbf{S}} Y_i(1) - \frac{1}{\left|\mathbf{C} \cap \mathbf{G}_j \cap \mathbf{S}\right|} \sum_{i \in \mathbf{C} \cap \mathbf{G}_j \cap \mathbf{S}} Y_i(0). \tag{5c}$$

4.2.2 Score definitions

With these definitions of the subgroups, we are now ready to define the calibration score:

$$\text{Cal-Score}(\mathbf{S};\mathbf{M}) := \sum_{j=1}^{K} \frac{\left|\mathbf{G}_j \cap \mathbf{S}\right|}{|\mathbf{S}|} \cdot \left|\overline{\mathbf{M}}_{\mathbf{G}_j \cap \mathbf{S}} - \hat{\tau}_{\mathbf{G}_j \cap \mathbf{S}}\right|, \tag{6a}$$

where we use absolute difference (and not squared difference) since the scale of the quantities $\{\overline{\mathbf{M}}_{\mathbf{G}_j \cap \mathbf{S}}, \hat{\tau}_{\mathbf{G}_j \cap \mathbf{S}}\}$ is pretty small for our dataset. Nonetheless, it is still hard to interpret the absolute scale of Cal-Score($\mathbf{M}$), and hence, we normalise these scores by a baseline to define a pseudo-R^2 score. More precisely, we consider a baseline calibration score Cal-Score($\mathbf{S}; \hat{\tau}_{\text{ATE}}$), obtained by replacing the CATE estimator average $\overline{\mathbf{M}}_{\mathbf{G}_j \cap \mathbf{S}}$ with that of the (constant) ATE estimate $\hat{\tau}_{\text{ATE}}$ in Equation (6a):

$$\text{Cal-Score}(\mathbf{S};\hat{\tau}_{\text{ATE}}) := \sum_{j=1}^{K} \frac{\left|\mathbf{G}_j \cap \mathbf{S}\right|}{|\mathbf{S}|} \cdot \left|\hat{\tau}_{\text{ATE}} - \hat{\tau}_{\mathbf{G}_j \cap \mathbf{S}}\right|. \tag{6b}$$

With Equations (6a) and (6b) in place, we define the $\mathcal{R}^2_{\text{C}}$ score as follows:

$$\mathcal{R}^2_{\text{C}}(\mathbf{S};\mathbf{M}) := 1 - \frac{\text{Cal-Score}(\mathbf{S};\mathbf{M})}{\text{Cal-Score}(\mathbf{S};\hat{\tau}_{\text{ATE}})}. \tag{6c}$$

Just like the usual R^2-score,[11] the score $\mathcal{R}^2_{\text{C}}(\mathbf{S};\mathbf{M})$ can take any value between $(-\infty, 1]$, and a model can be deemed a good fit if this score is close to 1. We interpret the score as measuring, conditioned on the partition of the feature space into bins, the degree to which the CATE estimator explains the variability of the CATE with respect to the partition, in comparison with the best constant model.

As different models induce different partitions, the scores are not necessarily comparable across models. Furthermore, similar to how calibrated classification algorithms need not have good prediction accuracy, it is possible for a CATE model to have a good $\mathcal{R}^2_{\text{C}}$ score and yet have poor overall prediction accuracy for the CATE. Nonetheless, having $\mathcal{R}^2_{\text{C}}$-scores that are reasonably close to 1 across a range of data perturbations is *necessary* albeit not sufficient for the CATE model to have good prediction performance. Moreover, the variability of the score between the choices $\mathbf{S} = \mathbf{S}_{\text{TF}}$ and $\mathbf{S} = \mathbf{S}_{\text{VF}}$ also provides a check on the *overfitting* of the CATE estimator.

To conclude, the $\mathcal{R}^2_{\text{C}}$ provides two predictive checks for the CATE estimators. On the one hand, when $\mathcal{R}^2_{\text{C}}(\mathbf{S}_{\text{TF}};\mathbf{M})$ is much smaller than 1, we conclude that the estimator $\mathbf{M}$ has a poor fit on the training data. On the other hand, a high value (close to 1) value for $\mathcal{R}^2_{\text{C}}(\mathbf{S}_{\text{TF}};\mathbf{M})$, and a lower value (close to 0 or negative) for $\mathcal{R}^2_{\text{C}}(\mathbf{S}_{\text{VF}};\mathbf{M})$ would necessarily indicate overfitting of the estimator $\mathbf{M}$.

4.3 Calibration-based Predictive Check on CATE Estimators for the VIGOR Study

We now compute the scores defined in the previous section for the 18 popular CATE estimators when applied to the VIGOR dataset. We use the evenly spaced quantile grid $\{0.2, 0.4, 0.6, 0.8\}$ and compute the $\mathcal{R}^2_{\text{C}}$-scores using the $K = 5$ bins it induces. We also consider a restricted $\mathcal{R}^2_{\text{C}}$-score to measure the predictive performance of the estimators for the bottom two bins for the GI event, and top two bins for the CVT event. To compute this *restricted* $\mathcal{R}^2_{\text{C}}$-score, we simply replace the sum over the index $j \in \{1, 2, \ldots, 5\}$ in Equations (6a)

and (6b) with $j \in \{1, 2\}$ for the GI event and $j \in \{4, 5\}$ for the CVT event, and then we plug this restricted sum in Equation (6c).

In the previous section, we described how, given a CATE estimator and a fixed fold ϝ, we obtain two (restricted) $\mathcal{R}^2_C$-scores—one on the training folds $\mathbf{S}_{\text{TRAIN}} \backslash \mathbf{S}_ϝ$ and one on the validation fold $\mathbf{S}_ϝ$. Repeating this over four-fold provides us with four pairs of such scores. And iterating over M different types of CATE estimators yields $M \times 4$ such pairs. Furthermore, if we consider L different four-folds splits, we get $M \times 4 \times L$ such pairs of scores.

We trained 18 different CATE estimators for both the outcomes, namely, the GI and CVT events. However, after fitting, the following estimators learned a zero CATE function: R-learner with XGBoost for the GI event, and S-learner with XGBoost, Causal Tree with a particular choice of hyperparameters, and R-learner with XGBoost for the CVT event. Thus, going forward, we report results for the remaining 17 CATE estimators for the GI event and 15 CATE estimators for the CVT event. See Section 4.1 for more details on all the estimators. We now first discuss the details of scores presented in various plots in Figure 2 and then discuss the conclusions in a separate paragraph.

4.3.1 Details of Figure 2

In Figure 2(a), we provide a scatter plot of $\mathcal{R}^2_C(\mathbf{S}_{\text{TF}}, \mathbf{M})$ (training score) and $\mathcal{R}^2_C(\mathbf{S}_{\text{VF}}, \mathbf{M})$ (validation score) for five different estimators for each fold of original CV split `cv_orig` on the VIGOR data both for GI and CVT events. These estimators are `t_rf, s_rf, x_rf, r_rfrf` and `cf_1`, which denote T, S, X and R-learners with RF as base learners and (one of the two) causal forest, respectively. In addition, in the two right figures in Figure 2(a), we also provide the scatter plot of the corresponding restricted $\mathcal{R}^2_C$-scores (see the first paragraph

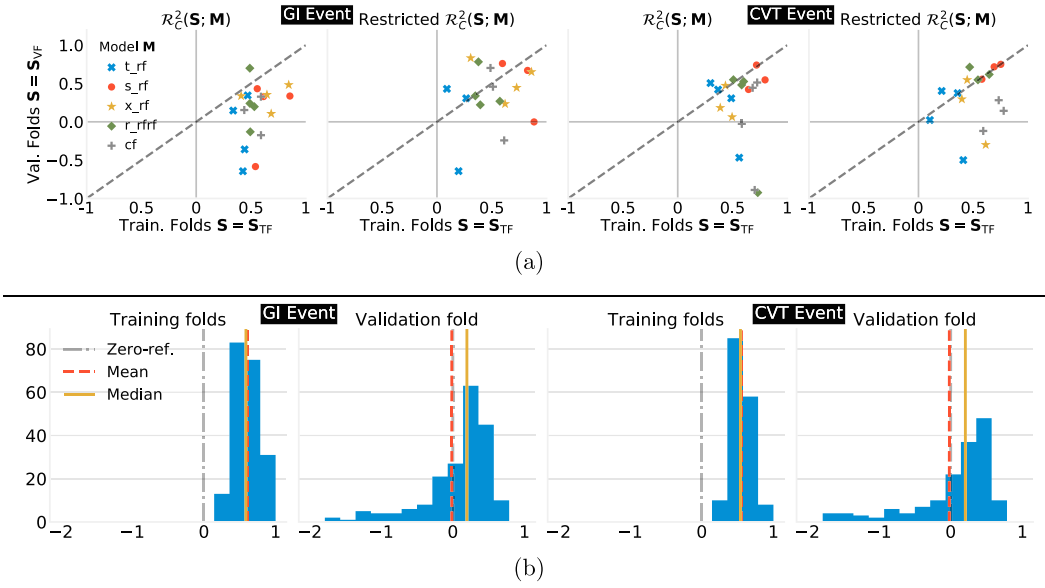


Figure 2. *Plots with the calibration-based $\mathcal{R}^2_C$-scores (6c) for various CATE estimators. (a) Scatter plot of $\mathcal{R}^2_C$-scores on the training and validation folds for five CATE estimators on the original four-fold split* `cv_orig` *on which hyperparameters were tuned via cross-validation. Refer to the text for definition of restricted $\mathcal{R}^2_C$-scores. (b) Histogram of the $\mathcal{R}^2_C$-scores on the 12 training and validation folds, four each from the three different CV splits, namely, {*`cv_orig, cv_0, cv_1`*} for 17 CATE estimators for GI event and for 15 CATE estimators for CVT event. [Colour figure can be viewed at* wileyonlinelibrary.com*]*

of this section for its definition) on the training and validation folds for the five estimators and both events.

Next, to check the *stability* of our conclusion, we compute these scores for all 17 CATE estimators for the GI event and all 15 CATE estimators for the CVT event on all three random CV splits {cv_orig,cv_0,cv_1}. That is, we obtain a total of 204 and 180 (training and validation) pairs of $\mathcal{R}^2_{\mathrm{C}}$-scores, respectively, for the GI and CVT events. In Figure 2(b), we plot the histogram of these scores.

4.3.2 Conclusions from Figure 2

Inspecting the scatter plots in Figure 2(a), we see clear evidence of overfitting, as the validation fold $\mathcal{R}^2_{\mathrm{C}}$-scores (computed as $\mathcal{R}^2_{\mathrm{C}}(\mathbf{S}_{\mathrm{VF}}, \mathbf{M})$ in equation (6c)) are systematically much smaller, and often negative, than those on the training folds (computed as $\mathcal{R}^2_{\mathrm{C}}(\mathbf{S}_{\mathrm{TF}}, \mathbf{M})$ in equation (6c)). Furthermore, there is substantial variability across different folds. For instance, one dot corresponding to s_rf for GI events was not even plotted because the validation fold $\mathcal{R}^2_{\mathrm{C}}$ score exceeded the lower y-limit of the plot. These findings are supported by the histograms in Figure 2(b), which show that the mean of the validation fold $\mathcal{R}^2_{\mathrm{C}}$-scores is in fact a negative number for both GI and CVT events. While we presented histograms of the aggregated scores over all the CATE estimators, the general behaviour was also true when looking at individual CATE estimators. Next, we also note that the bottom two-restricted $\mathcal{R}^2_{\mathrm{C}}$-score for the GI event and top two restricted $\mathcal{R}^2_{\mathrm{C}}$-score have slightly better generalisation because the validation scores are generally positive albeit with the caveat of larger variability across the training folds. (We revisit this aspect in more detail in Section 4.4.)

The poor performance on average as well as the high variability of performance both leads us to be sceptical of the conclusions from any CATE estimator on the VIGOR study data. Here, we remark that the variability of the scores stems from both fluctuations in the trained model and low signal-to-noise ratio in the validation fold (leading to Cal-Score deviating from its expected value). We remind the reader that in total there are 177 GI events and 59 total CVT events, and this fact implies that for each quantile-based subgroup, we should expect to see around 7.1 and 2.3 GI and CVT events, respectively, in the validation fold, under the assumption of no heterogeneity. The poor performance is hence entirely to be expected, and in fact could be a general theme for RCTs, as they are often sufficiently powered for only computing the ATE.

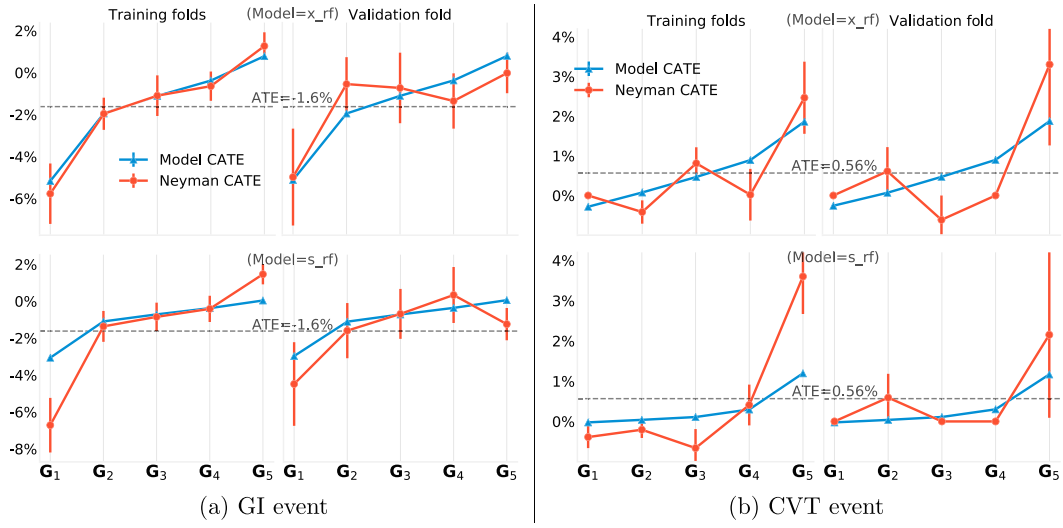
4.4 Extracting Data Conclusions that Can Be Relied Upon

While we conclude that we cannot trust the CATE models in their entirety, it remains to be seen if we can isolate data conclusions from them that we can rely on. To this end, we take a closer look the relative ordering of scores $\overline{\mathbf{M}}_{\mathbf{G}_j \cap \mathbf{S}}$ (5b) and $\hat{\tau}_{\mathbf{G}_j \cap \mathbf{S}}$ equation (5c) across the quantile-based subgroups $\{\mathbf{G}_j\}_{j=1}^5$ considered in the previous section. Given the quantile-based definition of the groups, it is natural to test whether we have

$$\overline{\mathbf{M}}_{\mathbf{G}_1 \cap \mathbf{S}} \leq \overline{\mathbf{M}}_{\mathbf{G}_2 \cap \mathbf{S}} \leq \ \ldots \ \leq \overline{\mathbf{M}}_{\mathbf{G}_5 \cap \mathbf{S}}, \ \text{(estimator CATEs) and} \tag{7a}$$

$$\hat{\tau}_{\mathbf{G}_1 \cap \mathbf{S}} \leq \hat{\tau}_{\mathbf{G}_2 \cap \mathbf{S}} \leq \ \ldots \ \leq \hat{\tau}_{\mathbf{G}_5 \cap \mathbf{S}}, \ \text{(subgroup CATE estimates)} \tag{7b}$$

for a set of individuals $\mathbf{S}$ comprising either the training folds or the validation fold. In Figure 3, we plot these estimates for two estimators x_rf and t_rf for the GI event in (a) and the CVT event in (b) for one set of training and validation folds from the original split. In each plot, the blue error bars denote the sample standard deviation estimate for the sample



(a) GI event (b) CVT event

Figure 3. *Investigating the monotonicty trend (equation (7)) for two CATE estimators* `x_rf` *and* `s_rf` *on one set of three training folds and one validation fold of the original four-fold split* `cv_orig`*, for (a) the GI event and (b) the CVT event. Here 'Model CATE' refers to the quantity* $\overline{\boldsymbol{M}}_{\boldsymbol{G}_j \cap \boldsymbol{S}}$*, and Neyman CATE refers to the quantity* $\hat{\tau}_{\boldsymbol{G}_j \cap \boldsymbol{S}}$*. In our notation, for training folds,* $\boldsymbol{S} = \boldsymbol{S}_{TF}$*, and for validation fold,* $\boldsymbol{S} = \boldsymbol{S}_{VF}$*. The error bars for Model CATE are the sample standard deviation for the estimated CATE values from the model, for each subgroup. For the Neyman CATE, the error bar denotes the square root of the estimated variance (11b). Note that the subgroups* $\{\boldsymbol{G}_j\}$ *are defined by the CATE estimator via the training folds. [Colour figure can be viewed at* wileyonlinelibrary.com*]*

mean $\overline{\mathbf{M}}_{\mathbf{G}_j \cap \mathbf{S}}$ computed from $\{\mathbf{M}(X_i), i \in \mathbf{G}_2 \cap \mathbf{S}\}$, and the red error bars denote the standard error estimate for $\hat{\tau}_{\mathbf{G}_j \cap \mathbf{S}_{\text{TF}}}$ given by Equation (11b). We observe that generally the model CATE estimates $\left\{\overline{\mathbf{M}}_{\mathbf{G}_1 \cap \mathbf{S}}\right\}_{j=1}^{5}$ are monotonic for both events on both training folds and validation fold. However, the story with the plug-in subgroup CATE estimates $\left\{\hat{\tau}_{\mathbf{G}_j \cap \mathbf{S}}\right\}_{j=1}^{5}$ is—not unexpectedly—mixed. For the GI event, while these estimates are monotonic on the training folds ($\mathbf{S} = \mathbf{S}_{\text{TF}}$), they are not monotonic on the validation fold ($\mathbf{S} = \mathbf{S}_{\text{VF}}$). For the rarer CVT event, the estimates $\left\{\hat{\tau}_{\mathbf{G}_j \cap \mathbf{S}}\right\}_{j=1}^{5}$ are not even monotonic on the training folds. This non-monotonic behaviour is far from unique to the two estimators presented here. Instead, the plots are representative of what we observe for all other estimators as well, even when using alternate data splits into training and validation folds.

