DATA INTEGRATION APPROACHES TO ESTIMATE HETEROGENEOUS TREATMENT EFFECTS

by

Carly Lupton Brantner

A dissertation submitted to Johns Hopkins University

in conformity with the requirements for the degree of

Doctor of Philosophy

Baltimore, Maryland

April, 2024

Abstract

Clinicians and practitioners are often motivated to determine which treatment would work best for a given individual based on their observed characteristics, but doing so can be challenging because sample sizes are typically not large enough, and the variables involved in the true treatment effect heterogeneity are often unknown. To better understand treatment effect heterogeneity, researchers can rely on combining information from multiple sources, e.g., multiple randomized controlled trials (RCTs), or RCTs in conjunction with observational datasets. However, combining data requires taking into account that the data comes from heterogeneous sources, and different sources might have different settings, potential biases, and site-level characteristics that can impact treatment effects. This dissertation discusses approaches for integrating multiple datasets to estimate heterogeneous treatment effects. Previous approaches are outlined, and new methods are developed and introduced to estimate the conditional average treatment effect function across multiple trials and in a target population. The methods used are primarily non-parametric but compared to parametric meta-analysis. Methods are applied to real data comparing treatments for major depression to investigate potential heterogeneity of the treatment effect in this setting.

Thesis Committee

Primary Readers

Elizabeth A. Stuart (Primary Advisor)
Professor
Department of Biostatistics
Johns Hopkins Bloomberg School of Public Health

Karen Bandeen-Roche
Professor
Department of Biostatistics
Johns Hopkins Bloomberg School of Public Health

Peter Zandi
Professor
Department of Psychiatry and Behavioral Sciences
Johns Hopkins School of Medicine

Catherine Lesko
Associate Professor
Department of Epidemiology
Johns Hopkins Bloomberg School of Public Health

David Roth (Non-voting member)
Professor
Department of Geriatric Medicine and Gerontology
Johns Hopkins School of Medicine

Alternate Readers

Marie Diener-West
Professor
Department of Biostatistics
Johns Hopkins Bloomberg School of Public Health

Trang Nguyen
Research Associate Professor
Department of Mental Health
Johns Hopkins Bloomberg School of Public Health

Acknowledgments

First, I want to thank my advisor, Liz Stuart, who truly made my experience what it was. Liz, I am so grateful that you responded to my email as an undergraduate and gave me my first exposure to biostatistical research; I am so lucky to have found you and your research lab so early on in my career. You have been an instrumental part of shaping my view on research and how I strive to emphasize effective communication, interdisciplinary teamwork, and mentorship. Thank you for giving me independence in my research but guidance when I needed it, for showing me patience and helping me learn how to problem solve when I was getting unexpected results or stuck on an issue. Thank you for providing a research environment where I was not afraid to make mistakes or ask questions, in both our one-on-one and lab group meetings. You are an incredible role model, and I can only hope to make an impact close to the one you have already made on biostatistics, public health, and beyond. I am so looking forward to our continued work together!

Thank you to all of the incredible individuals with whom I have collaborated throughout my time at Johns Hopkins. Thank you to the moderation grant team for all that I learned about working on grants, using electronic health record data, and maintaining a connection between the analysis and the

real world implications through discussions with practitioners and stakeholders. Thank you to Hwanhee Hong for your research and career mentorship – you are another incredible role model for me. Thank you to my amazing team of telehealth EHR researchers – Peter Zandi, Catherine Ettman, Jason Straub, Grace Ringlein, Elena Badillo Goicoechea, and others. Peter, thank you for teaching me so much about psychiatric care, for being a mentor in my dissertation work as well, and for your endless patience as we worked to define cohorts and deal with the challenges of EHR data. Catherine, thank you for modeling your ambition, organization, and leadership – I hope to emulate your drive in many ways in my future career. Thank you for valuing what I can bring to a project and for treating me as a partner in our work together.

To members of the Stuart lab, thank you for providing a safe space where we could discuss both research and career development. Thank you to Leon Di Stefano and Harsh Parikh for your amazing statistical knowledge that you were kind enough to share with me. Thank you to Nick Seewald for your generosity in sharing your brilliance in website creation, presentation design, communication, and job searching. And a huge thank you to Trang Nguyen – I am forever grateful for the time that we spent brainstorming, diving into papers together, and thinking through the bigger picture.

Thank you to the Epidemiology and Biostatistics of Aging training grant team for all that I learned about aging research. Thank you to Karen Bandeen-Roche and Brian Buta for creating and maintaining this flourishing group of researchers answering important questions for older adults. Thank you to David Roth for being much more than just my assigned mentor; I learned so

much from collaborating with you and John Bentley, and I appreciate your investment in my growth and career.

Thank you to all of the amazing individuals in the Department of Biostatistics for helping me grow and succeed as a new biostatistician. Thank you to Mary Joy Argo for making sure I always met my requirements and for showing me patience, kindness, and support. Thank you to Maria Beeson, Fallon Bachman, Nanette Bell, and many others for their help in organizing funding, planning events, and facilitating student engagement. Thank you to Hongkai Ji for leading the PhD program so well and for always being receptive to ideas and questions. Thank you to all of my professors throughout the PhD program; I learned so much from each of you, all while we navigated a global pandemic and virtual learning. And thank you to Marie Diener-West, Karen Bandeen-Roche, and Liz Stuart – I feel very grateful to have been able to serve as a teaching assistant in your classes, and you spurred on my love for teaching statistics and making an impact on students. Karen, thank you also for leading our department with such grace. From the moment I met you during my visit day, I knew you would be someone I could come to with questions and for advice, and that has been true to this day and beyond. We are so lucky as a department to have had you as a leader and I am lucky to have you as a mentor. Thank you again to Liz Stuart for your seamless transition in as our new department chair – I have loved being able to continue with you as my advisor and come to you with ideas for student events and ways to push our department forward.

Thank you to the entire Biostatistics student body for supporting me to

where I am today. Thank you to my cohort of brilliant biostatisticians, Gege Gui, Jennifer Xu, Yuzheng Dun, and Wentao Zhan. I will never forget our late night US-time, early morning China-time Zoom calls throughout first year. I am so glad we have now had our in-person time together, and I have cherished our memories cooking together, making gingerbread houses, and playing Spike Ball. Thank you to my other friends in the department, Claire Heffernan, Charlotte Clapham, Kinnary Shah, Martina Fu, Yi Wang, Emily Scott, Grant Schumock, Elizabeth Sarker, and others. Thank you for your patience with me when I did not make the long commute up to Baltimore and for being there when I did. Thank you especially to Claire for being my go-to with any and all questions and a person who I knew would always support me in and out of school. Thank you to those of you who were involved in the Biostatistics Student Organization with me and worked to plan events for students and make everyone feel valued in the department.

Thank you to all those outside of Hopkins who supported me throughout the past four years. Thank you to my friends from college and from DC for keeping my life outside of school fun and entertaining. Thank you to the wonderful people at the Kynisca Innovation Center and Mathematica for exciting external collaborations and for showing me the potential projects that this degree could lead to in the future.

Thank you so much to all of my committee members, past and present. I am grateful to have such a supportive group of incredible faculty providing guidance. Thank you to Liz Stuart, Karen Bandeen-Roche, Peter Zandi, Katie Lesko, and David Roth for providing feedback on this dissertation, and thank

you to Marie Diener-West and Trang Nguyen for being a part of my committee and for your guidance along the way. Thank you so much also to Scott Zeger for being a part of my committee previously and for being a continual mentor from me. I remember meeting you once as an undergraduate and knowing that your advice was always going to be very important for me, and that has absolutely been the case. You are a brilliant statistician and brilliant mentor, and I am lucky to have been able to be impacted by both.

Finally, thank you to my amazing family. Thank you to my parents and my sister for their endless love and support, and for countless visits to Baltimore and DC over the years. Thank you for instilling confidence in me that I could pursue this degree and make an impact in this field, and thank you for modeling what it means to work hard and help others. Thank you to my new second family whom I was so lucky to marry into during my time in this PhD. And thank you to my husband, Wes, for being there every step of the way. You have always supported me in achieving my goals and have made those goals more achievable by always providing perspective, showing love and kindness, and being an all-around amazing human being. I love you all.

Dedication

I would like to dedicate this dissertation to my parents, Helen and Craig Lupton-Smith, my sister, Laura Lupton-Smith, and my husband, Weston Brantner.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

In a clinical setting, an ultimate goal is for every patient to achieve the best outcome possible based on the treatments available to them. This goal expands beyond the clinical space and can be a priority in health policy, education, and other fields that involve intervention implementation and assessment. With a variety of potentially effective treatments available for conditions like depression (Sampogna et al., 2024; Fisher and Bosley, 2015), diabetes (Bertsimas et al., 2017), and cancer (Duffy and Crown, 2008), the question becomes determining which of the available treatments the patient should receive. To answer this question in the most accurate way, it is often necessary to estimate not just which treatment is the best on average in a given population, but which treatment is the best based on the patient's characteristics. This field of work is sometimes labeled as precision medicine, where the goal is to make more personalized treatment decisions based on observable information about the patients.

In randomized controlled trials (RCTs), as well as observational studies, researchers often estimate this average effectiveness of a treatment; however,

less information is available regarding whether the treatment might be more beneficial for certain groups of patients rather than others. RCTs can sometimes be used to answer this question, but they are typically under-powered to detect effect moderation, or differences in the treatment effect across patient characteristics (Fleiss, 2011). The small sample sizes in single trials can also be problematic when effect moderation is unknown a priori or involves complex and non-linear relationships between characteristics (Yusuf et al., 1991). On the other hand, observational datasets are larger and might better represent the target population for which decisions are being made, but treatment assignment is not random in those datasets, so estimating the treatment effect is less straightforward.

A growing field of literature is focusing on data integration (also called data fusion), where individual-level data from multiple sources can be combined to provide more information about an estimand of interest. Data integration can potentially help with precision medicine, as bringing in more data from different sources can aid in investigating heterogeneity of the treatment effect across patient characteristics. However, integrating data from different sources brings up new challenges, including heterogeneity across studies in measures, exposures, and outcomes.

This dissertation explores data integration approaches to estimate treatment effect heterogeneity. The methods discussed are agnostic to application area, but the data explored in the following chapters focus on treatments for major depressive disorder (MDD) and how they affect depression severity. Across all chapters, the target estimand is the conditional average treatment

effect (CATE). The CATE is defined here as the expected difference in potential outcomes under treatment versus control, conditional on observed characteristics. In other words, the CATE allows for the treatment effect to vary depending on values of patient covariates. This estimand is common when the goal is estimating these heterogeneous treatment effects and using those to inform treatment decisions.

Chapter 2 provides an extensive overview of data integration methods, specifically focusing on estimating heterogeneous treatment effects. The approaches are broken down based on the sources being combined and the level of data access, which can include aggregate-level data, federated learning (where individual participant-level data is available within studies but cannot be shared across sites), or individual participant-level data. The sources can also vary in terms of study type, including RCTs and observational data. Finally, the analysis approach could be parametric (i.e., meta-analysis) or non-parametric. In Chapter 2, we discuss approaches and reveal openings for further work in this field. We also introduce the conditional average treatment effect (CATE) estimand, and detail the relevant assumptions for this type of causal inference. This paper has been published in *Statistical Science* (Brantner et al., 2023).

In Chapter 3, we focus on integrating multiple RCTs to estimate heterogeneous treatment effects. We discuss the limitations of individual participant-level data (IPD) parametric meta-analysis, a common approach for combining trials but not commonly used to assess effect heterogeneity, and we explore the application of non-parametric machine learning methods instead. We develop

methods to aggregate information across trials and assess their performance in simulations, and we ultimately apply the methods to data comparing treatments for major depressive disorder, duloxetine and vortioxetine. We discuss the results after combining four RCTs and examine potential heterogeneity of the treatment effect across patient characteristics and across trials (Brantner et al., 2024).

Chapter 4 directly extends the methods explored in Chapter 3 to predict treatment effects in a new target population that we can examine through patient electronic health records (EHR) from a health care system. The goal of this chapter is to develop methods for obtaining clinically relevant results in a given population, based on previously conducted trials. Specifically, we utilize multiple RCTs to generate a model for the CATE as in Chapter 3, and we then define prediction intervals for the treatment effect conditional on observed covariates in the new target population, using both parametric and non-parametric methods. In an application to real data, we estimate the CATE in three of the same RCTs from Chapter 3 comparing duloxetine and vortioxetine for treatment of depression, and we subsequently estimate CATE prediction intervals in EHR data from patients in the Duke Health Care System who have major depression or other similar mood diagnoses.

Each chapter of this dissertation pushes the field further in data integration to estimate heterogeneous treatment effects, with a goal of moving towards relevance for clinical practice. We emphasize the challenges of interpreting results from non-parametric methods and explore options to do so more

effectively, and we focus on estimating uncertainty appropriately when integrating data and applying models to new populations. The final chapter of the dissertation (Chapter 5) discusses key takeaways and future directions.

References

Sampogna, Gaia, Claudia Toni, Pierluigi Catapano, Bianca Della Rocca, Matteo Di Vincenzo, Mario Luciano, and Andrea Fiorillo (2024). "New trends in personalized treatment of depression". In: *Current Opinion in Psychiatry* 37.1, pp. 3–8.

Fisher, Aaron J and Hannah G Bosley (2015). "Personalized assessment and treatment of depression". In: *Current Opinion in Psychology* 4, pp. 67–74.

Bertsimas, Dimitris, Nathan Kallus, Alexander M Weinstein, and Ying Daisy Zhuo (2017). "Personalized diabetes management using electronic medical records". In: *Diabetes care* 40.2, pp. 210–217.

Duffy, Michael J and John Crown (2008). "A personalized approach to cancer treatment: how biomarkers can help". In: *Clinical chemistry* 54.11, pp. 1770–1779.

Fleiss, Joseph L (2011). *Design and analysis of clinical experiments*. John Wiley & Sons.

Yusuf, Salim, Janet Wittes, Jeffrey Probstfield, and Herman A Tyroler (1991). "Analysis and interpretation of treatment effects in subgroups of patients in randomized clinical trials". In: *Jama* 266.1, pp. 93–98.

Brantner, Carly Lupton, Ting-Hsuan Chang, Trang Quynh Nguyen, Hwanhee Hong, Leon Di Stefano, and Elizabeth A. Stuart (2023). "Methods for integrating trials and non-experimental data to examine treatment effect heterogeneity". In: *Statistical Science* 38.4, pp. 640–654.

Brantner, Carly Lupton, Trang Quynh Nguyen, Tengjie Tang, Congwen Zhao, Hwanhee Hong, and Elizabeth A Stuart (2024). "Comparison of methods that combine multiple randomized trials to estimate heterogeneous treatment effects". In: *Statistics in Medicine*.

Chapter 2

Methods for Integrating Trials and Non-Experimental Data to Examine Treatment Effect Heterogeneity

Abstract:[1] Estimating treatment effects conditional on observed covariates can improve the ability to tailor treatments to particular individuals. Doing so effectively requires dealing with potential confounding, and also enough data to adequately estimate effect moderation. A recent influx of work has looked into estimating treatment effect heterogeneity using data from multiple randomized controlled trials and/or observational datasets. With many new methods available for assessing treatment effect heterogeneity using multiple studies, it is important to understand which methods are best used in which setting, how the methods compare to one another, and what needs to be done to continue progress in this field. This paper reviews these methods broken down by data setting: aggregate-level data, federated learning, and individual

[1]This chapter has undergone peer review and is published in *Statistical Science*: Brantner, C. L., Chang, T., Nguyen, T. Q., Hong, H., Di Stefano, L., and Stuart, E. A. (2023). Methods for integrating trials and non-experimental data to examine treatment effect heterogeneity. *Statistical Science, 38*(4), 640–654.

participant-level data. We define the conditional average treatment effect and discuss differences between parametric and nonparametric estimators, and we list key assumptions, both those that are required within a single study and those that are necessary for data combination. After describing existing approaches, we compare and contrast them and reveal open areas for future research. This review demonstrates that there are many possible approaches for estimating treatment effect heterogeneity through the combination of datasets, but that there is substantial work to be done to compare these methods through case studies and simulations, extend them to different settings, and refine them to account for various challenges present in real data.

2.1 Introduction

Identifying the right treatment for the right patient can improve quality of healthcare for individuals and populations. Treatments for disorders and diseases like depression (Trivedi et al., 2006), schizophrenia (Samara et al., 2019), and diabetes (Xie, Chan, and Ma, 2018) can exhibit differential treatment effects across individuals due to *effect moderators*, defined as known and unknown individual, genetic, environmental, and other characteristics that are associated with the effectiveness of medical treatments (Baron and Kenny, 1986). Finding ways to identify and leverage effect moderators at the point of care to facilitate clinical decision-making can improve efficiency, quality and outcomes of healthcare.