4.4.1 Pairwise comparisons

To summarise this phenomenon, we do a pairwise comparison of successive quantile-based subgroups and measure the frequency with which the ordering of their CATE values generalises to the validation fold, and we summarise our results in Figure 4(a). More precisely, for a given estimator $\mathbf{M}$, we define the Boolean indicators:

$$A_{j,j+1} = \mathbb{I}(\hat{\tau}_{\mathbf{G}_j \cap \mathbf{S}_{\text{VF}}} \leq \hat{\tau}_{\mathbf{G}_{j+1} \cap \mathbf{S}_{\text{VF}}}) \text{ for } j = 1, 2, 3, 4. \tag{8a}$$

We then compute how often we have $A_{j,j+1} = 1$ over the 12 validation folds four each from the three CV splits {`cv_orig`,`cv_0`,`cv_1`}, and we denote this value by $\overline{A}_{j,j+1}$. Finally, we provide a box plot of the distribution of the values $\left\{\overline{A}_{j,j+1}, j = 1, 2, 3, 4\right\}$ across all 17 CATE estimators for the GI event and 15 CATE estimators for the CVT event in Figure 4(a). A value close to 1 suggests good generalisation, and conversely, a value close to 0 reflect poor

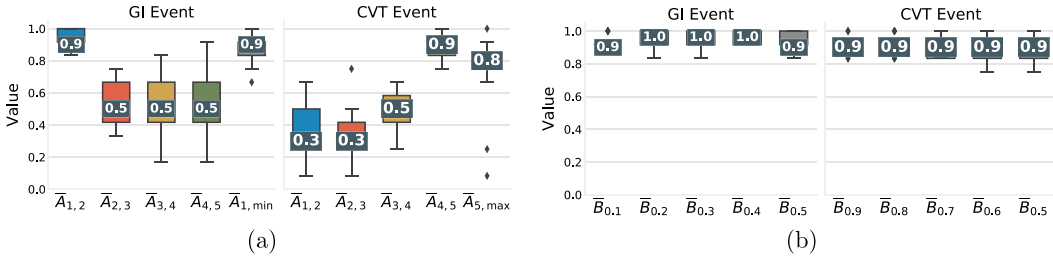


Figure 4. *FIGBox plots for pairwise comparisons of the subgroup CATE estimates for the five quantile-based subgroups based on the quantile grid* $\{0.2, 0.4, 0.6, 0.8\}$. *The boxplots in (a) denote the distribution for the mean fraction* $\overline{A}_{j,j+1}$ *(8a) (where the mean is computed over the 12 validation folds, four each from the three random CV splits* {`cv_orig`,`cv_0`,`cv_1`}*) across various CATE estimators, for the GI event on the left, and CVT event on the right. In addition, we also show the boxplot of the distribution of the Boolean variables* $\overline{A}_{1,\min}$ *(8b) for the GI event, and* $\overline{A}_{5,\max}$ *(8c) in the rightmost column of respective plot. In (b), we provide boxplots for the distribution of the mean value of Boolean indicators* $\{\overline{B}_{\mathfrak{q}}$ *(10) across all CATE estimators, for* $\mathfrak{q} \in \{0.1, 0.2, \ldots, 0.5\}$ *for the GI event, and* $\mathfrak{q} \in \{0.9, 0.8, \ldots, 0.5\}$ *for the GI event, where the mean is computed over the and the distribution is plotted across all the CATE estimators. Refer to Table A1 for estimator-wise results. [Colour figure can be viewed at* wileyonlinelibrary.com*]*

generalisation. On the one hand, we see that the pairwise ordering does not generalise well for most pairs of successive quantile-based subgroups as the frequency of generalisation $\overline{A}_{j,j+1}$ concentrates around values ≤ 0.5 for $j = 2, 3, 4$ for the GI event and $j = 1, 2, 3$ for the CVT event. On the other hand, we see that values of $\overline{A}_{1,2}$ for the GI event, and those of $\overline{A}_{4,5}$ for the CVT event are pretty close to 1 (we present more precise numerical values in Table A1.) This observation suggests that the ordering does generalise well for the subgroup with the strongest negative treatment effect for the GI event and the strongest positive treatment effect for the CVT event.

4.4.2 Investigating the quantile-based 'top' subgroups

We call the subgroups induced by $\mathbf{G}_1$ for the GI event, and $\mathbf{G}_5$ for the CVT event, the *quantile-based top subgroup*. Note that each subgroup is specific to a choice of estimator, a choice of training-validation split, and a choice of quantile grid. To further analyse the good generalisation of ordering for these top subgroups, we also compare them to the other quantile-based subgroups via two Boolean variables, as follows:

$$\text{for GI event: } A_{1,\min} := \mathbb{I}(\hat{\tau}_{\mathbf{G}_1 \cap \mathbf{S}_{\text{VF}}} = \min_j \hat{\tau}_{\mathbf{G}_j \cap \mathbf{S}_{\text{VF}}}), \text{ and} \tag{8b}$$

$$\text{for CVT event: } A_{5,\max} := \mathbb{I}(\hat{\tau}_{\mathbf{G}_5 \cap \mathbf{S}_{\text{VF}}} = \max_j \hat{\tau}_{\mathbf{G}_j \cap \mathbf{S}_{\text{VF}}}). \tag{8c}$$

We report the distribution of the frequency of generalisation $\overline{A}_{1,\min}$ (mean computed over the 12 validation folds) across the 17 CATE estimators for the GI event, and $\overline{A}_{5,\max}$ across the 15 CATE estimators for the CVT event as the rightmost entry of the corresponding figure in Figure 4(a). The plots show that, on the validation fold, the *quantile-based top subgroup* has the strongest treatment effect 90% of the time for the GI outcome, and about 80% of the time for the CVT outcome.

Next, to better investigate the performance of quantile-based top subgroups, we compare these top subgroups directly against their complement, reporting the results in Figure 4(b). In this plot, we also vary the $\mathfrak{q}$-value threshold used to define the quantile-based top subgroup. In particular, we consider groups of the form

$$\widetilde{\mathbf{G}}_{\mathfrak{q}} = \left\{x \in \mathcal{X} | \mathbf{M}(x) \in (-\infty, \mathfrak{m}_{\mathfrak{q}}]\right\}, \tag{9}$$

where $\mathfrak{m}_\mathfrak{q}$ denotes the q-th quantile of the CATE estimator $\mathbf{M}$ on the training folds [see Equation (4) for the mathematical expression]. Note that with this notation, $\widetilde{\mathbf{G}}_\mathfrak{q}^c = \{x \in \mathcal{X} | \mathbf{M}(x) \in (\mathfrak{m}_\mathfrak{q}, \infty\}$. In simple words, the subgroup $\widetilde{\mathbf{G}}_\mathfrak{q}$ is based on the quantile range $[0, \mathfrak{q}]$, and its complement subgroup $\widetilde{\mathbf{G}}_\mathfrak{q}^c$ is based on the quantile range $[\mathfrak{q}, 1]$. Then we check the ordering for between these subgroups via the following Boolean indicators:

$$B_\mathfrak{q} = \mathbb{I}\left(\hat{\tau}_{\widetilde{\mathbf{G}}_\mathfrak{q} \cap \mathbf{S}_{\text{VF}}} \leq \hat{\tau}_{\widetilde{\mathbf{G}}_\mathfrak{q}^c \cap \mathbf{S}_{\text{VF}}}\right), \text{ for } \begin{cases} \mathfrak{q} \in \{0.1, 0.2, \ldots, 0.5\} & \text{for GI event} \\ \mathfrak{q} \in \{0.9, 0.8, \ldots, 0.5\} & \text{for CVT event.} \end{cases} \tag{10}$$

Note that the subgroup of interest is $\widetilde{\mathbf{G}}_\mathfrak{q}$ for the GI event and $\widetilde{\mathbf{G}}_\mathfrak{q}^c$ for the CVT event. Moreover, in this new notation, the earlier subgroups (from Figure 4(a)) would be represented as $\mathbf{G}_1 = \widetilde{\mathbf{G}}_{0.2}$ and $\mathbf{G}_5 = \widetilde{\mathbf{G}}_{0.8}^c$. We notice that the ordering (10) holds much more frequently (compared with the pairwise ordering in Figure 4(a)). We also note from this figure that $\mathfrak{q} = 0.2$ and $\mathfrak{q} = 0.8$ provide the best generalisation performance for the GI and CVT events, respectively.

In summary, we have found that at least some of the CATE estimators yield quantile-based top subgroups that have subgroup CATE that is demonstrably stronger than that of the rest of the population. Thus, in the following sections, we use these quantile-based top subgroups, namely, the subgroups $\{\mathbf{G}_\mathfrak{q}, \mathfrak{q} = 0.1, 0.2, \ldots, 0.5\}$ for the GI event, and $\{\mathbf{G}_\mathfrak{q}^c, \mathfrak{q} = 0.9, 0.8, \ldots, 0.5\}$ for the CVT event for further analysis.

5 Stability-driven Ranking and Aggregation of CATE Estimators

Based on the discussion at the end of the last section, we believe that we can use a sub-collection of the CATE estimators to find subgroups with highly negative (in the case of the GI outcome) or positive (in the case of the CVT outcome) subgroup CATE, in the form of a quantile-based top subgroup. This observation brings us back to the question of estimator screening and choice: We seek to define a more stringent predictive test, and furthermore, out of all CATE estimators we considered, we would like to select those that are able to give us the best subgroups. While the overall goal of StaDISC is to find subgroups that are both statistically significant and interpretable, we focus in this part of paper on selecting estimators that yield the most significant subgroups, and we only address interpretability in Section 6.

5.1 Comparing Estimators Using t-statistics

We compare different CATE estimators using the statistical significance of their quantile-based top subgroup, measured via using standardised scores, namely, t-statistics. Given a subgroup $\mathbf{G}$, its corresponding t-statistic is given by

$$\mathbb{T}_\mathbf{G} := \frac{\hat{\tau}_\mathbf{G} - \hat{\tau}_{\text{ATE}}}{\sqrt{\hat{\text{Var}}\left[\hat{\tau}_\mathbf{G} - \hat{\tau}_{\text{ATE}} \mid \mathcal{F}\right]}}. \tag{11a}$$

Here, the term in the denominator is a plug-in estimate of a conditional variance, where the conditioning is over a σ-algebra $\mathcal{F}$ comprising knowledge of the group labels and treatment labels for all individuals in the sample population. More precisely, the variance estimate

is given by

$$\hat{\mathrm{Var}}\left[\hat{\tau}_{\mathbf{G}} - \hat{\tau}_{\mathrm{ATE}} \mid \mathcal{F}\right] := \left(\frac{|\mathbf{G}^c \cap \mathbf{C}|}{|\mathbf{C}|}\right)^2 \cdot \left(\frac{\hat{\mathrm{Var}}\left[Y(0) \mid \mathbf{G} \cap \mathbf{C}\right]}{|\mathbf{G} \cap \mathbf{C}|} + \frac{\hat{\mathrm{Var}}\left[Y(0) \mid \mathbf{G}^c \cap \mathbf{C}\right]}{|\mathbf{G}^c \cap \mathbf{C}|}\right) + \left(\frac{|\mathbf{G}^c \cap \mathbf{T}|}{|\mathbf{T}|}\right)^2 \cdot \left(\frac{\hat{\mathrm{Var}}\left[Y(1) \mid \mathbf{G} \cap \mathbf{T}\right]}{|\mathbf{G} \cap \mathbf{T}|} + \frac{\hat{\mathrm{Var}}\left[Y(1) \mid \mathbf{G}^c \cap \mathbf{T}\right]}{|\mathbf{G}^c \cap \mathbf{T}|}\right), \tag{11b}$$

where for a given set $\mathcal{A} \subset \mathcal{S}$, the quantity $\hat{\mathrm{Var}}\left[Y(t) \mid \mathcal{A}\right]$ denotes the sample variance:

$$\hat{\mathrm{Var}}\left[Y(t) \mid \mathcal{A}\right] = \frac{1}{|\mathcal{A}| - 1} \sum_{i \in \mathcal{A}} \left(Y_i(t) - \frac{1}{|\mathcal{A}|} \sum_{j \in \mathcal{A}} Y_j(t)\right)^2 \text{ for } t = 0, 1. \tag{11c}$$

We show in Appendix B1 that the estimator (11b) is an unbiased estimator of the conditional variance of $\hat{\tau}_{\mathbf{G}} - \hat{\tau}_{\mathrm{ATE}}$, and from the proof, it also easily follows that the estimator is consistent. As such, under the null hypothesis that $\tau_{\mathbf{G}} - \tau_{\mathrm{ATE}} = 0$, the t-statistic yields an asymptotically valid p-value.

In this paper, we deliberately choose not to use p-values to report the results, so as to avoid their susceptibility to misinterpretation. For interested readers, however, we mention the mapping between p-values and t-statistics ($\mathbb{T}$). The t-statistics presented throughout this work can be associated with one-sided p-values. In particular, a negative t-statistic with magnitude 1.65, 1.96 and 2.33 can be mapped to a left one-sided p-value of 0.05, 0.025 and 0.01, respectively. The same mapping exists between positive t-statistics and right one-sided p-values.

5.2 Defining Appropriate Perturbations

In order to guard against spurious and unreliable discoveries, the stability principle of the PCS framework requires conclusions to be stable to reasonable or appropriate perturbations at various stages of the data science life cycle. These include modelling and data perturbations familiar to statisticians which are appropriate under the Neyman–Rubin model assumptions, and also 'judgement call' perturbations where we reproduce or at least approximate the conclusions that would have been reached had various contingent choices been made differently. Examples of these choices include those made during data cleaning and feature engineering.[12]

As mentioned earlier in the paper, we have used a random CV split in order to fit and analyse our CATE models for the VIGOR data. In line with our prior discussion, we do not just evaluate each estimator based on the three CV splits {`cv_orig`,`cv_0`,`cv_1`}, but we also perform concurrent analyses of the estimator fitted and validated using four-fold splits of the data under four additional perturbations. Overall, we denote the set of all seven perturbations by {`cv_orig, cv_0, cv_1, cv_time, elderly_60, overweight, pert_outcome`}, where the three (random) CV splits {`cv_orig`,`cv_0`,`cv_1`} have already been used multiple times in the previous results of our paper. For completeness and to put them in context here, we revisit them while introducing the *new* perturbations {`cv_time, elderly_60, overweight, pert_outcome`} that we make use of in our subsequent analysis of the VIGOR dataset. We remind the reader that for each perturbation, we perform the same four-fold split for all the CATE estimators. Moreover, we continue to use the tuned hyperparameters from `cv_orig` for all other perturbations.