Although crucial for delivery of treatment and preventative medicine, detecting treatment effect heterogeneity is challenging with common study

designs. Randomized trials yield comparable treatment groups on average but are typically under-powered to detect moderation. One rule-of-thumb is that study samples need to be four times larger to test an effect moderator than to detect the overall average effect (Enderlein, 1988). In addition, randomized trial samples are also often not representative of the target population for which treatment decisions will be made; for instance, Black individuals are on the whole underrepresented in pivotal clinical trials (Green et al., 2022). Therefore, conclusions from one particular trial might not reflect conclusions for a target population, and different trials might give conflicting results due to differences in their enrolled participants. On the other hand, large-scale non-experimental studies can have improved external validity, but these studies can suffer from confounding bias. Given power concerns in single randomized trials and bias concerns in non-randomized studies, much can be gained by combining multiple trials, or combining experimental and non-experimental studies, to examine effect moderation (Berlin et al., 2002; Brown et al., 2013).

Many methods have been proposed to examine effect moderation in a single study. One of the popular approaches is to prespecify a few key subgroups and fit models with treatment-subgroup interactions. This approach is limited in that data analysts could explore a range of possible subgroups and report only those that are statistically significant (Kent et al., 2010); additionally, this approach does not allow the contribution of multivariate factors in effect moderation. Another approach is "risk modeling" (Kent et al., 2010; Kent et al., 2020), where a risk score is created using the covariates to predict the outcome (usually outcome under the comparison/control condition), and the

treatment effect is assessed based on the interaction between treatment and this risk score in a regression model of the outcome. This review focuses on what is sometimes called "effect modeling". Effect modeling spans a spectrum that includes parametric approaches in which a few effect moderators are pre-specified, and nonparametric approaches where effect moderation is assumed to be via some potentially complex function of a large set of covariates. Regression analyses and variable selection are common approaches for the former; machine learning methods for the latter.

In order to examine treatment effect heterogeneity based on observed characteristics, the target estimand in the present work is the conditional average treatment effect (CATE). Notation for this estimand is presented in the following section. The CATE is a general function of covariates that could be quite complex and so requires large sample sizes to estimate reliably. A key assumption when combining studies to estimate the conditional average treatment effect is that the CATE function is substantially similar across studies. When discussing the CATE, it is relevant to note that the CATE function is related to subgroup average treatment effects and identification of groups who benefit from treatment; these similar goals are mostly outside of the scope of this review. We therefore focus on the CATE and mention subgroup treatment effects and other similar topics briefly when relevant.

There have been recent statistical advances in modeling heterogeneous treatment effects and a separate burgeoning interest in combining data from multiple sources. A select few works have done both – simultaneously leveraging data from multiple studies to assess treatment effect heterogeneity.

Methods like these are needed to best harness the available data to optimize and individualize treatments, and to leverage information from multiple studies to provide more systematic, comprehensive, and generalizable conclusions. This paper reviews these novel methods of assessing treatment effect heterogeneity using multiple studies in the form of multiple randomized trials, or one randomized trial with a large observational dataset. We focus on methods identifying which of two treatments is more likely to improve outcomes for an individual or subgroup – a causal question that sits at the core of clinical practice. In this review, we consider the situation where the variables are similarly defined and available from all studies. It is common though that different studies may have different sets of variables. In this more complicated case, either harmonization is needed on the variables or some shared structure is required on conceptually related variables. We will return to this point in the Discussion section (2.6).

Methods discussed in this paper are broken down based on data setting: aggregate-level data, federated learning, and individual participant-level data (IPD). The aggregate-level data setting occurs when researchers only have access to summary information from each study. With aggregate-level data, individual-level effect heterogeneity can only be truly assessed if each study estimated treatment-covariate interactions using the same statistical models (e.g., same link function, same set of covariates), which is not often feasible. In the federated learning setting, sensitive individual-level data are distributed across decentralized studies and cannot be shared beyond their original storage location (Vo et al., 2021). Finally, the IPD setting is the

most straightforward and powerful scenario for assessing treatment effect heterogeneity, as individual-level covariates are available from all studies simultaneously. With IPD, we can harmonize covariates, estimate effect moderation by using the same statistical models in each study, and assess model assumptions consistently.

Within each of these data settings, methods are primarily geared towards either combining multiple RCTs or one RCT with one observational dataset. We discuss the use of meta-analysis models with multiple RCTs (Debray et al., 2015; Burke, Ensor, and Riley, 2017), along with the opportunity to employ variable selection approaches to identify effect moderators (Seo et al., 2021). When combining an RCT with observational data, we consider various methods that allow for complicated relationships to be included in the treatment effect function and account for potential bias from the observational data. These methods can involve estimating the CATE in the RCT and observational data separately and then combining them through an estimated weighting factor (Rosenman et al., 2022; Rosenman et al., 2020; Cheng and Cai, 2021; Yang, Zeng, and Wang, 2020), or estimating the observational CATE and the confounding effect in the observational dataset (Kallus, Puli, and Shalit, 2018; Yang, Zeng, and Wang, 2020; Wu and Yang, 2021; Hatt et al., 2022). Colnet et al., 2021b reviewed some methods that combine RCT and observational data, and we extend upon this review by focusing on this combination explicitly for treatment effect heterogeneity. We also add in more methods that combine RCT with observational data along with methods that focus on combining multiple RCTs. In general, there are many approaches outside of those we

reference here that focus on estimating the *average* treatment effect by combining datasets, some of which are discussed by Colnet et al. (2021); we choose to primarily focus on efforts to examine treatment effect heterogeneity in the present review.

To provide context to the methods discussed in this review, we can consider a few example scenarios. We first consider an assessment of the efficacy of surgery in stage IV breast cancer according to 15 studies where researchers combining the studies only had access to aggregate-level data (Petrelli and Barni, 2012). We also discuss a comparison of outcomes for veterans who received the Moderna versus the Pfizer vaccination for COVID-19 in five different sites where IPD was available within each site but could not be shared across sites, known as a "federated learning" situation (Han et al., 2021). Another setting investigates a diabetes medication, pioglitazone, versus placebo for individuals coming from one of six RCTs, where IPD was available in each trial (Hong et al., 2015). And finally, we discuss data assessing the treatment effect comparing two active treatments for major depression, duloxetine and vortioxetine, wherein we have access to IPD from a combination of RCT data and electronic health records (EHR) from a hospital system (Brantner et al., 2024). These scenarios all could clearly benefit from combining data to examine heterogeneity in treatment effects, but they each require distinct considerations and statistical approaches to best integrate information. We will use these examples throughout the paper to ground the methods in specific applications.

Importantly, to effectively combine information from multiple datasets, the

original studies need to have high transparency and reproducibility. Whether data are reported in aggregate or at the individual participant level, researchers using the data for additional analyses – such as those discussed here – need extensive information about how the data were collected, analyzed, and presented to be able to determine if and how to combine the information with other datasets. It is therefore vital to keep these ideas of transparency and reproducibility of data, code, and results at the forefront when applying these methods. Movements towards data sharing and reproducible research will greatly facilitate the types of research discussed here, which can lead to important new insights regarding effect heterogeneity that cannot be answered from single studies alone due to generalizability, sample size, or confounding concerns.

In the following section (2.2), we introduce the estimand and assumptions. The next sections are then organized based on the level of data access so that researchers can determine available methods in their given data setting. Specifically, Section 2.3 discusses aggregate-level data; Section 2.4, federated learning; and Section 2.5, individual participant-level data (IPD). Finally, Section 2.6 compares methods and provides an overview of potential future areas for research.

2.2 Notation

2.2.1 Target Estimand

Our target estimand to assess effect heterogeneity is the conditional average treatment effect (CATE), defined using the potential outcomes framework

under the Stable Unit Treatment Value assumption (Rubin, 1974). Suppose S is the categorical variable indicating study membership, $A = 0, 1$ is a binary treatment variable, Y is the observed outcome, $Y(1)$ and $Y(0)$ are the potential outcomes under treatment and control respectively, $\boldsymbol{X}$ is a set of covariates, and $\boldsymbol{Z}$ is a subset of $\boldsymbol{X}$ containing the proposed effect moderators.

The CATE can be formally defined as a function of $\boldsymbol{X}$:

$$\tau(\boldsymbol{X}) = \ g(E[Y(1)|\boldsymbol{X}]) - g(E[Y(0)|\boldsymbol{X}])$$

(Abrevaya, Hsu, and Lieli, 2015; Künzel et al., 2019), where $E[.|.]$ denotes conditional expectation in the target population of interest and $g(.)$ is a link function that defines the scale on which the interactions occur, whether additive (mean or risk difference) or multiplicative (risk, rate, or odds ratio). In this paper we primarily discuss a continuous outcome, in which case we use the identity link function and write the CATE as

$$\tau(\boldsymbol{X}) = \ E[Y(1) - Y(0)|\boldsymbol{X}]. \tag{2.1}$$

This $\tau(.)$ can often be assumed to be a flexible function in which all covariates are considered as potential moderators, so we do not have to a priori differentiate $\boldsymbol{Z}$ and $\boldsymbol{X}$ when methods allow for this flexibility.

One can also consider study-specific CATE functions. This is often the case when researchers are interested in assessing heterogeneity of the treatment effect functions across trials/datasets, or when this heterogeneity is high and it is potentially unreasonable to combine information across studies. We can denote study by S: in the case where data is being combined from one RCT and

one observational dataset, $S = 0$ will indicate RCT and $S = 1$ observational data; otherwise, S will be a categorical variable ranging from 1 to K, where K is the number of RCTs. The above Equation (2.1) defines a general CATE that is not study-specific. When estimating study-specific CATEs, Equation (2.1) can be rewritten as

$$\tau_s(\boldsymbol{X}) = E[Y(1) - Y(0)|\boldsymbol{X}, S = s]. \tag{2.2}$$

In most of the methods to follow, the CATE is defined by conditioning on a set of available covariates, $\boldsymbol{X}$. An alternative is to a priori define subgroups of interest and estimate subgroup-specific treatment effects. This approach is similar to the methods discussed in this review but somewhat distinct because subgroups must be specified first. The form of the estimand when examining subgroup-specific effect estimates is instead

$$\tau_k = E[Y(1) - Y(0)|K = k]$$

where K represents subgroup membership (Rosenman et al., 2020; Rosenman et al., 2022).

2.2.2 Assumptions

Across many methods, the key assumption that allows pooling data from multiple studies to estimate the treatment effect is that either entire or partial components of the treatment effect function $\tau(\boldsymbol{X})$ is shared across studies. This review also focuses solely on the case when there are only two treatments (or one treatment and one control/placebo) being compared. If there are more

than two conditions being compared, different approaches would need to be used (i.e., network meta-analysis; Efthimiou et al., 2016; Debray et al., 2018; Hong et al., 2015). Aside from these overarching assumptions, individual methods employ their own specific assumptions. When multiple RCTs are included in meta-analyses, they are often assumed to have similar eligibility criteria (specifically in terms of the covariates thought to be effect modifiers) (Dahabreh et al., 2020), and distributional assumptions are made for model parameters (Debray et al., 2015).

Broadly, parametric approaches require the assumption of a parametric relationship between covariates (including treatment, effect moderators, and interactions between the two) and outcomes; further, this parametric relationship is assumed to be approximately correctly specified (Debray et al., 2015; Yang, Zeng, and Wang, 2022; Yang, Zeng, and Wang, 2020). Specifically in the meta-analytic framework when combining multiple RCTs, effect moderation is often assessed using treatment-covariate interaction terms. This approach typically uses an outcome model of the form

$$h(E(Y)) = \mu(\boldsymbol{X}) + A \times \tau(\boldsymbol{Z}),$$

where $h(.)$ is a link function, $\mu(\boldsymbol{X})$ is the modelled mean of the outcomes under control, $\boldsymbol{Z}$ contains a subset of the variables in $\boldsymbol{X}$ that often needs to be pre-specified, and $\tau(\boldsymbol{Z})$ is the the CATE function:

$$\tau(\boldsymbol{Z}) = \delta + \boldsymbol{\theta}^T \boldsymbol{Z}. \tag{2.3}$$

In this expression for $\tau(\boldsymbol{Z})$, δ corresponds to the effect of treatment A when

$\boldsymbol{Z} = 0$ (or when the covariates in $\boldsymbol{Z}$ equal their means if they have been centered), and $\boldsymbol{\theta}$ corresponds to the coefficients of treatment-moderator interaction terms $A\boldsymbol{Z}$ in the $h(E(Y))$ model. Similarly to the general format of the CATE in Equation (2.1), this parametric form of $\tau(\boldsymbol{Z})$ can be expressed as multiple study-specific functions:

$$\tau_s(\boldsymbol{Z}) = \delta_s + \boldsymbol{\theta}_s^T \boldsymbol{Z}. \tag{2.4}$$

When combining an RCT with an observational dataset, there are a few within-study assumptions, including unconfoundedness (Assumption 2.1), positivity (Assumption 2.2), and consistency (Assumption 2.3) (Colnet et al., 2021b; Cheng and Cai, 2021):

Assumption 2.1 $\{Y(0), Y(1)\} \perp\!\!\!\perp A|\boldsymbol{X}$ *within each study.*

Assumption 2.2 *For almost all* $\boldsymbol{X}$ *with* $\pi(\boldsymbol{X}) = P(A = 1|\boldsymbol{X})$ *(the propensity score), there exists a constant* $c > 0$ *such that* $c < \pi(\boldsymbol{X}) < 1 - c$ *within each study.*

Assumption 2.3 $Y = AY(1) + (1 - A)Y(0)$ *almost surely.*

The unconfoundedness assumption (2.1) is satisfied by design in an RCT. Assumption 2.2 also holds by design in an RCT since the probability of treatment is independent of observed covariates and is pre-specified.

When combining datasets, we expand upon the previous assumptions. In the setting where observational data is being combined with an RCT, the unconfoundedness assumption (2.1) can be relaxed in the observational data. This is because there are analysis possibilities with multiple datasets that

include assessing whether this assumption is met or not and using the RCT to account for any confounding in the observational data (Cheng and Cai, 2021; Yang, Zeng, and Wang, 2020; Yang, Zeng, and Wang, 2022). Assumption 2.3 in the multi-study setting implies that the treatments being compared are the same across all studies (since there is no s subscript) to ensure that the potential outcomes $Y(0)$ and $Y(1)$ are well-defined. We also can introduce two other assumptions that are involved at some level in methods that combine an RCT with observational data; these assumptions include study membership positivity (Assumption 2.4) (Colnet et al., 2021b; Cheng and Cai, 2021) and unconfounded study membership (Assumption 2.5) (Hatt et al., 2022; Cheng and Cai, 2021; Kallus, Puli, and Shalit, 2018).

Assumption 2.4 *For almost all* $\boldsymbol{X}$ *there exists a constant* $d > 0$ *such that* $d < P(S = s|\boldsymbol{X} = \boldsymbol{x}) < 1 - d$.

Assumption 2.5 $\{Y(0), Y(1)\} \perp\!\!\!\perp S|\boldsymbol{X}$.

The following sections break down methods based on available data.

2.3 Aggregate-Level Data

The broadest level of data access is in the form of aggregate-level data (AD), where individual studies have been carried out and analyzed, and only summary data (e.g., sample mean, standard deviation, or regression model coefficient estimates) are available. AD are often used in meta-analyses when IPD are unavailable. Meta-analysis with AD can estimate average effects effectively and provide similar results as meta-analysis with IPD (Burke, Ensor,

and Riley, 2017; Hong et al., 2015). However, aggregation bias (also known as the ecological fallacy), which occurs when conclusions are incorrectly drawn about individuals when the relationship is found at the group level, can easily be introduced if researchers want to make a conclusion about individual-level effect moderation when only AD is available (Berlin et al., 2002; Debray et al., 2015; Teramukai et al., 2004). This aggregation bias will not be present if each paper reports subgroup-specific outcomes for all necessary subgroups; however, this is rare in practice because subgroups are often defined by more than one covariate. AD therefore has limited power for detecting effect moderation (Lambert et al., 2002). However, IPD is not always easy to access or use, so the following section discusses what can be done with AD. In framing this discussion, one can think of the example assessing the effects of tumor-removal surgery in individuals with breast cancer (Petrelli and Barni, 2012) using aggregate data from several relevant studies.