5.2.1 Sampling perturbations (cv_0, cv_1, cv_time)

The additional CV (random) splits {`cv_0, cv_1`}, used earlier and also in the sequel, help to account for sampling variability and are pretty commonly used in statistics and machine learning. Nonetheless, we also share Efron's concern that the use of random splits (Efron, 2020) does not play well with possible covariate shift and may lead researchers to be overly optimistic about conclusions that do not have external validity. To address this, we also split the training data into four equally sized folds by binning based on enrolment time, denoted by {`cv_time`}. This simulates possible variability in the sample population due to human choices (i.e. the date of the RCT),[13] and can also be seen more generally as making use of an a priori irrelevant variable to create heterogeneous folds and thus penalise ephemeral predictors.

5.2.2 Feature engineering perturbations (elderly_60, overweight, pert_outcome)

We use alternative thresholds to create perturbed versions of the ELDERLY and OBESE features. Instead of thresholding the adjusted age at 65, we create an ELDERLY_60 feature by thresholding it at 60, and instead of thresholding body mass index at 30, we instead threshold it at 25 to define the feature OVERWEIGHT. In this way, we create two perturbed datasets, denoted by {`elderly_60, overweight`}. Finally, for both the GI and CVT outcomes, the VIGOR study recorded for each patient both whether an event occurred, and also whether the occurred event was confirmed (meaning that it met the stringent criteria of an independent panel). In the original study, and thus far in our paper, we have used the confirmed events as the response of interest, but we now make use of the unconfirmed events to create a new response variable tracking all events. This increases the number of GI events from 177 to 190 and the number of CVT events from 59 to 84. Replacing the original responses with these one creates a further perturbed dataset for each outcome, which we denote by {`pert_outcome`}. For the three perturbations {`elderly_60, overweight, pert_outcome`}, we use the original four-fold split `cv_orig` of the patients (albeit with the perturbed features or outcomes in the data).

Performing our analyses on these perturbed datasets reveals to us what would have happened had we, or the original study authors, made different contingent decisions in feature engineering or problem formulation. Although models fit on these datasets no longer have exactly the same meaning as those fit on the original data, we still expect the estimators that perform well on the original data to also perform well on these perturbed datasets.

5.3 Ranking and Aggregation of CATE Estimators

In this section, we first rank the CATE estimators based on their performance across all data perturbations elaborated in the previous section. And then we select the estimators that are ranked in top 10 estimators across all the perturbations. Finally, we build a single 'ensemble CATE estimator' by taking a simple average (equal weights) of all the selected CATE estimators. Quantile-based top subgroups of the ensemble estimator form the starting point of finding interpretable subgroups in Section 6. We now describe the details of our ranking procedure.

5.3.1 Mean t-statistic per data perturbation

For a CATE estimator **M**, for each data perturbation $\mathfrak{D} \in$ {`cv_orig, cv_0, cv_1, cv_time, elderly_60, overweight, pert_outcome`}, we compute the mean *t*-statistic averaged across all quantiles across the corresponding four validation folds. In our

notation, for the GI event, this mean *t*-statistic is given by

$$\overline{\mathbb{T}}_{\text{GI}}(\mathfrak{D}) = \frac{1}{20} \sum_{\mathfrak{q} \in \mathcal{Q}} \sum_{\mathbf{S}_{\text{VF}} \in \mathcal{F}} \mathbb{T}_{\widetilde{\mathbf{G}}_{\mathfrak{q}} \cap \mathbf{S}_{\text{VF}}} \text{ where } \mathcal{Q} = \{0.1, 0.2, \ldots, 0.5\}, \mathcal{F} = \left\{\mathbf{S}_{\mathfrak{f}}, \mathfrak{f} = 1, 2, 3, 4\right\}, \tag{12a}$$

where the quantile-based top subgroup $\widetilde{\mathbf{G}}_{\mathfrak{q}}$ was defined in Equation (9). Moreover, we remind the reader that the quantiles that define the subgroup $\widetilde{\mathbf{G}}_{\mathfrak{q}}$ (see equations (4) and (5a)) are computed based on the CATE estimates from the fitted $\mathbf{M}$ on its training folds $\mathbf{S}_{\text{TF}} = \mathbf{S}_{\text{TRAIN}} \backslash \mathbf{S}_{\text{VF}}$. On the other hand, the *t*-statistic on the right-hand side of Equation (12a) is computed on the

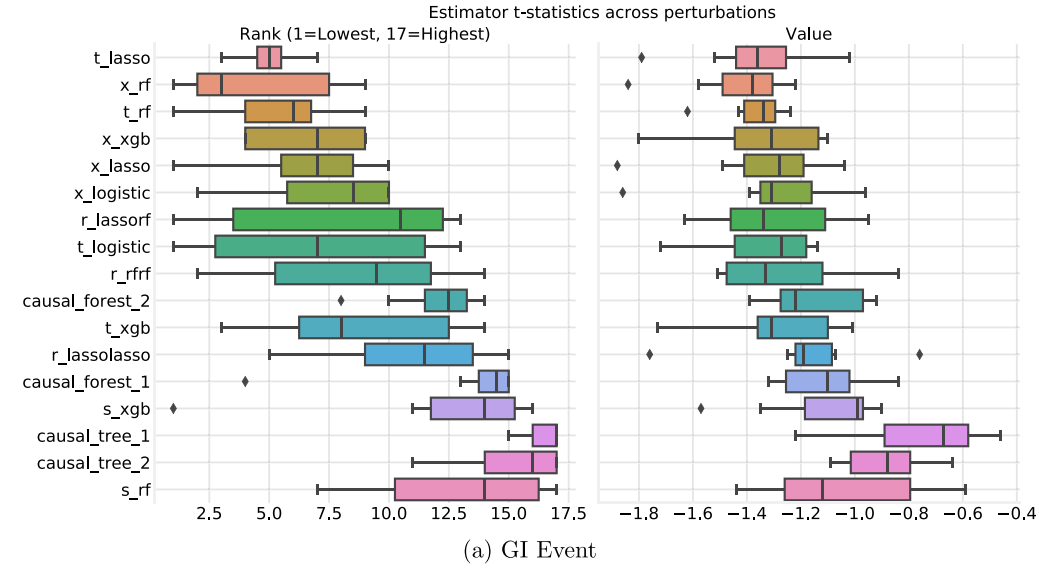
Table 2. *Estimator-wise and perturbation-wise t-statistic* $\overline{\mathbb{T}}_{GI}(\mathfrak{D})$ *(12a) for the GI event in (a) and* $\overline{\mathbb{T}}_{CVT}(\mathfrak{D})$ *(12b) for the CVT event in (b). In each column, the best (the lowest for GI event and highest for CVT event) t-statistic is highlighted in bold. The order of the estimators in (a) and (b) is the same order as that in Figure 5(a) and (b), respectively.*

Perturbation $\mathfrak{D}$ / Estimator $\mathbf{M}$	cv_orig	cv_0	cv_1	cv_time	elderly_60	overweight	pert_outcome
				$\overline{\mathbb{T}}_{\text{GI}}(\mathfrak{D})$			
t_lasso	−1.27	−1.79	**−1.52**	−1.36	−1.36	−1.02	−1.24
x_rf	−1.24	−1.84	−1.37	**−1.58**	−1.40	−1.22	−1.38
t_rf	−1.25	−1.62	−1.39	−1.34	−1.34	**−1.24**	**−1.43**
x_xgb	−1.16	−1.80	−1.44	−1.45	−1.31	−1.11	−1.10
x_lasso	−1.23	**−1.88**	−1.49	−1.33	−1.28	−1.04	−1.15
x_logistic	−1.31	−1.86	−1.39	−1.26	−1.31	−0.96	−1.06
r_lassorf	−1.26	−1.34	−1.36	−1.56	**−1.63**	−0.95	−0.96
t_logistic	**−1.33**	−1.72	−1.56	−1.14	−1.27	−1.17	−1.19
r_rfrf	−1.24	−1.45	−1.33	−1.51	−1.50	−1.00	−0.84
causal_forest_2	−1.00	−1.32	−1.39	−1.23	−1.22	−0.94	−0.92
t_xgb	−1.02	−1.73	−1.18	−1.31	−1.38	−1.01	−1.34
r_lassolasso	−1.10	−1.76	−1.25	−1.19	−1.19	−1.07	−0.76
causal_forest_1	−0.97	−1.26	−1.25	−1.10	−1.07	−0.84	−1.32
s_xgb	−0.95	−1.35	−1.57	−0.99	−1.02	−0.90	−0.99
causal_tree_1	−0.67	−1.22	−0.98	−0.50	−0.66	−0.80	−0.46
causal_tree_2	−1.07	−0.87	−0.72	−0.96	−1.09	−0.88	−0.64
s_rf	−0.78	−1.44	−0.81	−1.19	−1.33	−0.59	−1.12

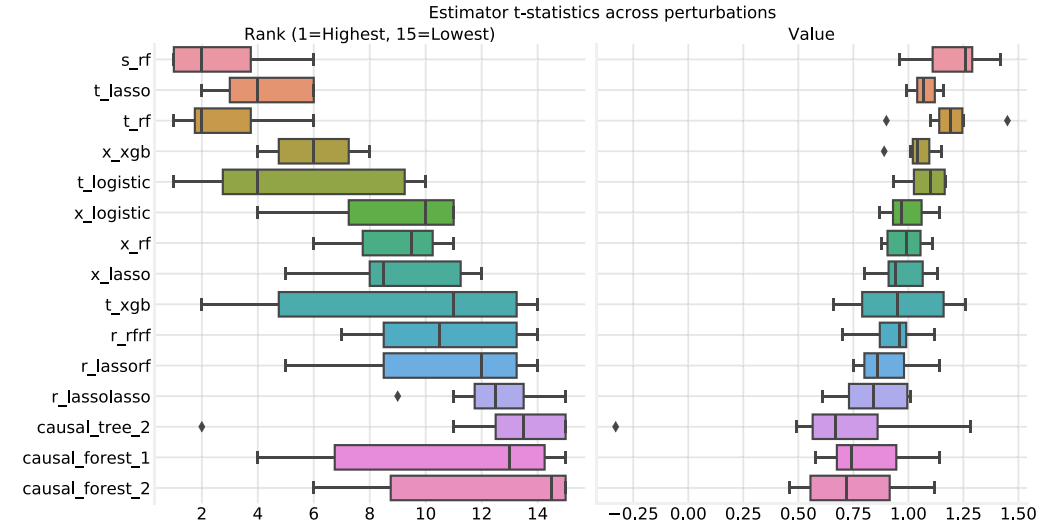
(a) GI event

Perturbation $\mathfrak{D}$ / Estimator $\mathbf{M}$	cv_orig	cv_0	cv_1	cv_time	elderly_60	overweight	pert_outcome
				$\overline{\mathbb{T}}_{\text{CVT}}(\mathfrak{D})$			
s_rf	0.96	**1.29**	**1.17**	**1.42**	**1.29**	1.05	1.26
t_lasso	1.06	1.16	0.99	1.02	1.10	1.07	1.14
t_rf	**1.10**	1.19	0.90	1.25	1.24	**1.18**	**1.45**
x_xgb	1.01	1.15	0.89	1.03	1.08	1.04	1.11
t_logistic	**1.10**	1.16	1.03	1.17	1.17	0.93	1.02
x_logistic	0.97	1.11	0.87	0.94	1.14	0.92	1.01
x_rf	0.90	1.11	0.88	0.91	1.09	0.99	1.02
x_lasso	0.92	1.13	0.80	0.90	1.10	0.94	1.03
t_xgb	0.66	1.06	0.92	1.26	0.95	0.66	1.26
r_rfrf	0.86	1.12	0.70	1.01	0.88	0.96	0.97
r_lassorf	0.79	1.14	0.75	0.93	0.86	1.03	0.81
r_lassolasso	0.81	1.01	0.65	0.61	1.01	0.84	0.98
causal_tree_2	0.67	0.88	0.84	−0.33	0.64	0.49	1.28
causal_forest_1	0.93	1.14	0.96	0.74	0.58	0.64	0.71
causal_forest_2	0.46	0.72	0.87	0.55	0.56	0.96	1.12

(b) CVT event



(a) GI Event



(b) CVT Event

Figure 5. *Box plots of the rank and value of mean t-statistic scores* $\overline{\mathbb{T}}_{GI}(\mathfrak{D})$ *(12a) and* $\overline{\mathbb{T}}_{CVT}(\mathfrak{D})$ *(12b), where the distribution is over the seven data perturbations* $\mathfrak{D} \in \{$`cv_orig, cv_0, cv_1, cv_time, elderly_60, overweight, pert_outcome`$\}$*. Here, the rank for the mean t-statistic score is computed per perturbation* $\mathfrak{D}$*, and all CATE estimators are ranked the lowest to the highest for the GI event, and the highest to the lowest for the CVT event. The estimator-wise and perturbation-wise numbers for both panels are reported in Table 2. [Colour figure can be viewed at* wileyonlinelibrary.com*]*

validation fold $\mathbf{S}_{\text{VF}}$. For the CVT event, the corresponding mean *t*-statistic is given by

$$\overline{\mathbb{T}}_{\text{CVT}}(\mathfrak{D}) = \frac{1}{20} \sum_{\mathfrak{q} \in \mathcal{Q}} \sum_{\mathbf{S}_{\text{VF}} \in \mathcal{F}} \mathbb{T}_{\widetilde{\mathbf{G}}^{c}_{\mathfrak{q}} \cap \mathbf{S}_{\text{VF}}} \text{ where } \mathcal{Q} = \{0.9, 0.8, \ldots, 0.5\}, \mathcal{F} = \left\{\mathbf{S}_{\mathfrak{f}}, \mathfrak{f} = 1, 2, 3, 4\right\}. \tag{12b}$$

We report the mean *t*-statistic $\overline{\mathbb{T}}(\mathfrak{D})$ for each CATE estimator and all seven data perturbations in Table 2(a) for the GI event and Table 2(b) for the CVT event. We also provide a visual summary

of the seven mean *t*-statistic for each estimator in the form of boxplot in Figure 5(a) for the GI event and (b) for the CVT event.

5.3.2 Ranking the CATE estimators

Next, for each category $\mathfrak{D}$, we rank the mean *t*-statistic from the lowest to the highest for the GI event, and the highest to the lowest for the CVT event. In accordance with the stability principle of the PCS framework, we screen for estimators that perform well across perturbations and thereby select all estimators that rank in top 10 across all data perturbations $\mathfrak{D}$. We provide the visual illustration of these ranks also in Figure 5 for the two events. In fact, the estimators in Figure 5 are sorted based on their worst rank across the perturbations. This criterion selects (i) two T-learners and four X-learners {`t_lasso, x_rf, t_rf, x_xgb, x_lasso, x_logistic`} for the GI event, and (ii) one S-learner, three T-learners and one X-learner {`s_rf, t_lasso, t_rf, x_xgb, t_logistic`} for the CVT event. The selected list can also be verified by a simple inspection of the rank plots from Figure 5.

5.3.3 Final step before interpreting

Keeping in mind the computational aspects of the next step (finding interpretable subgroups), and to increase stability, we decided to build an ensemble CATE estimator by using a simple average of the selected CATE estimators. Moreover, we also investigate the performance of the quantile-based top subgroups for this ensemble, and we report the mean *t*-statistic across the 12 validation folds from {`cv_orig,cv_0,cv_1`} for $\widetilde{\mathbf{G}}_{\mathfrak{q}}$ (9) for the GI event, and $\widetilde{\mathbf{G}}^{c}_{\mathfrak{q}}$ for the CVT event in Table 3. We report the standard deviation of the *t*-statistic across these folds in parentheses. In addition, we also report the mean percentage overlap computed pairwise across the entire training set $\mathbf{S}_{\text{TRAIN}}$ for the 12 ensemble estimators, four each from the three CV splits {`cv_orig,cv_0,cv_1`}. We observe that for the GI event, the subgroups corresponding to $\mathfrak{q} \in \{0.2, 0.3\}$ have higher $\mathbb{T}$, and for the CVT event, $\mathfrak{q} \in \{0.9, 0.8\}$ are the top two choices. The trends for overlap are as expected: with the increase in size of the group, the overlap generally increases and remains $>70\%$ across all choices. In the next section, we discuss our methodology to find an interpretable representation of the quantile-based top subgroups using the ensemble CATE estimator. As a final decision before that step, we choose the groups $\widetilde{\mathbf{G}}_{0.2}$ and $\widetilde{\mathbf{G}}_{0.3}$ for the GI event, and $\widetilde{\mathbf{G}}^{c}_{0.9}$ for the CVT event, based on their high *t*-statistic. We also include the group $\widetilde{\mathbf{G}}^{c}_{0.8}$ for the CVT event, keeping in mind the fact that the CVT event is very rare, and thus the low signal in the subgroup $\mathbf{G}^{c}_{0.9}$ (having only 10% of the training data) may become a bottleneck for any reasonable inference task.