2.3.1 Meta-Analysis of Interaction Terms

If AD is all that is available for a question of interest, there is still an opportunity to estimate individual-level effect moderation under specific circumstances. If all previous studies have performed similar analyses and have included a particular treatment-covariate interaction term using the IPD from that given study, then these interaction terms can be pooled at the aggregate level (Simmonds and Higgins, 2007; Kovalchik, 2013). For instance, although this approach was not taken by Petrelli and Barni (2012), if a treatment-age interaction term was estimated in each of the individual studies assessing the

effect of surgery on mortality in individuals with stage IV breast cancer, then these interaction terms could be pooled together. In this way, researchers can estimate an individual-level effect moderation term across multiple studies and can combine such terms to estimate $\tau(\boldsymbol{Z})$ as in Equation (2.3). However, this requires that the studies assess and report the interactions of interest consistently. Similarly, the aggregate data could include subgroup-specific treatment effects rather than interactions, which could also be pooled to describe effect moderation if the effects are reported in each study (Godolphin et al., 2022).

2.3.2 Meta-Regression

If such study-specific interaction coefficients are not available across all studies, AD can be also modeled through meta-regression with treatment-covariate interaction terms, where importantly only aggregate level covariates (e.g., mean age, proportion female) are available. For example, the individual-level covariate of interest might be whether the person has severe disease or not; in an AD meta-regression, this covariate would become the percentage of individuals in the study who have severe disease. Meta-regression was the approach taken by Petrelli and Barni (2012) in their assessment of surgery efficacy. Specifically, they investigated hazard ratios of overall survival according to the fifteen different studies and did so while including covariates such as median age and mastectomy rate.

AD analyses can handle study-level effect moderators well. However, the ability to assess individual-level moderators depends on the level of detail

available in the AD. Multiple papers have assessed the differences between AD and IPD meta-regressions for estimating treatment effect heterogeneity. In an analysis by Berlin and colleagues (2002), models using IPD picked up on a key effect moderator that had been found in previous literature, but all models using AD missed this effect moderator at the group level. Extensive simulation studies also have shown that the power for detecting treatment effect moderation is much lower in meta-regression using AD; in these simulations, effect moderation was only effectively discovered in AD analyses when there were a large number of trials with large sample sizes (Lambert et al., 2002). Again, relationships that are picked up in an AD meta-regression cannot be immediately interpreted as individual-level effects; for example, if the percentage of individuals with severe disease is an effect moderator in the AD model, researchers cannot immediately conclude that the individual-level presence of severe disease is an effect moderator at the individual level.

Furthermore, the aggregate-level covariates also often do not vary much across studies. Since studies included in meta-regressions require similar eligibility criteria, they likely will have somewhat similar covariate distributions. For instance, the percentage of individuals with severe disease is likely to be similar across trials; in this case, the interpretation of effect moderation cannot be extrapolated beyond the aggregate-level range of the covariates.

The estimand in meta-regression can still be considered to be a version of the CATE, but it is the CATE according to group-level effect moderators; for example, it could be written like Equation (2.3) but as $\tau(\bar{\boldsymbol{Z}})$ where $\bar{\boldsymbol{Z}}$ consists

of aggregations of $\boldsymbol{Z}$ at the study-level. Such an estimand assumes that the included studies are representative of the target population of studies.

2.4 Federated Learning

Federated learning (similar to distributed modeling) uses a combination of IPD and AD; namely, IPD exists across decentralized studies but can only be accessed in the study in which it is stored (Yang et al., 2019). An example of this is a study of the efficacy of two COVID-19 vaccinations (developed by Moderna and Pfizer) for preventing COVID-19 in veterans in five Veterans Affairs sites (Han et al., 2021). This data setup is increasingly common in fields where there is interest in combining multiple cohorts ("cohort consortia"), but where data privacy concerns prohibit full direct data sharing. Therefore, the IPD data must be turned into AD or aggregated models so that information can be shared across studies.

We discuss two approaches for CATE estimation in federated learning in this section. Other approaches exist that focus on estimating the average treatment effect (ATE) (Han et al., 2021), and those can be extended to CATE estimation but must provide sufficient information about the parameters of effect moderation. Depending on the ATE approach, it is unclear how easily the method can be extended to CATE estimation; we focus instead on methods explicitly focused on CATE estimation.

2.4.1 Meta-Analysis after Local Model Formulation

There are three steps in meta-analysis within the federated learning setting: (1) fit models within studies, (2) aggregate the model coefficients, and then (3) conduct a meta-analysis (Silva et al., 2019). This is similar to the meta-analyses of interaction terms using aggregate data discussed in Section 2.3.1. A key difference here is that federated learning models apply a pre-determined statistical model including desired interaction terms so that the interaction effects are assessed consistently across all studies, while the traditional meta-analysis with AD has access to model coefficient estimates but not the model fitting process. Here, the estimand of interest is the common CATE function as in Equation (2.3) that is calculated by summarizing model coefficients corresponding to interaction terms $A\boldsymbol{Z}$ (treatment-moderator) and A (treatment) from each study-specific regression.

2.4.2 Tree-Based Ensemble

Another option within federated learning would be to still create study-specific models first, but to use information from other studies to improve those individual models. Tan, Chang, and Tang (2021) use tree-based ensemble methods to combine information about treatment effect heterogeneity from multiple separate studies. Specifically, they allow for study-level heterogeneity as well as heterogeneity due to individual-level covariates.

Their procedure involves first fitting models to estimate the CATE in each of K individual studies, using single-study machine learning methods like causal forests (Athey, Tibshirani, and Wager, 2019). These K study-specific

models are then applied to a single "coordinating study", so that each individual in the coordinating study has K estimates of the CATE. In other words, if there are n individuals in the coordinating study, there will be $n * K$ CATE estimates. Finally, these $n * K$ estimates are used as outcomes in an ensemble regression tree or random forest, in which the predictors are the individual-level covariates and an indicator of the study model from which the specific CATE estimate was estimated. Ultimately, this method provides study-specific CATE functions (Equation (2.2)) that have hopefully been made more accurate because they have been adjusted to incorporate information from other studies. Tan, Chang, and Tang (2021) applied this approach to investigate the effects of oxygen saturation on hospital mortality across 20 hospitals and found effects that varied across sites but did not have high levels of within-site heterogeneity based on covariates like age or gender.

2.5 Individual Participant-Level Data

Finally, when individual participant-level data (IPD) is available from all studies, treatment effect heterogeneity can be estimated through a wide variety of methods. Recently, many novel methods have been proposed and are actively being developed. While the previous two settings of AD and federated learning are more restrictive, estimating individual-level effect moderation in this setting with all IPD available is much more feasible and flexible. The methods to follow are broken down based on whether the data being combined is from multiple RCTs or from one RCT and one observational dataset. Many of the methods in this multi-study setting build upon single-study methods, which

are discussed in depth in the Supplementary Materials B.

2.5.1 Combining Multiple RCTs

As mentioned when discussing aggregate data, meta-analyses are an effective and widely used parametric approach for combining information from multiple RCTs (Riley, Stewart, and Tierney, 2021). Recently, more and more IPD has become accessible to researchers, allowing them to go a step further from AD and more effectively assess effect moderation. Having IPD available, such as in the example of assessing the effects of pioglitazone for individuals with diabetes (Hong et al., 2015), allows for baseline individual-level covariates to be used to study subgroup effects and effect moderation at the individual level.

2.5.1.1 Types of IPD Meta-Analyses

There are two commonly discussed IPD meta-analysis estimation methods: two-stage and one-stage. In two-stage IPD meta-analysis, aggregate statistics are calculated within each study (e.g., overall treatment effects, effects for each subgroup, interaction terms), and then these results are combined in a between-study model. In one-stage IPD meta-analysis, all individual-level data are put directly into a hierarchical or multilevel model (Burke, Ensor, and Riley, 2017). Although results with respect to average treatment effects are often similar between the two approaches (Burke, Ensor, and Riley, 2017; Debray et al., 2015; Tierney et al., 2015), model assumptions do differ, and choosing the approach that seems best fit to a specific research question is an

important decision. In this paper, we focus on one-stage IPD meta-analysis because of its flexibility (Debray et al., 2015).

2.5.1.2 One-Stage IPD Meta-Analysis

In one-stage IPD meta-analysis, a common technique is to use a generalized linear mixed model (GLMM) to estimate the mean outcome given covariates. The model can have the form

$$g(E(Y_{is})) = \alpha_s + \delta_s A_{is} + \boldsymbol{\beta}_s^T \boldsymbol{X}_{is} + \boldsymbol{\theta}_s^T A_{is} \boldsymbol{Z}_{is}, \tag{2.5}$$

where Y_{is} is the outcome for individual i from study s, $\alpha_s \sim N(\alpha, \sigma_\alpha^2)$ is a study-specific intercept, $\delta_s \sim N(\delta, \sigma_\delta^2)$ is the vector of study-specific treatment effects when the covariates are set to 0 (or their means, if centered), $\boldsymbol{\beta}_s \sim N(\boldsymbol{\beta}, \boldsymbol{\Sigma}_{\boldsymbol{\beta}})$ is the study-specific vector of main effects of covariates on the outcome, and $\boldsymbol{\theta}_s \sim N(\boldsymbol{\theta}, \boldsymbol{\Sigma}_{\boldsymbol{\theta}})$ is the study-specific vector of effect moderation terms (Seo et al., 2021). Here, σ_α^2, σ_δ^2 and the diagonal elements of $\boldsymbol{\Sigma}_{\boldsymbol{\beta}}$ and $\boldsymbol{\Sigma}_{\boldsymbol{\theta}}$ measure the between-study variability of the effects. β_s and θ_s are often assumed to be uncorrelated in the literature; however, we can extend this model to allow for correlation between β_s and θ_s.

If the outcome is continuous (as assumed in this paper), $g(.)$ is often set to be the identity function; if the outcome is binary, $g(.)$ could be the logit link function. Key parameters of interest are δ, which indicates an overall measure of the treatment effect when the moderators are set to 0, and $\boldsymbol{\theta}$, which indicates the magnitude of the effect moderation. For easy interpretation, covariates can be centered at zero so that the treatment effects, δ_s represent the treatment

effects at the mean value of each covariate (Dagne et al., 2016; Gelman, Hill, and Vehtari, 2020).

The model above includes random effects for all coefficients, and so explicitly models between-study heterogeneity for each coefficient (the $\boldsymbol{\beta}_s$'s and $\boldsymbol{\theta}_s$'s). This approach can be thought of as interpolating between two extremes. The first of these is a "no-pooling" model, with the same structure as Equation (2.5) but with study-specific coefficients fit as fixed effects independently to the data from each study. Such a model avoids the sharing of information across studies, but also includes more free parameters, which may be less stably estimated. This approach also does not ultimately provide a global treatment effect estimate across studies, as all studies are given their own fixed coefficients.

A simpler model would treat some coefficients as shared across studies. This might take the form of assuming a common intercept or slope (Thomas, Radji, and Benedetti, 2014); for example, in Equation (2.5), if between-study variability of the main covariate effects (represented by $\boldsymbol{\Sigma}_{\beta}$) were small, a common coefficient could be estimated instead by replacing $\boldsymbol{\beta}_s$ with $\boldsymbol{\beta}$. In practice, $\boldsymbol{\theta}$ is often assumed to be shared across studies. GLMMs can quickly become too complicated if many effects are allowed to vary across studies (especially when study sample sizes are small); on the other hand, the model might be misspecified if it ignores important variation that does exist. Therefore, each coefficient – and whether it should be treated as common across studies, modelled as random, or estimated independently within each study – should be considered carefully to ensure that the model effectively represents

between-study variability while still being sufficiently simple.

GLMMs can be fit under both frequentist and Bayesian frameworks (Debray et al., 2015). If a Bayesian framework is used, prior distributions need to be assigned to each parameter; an option for this is non-informative priors to all parameters of interest (McCandless, 2009). Informative priors can be used when information about the parameters is available from expert opinion or historical data analysis. Hong et al. (2015) utilize a Bayesian framework for their analysis of diabetes medication; however, they compare more than just two treatments and perform network meta-analysis, which is not the focus of this paper.

One other consideration in one-stage IPD meta-analysis is the option to decompose between-study and within-study variability. To avoid aggregation bias, some researchers (Hua et al., 2017; Debray et al., 2015; Donegan et al., 2012; Hong et al., 2015) suggest decomposing the interactions into two sources: individual-level (i.e., within-study effect) and aggregate-level (i.e., between-study effect) interactions. This model can be written by extending Equation (2.5):

$$g(E(Y_{is})) = \alpha_s + \delta_s A_{is} + \boldsymbol{\beta}^T_{s,\text{within}}(\boldsymbol{X}_{is} - \bar{\boldsymbol{X}}_s) +$$

$$\boldsymbol{\beta}^T_{\text{across}}\bar{\boldsymbol{X}}_s + \boldsymbol{\theta}^T_{s,\text{within}} A_{is}(\boldsymbol{Z}_{is} - \bar{\boldsymbol{Z}}_s) + \boldsymbol{\theta}^T_{\text{across}} A_{is}\bar{\boldsymbol{Z}}_s.$$

Here, we have broken up the covariate and treatment-covariate interaction terms into within-study effect and between-study components so that we can separately assess the associations of individual covariates and their study-level summaries with the outcome. This is especially helpful when specific

effect moderators vary significantly both within studies and across studies (Debray et al., 2015). Equation (2.5) is a special case of this model when β_{across} and θ_{across} are equal to the average of the $\beta_{s,\text{within}}$'s and the $\theta_{s,\text{within}}$'s, respectively (Hua et al., 2017).

Standard implementations of meta-analysis techniques to assess effect heterogeneity assume that a set of potential moderators has already been identified and observed in all included studies. Because studies measure several variables that could plausibly serve as effect moderators, selecting which terms to include in the model is an important and challenging decision. Furthermore, testing a high number of potential effect moderators can increase the risk of false positives (Hayward et al., 2020). When many potential moderators exist, variable selection or shrinkage methods can help overcome these challenges and identify meaningful moderators while controlling for overfitting. Seo et al. (2021) compared one-stage IPD meta-analysis methods that identified effect moderators and estimated their effect size. They compared various variable selection methods under both frequentist and Bayesian frameworks including stepwise selection, Lasso regression, Ridge regression, adaptive Lasso, Bayesian Lasso, and stochastic search variable selection (SVSS). In extensive simulation studies, the shrinkage methods (Lasso, Ridge, adaptive Lasso, Bayesian Lasso, and SVSS) performed best, supporting the usage of such methods in IPD meta-analysis to enhance performance (Seo et al., 2021). Especially in settings in which large numbers of variables are available and many could plausibly serve as treatment effect moderators, these methods could be useful to efficiently estimate the conditional average treatment effect.

2.5.1.3 Integrating IPD with AD

If data are available at the individual level in some studies but at the aggregate level in others, both levels of data can still be combined to estimate treatment effects. One straightforward way to do so is through two-stage meta-analysis, as introduced in 2.5.1.1, where models are fit to each study with IPD to calculate aggregate statistics, and then these statistics can be combined with those reported in the AD (Riley et al., 2008). Another more complicated but effective approach is to combine the IPD and AD simultaneously in one-stage meta-analysis: Riley et al. (2008) describe a method for doing this where the outcome for each trial with only AD is simply the estimate of the treatment effect and there is just one observation. They also incorporate an indicator of IPD versus AD.

Bayesian methodology can also be incorporated to combine IPD with AD and allow for adaptive borrowing of information. In such a setting, Hong, Fu, and Carlin (2018) recommend treating the AD as auxiliary data and utilizing a power prior to adaptively incorporate the AD and a commensurate prior to borrow from the AD to estimate treatment effects. In another Bayesian approach, Saramago et al. (2012) incorporate IPD-level covariates to improve estimation of treatment-covariate interactions over that available by AD alone.

2.5.2 Combining an RCT with Observational Data

Another usage for IPD in estimating treatment effect heterogeneity is through combining data from an RCT with an observational dataset. For example, we can consider the scenario introduced earlier where we are interested in

comparing two treatments for major depression, duloxetine and vortioxetine, and we have access to RCT data and a large observational dataset containing electronic health records (Brantner et al., 2024). This scenario requires attention to potential confounding in the observational dataset; notably, the individuals are not randomly assigned to treatment in the observational data unlike in the RCTs. In this setting, the approaches are often nonparametric, with some exceptions, and they include some approach for accounting for confounding in the observational dataset. We use $\hat{\tau}^r(\boldsymbol{X})$ and $\hat{\tau}^o(\boldsymbol{X})$ to represent the estimated CATE function based on data from the RCT and observational study, respectively.

Colnet et al., 2021b provides a literature review of methods that combine RCT and observational data. They touch on many different purposes of combination, one of which is CATE estimation. Their review includes some of the nonparametric approaches listed in this section (Kallus, Puli, and Shalit, 2018; Yang, Zeng, and Wang, 2022; Yang, Zeng, and Wang, 2020) and discusses key assumptions, code, and implementation of methods. Our review incorporates some of the same papers but includes other recent and related approaches as well.