TABLE 3. *t-statistic for different quantile-based top subgroups of the ensemble CATE estimator. 'Overlap' column reports the average % pairwise overlap between the 12 quantile-based top subgroups on the entire training data, namely, $\widetilde{\boldsymbol{G}}_{\mathfrak{q}} \cap \boldsymbol{S}_{TRAIN}$ for the GI event, and $\widetilde{\boldsymbol{G}}^{c}_{\mathfrak{q}} \cap \boldsymbol{S}_{TRAIN}$ for the CVT event. The 12 subgroups correspond to four each to the three CV splits {cv_orig, cv_0, cv_1}.*

Bottom quantile based subgroup $\widetilde{\mathbf{G}}_{\mathfrak{q}}$	GI event $\mathbb{T}_{\widetilde{\mathbf{G}}_{\mathfrak{q}}}$	Overlap (%)	**Top quantile** based subgroup $\widetilde{\mathbf{G}}^{c}_{\mathfrak{q}}$	CVT event $\widetilde{\mathbf{G}}^{c}_{\mathfrak{q}}$	Overlap (%)
$\mathfrak{q} = 0.9$	**1.28** (0.22)	77			
$\mathfrak{q} = 0.1$	−1.32 (0.20)	73	$\mathfrak{q} = 0.8$	1.03 (0.12)	75
$\mathfrak{q} = 0.2$	**−1.58** (0.19)	77	$\mathfrak{q} = 0.7$	0.85 (0.12)	77
$\mathfrak{q} = 0.3$	−1.47 (0.16)	82	$\mathfrak{q} = 0.6$	0.71 (0.09)	79
$\mathfrak{q} = 0.4$	−1.02 (0.12)	83	$\mathfrak{q} = 0.5$	0.57 (0.13)	**82**
$\mathfrak{q} = 0.5$	−0.81 (0.12)	**87**			

6 Finding Interpretable Subgroups

The next and final step of our investigation is to make our findings interpretable. Recall that the end goal in investigating the heterogeneous treatment effects in the VIGOR study is to inform treating physicians which subgroups of patients are likely to benefit from the reduced risk of GI events, without simultaneously incurring an increased risk of CVT events. Physicians may then favour prescribing the drug for patients in this subgroup. In situations involving high stakes decision making such as this one, decision makers are usually not comfortable with black-box decision rules but instead ideally require rules to be transparent and interpretable, so as to align them with their own knowledge base and justify them to patients and regulators.

6.1 Interpreting Using 'Cells'

In the work by Murdoch *et al*. (2019), one of us has argued that a key element of interpretability is the notion of relevance. Interpretations need to provide 'insight for a particular audience into a chosen domain problem'. Because clinical decision rules usually take the form of decision trees, a decision tree is the gold standard for our problem at hand. Each leaf of a decision tree constitutes a subset of the feature space defined by constraining the values of the features occurring along the root-to-leaf path. We call such a subset of a feature space a *cell*,[14] and propose to make our quantile-based top subgroups interpretable by approximating it with a union of a few cells, which we call a *cell cover*.[15]

Two remarks are in order. First, we find empirically that no single cell gives a good approximation of quantile-based top subgroups, so we require the additional flexibility of a union of multiple cells. Furthermore, reporting a union of cells is more flexible than reporting a decision tree, because it is not always possible to construct a tree with a given collection of cells as its leaf nodes.[16] Second, by focusing on cells, we recognise the importance of interactions, or in other words, nonlinear dependence of treatment effect on the covariates. Chernozhukov *et al*. (2018) proposed interpreting quantile-based top subgroups by estimating the differences in the 'observed characteristics' between the quantile-based top subgroup and the subgroup that is defined to be least affected by the treatment, but this only considers the marginal importance of each feature.

6.2 Cell Search Methodology

In this section, we demonstrate a general framework for how to search for a cell cover that contains most of the individuals in the quantile-based top subgroup but does not include too many individuals from outside it.

6.2.1 Feature selection

We start by selecting up to 10 features from the original list of 16 features. This is both to make the subsequent steps of cell search more computationally tractable and also to act as a form of regularisation.[17] To do this, we compute feature importance scores in two different ways. (i) Following Chernozhukov *et al*.(2018), we make use of the difference between the mean of the feature values over the quantile-based top subgroup and that over its complement. We refer to this score as the 'Logistic' feature importance score. (ii) We train a logistic classifier to predict membership in the quantile-based top subgroup and make use of the coefficients. In either case, we normalise so that the absolute values of the scores sum to one. We refer to this score as the 'difference' feature importance score. We compute these two types of scores for the ensemble CATE estimators' quantile-based top subgroups selected at the end of Section 5.3, namely, $\widetilde{\mathbf{G}}_{0.2}$ and $\widetilde{\mathbf{G}}_{0.3}$ for the GI outcome, and $\widetilde{\mathbf{G}}^{c}_{0.9}$ and $\widetilde{\mathbf{G}}^{c}_{0.8}$, across the 12 random

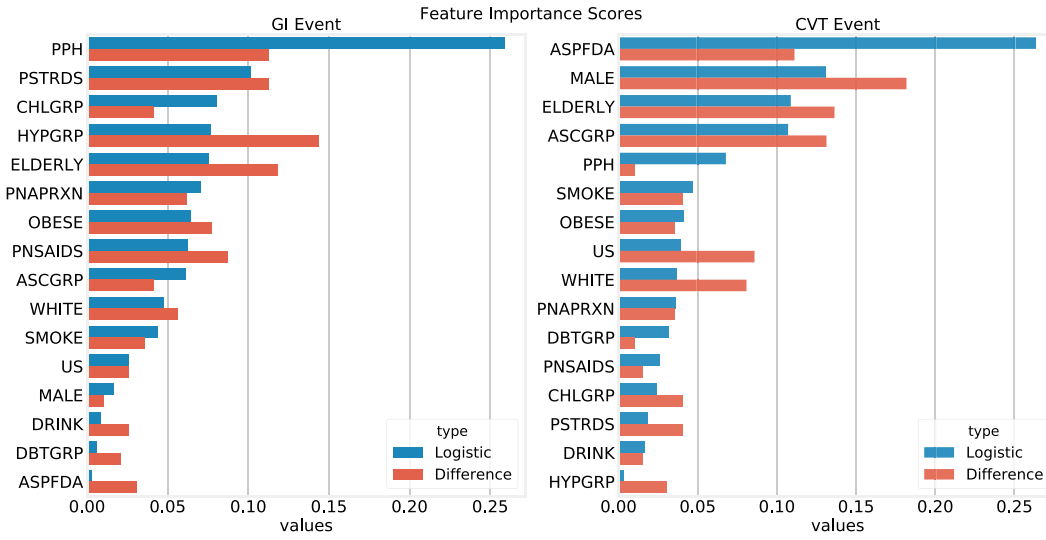


Figure 6. *Mean feature importance scores for the quantile-based top subgroups from the ensemble CATE estimator. Best seen in colour. We plot both the scores next to each other for each feature with the order (top, bottom) = (logistic, difference) but separately for each outcome. The blue bars and red bars respectively denote the 'logistic' and 'difference' feature importance scores described in the text. [Colour figure can be viewed at* wileyonlinelibrary.com*]*

training-validation splits ({`cv_orig`, `cv_0`, `cv_1`}). For each outcome, we average the feature importance scores across the different splits as well as both choices of the quantile-based top subgroups. The final results are shown in Figure 6.

Ranking the 16 features according to the two measures of feature importance, we select the features that rank among the top eight under either measure. Note that we choose to make use of both feature importance measures because they have different meanings: while the first score measures the marginal importance of each feature, the second measures its conditional importance. However, the choice of 'top eight' was also selected keeping in mind the fact that the top features for the two measures have a high overlap, and we end up selecting nine and 10 features, respectively, for the GI and CVT events listed (alphabetically) below:

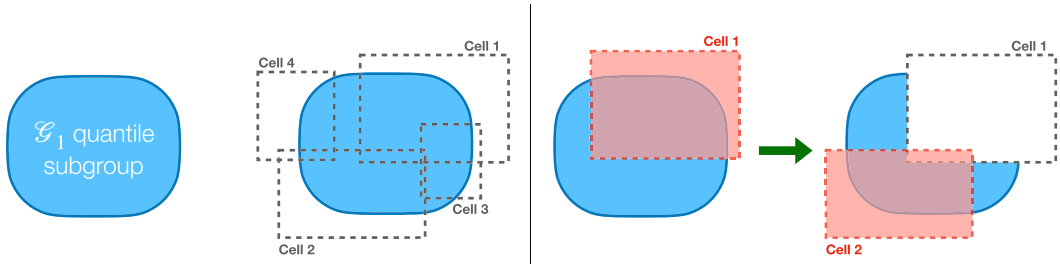
GI event: CHLGRP, HYPGRP, PNAPRXN, PNSAIDS, PSTRDS, PPH, ELDERLY, OBESE, WHITE

CVT event: ASCGRP, ASPFDA, CHLGRP, PPH, US, ELDERLY, MALE, OBESE, SMOKE, WHITE

Readers may refer to Table 1 to remind themselves about the definitions of all the features.

6.2.2 Iterative procedure

We now describe the `CellSearch` procedure for finding the cell cover for a quantile-based top subgroup one cell at a time, with Figure 7 also providing a pictorial explanation. For clarity, we introduce some notation, denoting the quantile-based top subgroup by $\mathbf{G}_{\text{top}}$, and the cell found at the i-th step by $\mathbb{C}_i$. For GI event, $\mathbf{G}_{\text{top}}$ takes the form $\widetilde{\mathbf{G}}_{\mathfrak{q}}$, and for the CVT event, $\widetilde{\mathbf{G}}^{c}_{\mathfrak{q}}$ for suitable choices of $\mathfrak{q}$. As before, we will abuse notation, using these symbols to refer to the subgroups and cells as subsets of the feature space, as well as the subpopulation of individuals that belong to them. At the first step, we consider every possible cell $\mathbb{C}$ defined with m features or less, where m is a user-specified tuning parameter, and we compute its 'true positive' (`TP`)



(a) Cover found by `CellSearch` (b) Illustration of one step of `CellSearch`

Figure 7. *A simplified illustration of* `CellSearch` *methodology for finding a cell-based cover for a given (quantile-based) subgroup. [Colour figure can be viewed at* wileyonlinelibrary.com*]*

and 'false positive' (`FP`) values with respect to $\mathbf{G}_{\text{top}}$, as follows:

$$\texttt{TP}(\mathbb{C}, \mathbf{G}_{\text{top}}) := \left|\mathbb{C} \cap \mathbf{G}_{\text{top}}\right|, \text{and } \texttt{FP}(\mathbb{C}, \mathbf{G}_{\text{top}}) := \left|\mathbb{C} \cap \mathbf{G}_{\text{top}}^{c}\right|.^{18} \tag{13}$$

Moreover, let $\Delta(\mathbb{C}, \mathbf{G}_{\text{top}}) := \texttt{TP}(\mathbb{C}, \mathbf{G}_{\text{top}}) - \texttt{FP}(\mathbb{C}, \mathbf{G}_{\text{top}})$ denote the difference of these values.[18]

We rank the cells based on their difference score $\Delta(\mathbb{C}, \mathbf{G}_{\text{top}})$, but instead of simply picking the cell to achieve the largest positive value $\Delta_{\max}$, we first create a candidate list of cells for which $\Delta(\mathbb{C}, \mathbf{G}_{\text{top}}) \geq \max(0, \Delta_{\max} - 0.05\left|\mathbf{G}_{\text{top}}\right|)$, we remove from cells any that are sub-cells[19] of other cells on this list, and then we choose one of remaining cells uniformly at random. The returns on adding this layer of complexity are to favour simpler, more interpretable cells and also (by running the procedure multiple times) to discover if two or more cells have comparable performance.[20]

In each subsequent step of the algorithm, to find the next cell in the cell cover, we first remove from the study population all individuals belonging to the cells already found, and then we repeat the above process. More rigorously, suppose cells $\mathbb{C}_1, \ldots, \mathbb{C}_{i-1}$ have already been determined. The true-positive and false-positive scores are now defined by

$$\begin{aligned} \texttt{TP}(\mathbb{C}, \mathbf{G}_{\text{top}}; \cup_{j=1}^{i-1}\mathbb{C}_j) &:= \left|\mathbb{C} \cap \mathbf{G}_{\text{top}} \setminus \cup_{j=1}^{i-1}\mathbb{C}_j\right|, \\ \text{and } \texttt{FP}(\mathbb{C}, \mathbf{G}_{\text{top}}; \cup_{j=1}^{i-1}\mathbb{C}_j) &:= \left|\mathbb{C} \cap \mathbf{G}_{\text{top}}^{c} \setminus \cup_{j=1}^{i-1}\mathbb{C}_j\right|, \end{aligned} \tag{14}$$

while $\Delta_{\max}$ and the threshold are also modified accordingly. Finally, the procedure terminates if $\Delta_{\max}$ at any iteration is less than or equal to 0 or if the number of iterations has reached a pre-specified threshold (default value 3).

6.2.3 Aggregating results over multiple runs

In accordance with the stability principle, we run `CellSearch` multiple times, and we check whether the same cell cover is found. In our case, we ran it five times on each top quantile subgroup arising from 12 random training-validation splits, for a total of 60 runs. While the cell cover did not turn out to be stable, we found that certain cells or their sub-cells frequently re-appeared within each run. We thus turn our focus to individual cells, and we aggregate the results over the multiple runs, calling this procedure `StabilizedCellSearch`.

To describe how we aggregate the results, we first use $\mathcal{B}$ to denote the collection of all 60 runs, and for each run $b \in \mathcal{B}$, we let $\mathfrak{C}_b$ denote the cover returned by the procedure, while the

collection of all cells found is denoted $\mathfrak{C} := \cup_{b\in\mathfrak{B}}\mathfrak{C}_b$. For each cell $\mathbb{C} \in \mathfrak{C}_b$, we define its stability score as follows:

$$\texttt{Stab}(\mathbb{C}) = \frac{1}{|\mathcal{B}|}\sum_{b\in\mathcal{B}}\sum_{\mathbb{C}'\in\mathcal{C}} \mathbf{1}(\mathbb{C}' \in \mathfrak{C}_b \text{ and } \mathbb{C}' \text{is sub-cell of } \mathbb{C})\frac{|\mathbb{C}'|}{|\mathbb{C}|}. \tag{15}$$

This score measures how frequently cell $\mathbb{C}$ and its proper sub-cells are found across the different runs, with each occurrence weighted by the relative size of the sub-cell.

Finally, we rank the cells according to their stability scores and output those for which the score exceeds a user-defined threshold. In our case, we chose the threshold to be 1/3, which results in finding three cells each for the GI and CVT outcomes. We discuss these cells in the next section, while the full results obtained by running `StabilizedCellSearch` on the VIGOR data with respect to both the GI and CVT outcomes are shown in Table A2.

6.3 Discussion of Cells Found and Performance on Test Set

In this section, we discuss the statistical significance of the cells found for both GI and CVT outcomes. First, we list the top three cells found for each outcome, where detailed results for top 20 cells (by `Stab`-scores) are reported in Table A2.

For the GI outcome, the top three stable cells are as follows:

(i) $\mathbb{C}_1$: patients with prior history of GI event denoted as $\{\text{PPH} = 1\}$;
(ii) $\mathbb{C}_2$: patients who (self) reported a prior (to the experiment) usage of steroids, and a history of hypertension denoted as $\{\text{PSTRDS} = 1, \text{HYPGRP} = 1\}$; and
(iii) $\mathbb{C}_3$: elderly patients who reported a prior usage of steroid drugs denoted as $\{\text{PSTRDS} = 1, \text{ELDERLY} = 1\}$.

For the CVT outcome, they are as follows:

(i) $\widetilde{\mathbb{C}}_1$: patients for which use of aspirin has been indicated as per FDA guidelines $\{\text{ASPFDA} = 11\}$;
(ii) $\widetilde{\mathbb{C}}_2$: male elderly patients $\{\text{MALE} = 1, \text{ELDERLY} = 1\}$; and
(ii) $\widetilde{\mathbb{C}}_3$: patients who have reported prior history $\{\text{ASCGRP} = 1\}$.

For further details on the features appearing above, please refer back to Section 2.2. In Figure 8, we plot the overlap between these cells.

6.3.1 Conclusions from Figure 8

As can be seen in Figure 8(a), there is little to moderate overlap among the cells $\mathbb{C}_1$ and $\mathbb{C}_3$, which shows that they are meaningfully different. On the other hand, there is significant overlap among the cells $\widetilde{\mathbb{C}}_1, \widetilde{\mathbb{C}}_3$ in Figure 8(b). In particular, $\widetilde{\mathbb{C}}_1$ is a subset (but not a sub-cell) of $\widetilde{\mathbb{C}}_3$. The reason we report both cells is because of the suspected multi-scale nature of treatment effect variation for the CVT outcome, with $\widetilde{\mathbb{C}}_1$ found more often for $\mathfrak{q} = 0.9$, and $\tilde{\mathbb{C}}_3$ found more often for $\mathfrak{q} = 0.8$.

We now compute and report several quantities for each of these six cells, finally making use of the holdout test dataset (20% of the study size) for the *very first time*. For cells $\mathbb{C}_1, \mathbb{C}_2$ and $\mathbb{C}_3$, as well as the union $\cup_{j=1}^3 \mathbb{C}_j$ of these three cells, the results are reported in Table 4. Similar results for the cells $\tilde{\mathbb{C}}_1, \tilde{\mathbb{C}}_2$, and $\widetilde{\mathbb{C}}_3$ and their union $\cup_{j=1}^3 \widetilde{\mathbb{C}}_j$ are reported in Table 5. We now discuss the results from Tables 4 and 5 one by one.

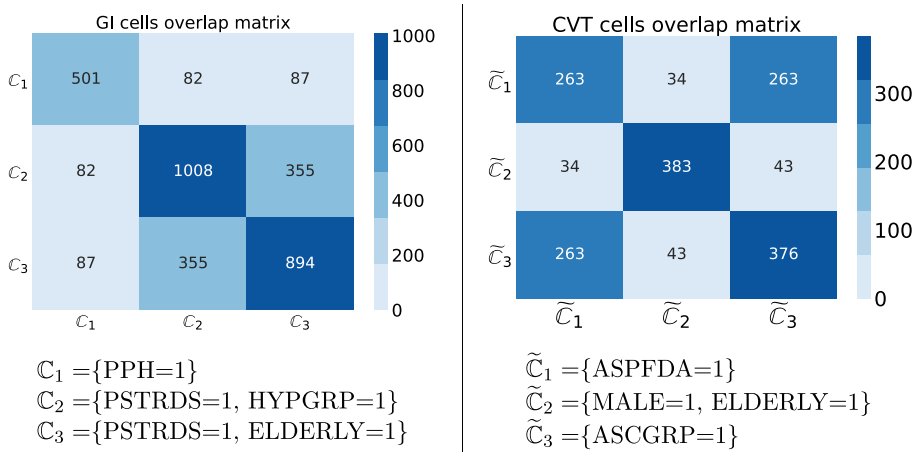


Figure 8. *Overlap matrix for final discovered cells on the training data* $\boldsymbol{S}_{\text{TRAIN}}$. *For (a), the data split is stratified on the treatment indicator and the GI outcome, and that for (b) is stratified on the treatment indicator and the CVT outcome. For instance, the number 82 for the entry corresponding to* $\mathbb{C}_1$ *and* $\mathbb{C}_2$ *in (a) represents that the two cells had 82 patients in common on the training data. [Colour figure can be viewed at* wileyonlinelibrary.com*]*

Table 4. *Results for the final cells selected after* `StabilizedCellSearch` *for the GI event, namely,* $\mathbb{C}_1 = \{PPH = 1\}$, $\mathbb{C}_2 = \{PSTRDS = 1, HYPGRP = 1\}$ *and* $\mathbb{C}_3 = \{PSTRDS = 1, ELDERLY = 1\}$ *from Section 6.3. We also report the results for the other outcome, namely, CVT event, on the entire data (all 8076 patients).* † *In column* $\boldsymbol{S}_{VAL}$, *we report the mean t-statistics and standard deviation in parentheses, across the 12 different folds of the training data* $\boldsymbol{S}_{TRAIN}$ *obtained from the three random CV splits {*`cv_orig`, `cv_0`, `cv_1`*}.*

	#evts/size		CATE Est. $\widehat{\tau}_{\mathbb{C}\cap\mathbf{S}}$ **(std)**		t-statistic $\mathbb{T}_{\mathbb{C}\cap\mathbf{S}}$		
Dataset S Cell $\mathbb{C}$	$\mathbf{S}_{\text{TRAIN}}$	$\mathbf{S}_{\text{TEST}}$	$\mathbf{S}_{\text{TRAIN}}$	$\mathbf{S}_{\text{TEST}}$	$\mathbf{S}_{\text{TRAIN}}$	$\mathbf{S}_{\text{TEST}}$	$^{\dagger}\mathbf{S}_{\text{VAL}}$
GI event (GI-stratified split)							
PPH = 1	36/501	8/129	−0.057 (0.023)	−0.055 (0.042)	−1.89	−1.01	−0.99 (0.27)
PSTRDS = 1, HYPGRP = 1	39/1008	6/238	−0.050 (0.012)	−0.037 (0.021)	−3.17	−1.06	−1.57 (0.22)
PSTRDS = 1, ELDERLY = 1	46/894	9/227	−0.051 (0.015)	−0.063 (0.026)	−2.74	−2.00	−1.38 (0.17)
Union	79/1905	19/471	−0.038 (0.009)	−0.047 (0.018)	−3.15	−2.22	−1.59 (0.20)
All	142/6460	35/1616	−0.016 (0.004)	−0.016 (0.007)	-	-	-
CVT event (entire data)							
PPH = 1	2/630		−0.006 (0.004)		−2.66		
PSTRDS = 1, HYPGRP = 1	11/1246		0.008 (0.005)		0.44		
PSTRDS = 1, ELDERLY = 1	16/1121		0.015 (0.007)		1.42		
Union	21/2376		0.007 (0.004)		0.55		
All	59/8076		0.006 (0.002)		-		

6.3.2 Results from Table 4

In the first three rows of Table 4, we examine the subgroup treatment effect for these cells with respect to the GI outcome. In the second and third columns, we report two versions of the Neyman estimate for the cell CATE $\hat{\tau}_{\mathbb{C}\cap\mathbf{S}}$, one computed on the training set $\mathbf{S}_{\text{TRAIN}}$ as well as one computed on the test set $\mathbf{S}_{\text{TEST}}$. Likewise, in the next two columns, we report the t-statistic $\mathbb{T}_{\mathbf{G}\cap\mathbf{S}}$, one computed on the training set $\mathbf{S}_{\text{TRAIN}}$ and on the test set $\mathbf{S}_{\text{TEST}}$. Finally, in the last column with header $^{\dagger}\mathbf{S}_{\text{VAL}}$, we report the mean (and standard deviation in parenthesis) of the t-statistics $\mathbb{T}$ computed on the 12 different folds of $\mathbf{S}_{\text{TRAIN}}$ from the three random CV

Table 5. *Results for the final cells selected after* `StabilizedCellSearch` *for the CVT event, namely,* $\widetilde{\mathbb{C}}_1 = \{ASPFDA = 1\}$, $\widetilde{\mathbb{C}}_2 = \{MALE = 1, ELDERLY = 1\}$ *and* $\widetilde{\mathbb{C}}_3 = \{ASCGRP = 1\}$ *from Section 6.3. We also report the results for the other outcome, namely, GI event, on the entire data (all 8076 patients).* †*In column* $\boldsymbol{S}_{VAL}$*, we report the mean t-statistics and standard deviation in parentheses, across the 12 different folds of the training data* $\boldsymbol{S}_{TRAIN}$ *obtained four each from the three random CV splits {*`cv_orig`, `cv_0`, `cv_1`*}.*

	#evts/size		CATE Est. $\widehat{\tau}_{\mathbb{C}\cap\mathbf{S}}$ **(std)**		t-statistic $\mathbb{T}_{\mathbb{C}\cap\mathbf{S}}$		
Dataset S **Cell** $\mathbb{C}$	$\mathbf{S}_{\text{TRAIN}}$	$\mathbf{S}_{\text{TEST}}$	$\mathbf{S}_{\text{TRAIN}}$	$\mathbf{S}_{\text{TEST}}$	$\mathbf{S}_{\text{TRAIN}}$	$\mathbf{S}_{\text{TEST}}$	$^{\dagger}\mathbf{S}_{\text{VAL}}$
CVT Event (CVT-stratified split)							
ASPFDA = 1	13/263	5/58	0.062 (0.025)	0.103 (0.074)	2.28	1.38	1.09 (0.20)
MALE = 1, ELDERLY = 1	12/383	0/111	0.040 (0.017)	0 (0)	2.09	−1.16	0.85 (0.24)
ASCGRP = 1	15/376	6/78	0.044 (0.020)	0.047 (0.060)	2.05	0.74	1.04 (0.23)
Union	24/716	6/175	0.042 (0.013)	0.024 (0.028)	3.09	0.77	1.55 (0.13)
All	**47/6460**	12/1616	0.006 (0.002)	0.005 (0.004)	-	-	-
GI Event (entire data)							
ASPFDA = 1	6/321		−0.027 (0.016)		−0.71		
MALE = 1, ELDERLY = 1	17/494		−0.045 (0.016)		−1.85		
ASCGRP = 1	8/454		−0.028 (0.013)		−0.96		
Union	25/891		−0.040 (0.011)		−2.27		
All	177/8076		−0.016 (0.003)		-		

splits {`cv_orig`, `cv_0`, `cv_1`}. Overall, the test set results are promising, with test set CATE estimates being much more negative than the estimated ATE and comparable with their training set counterparts. While we do not report p-values because they can be easily misunderstood, we note that the test set t-statistic values for the GI outcome are $\mathbb{C}_3$, and the union $\cup_{j=1}^{3}\mathbb{C}_j$, are both significant at the 0.025 level for a one-sided z-test.

The starting point of our investigation of the VIGOR dataset was the hope to identify a subgroup for which Vioxx simultaneously has a strong negative treatment effect for GI risk and a low positive treatment effect for CVT risk. Consequently, in the last three rows of Table 4, we report the treatment effect results for the cells $\{\mathbb{C}_j\}_{j=1}^{3}$ and their union, with respect to the CVT outcome. While $\mathbb{C}_2$ and $\mathbb{C}_3$ experience increased CVT risk, $\mathbb{C}_1 = \{\text{PPH} = 1\}$ in fact shows reduced CVT risk, which makes it especially promising for further clinical investigation. We note that for the CVT outcome, we report the CATE estimates and the t-statistic on the entire data as this outcome had no role to play in the entire StaDISC pipeline with the GI outcome, and hence the entire data can be treated as a 'valid' test set for estimating heterogeneous treatment effect of Vioxx with the CVT outcome.

6.3.3 Results from Table 5

In Table 5, we report the analogous results for cells $\widetilde{\mathbb{C}}_1, \widetilde{\mathbb{C}}_2$, and $\widetilde{\mathbb{C}}_3$, and their union $\cup_{j=1}^{3}\widetilde{\mathbb{C}}_j$, first for the CVT outcome, and then the GI outcome. For these cells, the generalisation to the holdout test set is weaker, with only $\widetilde{\mathbb{C}}_1$ and $\widetilde{\mathbb{C}}_3$ having test set CATE values that remain substantially positive. Furthermore, the test set t-statistic values are smaller. All these observations are unsurprising given the rarity of the CVT outcome—in particular, only 12/1616 individuals in the test set $\mathbf{S}_{\text{TEST}}$ experienced an event. Nonetheless, the test set CVT-CATE estimates for $\widetilde{\mathbb{C}}_1$ and $\widetilde{\mathbb{C}}_3$ support the view that the treatment effect is stronger on these subgroups, while the GI-CATE estimates do not suggest that these subgroups benefit especially strongly from the treatment with Vioxx.

7 Complementary Analysis with the APPROVe Study

It is well documented that RCTs have problems with *external validity* (Jüni *et al*., 2001; Rothwell, 2005; Fortin *et al*., 2006; Krauss, 2018), which is defined by Rothwell to be 'whether the results can be reasonably applied to a definable group of patients in a particular clinical setting in routine practice' (Rothwell, 2005). This phenomenon arises primarily because RCTs have carefully defined enrolment criteria, so conclusions in such studies may not apply to patients who do not conform to these criteria. In more mathematical language, the ATE, subgroup CATE and other estimands of interest are all defined in terms of expectations with respect to a particular distribution of patients, a particular outcome and a particular treatment and hence do not directly apply when any of these change. We refer the interested reader to the excellent articles by Rothwell (2005, 2006) for a further discussion on these topics.

Despite its importance for clinical relevance, external validity has been relatively neglected by researchers and institutions overseeing the conduct of RCTs (Rothwell, 2005; Krauss, 2018). One way to argue for external validity is to attempt *external validation*, that is, to reproduce the results obtained on one dataset on a different but related dataset. Recent voices that urge the community to give external validation a higher priority across many domains (Debray *et al*., 2015; Krauss, 2018; Norgeot *et al*., 2020) are very much in accordance with Yu and Kumbier's (2020) call to statisticians to broaden the scope of their concern from data modelling to the entire data science life cycle as part of the PCS framework. This can be seen not only as one more predictive and stability check under the PCS framework but also as a special case of 'transfer learning' where the desiderata is the transferability of the conclusions or findings from one dataset to other related datasets.

These reasons motivate the following complementary analysis of the APPROVe study (Baron *et al*., 2008), another RCT investigating Vioxx. More precisely, we compute the subgroup CATEs with respect to both the GI and CVT outcomes over this new dataset, and we show that the *qualitative* conclusions obtained by applying StaDISC to the VIGOR study also generalise to this dataset for four out of the six subgroups from Figure 8; the other two subgroups were too small in size and did not have any GI events. We now start with a background on the APPROVe study followed by a discussion of the results on subgroup CATEs.

7.1 Background for the APPROVe Study

In this section, we provide only a brief background for the APPROVe study and refer the readers to the original paper (Baron *et al*., 2008) for additional details.

The Adenomatous Polyp Prevention on Vioxx (APPROVe) study was another randomised trial sponsored by Merck, but unlike VIGOR, it was placebo controlled. Conducted in 2001–2004, it was designed to assess whether Vioxx could 'reduce the risk of adenomatous polyps in individuals with a recent history of these tumours' (Baron *et al*., 2008). The study population comprised 2587 patients who had colon adenomatous polyps removed during a 12-week period before being entered into the study and who had no known polyps remaining. After it was discovered that Vioxx had significant cardiovascular toxicity, the study was terminated 2 months early in September 2004, but all individuals were followed up for at least a year afterwards off-treatment.