Existing methods for combining RCT and observational data first involve estimating the CATE in either the randomized trial data, the observational data, or both, using single-study methods. These estimators are then combined in one of multiple different ways.

2.5.2.1 Combining Separate CATE Estimates from RCT and Observational Studies

When combining one RCT with one large observational dataset (the usual approach in the methods to follow), one category of approaches involves estimating the CATE in both datasets. In several of these approaches, the final CATE estimate is a weighted combination of the two study-specific CATE estimates, where the weight is derived based on a method-specific estimate of bias in the observational data. This is the approach taken by Rosenman et al. in two papers (2022; 2020). In each paper, Rosenman and colleagues discuss the CATE in terms of average treatment effects within "strata", or subgroups that can be defined as a complex function of covariates (Rosenman et al., 2022). The authors construct strata based on effect moderators and propensity score estimates from the observational data. They assume that within each stratum, the true average treatment effect is the same for both the observational and RCT data; however, the observational data may yield a biased estimate due to unobserved confounding. The base estimator used in their papers is a difference in mean outcomes between the treatment and control group within stratum k:

$$\hat{\tau}_k^o = \frac{\sum_{i \in O_k} A_i Y_i}{\sum_{i \in O_k} A_i} - \frac{\sum_{i \in O_k} (1 - A_i) Y_i}{\sum_{i \in O_k} (1 - A_i)} \tag{2.6}$$

where o indicates observational study, k indexes strata, and O_k is the set of individuals in the observational study belonging to stratum k. The same estimator can be established for the RCT by replacing o and O_k with r and R_k, respectively. From this, Rosenman et al. (2022) construct a "spiked-in" estimator, in which individuals from the RCT are assigned to their corresponding

strata with individuals from the observational data. Then the stratum-specific treatment effects are estimated as in Equation (2.6) but including both RCT and observational data. They compare this "spiked-in" estimator with a dynamic weighted average in which stratum-specific treatment effects are estimated separately in the RCT and observational data, and then the weight for combining the RCT and observational stratum-specific treatment effects is constructed based on the variance of the RCT estimator and the mean squared error (MSE) of the observational data estimator. Ultimately, they discover that the "spiked-in" estimator is only effective when the covariate distributions are very similar across datasets and that their dynamic weighted average has low bias regardless of whether the covariate distributions are similar or not.

In their second paper in this stratum-specific treatment effect framework, Rosenman et al. (2020) utilize shrinkage estimation to combine CATE estimators from the RCT and observational dataset. They first determine a structure for a given shrinkage factor, λ, and then optimize an unbiased risk estimate to solve for this λ. They again define stratum-specific average treatment effects under the assumption that treatment effect heterogeneity can be assessed by dividing up the dataset into strata. For example, they define a common shrinkage factor λ selected by minimizing the unbiased risk estimate such that

$$\hat{\tau}_k(\lambda) = \hat{\tau}_k^r - \lambda(\hat{\tau}_k^r - \hat{\tau}_k^o) \tag{2.7}$$

where r indexes the RCT estimator, o the observational estimator, k indexes strata, and $\hat{\tau}_k^r$ and $\hat{\tau}_k^o$ can be estimated as specified in Equation (2.6). They also discuss an estimator that is the same as Equation (2.7) but multiplies the

difference $\lambda(\hat{\tau}_k^r - \hat{\tau}_k^o)$ by the variance matrix from the RCT. Note that both of these approaches by Rosenman and colleagues are technically at the subgroup-level; however, these subgroups can be complex functions of covariates, so the approach can be easily discussed in terms of covariates, $\boldsymbol{X}$, instead of stratum membership.

A recent paper by Cheng and Cai (2021) incorporates a similar approach to the shrinkage estimation by Rosenman et al. (2020) by adaptively combining CATE functions between an RCT and observational dataset based on the estimated degree of bias in the observational estimator to yield study-specific CATE estimates that minimize MSE. Cheng and Cai (2021) also use a weighted linear combination of CATE estimators from the RCT, $\hat{\tau}_s^r(\boldsymbol{X})$ and the observational data, $\hat{\tau}_s^o(\boldsymbol{X})$:

$$\hat{\tau}_s(\boldsymbol{X}) = \hat{\tau}_s^r(\boldsymbol{X}) + \nu_{\boldsymbol{X}}\{\hat{\tau}_s^o(\boldsymbol{X}) - \hat{\tau}_s^r(\boldsymbol{X})\}$$

where $s = 0, 1$ denotes RCT and observational data, respectively and $\nu_{\boldsymbol{X}}$ is a weight function. To estimate CATE functions in each study separately, the authors use doubly-robust pseudo-outcomes (Kennedy, 2020) that are defined as influence functions for the average treatment effect (see more in the Supplementary Material B). These influence functions are then regressed on the potential effect moderators, $\boldsymbol{X}$, to estimate the CATE in both the RCT ($\hat{\tau}_s^r(\boldsymbol{X})$) and observational data ($\hat{\tau}_s^o(\boldsymbol{X})$) separately. The weight $\nu_{\boldsymbol{X}}$ is estimated by minimizing a decomposition of an estimate of the mean squared error (MSE) for the CATE function and varies based on $\boldsymbol{X}$. This strategy allows for the weight to heavily favor the RCT estimator when the observational data

is biased and to combine both estimators efficiently to minimize asymptotic variance in the presence of insignificant bias in the observational data.

Cheng and Cai's method of estimating ν_X is similar to Rosenman et al. (2020) approach of estimating λ using an unbiased risk estimate. An important distinction between the two approaches is that Rosenman et al. (2020) represent treatment effect heterogeneity through K distinct strata within which they assume that the treatment effect is common across the RCT and observational datasets. Cheng and Cai (2021) instead use individual covariates as part of their CATE estimation, and they do not require the treatment effects to be equivalent between the RCT and observational datasets. Cheng and Cai (2021) also use a different base estimation procedure for the initial estimates of τ in the RCT and observational data.

Finally, Yang, Zeng, and Wang (2020) also combine separate estimates of the CATE from the RCT and observational data to minimize MSE under the assumptions of unconfoundedness in the RCT (Assumption 2.1 in the RCT; satisfied via randomization) and a structural model for the CATE ($\tau(\boldsymbol{X}) = \tau_{\psi_0}(\boldsymbol{X})$). This approach uses elastic integration to combine the estimates based on a hypothesis test that determines whether the assumption of unconfoundedness in the observational data (Assumption 2.1 in the observational data) is sufficiently met or not (Yang, Zeng, and Wang, 2020). To construct this test, Yang et al. (2020) introduce

$$H_{\psi_0}(\boldsymbol{X}) = Y - \tau_{\psi_0}(\boldsymbol{X})A \tag{2.8}$$

such that $E(H_{\psi_0}|A, \boldsymbol{X}, S) = E(Y(0)|A, \boldsymbol{X}, S)$. From here, they introduce a

semiparametric efficient score of the parameters ψ_0 which we will call SES_{ψ_0}. This semiparametric efficient score is used in their hypothesis test with a null hypothesis of $E(\text{SES}^o_{\psi_0}) = 0$ where $\text{SES}^o_{\psi_0}$ is the score in the observational data. If this null hypothesis is rejected, the ultimate parameters for the CATE are determined solely from the RCT data; if not, parameters are solved for using an elastic integration of both the RCT and observational data. Estimating the parameters is discussed in more detail in Yang et al.'s (2020) paper; briefly, they solve

$$\frac{\sum_{i=1}^{N} \widehat{\text{SES}}_{\psi}}{N} = 0$$

by plugging in estimators of unknown quantities and solving for ψ.