The data files of the APPROVe study followed a very similar format to that of the VIGOR study albeit with two major differences: (i) GI event was not directly labelled in the dataset, and (ii) the risk factor file was not available. As a result, outcomes related to the GI event, and features (including but not limited to) ASPFDA, ASCGRP, HYPGRP and PSTRDS—which were used to define the final subgroups obtained in the previous section—were not directly

Table 6. *Results for the subgroups found with StaDISC on VIGOR, for the APPROVe dataset. Note that unlike VIGOR, the patients in the control arm for the APPROVe study were treated with a placebo, which makes the quantitative results reported here not directly comparable with those reported in Tables 4 and 5. Refer to the text for further discussion. The armwise statistics of the features and outcomes for the APPROVe study are provided in Table 9.*

Cell $\mathbb{C}$	**#evts/size**	**CATE Est.** $\widehat{\tau}_{\mathbb{C}\cap\mathbf{S}}$ **(std)**	t-statistic $\mathbb{T}_{\mathbb{C}\cap\mathbf{S}}$
GI Event with $\mathbf{S} =$ *all data*			
PPH = 1	6/184	0.066 (0.026)	2.012
PSTRDS = 1, HYPGRP = 1	0/30	-	-
PSTRDS = 1, ELDERLY = 1	0/21	-	-
All	33/2587	0.016 (0.004)	-
CVT Event with $\mathbf{S} =$ *all data*			
ASPFDA = 1	13/151	0.107 (0.043)	2.128
MALE = 1, ELDERLY = 1	30/416	0.069 (0.025)	2.251
ASCGRP = 1	17/250	0.068 (0.031)	1.664
Union (of 3 cells above)	41/588	0.065 (0.021)	2.650
PPH = 1	4/184	0.022 (0.022)	0.119
All	89/2587	0.020 (0.007)	-

available for APPROVe. However, with the data available to us, we were able to impute the GI outcome and the missing relevant features (used for the cells reported in Tables 4 and 5). The data cleaning and imputation were done before looking at the final results. The details for this data cleaning are provided in Appendix C1, and the distribution of the selected features and the two outcomes is reported in Table B1 GI outcome, the top three. Once we have the features and the outcomes, we compute the subgroup CATE (3b) and t-statistics (11a) and report the results in Table 6.

7.2 Results with the APPROVe Study

Before presenting the quantitative results, we make a few remarks. In direct analogy with the problems with external validity mentioned earlier, there are several ways in which the causal estimands in APPROVe differ from those in VIGOR. First, the 'control' arm of both studies was of entirely different nature: while VIGOR was a comparison between Vioxx and naproxen, APPROVe compared Vioxx with a placebo. Second, the lengths of both studies were different, which is important because our estimands are defined in terms of accumulated risk over the duration of the study. Patients in VIGOR were followed up for a median time of 9 months, whereas most patients in APPROVe were tracked for at least 4 years. Furthermore, while GI events were adjudicated in VIGOR, this was not the case for APPROVe. Lastly, the study populations are different. As elaborated earlier in Section 2, the VIGOR study comprised patients who were diagnosed with rheumatoid arthritis. On the other hand, APPROVe comprised patients with a recent history of colon polyps. Furthermore, unlike VIGOR, APPROVe excluded patients likely needing regular NSAID treatment but allowed for concomitant low-dose aspirin therapy.

Table 6 describes the quantitative results for the final subgroups (from Section 6.3) for the APPROVe study. For the reasons explained in the previous paragraph, we do not expect the subgroup CATE estimands to be the same across the two studies. However, comparing the results across Tables 4,5 and 6, it is reassuring that the subgroups we found for the VIGOR study continue to be meaningful for APPROVe in illustrating the heterogeneity of treatment effects. We now discuss the results first for the CVT outcome followed by those for the GI outcome, as the interpretation of the results for the latter is a bit more subtle.

7.2.1 Results for the CVT outcome

We note that the three subgroups {ASPFDA = 11}, {MALE = 1, ELDERLY = 1} and {ASCGRP = 1} and the union of these three subgroups all had subgroup CATEs that were much larger than the ATE, with *t*-statistics that were significant at the 0.05 level for a one-sided *z*-test, even after accounting for multiple testing (refer to end of this section for further discussions related to multiple testing.) Overall, these results provide evidence for the heterogeneous treatment effects of Vioxx for the CVT outcomes over these subgroups, namely, that Vioxx disproportionately increases the CVT event risk for these subgroups when compared with either naproxen or a placebo. To be consistent with the earlier results in Table 4, we also computed the subgroup CATE for {PPH = 1} for the CVT outcome and (like the VIGOR study) did not find any evidence for a disproportionate increase in the risk for the CVT event compared with the entire population.

Recall that the found increase in risk for VIGOR was relative to naproxen. This observation alone may suggest a possibility that Vioxx was not the cause of the observed increase in CVT events, and the positive ATE could have resulted owing to a protective effect of naproxen reducing them. Merck, the manufacturer of Vioxx, interpreted the CVT signal in VIGOR as being a consequence of a hitherto unknown protective effect of naproxen, rather than a deleterious consequence of Vioxx. The CVT signal in the APPROVe study associated with Vioxx relative to placebo conclusively confirmed that Vioxx can have deleterious consequences. Moreover, both VIGOR and APPROVe study suggest that Vioxx has significant heterogeneity in how it increases the risk for CVT events for different subgroups.

7.2.2 Results for the GI outcome

As noted above, additional care is required to interpret the CATE results for the GI outcome. Whereas naproxen was known to have GI toxicity and was shown in VIGOR to increase the risk of GI events more than Vioxx, a placebo by definition does not have any toxicity. As such, our finding that treatment with Vioxx had a positive estimated ATE (1.6%) with the GI outcome in the APPROVe study does not contradict our earlier reporting of a negative ATE with respect to the GI outcome (−1.6%) in the VIGOR study. In fact, this discovery is surprising insofar as Vioxx was initially believed to have minimal if any, GI toxicity whatsoever (Laine *et al*., 1999).

We found the subgroup {PPH = 1} to have a large positive estimated subgroup CATE (6.6%) resulting in a *t*-statistic score significant at the 0.025 level for a one-sided *z*-test (without correcting for multiple testing.) As discussed above, this result does not contradict the negative CATE value of −5.7% (or −5.5% for the test set) estimated for the VIGOR study (see Table 4). We furthermore note that the GI event rates over both arms in VIGOR and the Vioxx arm in APPROVe were all elevated compared with the entire population. The corresponding rates for the placebo in the APPROVe study were fairly similar (0% for {PPH = 1} and 0.4% on average.)

We summarise our finding across the two studies as follows. (i) VIGOR study: Vioxx, in comparison with naproxen, reduced the GI toxicity disproportionately for the subgroup {PPH = 1} when compared with the average. (ii) APPROVe study: Vioxx, in comparison with the placebo, increases the GI toxicity disproportionately for the subgroup {PPH = 1} when compared with the average. Nonetheless, the conclusion that the estimated subgroup CATE for {PPH = 1} was significantly different than the estimated ATE is *consistent* across the two studies.

Finally, owing to the difference in the study population, two out of the three subgroups for the GI event reported in Table 4, namely, {PSTRDS = 1, HYPGRP = 1} and {PSTRDS = 1, ELDERLY = 1}, were too small in size and had no GI events.[21] Consequently, it does not make sense to quantify the subgroup CATE for these subgroups.

7.2.3 Multiple testing with FWER control

Given enough data points in the APPROVe study, we also perform corrected multiple hypothesis testing using Holm–Bonferroni procedure controlling familywise error rate (FWER) at level 0.05. Overall, we test five null hypotheses that the subgroup CATE is equal to the average treatment effect for the following cases: (i) $\mathbb{C}_1 = \{\text{PPH} = 1\}$ for the GI event, (ii) $\widetilde{\mathbb{C}}_1 = \{\text{ASPFDA} = 1\}$, (iii) $\widetilde{\mathbb{C}}_2 = \{\text{MALE} = 1, \text{ELDERLY} = 1\}$, (iv) $\widetilde{\mathbb{C}}_3 = \{\text{ASCGRP} = 1\}$ and (v) the union $\cup_{i=1}^{3} \widetilde{\mathbb{C}}_i$—where the treatment effect in subgroups (ii)–(v) corresponds to the CVT event. The *t*-statistics for these hypotheses (sorted by magnitude) as reported in Table 6 are 2.650, 2.251, 2.128, 2.012 and 1.664, and thereby the corresponding one-sided *p*-values are 0.004, 0.012, 0.0167, 0.022 and 0.048. The corrected procedure for significance level 0.05 compares these sorted *p*-values with the cut-offs 0.01, 0.0125, 0.0167, 0.025 and 0.05. In doing so, we find that all five hypotheses are rejected, and thus we conclude all the subgroups (i)–(v) have statistically significant heterogeneous treatment effect.[22]

8 Discussion

In this work, we have made three major contributions: (i) We have re-analysed a dataset from the 1999–2000 VIGOR study, an RCT of 8076 patients, and found three clinically relevant subgroups each for the GI outcome (total size 29.4%), and the CVT outcome (total size 11.0%), for which the treatment drug Vioxx has significantly large estimated treatment effect when compared with that from the estimated ATE. We provided external evidence for the significance of the heterogeneous treatment effects for four out of the six subgroups through a complementary analysis of the 2001–2004 APPROVe study, another RCT of 2587 patients. (ii) Our work is an illustration of how clinical trial data can be analysed to provide a basis for differential treatment decisions in subgroups in order to optimise outcomes, and how the findings can be validated with another study. We call this novel methodology StaDISC, and we develop it by building on the PCS framework (Yu & Kumbier, 2020), the calibration literature and recent developments in CATE estimation. (iii) Our work introduces the PCS framework to the causal inference community and provides a template for a more informative understanding of heterogeneous treatment effects.

An important point to note is that the notions of estimated treatment effects ATE, CATE and subgroup CATE (defined in Equation (1)) used in this work and more broadly in CATE estimation measure the *difference* in the adverse event risk in the treatment group with that in the control group. However, when investigating the efficacy of medical interventions, medical professionals are often more interested in relative risk, which measures the *ratio* of the two risks. This alternate conception of treatment effect in terms of relative risk changes the meaning of heterogeneity. For instance, the subgroup $\mathbb{C}_1$ $\{\text{PPH} = 1\}$ has a relative risk of 0.43 with respect to GI events, which is barely any different than the population relative risk of 0.46. On the other hand, because the baseline risk of individuals in this subgroup is far higher than that of the rest of the population, the subgroup CATE is similarly inflated.

We do not attempt to debate which notion of heterogeneity is better since it is context dependent. Nevertheless, given the popularity of relative risk in the medical literature, in our future work, we plan to develop a formal framework for subgroup discovery with respect to relative risk by adapting generic CATE estimation methods and consequently extend StaDISC for relative risk estimation.

There are several other extensions of StaDISC that remain interesting future directions. First, StaDISC is currently motivated and defined for randomised experiments. We intend to

formulate a statistical framework that would also make it applicable to observational studies. Second, the cell search step of StaDISC only works with binary features. One can either propose to incorporate continuous features through either careful binary encoding using quantile-thresholding or through amending the cell search procedure. Third, we have thus far applied StaDISC to the GI and CVT outcomes in the VIGOR study one at a time and a joint investigation with multiple outcomes, even more generally, is an interesting future direction.

Acknowledgements

BY acknowledges the support from ARO Army Research Office grant W911NF-17-1-0005; Office of Naval Research Grant N00014-17-1-2176; the Center for Science of Information (CSoI), a US NSF Science and Technology Center, under grant agreement CCF-0939370; UCSF fund N7251, National Science Foundation grants NSF-DMS-1613002, 1953191, and IIS 1741340; and a research award from the Amazon Web Services (AWS Servers were used for running all our Jupyter Notebooks). BY is a Chan Zuckerberg Biohub investigator. The research was also partly supported by National Science Foundation grant NSF-CCF-1740855. The authors would like to thank Peng Ding, Sam Pimental, Chandan Singh, Tiffany Tang and our anonymous reviewer for their helpful comments.

Notes

[1]With important extensions also by Cox (1958).

[2]Budget constraints would dictate that they be only sufficiently powered to detect the ATE.

[3]More importantly, investigating subgroups in this manner is particularly sensitive to human failures. It opens the door to *p*-value hacking (Zettler, 2020), while Gelman has argued that even when researchers try to be honest, they nonetheless have a hard time accounting for 'researcher degrees of freedom' (Gelman & Loken, 2013).

[4]In fact, different methods and research groups sometimes reach different conclusions on the same datasets; see the paper of Carvalho *et al*. (2019) and the references therein.

[5]In Professor Efron's timely and thought-provoking revisiting (Efron, 2020) of the *Two Cultures* debate (Breiman, 2001), it is argued that contrasting philosophies on scientific truth is a clear line that separates traditional regression methods from modern machine learning methods (or pure prediction algorithms). While the former aims at an eternal scientific truth, the latter is truth-agnostic and instead content to exploit contingent and ephemeral patterns.

[6]Owing to the low signal in data, we decided not to split the data into training and validation sets, and we instead use four-fold cross-validation on the training data.

[7]However, the study was conducted with a safety monitoring board: an independent committee whose purpose is to monitor the results of an ongoing trial to ensure the safety of trial participants).

[8]This estimate and the other estimates reported in this paper are based on an intention-to-treat analysis. The study also performed per-protocol and sensitivity analyses and obtained similar results.

[9]https://www.cdc.gov/obesity/adult/defining.html, last accessed on 11 August 2020.

[10]Note that the standard variance estimates reported using this perspective can be taken as conservative estimates of the finite-sample variances defined in Neyman's repeated sampling framework (Ding *et al*., 2017).

[11]While R^2-score was originally introduced for linear regression, several similar measures have been proposed for providing an interpretable scale to measure the model fit. The R^2 for linear regression takes value in [0, 1] for training data and $(-\infty, 1]$ for test data. Close to 1 value suggests a good fit, and a smaller score implies a poor fit. Note that unlike the R^2 for

linear regression, for CATE estimators, the pseudo-score $\mathcal{R}^2_C$ is not guaranteed to take value in [0, 1] even on the training data, that is, $\mathcal{R}^2_C(\mathbf{S}_{TF}; \mathbf{M}) \in (-\infty, 1]$. Nonetheless, in Figure 2, we observe that for all the CATE estimators, this score lies in [0, 1] on the training folds, that is, $\mathcal{R}^2_C(\mathbf{S}_{TF}; \mathbf{M}) \in [0, 1]$

[12]This concern is similar to that expressed by Gelman in his influential paper on *The Garden of Forking Paths* (Gelman & Loken, 2013).

[13]In fact, such a time-based split would be even more relevant for studies based on RCTs that are *online* in nature, meaning that during the trial, results from earlier stages of the trial are used to guide whether the trial would be continued further or concluded.

[14]This term is motivated by the geometric interpretation of such subsets as subcubes of the hypercube that comprises the entire feature space.

[15]One may also think of this as a disjunction of conjunctions.

[16]For instance, leaf nodes will always involve the feature that splits the root node.

[17]The iterative RF (Basu *et al*., 2018) algorithm for finding higher-order interactions in genomics data does soft feature selection for precisely these reasons.

[18]We are able to compute this efficiently using the `FPGrowth` algorithm (Han *et al*., 2000).

[19]We say that Cell A is a sub-cell of Cell B if it is contained in Cell A when both are though as subsets of the feature space.

[20]A user may wish to simply follow the greedy procedure.

[21]Indeed, comparing Tables 1 and 9, we can attribute the discrepancy in these subgroups' sizes between the two studies to the smaller population of patients (74/2587) with a history of using glucocorticoids (PSTRDS = 1) in the APPROVe study versus that of the much larger population of such patients (4479/8076) in the VIGOR study.

[22]Note that for the APPROVe study, we did not test for heterogeneity in the subgroups {PSTRDS = 1, HYPGRP = 1}, and {PSTRDS = 1, ELDERLY = 1} owing to their small size in this study. As the size of the subgroup can only be observed once we know the group membership of the patients, our testing procedure and the associated discoveries can be considered as being conditional on observing the group membership, and treatment variable for all the patients in the APPROVe study. In other words, the statistical significance is over the randomness in the outcome, and the conditional randomness in the covariates given the group membership indicators.

References

Alaa, A. & Van Der Schaar, M. (2019). Validating causal inference models via influence functions, Proceedings of the 36th International Conference on Machine Learning, pp. 191–201.

Angrist, J.D. & Pischke, J.S. (2008). *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press: Princeton.

Athey, S. (2018). The impact of machine learning on economics. In *The Economics of Artificial Intelligence: An Agenda*, pp. 507–547. Chicago: University of Chicago Press.

Athey, S. & Imbens, G. (2016). Recursive partitioning for heterogeneous causal effects. *Proc. Natl. Acad. Sci. U. S. A.*, **113**(27), 7353–7360.

Ballarini, N.M., Rosenkranz, G.K., Jaki, T., König, F. & Posch, M. (2018). Subgroup identification in clinical trials via the predicted individual treatment effect. *PloS One*, **13**(10), e0205971.

Baron, J.A., Sandler, R.S., Bresalier, R.S., Lanas, A., Morton, D.G., Riddell, R., Iverson, E.R. & DeMets, D.L. (2008). Cardiovascular events associated with rofecoxib: final analysis of the APPROVe trial. *Lancet*, **372**(9651), 1756–1764.

Basu, S., Kumbier, K., Brown, J.B. & Yu, B. (2018). Iterative random forests to discover predictive and stable high-order interactions. *Proc. Natl. Acad. Sci.*, **115**(8), 1943–1948.

Bloniarz, A., Liu, H., Zhang, C.-H., Sekhon, J.S. & Yu, B. (2016). Lasso adjustments of treatment effect estimates in randomized experiments. *Proc. Natl. Acad. Sci.*, **113**(27), 7383–7390.

Bombardier, C., Laine, L., Reicin, A., Shapiro, D., Burgos-Vargas, R., Davis, B., Day, R., Ferraz, M.B., Hawkey, C.J., Hochberg, M.C., Kvien, T.K. & Schnitzer, T.J. (2000). Comparison of upper gastrointestinal toxicity of rofecoxib and naproxen in patients with rheumatoid arthritis. VIGOR study group. *N. Engl. J. Med.*, **343**(21), 1520–1528.
Breiman, L. (2001). Statistical modeling: the two cultures (with discussion). *Statist. Sci*, **16**(3), 16199–231.
Brier, G.W. (1950). Verification of forecasts expressed in terms of probability. *Mon. Weather Rev.*, **78**(1), 1–3.
Cai, T., Tian, L., Wong, P.H. & Wei, L.J. (2011). Analysis of randomized comparative clinical trial data for personalized treatment selections. *Biostatistics*, **12**(2), 270–282.
Carini, C., Menon, S.M. & Chang, M. (2014). *Clinical and Statistical Considerations in Personalized Medicine*. Boca Raton: CRC Press.
Carvalho, C., Feller, A., Murray, J., Woody, S. & Yeager, D. (2019). Assessing treatment effect variation in observational studies: results from a data challenge.
Chen, H., Harinen, T., Lee, J.-Y., Yung, M. & Zhao, Z. (2020). Causalml: Python package for causal machine learning.
Chernozhukov, V., Demirer, M., Duflo, E. & Fernandez-Val, I. (2018). Generic machine learning inference on heterogenous treatment effects in randomized experiments, National Bureau of Economic Research.
Collins, F.S. & Varmus, H. (2015). A new initiative on precision medicine. *N. Engl. J. Med.*, **372**(9), 793–795. PMID: 25635347.
Cox, D.R. (1958). Planning of experiments.
Dawid, A.P. (1982). The well-calibrated Bayesian. *J. Am. Stat. Assoc.*, **77**(379), 605–610.
DeGroot, M.H. & Fienberg, S.E. (1983). The comparison and evaluation of forecasters. *J. J. R. Stat. Soc. Ser. D (Statistic.)*, **32**(1-2), 12–22.
Debray, T.P., Vergouwe, Y., Koffijberg, H., Nieboer, D., Steyerberg, E.W. & Moons, K.G. (2015). A new framework to enhance the interpretation of external validation studies of clinical prediction models. *J. Clin. Epidemiol.*, **68**(3), 279–289.
Ding, P., Li, X. & Miratrix, L.W. (2017). Bridging finite and super population causal inference. *J. Causal Inf.*, **5**(2).
Dusseldorp, E. & Van Mechelen, I. (2014). Qualitative interaction trees: a tool to identify qualitative treatment–subgroup interactions. *Stat. Med.*, **33**(2), 219–237.
Efron, B. (2020). Prediction, estimation, and attribution. *J. Am. Stat. Assoc.*, **115**(530), 636–655.
Feller, A. & Holmes, C.C. (2009). *Beyond Toplines: Heterogeneous Treatment Effects in Randomized Experiments*. Unpublished manuscript, Oxford University.
Fisher, R.A. (1936). Design of experiments. *Br. Med. J.*, **1**(3923), 554–554.
Fortin, M., Dionne, J., Pinho, G., Gignac, J., Almirall, J. & Lapointe, L. (2006). Randomized controlled trials: do they have external validity for patients with multiple comorbidities? *Ann. Fam. Med.*, **4**(2), 104–108.
Foster, J.C., Taylor, J.M. & Ruberg, S.J. (2011). Subgroup identification from randomized clinical trial data. *Stat. Med.*, **30**(24), 2867–2880.
Gelman, A. & Loken, E. (2013). The garden of forking paths: why multiple comparisons can be a problem, even when there is no 'fishing expedition' or 'p-hacking'. Unpublished draft.
Gerber, A.S., Green, D.P. & Larimer, C.W. (2008). Social pressure and voter turnout: evidence from a large-scale field experiment. *Am. Polit. Sci. Rev.*, **102**, 33–48.
Guo, C., Pleiss, G., Sun, Y. & Weinberger, K.Q. (2017). On calibration of modern neural networks. In *34th International Conference on Machine Learning, ICML 2017*, Vol. **3**, pp. 2130–2143.
Han, J., Pei, J. & Yin, Y. (2000). Mining frequent patterns without candidate generation. *SIGMOD Rec.*, **29**(2), 1–12.
Hernández-Díaz, S. & Rodríguez, L.A.G. (2001). Steroids and risk of upper gastrointestinal complications. *Am. J. Epidemiol.*, **153**(11), 1089–1093.
Holland, P.W. (1986). Statistics and causal inference. *J. Am. Stat. Assoc.*, **81**(396), 945–960.
Huber, C., Benda, N. & Friede, T. (2019). A comparison of subgroup identification methods in clinical drug development: simulation study and regulatory considerations. *Pharm. Stat.*, **18**(5), 600–626.
Imai, K. & Ratkovic, M. (2013). Estimating treatment effect heterogeneity in randomized program evaluation. *Ann. Appl. Stat.*, **7**(1), 443–470.
Imbens, G.W. & Rubin, D.B. (2015). *Causal Inference in Statistics, Social, and Biomedical Sciences*. Cambridge: Cambridge University Press.
Imbens, G.W. & Wooldridge, J.M. (2009). Recent developments in the econometrics of program evaluation. *J. Econ. Lit.*, **47**(1), 5–86.
Jüni, P., Altman, D.G. & Egger, M. (2001). Assessing the quality of controlled clinical trials. *Bmj*, **323**(7303), 42–46.
Künzel, S.R., Sekhon, J.S., Bickel, P.J. & Yu, B. (2019). Metalearners for estimating heterogeneous treatment effects using machine learning. *Proc. Natl. Acad. Sci. U. S. A.*, **116**(10), 4156–4165.
Kennedy, E.H. (2020). Optimal doubly robust estimation of heterogeneous causal effects.
Krauss, A. (2018). Why all randomised controlled trials produce biased results. *Ann. Med.*, **50**(4), 312–322.

Krumholz, H.M., Ross, J.S., Presler, A.H. & Egilman, D.S. (2007). What have we learnt from Vioxx? *Bmj*, **334**(7585), 120–123.

Laine, L., Harper, S., Simon, T., Bath, R., Johanson, J., Schwartz, H., Stern, S., Quan, H., Bolognese, J. & Group, R.O.E.S. (1999). A randomized trial comparing the effect of rofecoxib, a cyclooxygenase 2-specific inhibitor, with that of ibuprofen on the gastroduodenal mucosa of patients with osteoarthritis. *Gastroenterology*, **117**(4), 776–783.

Lin, W. (2013). Agnostic notes on regression adjustments to experimental data: reexamining Freedman's critique. *Ann. Appl. Stat.*, **7**(1), 295–318.

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. *Stat. Med.*, **30**(21), 2601–2621.

Michel, R., Schnakenburg, I. & von Martens, T. (2019). *Targeting Uplift: An Introduction to Net Scores*. New York City: Springer International Publishing.

Miller, R.G. (1962). Statistical prediction by discriminant analysis. In *Statistical Prediction by Discriminant Analysis*, pp. 1–54. New York City: Springer.

Molina, M. & Garip, F. (2019). Machine learning for sociology. *Annu. Rev. Sociol.*, **45**(1), 27–45.

Murdoch, W.J., Singh, C., Kumbier, K., Abbasi-Asl, R. & Yu, B. (2019). Definitions, methods, and applications in interpretable machine learning. *Proc. Natl. Acad. Sci.*, **116**(44), 22071–22080.

Murphy, A.H. (1973). A new vector partition of the probability score. *J. Appl. Meteorol.*, **12**(4), 595–600.

Naeini, M.P., Cooper, G. & Hauskrecht, M. (2015). Obtaining well calibrated probabilities using Bayesian binning. *AAAI Conference on Artificial Intelligence*. pp. 2901-2907.

Negassa, A., Ciampi, A., Abrahamowicz, M., Shapiro, S. & Boivin, J.-F. (2005). Tree-structured subgroup analysis for censored survival data: Validation of computationally inexpensive model selection criteria. *Stat. Comput.*, **15**(3), 231–239.

Neyman, J. & Iwaszkiewicz, K. (1935). Statistical problems in agricultural experimentation. *Suppl. J. R. Stat. Soc.*, **2**(2), 107–180.

Niculescu-Mizil, A. & Caruana, R. (2005). Predicting good probabilities with supervised learning. In *Proceedings of the 22nd International Conference on Machine Learning*, pp. 625–632.

Nie, X. & Wager, S. (2017). Quasi-oracle estimation of heterogeneous treatment effects. arXiv preprint arXiv:1712.04912.

Norgeot, B., Quer, G., Beaulieu-Jones, B.K., Torkamani, A., Dias, R., Gianfrancesco, M., Arnaout, R., Kohane, I.S., Saria, S., Topol, E. & Obermeyer, Z. (2020). Minimum information about clinical artificial intelligence modeling: the MI-CLAIM checklist. *Nat. Med.*, **26**(9), 1320–1324.

Olson, R.S., Cava, W.L., Mustahsan, Z., Varik, A. & Moore, J.H. (2018). Data-driven advice for applying machine learning to bioinformatics problems. *Pacific Symposium on Biocomputing* , **23**, 192–203.

Ondra, T., Dmitrienko, A., Friede, T., Graf, A., Miller, F., Stallard, N. & Posch, M. (2016). Methods for identification and confirmation of targeted subgroups in clinical trials: a systematic review. *J. Biol. Stand.*, **26**(1), 99–119.

Peck, L.R. (2003). Subgroup analysis in social experiments: measuring program impacts based on post-treatment choice. *Am. J. Eval.*, **24**(2), 157–187.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M. & Duchesnay, E. (2011). Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.*, **12**, 2825–2830.

Popper, K.R. (1959). *The Logic of Scientific Discovery*. Basic Books: Oxford, England.

Rolling, C.A. & Yang, Y. (2014). Model selection for estimating treatment effects. *J. R. Stat. Soc. Ser. B (Stat Methodol.)*, **76**(4), 749–769. https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/rssb.12043

Ross, J.S., Madigan, D., Hill, K.P., Egilman, D.S., Wang, Y. & Krumholz, H.M. (2009). Pooled analysis of rofecoxib placebo-controlled clinical trial data: lessons for postmarket pharmaceutical safety surveillance. *Arch. Intern. Med.*, **169**(21), 1976–1985.

Rothwell, P.M. (2005). External validity of randomised controlled trials: "to whom do the results of this trial apply?". *Lancet*, **365**(9453), 82–93.

Rothwell, P.M. (2006). Factors that can affect the external validity of randomised controlled trials. *PLOS Clin Trial*, **1**(1), e9.

Rubin, D.B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *J. Educ. Psychol.*, **66**(5), 688.

Schuler, A., Baiocchi, M., Tibshirani, R. & Shah, N. (2018). A comparison of methods for model selection when estimating individual treatment effects.

Shahn, Z. & Madigan, D. (2017). Latent class mixture models of treatment effect heterogeneity. *Bayesian Anal.*, **12**(3), 831–854.

Splawa-Neyman, J., Dabrowska, D.M. & Speed, T.P. (1990). On the application of probability theory to agricultural experiments. Essay on principles. Section 9. *Statistical Science*, **5**, 465–472.
Su, X., Tsai, C.-L., Wang, H., Nickerson, D.M. & Li, B. (2009). Subgroup analysis via recursive partitioning. *J. Mach. Learn. Res.*, **10**(2), 141–158.
Tian, L., Alizadeh, A.A., Gelman, A. & Tibshirani, R. (2014). A simple method for estimating interactions between a treatment and a large number of covariates. *J. Am. Stat. Assoc.*, **109**(508), 1517–1532.
Wager, S. & Athey, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. *J. Am. Stat. Assoc.*, **113**(523), 1228–1242.
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. *N. Engl. J. Med.*, **357**(21), 2189–2194. https://doi.org/10.1056/NEJMsr077003, PMID: 18032770.
Yu, B. & Barter, R. (2020). The data science process: one culture. *J. Am. Stat. Assoc.*, **115**(530), 672–674. https://doi.org/10.1080/01621459.2020.1762615
Yu, B. & Kumbier, K. (2020). Veridical data science. *Proc. Natl. Acad. Sci.*, **117**(8), 3920–3929.
Zettler, P. (2020). U.S. v. Harkonen: should scientists worry about being prosecuted for how they interpret their research results?.

[*Received 20 August 2020, Revised September 2020, Accepted October 2020*]

Appendix A : Derivation of Variance Formula in *t*-statistic

In this section, we derive a formula for the variance of $\hat{\tau}_{\mathbf{G}} - \hat{\tau}_{\text{ATE}}$, thereby justifying the formula for the plug-in estimator used in the definition of the *t*-statistic, which we repeat here for convenience.

$$\mathbb{T}_{\mathbf{G}} := \frac{\hat{\tau}_{\mathbf{G}} - \hat{\tau}_{\text{ATE}}}{\sqrt{\hat{\text{Var}}(\hat{\tau}_{\mathbf{G}} - \hat{\tau}_{\text{ATE}})}}. \tag{A1}$$

We first group terms to get

$$\begin{aligned}\hat{\tau}_{\mathbf{G}} - \hat{\tau}_{\text{ATE}} &= \left(\frac{1}{|\mathbf{G}\cap\mathbf{T}|}\sum_{i\in\mathbf{G}\cap\mathbf{T}} Y_i(1) - \frac{1}{|\mathbf{G}\cap\mathbf{C}|}\sum_{i\in\mathbf{G}\cap\mathbf{C}} Y_i(0)\right) - \left(\frac{1}{|\mathbf{T}|}\sum_{i\in\mathbf{T}} Y_i(1) - \frac{1}{|\mathbf{C}|}\sum_{i\in\mathbf{C}} Y_i(0)\right)\\ &= \alpha_1 \sum_{i\in\mathbf{G}\cap\mathbf{T}} Y_i(1) + \alpha_0 \sum_{i\in\mathbf{G}\cap\mathbf{C}} Y_i(0) + \beta_1 \sum_{i\in\mathbf{G}^c\cap\mathbf{T}} Y_i(1) + \beta_0 \sum_{i\in\mathbf{G}^c\cap\mathbf{C}} Y_i(0),\end{aligned}$$

where

$$\alpha_1 = \left(\frac{1}{|\mathbf{G}\cap\mathbf{T}|} - \frac{1}{|\mathbf{T}|}\right),\ \alpha_0 = -\left(\frac{1}{|\mathbf{G}\cap\mathbf{C}|} - \frac{1}{|\mathbf{C}|}\right),\ \beta_1 = -\frac{1}{|\mathbf{T}|},\ \text{and}\ \beta_0 = \frac{1}{|\mathbf{C}|}.$$

Next, observe that even after we condition on $\mathcal{F}$, the collection of random variables $\{Y_i(1), Y_i(0) : 1 \le i \le N\}$ are fully independent, and furthermore, the terms within each sum are identically distributed. Applying the linearity of variance thus gives us

$$\begin{aligned}\text{Var}\left[\hat{\tau}_{\mathbf{G}} - \hat{\tau}_{\text{ATE}} \,|\, \mathcal{F}\right] = {} & \alpha_1^2\,|\mathbf{G}\cap\mathbf{T}|\cdot\text{Var}\left[Y(1) \mid \mathbf{G}\cap\mathbf{T}\right] + \alpha_0^2\,|\mathbf{G}\cap\mathbf{C}|\cdot\text{Var}\left[Y(0) \mid \mathbf{G}\cap\mathbf{C}\right]\\ & + \beta_1^2\,|\mathbf{G}^c\cap\mathbf{T}|\cdot\text{Var}\left[Y(1) \mid \mathbf{G}^c\cap\mathbf{T}\right] + \beta_0^2\,|\mathbf{G}^c\cap\mathbf{C}|\cdot\text{Var}\left[Y(0) \mid \mathbf{G}^c\cap\mathbf{C}\right],\end{aligned}$$

where $\text{Var}\left[Y(1) \mid \mathbf{G}\cap\mathbf{T}\right]$ denotes the variance of $Y(1)$ when conditioned on $X \in \mathbf{G}$ (recall our abuse of notation described in Section 3) and $T = 1$, with the other terms defined similarly.

Simplifying this formula leads to

$$\begin{aligned}\mathrm{Var}\left[\hat{\tau}_{\mathbf{G}} - \hat{\tau}_{\mathrm{ATE}} \,|\, \mathcal{F}\right] &= \left(1 - \frac{|\mathbf{G}\cap\mathbf{C}|}{|\mathbf{C}|}\right)^2 \cdot \frac{\mathrm{Var}\left[Y(0) \mid \mathbf{G}\cap\mathbf{C}\right]}{|\mathbf{G}\cap\mathbf{C}|} + \left(1 - \frac{|\mathbf{G}\cap\mathbf{T}|}{|\mathbf{T}|}\right)^2 \cdot \frac{\mathrm{Var}\left[Y(1) \mid \mathbf{G}\cap\mathbf{T}\right]}{|\mathbf{G}\cap\mathbf{T}|} \\ &+ \left(\frac{|\mathbf{G}^c\cap\mathbf{C}|}{|\mathbf{C}|}\right)^2 \cdot \frac{\mathrm{Var}\left[Y(0) \mid \mathbf{G}^c\cap\mathbf{C}\right]}{|\mathbf{G}^c\cap\mathbf{C}|} + \left(\frac{|\mathbf{G}^c\cap\mathbf{T}|}{|\mathbf{T}|}\right)^2 \cdot \frac{\mathrm{Var}\left[Y(1) \mid \mathbf{G}^c\cap\mathbf{T}\right]}{|\mathbf{G}^c\cap\mathbf{T}|} \\ &= \left(\frac{|\mathbf{G}^c\cap\mathbf{C}|}{|\mathbf{C}|}\right)^2 \cdot \left(\frac{\mathrm{Var}\left[Y(0) \mid \mathbf{G}\cap\mathbf{C}\right]}{|\mathbf{G}\cap\mathbf{C}|} + \frac{\mathrm{Var}\left[Y(0) \mid \mathbf{G}^c\cap\mathbf{C}\right]}{|\mathbf{G}^c\cap\mathbf{C}|}\right) \\ &+ \left(\frac{|\mathbf{G}^c\cap\mathbf{T}|}{|\mathbf{T}|}\right)^2 \cdot \left(\frac{\mathrm{Var}\left[Y(1) \mid \mathbf{G}\cap\mathbf{T}\right]}{|\mathbf{G}\cap\mathbf{T}|} + \frac{\mathrm{Var}\left[Y(1) \mid \mathbf{G}^c\cap\mathbf{T}\right]}{|\mathbf{G}^c\cap\mathbf{T}|}\right).\end{aligned}$$

Appendix B : Details on Data Cleaning with VIGOR and APPROVe

Here, we collect additional details deferred from the main paper. First, we provide the details on how we identified the patients with prior history of GI event (PPH = 1) for the VIGOR study. Although this subgroup was analysed in the original study, the data files we had did not contain a membership indicator, nor were there specific constructions on how to construct this subgroup. We applied a similar procedure to determine the patients with PPH = 1 for the APPROVe study. Following that, we describe the steps we followed to impute the GI outcome as well as the features {ASPFDA, ASCGRP, HYPGRP, PSTRDS} for the APPROVe study. We also note that the other features, namely, MALE and ELDERLY, reported in Table 6, could be readily identified from the demographics dataset for the APPROVe study, where the ELDERLY feature uses normalised age as detailed in Section 2.2.

Table A1. *Estimator-wise values of the mean scores* $\overline{A}_{j,j+1}$ *(8a) for* $j = 1, 2, 3, 4$ *for both GI and CVT events,* $\overline{A}_{1,\min}$ *(8b) for the GI event, and* $\overline{A}_{5,\max}$ *(8c) for the CVT event, where the mean was taken over the 12 validation folds, four each from the three random CV splits {*`cv_orig`,`cv_0`,`cv_1`*}.*

Estimator **M**	$\overline{A}_{1,2}$	$\overline{A}_{2,3}$	$\overline{A}_{3,4}$	$\overline{A}_{4,5}$	$\overline{A}_{1,\min}$
`t_logistic`	**1.00**	0.67	**0.83**	0.25	**1.00**
`causal_forest_2`	**1.00**	0.50	**0.83**	0.17	**1.00**
`x_lasso`	**1.00**	0.50	0.58	0.67	**1.00**
`x_rf`	**1.00**	0.42	0.42	0.67	**1.00**
`t_lasso`	**1.00**	0.42	0.50	0.58	0.92
`x_logistic`	**1.00**	0.33	0.50	0.75	0.92
`s_xgb`	**1.00**	0.67	0.58	0.58	0.92
`r_lassolasso`	0.92	0.42	0.42	**0.92**	0.92
`r_rfrf`	0.92	0.50	0.42	0.50	0.92
`r_lassorf`	0.92	0.42	0.42	0.42	0.92
`causal_forest_1`	0.92	0.67	0.75	0.50	0.83
`x_xgb`	0.92	0.33	0.50	0.83	0.83
`t_xgb`	0.92	0.42	0.67	0.17	0.83
`t_rf`	0.92	**0.75**	0.50	0.33	0.83
`causal_tree_2`	0.92	**0.75**	0.25	0.42	0.75
`s_rf`	0.83	0.58	0.67	0.42	0.75
`causal_tree_1`	0.83	0.58	0.17	0.67	0.67

(a) GI event

Estimator **M**	$\overline{A}_{1,2}$	$\overline{A}_{2,3}$	$\overline{A}_{3,4}$	$\overline{A}_{4,5}$	$\overline{A}_{5,\max}$
`t_lasso`	0.33	0.42	0.42	**1.00**	**1.00**
`x_xgb`	0.33	0.50	0.58	0.92	0.92
`x_logistic`	0.50	0.50	0.42	0.92	0.92
`r_rfrf`	0.25	0.42	0.50	0.92	0.83
`s_rf`	0.42	0.42	0.42	0.92	0.83
`x_lasso`	0.50	0.33	0.50	0.83	0.75
`t_rf`	0.33	0.25	**0.67**	0.83	0.75
`x_rf`	0.50	0.33	0.58	0.83	0.75
`t_logistic`	0.33	0.25	0.58	0.83	0.75
`r_lassorf`	0.17	0.42	0.42	0.92	0.75
`causal_forest_1`	**0.67**	0.33	**0.67**	0.92	0.75
`causal_forest_2`	0.50	0.08	0.33	0.92	0.75
`r_lassolasso`	0.17	**0.75**	0.50	0.75	0.67
`causal_tree_2`	0.25	0.08	0.33	0.83	0.25
`t_xgb`	0.08	0.08	0.25	0.75	0.08

(b) CVT event

In each column, the maximum score is highlighted in bold. The estimators are listed in the order sorted by the value in last column. Recall that each column was plotted earlier as a boxplot in Figure 4(a).

Table A2. `Stab`($\mathbb{C}$)*-scores (in % rounded to nearest integer) for the top 20 cells* $\mathbb{C}$ *found by* `CellSearch`*-methodology for quantile-based top subgroups* $\boldsymbol{G}_{top}$ *of the ensemble CATE estimator. The cells are sorted by the 'Mean' column of* `Stab`($\mathbb{C}$)*-scores, which in turn denote the average of the scores in second and third columns. For each score column, cells corresponding to top 3 scores are displayed in bold. The choices* $\mathfrak{q} = 0.2, 0.3$ *for the GI event in (a), and* $\mathfrak{q} = 0.8, 0.9$ *for the CVT event in (b) were made based on the results reported in Table 3 and the discussion around it.*

	`Stab`($\mathbb{C}$)-score in % with $\mathbf{G}_{\text{top}} = \widetilde{\mathbf{G}}_{\mathfrak{q}}$				`Stab`($\mathbb{C}$)-score in % with $\mathbf{G}_{\text{top}} = \widetilde{\mathbf{G}}^{c}_{\mathfrak{q}}$		
Cell $\mathbb{C}$ for GI event	$\mathfrak{q} = 0.2$	$\mathfrak{q} = 0.3$	**Mean**	Cell $\mathbb{C}$ for CVT event	$\mathfrak{q} = 0.9$	$\mathfrak{q} = 0.8$	**Mean**
{PPH = 1}	**92**	**92**	**92**	{ASPFDA = 1}	**82**	50	**66**
{PSTRDS = 1, HYPGRP = 1}	**36**	**54**	**45**	{MALE = 1, ELDERLY = 1}	**70**	**57**	**64**
{PSTRDS = 1, ELDERLY = 1}	**37**	**48**	**42**	{ASCGRP = 1}	**32**	**54**	**43**
{PNAPRXN = 0, PSTRDS = 1, ELDERLY = 1}	23	18	21	{MALE = 1}	0	**62**	31
{PNAPRXN = 0, HYPGRP = 1, PSTRDS = 1}	25	8	17	{ELDERLY = 1, SMOKE = 1}	22	27	25
{PSTRDS = 1, PNSAIDS = 0}	8	23	15	{MALE = 1, ELDERLY = 1, US = 1}	30	0	15
{WHITE = 0, PSTRDS = 1, ELDERLY = 1}	18	3	11	{MALE = 1, US = 1}	0	26	13
{CHLGRP = 1, HYPGRP = 1}	17	2	10	{OBESE = 1, ELDERLY = 1}	0	21	10
{OBESE = 1, WHITE = 0, PSTRDS = 1}	10	8	9	{MALE = 1, WHITE = 1, ELDERLY = 1}	20	0	10
{PNAPRXN = 0, ELDERLY = 1}	0	18	9	{MALE = 1, ASCGRP = 1}	18	0	9
{OBESE = 1, WHITE = 0}	0	17	8	{WHITE = 1, OBESE = 1, ELDERLY = 1}	0	15	8
{HYPGRP = 1, PNSAIDS = 0}	16	0	8	{MALE = 1, PPH = 0, ELDERLY = 1}	13	0	7
{WHITE = 0, PNSAIDS = 0}	14	0	7	{MALE = 1, WHITE = 1}	0	12	6
{OBESE = 1, WHITE = 0, PNAPRXN = 0}	3	10	7	{PPH = 0, US = 1, ASCGRP = 1}	2	8	5
{OBESE = 1, PSTRDS = 1, HYPGRP = 1}	5	8	7	{WHITE = 1, ELDERLY = 1, SMOKE = 1}	7	3	5
{PSTRDS = 1, HYPGRP = 1, ELDERLY = 1}	12	0	6	{ELDERLY = 1, US = 1, SMOKE = 1}	7	3	5
{WHITE = 0, PSTRDS = 1, PNSAIDS = 0}	10	2	6	{MALE = 1, PPH = 0}	0	9	4
{CHLGRP = 1}	0	11	6	{ELDERLY = 1, US = 1, CHLGRP = 1}	0	8	4
{PNAPRXN = 0, HYPGRP = 1}	0	10	5	{CHLGRP = 1, ASCGRP = 1}	8	0	4
{OBESE = 1, PNSAIDS = 0}	4	6	5	{MALE = 1, ELDERLY = 1, SMOKE = 1}	7	0	3
(a) GI event				(b) CVT event			

B.1 Defining PPH for both studies

To identify patients with a history of GI events, we identified a list of medical terms associated with such events, namely, gastroduodenal perforation, obstruction, ulcer or upper GI bleeding, from the medical history file. (We used PREFTERM field for this part, as PREFTERM was not available in the medical history file for the VIGOR dataset.) Using this procedure, we identified 313 patients in the control arm and 317 patients in the treatment arm who had a prior history of GI events (identified as PPH = 1). These number are off by 1 when compared with the 314 and 316 patients reported with PPH = 1 for the control and treatment arms, respectively, by Bombardier *et al*. in their paper on the VIGOR study (Bombardier *et al*., 2000).

To identify the patients with PPH = 1 for the APPROVe study, we used the medical terms identified above, with some adjustment for different spellings. Doing this gives us a subgroup of 184 patients. For this dataset, we used the PREFTERM in the medical history file for identification (since PREFTERM uses standardised terminology). Note that the paper on APPROVe study by Baron *et al*. (2008) does not report any information about the PPH feature.

B.2 Defining GI outcome and other features for APPROVe study

On the VIGOR dataset, we identified all possible medical terms (PREFTERM field in the adverse event file) that were relevant and possibly associated with GI events during the treatment period. To be consistent with our procedure on VIGOR dataset, we excluded pre-treatment events and included events that occurred during the treatment and post-study periods. A confirmed CVT event was a designated end point of the study, so these labels were directly provided to us in the study's data files. Such a process, that is, using only a relevant list of medical terms in the adverse event file, correctly identified 166 out of 177 patients with GI events. Despite our best efforts, this procedure also falsely identified 12 out of the remaining 7899 patients who did not have a confirmed GI event. Next, we found that 33 patients in APPROVe had recorded adverse events with PREFTERM contained in the list of terms identified above (with some adjustment of different spellings) during the treatment or the post-study periods. We declared these 33 patients to have had a GI event. Because the APPROVe study did not aim to study GI

Table B1. *Overview of the selected baseline covariates in the control and treatment arm of the APPROVe study. The treatment arm was given Vioxx, while the control arm was given placebo. PUB stands for perforations, ulcers and bleeding.* † *Adjusted age denotes age multiplied by the ratio of the life expectancy in the USA to that in the individual's country of residence.*

Covariate (ABBRV)	**Control no.** (%)	**Treatment no.** (%)
Overall population	1300 (50.3)	1287 (49.7)
Demographics		
Whether *gender* is male (MALE = 1)	805 (61.9)	804 (62.4)
Whether *adjusted age*† >65 (ELDERLY = 1)	338 (26.0)	329 (25.6)
Prior medical history		
Of *GI PUB events* (PPH = 1)	93 (7.2)	91 (7.1)
Of *hypertension* (HYPGRP = 1)	446 (34.3)	463 (36.0)
Of *atherosclerotic cardiovascular disease* (ASCGRP = 1)	121 (9.3)	129 (10.0)
indicating use of *aspirin* under FDA guidelines (ASPFDA = 1)	70 (5.4)	81 (6.3)
Prior usage of drugs		
Whether used *glucocorticoids/steroids* (PSTRDS = 1)	40 (3.1)	34 (2.6)
Outcomes		
Whether *GI event* occurred (GI = 1)	6 (0.46)	27 (2.1)
Whether *CVT event* occurred (CVT = 1)	32 (2.5)	57 (4.4)

toxicity, the paper on the study by Baron *et al*. (2008) does not report any information about the GI event as well as the risk factor features that we discuss next.

We followed a similar strategy to develop a mapping using the medical terms from the medical history file to the risk factor indicators for {ASPFDA, ASCGRP, HYPGRP} for the VIGOR study. Doing so, we correctly identified (i) 320/321 patients with ASPFDA = 1 (indication of aspirin usage by FDA due to their medical history), (ii) 453/454 patients with ASCGRP = 1 (history of atherosclerosis) and (iii) all 2385/2385 patients with HYPGRP = 1 (history of hypertension). For all three features ({ASPFD, ASCGRP, HYPGRP}), we did not have any false inclusion, that is, using just the selected list of medical terms did not incorrectly impute a value of 1 for any patient. Finally, to identify the patients with prior usage of glucocorticoids (PSTRDS = 1), we developed a mapping between the information from the concomitant therapy file and the PSTRDS indicator from the risk factor file. Our mapping correctly identified all 4479/4479 patients with PSTRDS = 1 but also falsely identified an additional 248 (out of the remaining 3597) patients.

The mappings described above were then used to impute the GI outcome and relevant missing features in the APPROVe study, thereby allowing us to report the 'transfer' results for the subgroups found by StaDISC on the VIGOR study (Tables 4 and 5) to the APPROVe study (Table 6).