2.5.2.2 Estimating and Accounting for the Confounding Bias in the Observational Data

Another category focuses on estimating the CATE – and the confounding bias, as estimated by bringing in the RCT data – in the observational data, rather than estimating the CATE in each dataset. Kallus and colleagues (2018) estimate the CATE in the observational data first and then estimate a correction term to adjust for confounding. They focus on deriving a CATE estimator that is consistent. The approach assumes unconfoundedness (Assumption 2.1) in the RCT, but does not assume that the observational data fully overlaps with the RCT data (Kallus, Puli, and Shalit, 2018; Colnet et al., 2021b). The authors note that the CATE function in the observational data, $\tau^o(\boldsymbol{X})$ does not equal the true CATE, $\tau(\boldsymbol{X})$ because of confounding, so they define the confounding effect to be

$$\eta(\boldsymbol{X}) = \tau(\boldsymbol{X}) - \tau^o(\boldsymbol{X})$$

and focus on estimating this η to correct the observational CATE estimator. The observational CATE is estimated using any single-study approach, such as a causal forest (Athey, Tibshirani, and Wager, 2019), and the confounding effect is estimated using the following equation. For the propensity score in the RCT, $\pi^r(\boldsymbol{X}) = P(A = 1|\boldsymbol{X}, S = 0)$, Kallus et al. define

$$q^r(\boldsymbol{X}_i) = \frac{A_i}{\pi^r(\boldsymbol{X}_i)} - \frac{1 - A_i}{1 - \pi^r(\boldsymbol{X}_i)}$$

for individuals in the RCT. This leads to the final equation to solve to estimate the confounding effect:

$$\hat{\boldsymbol{\theta}} = \text{argmin}_{\boldsymbol{\theta}} \sum_{i=1}^{n^r} (q^r(\boldsymbol{X}_i) Y_i - \hat{\tau}^o(\boldsymbol{X}_i) - \boldsymbol{\theta}^T \boldsymbol{X}_i)^2$$

again applied to only individuals in the RCT, where n^r is the total number of individuals in the RCT. Finally, they set $\hat{\eta}(\boldsymbol{X}) = \hat{\boldsymbol{\theta}}^T \boldsymbol{X}$ and ultimately define

$$\hat{\tau}(\boldsymbol{X}) = \hat{\tau}^o(\boldsymbol{X}) + \hat{\eta}(\boldsymbol{X}).$$

Yang, Zeng, and Wang (2022) also estimate confounding in the observational study directly. They focus on the conditional average treatment effect on the treated (CATT), $\tau(\boldsymbol{X}) = E[Y(1) - Y(0)|\boldsymbol{X}, A = 1]$, and define a confounding function to estimate the effect of unobserved confounding in the observational data. They assume unconfoundedness in the RCT (Assumption 2.1), a structural model for both the CATT and the confounding function, ζ, and that the RCT and observational data come from the same target population, though their covariate distributions need not overlap. Their confounding function is defined in the observational study as the difference in potential

outcome means between treatment groups:

$$\zeta(\boldsymbol{X}) = E[Y(0)|A = 1, \boldsymbol{X}, S = 1] -$$

$$E[Y(0)|A = 0, \boldsymbol{X}, S = 1].$$

When all confounders are measured, $\zeta(\boldsymbol{X}) = 0$, but in reality, unobserved confounders will lead the function to be non-zero. Yang, Zeng, and Wang (2022) show that this function is only identifiable when the RCT data is used with the observational data.

To estimate the parameters for the CATT and the confounding function, Yang, Zeng, and Wang (2022) utilize estimating equations and semiparametric efficiency theory, similar to the approach taken by Yang, Zeng, and Wang (2020). Specifically, they define an equation similar to that of their previous work (Yang, Zeng, and Wang, 2020) shown in Equation (2.8):

$$H_{\psi_0} = Y - \tau_{\varphi_0}(\boldsymbol{X})A - S\zeta_{\phi_0}(\boldsymbol{X})(A - e(\boldsymbol{X}, S))$$

where $\psi_0 = (\varphi_0, \phi_0)$ are parameters and such that the final term in the equation will only come into play when $S = 1$, i.e., in the observational data. They solve an estimating equation based around this H to get a preliminary estimator of the parameters for τ and ζ; next, they update this solution based on a semiparametric efficient score. The authors finally show that their estimator of the CATT, which integrates both datasets, is more efficient than the CATT from the RCT data when the predictors from the CATT function and confounding function are linearly independent.

The "integrative R-learner" falls in a similar category of methods and

is based on adapting the original R-learner by Nie and Wager (2021) (see Supplementary Material B) to the setting with one RCT and one observational dataset (Wu and Yang, 2021). This approach minimizes loss and is consistent and asymptotically efficient compared to an RCT-only estimator. The authors use a very similar definition of the confounding function as in Yang, Zeng, and Wang, 2022, with a slight adjustment:

$$c(\boldsymbol{X}) = E(Y|\boldsymbol{X}, A = 1, S = 1) - E(Y|\boldsymbol{X}, A = 0, S = 1) - \tau(\boldsymbol{X})$$

where $c(X) = 0$ when there is no unobserved confounding in the observational dataset (Assumption 2.1). Wu and Yang (2021) estimate this confounding function and $\tau(\boldsymbol{X})$ by minimizing an empirical loss function that has the Neyman orthogonality property, as found in the original R-learner (Nie and Wager, 2021).

Finally, Hatt et al. (2022) propose a method that utilizes the estimated confounding effect in the observational data through a representation learning approach. Under similar assumptions to previous methods such as consistency (Assumption 2.3), common support across the RCT and observational data (Assumption 2.4), and unconfoundedness in the RCT (Assumption 2.1) among others, Hatt et al. (2022) define ϕ^* to be a representation of the shared structure of covariates in both the RCT and the observational data. They also define h_a^r and h_a^o as "hypotheses" in the RCT and observational data, respectively, for $a = 0, 1$ indicating control or treatment. These so-called hypotheses are functions meant to be applied to the representation, ϕ^* where

for r representing membership in the RCT and o in the observational data,

$$E(Y^r|A^r = a, \boldsymbol{X}^r = \boldsymbol{x}) - E(Y^o|A^o = a, \boldsymbol{X}^o = \boldsymbol{x})$$

$$= h_a^r(\phi^*(\boldsymbol{x})) - h_a^o(\phi^*(\boldsymbol{x})).$$

Similarly to previous methods, Hatt et al. (2022) use a confounding function to represent the bias, defined as $\gamma_a = h_a^r - h_a^o$. Their algorithm starts by estimating $\hat{\phi}$ and $\hat{h}_a^o$ for $a = 0, 1$ from the observational data by minimizing an empirical loss. Next, these estimates are applied to the RCT data and the empirical loss in this dataset is minimized to derive an estimate for the bias $\hat{\gamma}_a$, $a = 0, 1$. Finally, these estimates are combined using the fact that $\gamma_a = h_a^r - h_a^o$ to solve for $\hat{h}_a^r = \hat{\gamma}_a + \hat{h}_a^o$ and to ultimately estimate the CATE as

$$\hat{\tau}(\boldsymbol{X}) = \hat{h}_1^r(\hat{\phi}(\boldsymbol{X})) - \hat{h}_0^r(\hat{\phi}(\boldsymbol{X})).$$

2.6 Discussion

2.6.1 Comparison of Approaches

The recent influx of interest in studying treatment effect heterogeneity has led to novel and adapted methods that strive to improve the identification of tailored interventions. Furthermore, with the increase of IPD availability and the simultaneous research interests of combining data sources, assessing treatment effect heterogeneity in a reproducible manner is more feasible than before. Table 2.1 summarizes the aforementioned approaches, with a focus on their data setting, modeling approach, and motivation.

Table 2.1: Comparison of approaches to estimate the CATE using multiple studies

Approach	Data Level	Data Types	Model	Estimand	Motivation
Meta-Analysis of Interactions	AD	RCTs	Parametric	Pooled	Pool treatment-covariate interactions
Meta-Regression	AD	RCTs	Parametric	Pooled	Model group-level treatment-covariate interactions
Meta-Analysis of Local Models	FL	RCTs	Parametric	Pooled	Pool treatment-covariate interactions
Tan, Chang, and Tang, 2021	FL	RCTs	Non-parametric	Study-specific	Borrow information from other studies to improve model
One-Stage Meta-Analysis	IPD	RCTs	Parametric	Pooled	Model individual-level treatment-covariate interactions
Meta-Analysis of IPD and AD	IPD/AD	RCTs	Parametric	Pooled	Adaptively incorporate AD as auxiliary data
Rosenman et al., 2022	IPD	RCT and OD	Parametric	Pooled	Weight combination of CATE estimators based on OD bias
Rosenman et al., 2020	IPD	RCT and OD	Parametric	Pooled	Weight combination of CATE estimators based on OD bias
Cheng and Cai, 2021	IPD	RCT and OD	Non-parametric	Study-specific	Weight combination of CATE estimators based on OD bias
Yang, Zeng, and Wang, 2020	IPD	RCT and OD	Parametric	Pooled	Weight combination of CATE estimators based on OD bias
Kallus, Puli, and Shalit, 2018	IPD	RCT and OD	Non-parametric	Pooled	Estimate confounding function
Yang, Zeng, and Wang, 2022	IPD	RCT and OD	Parametric	Pooled	Estimate confounding function
Wu and Yang, 2021	IPD	RCT and OD	Non-parametric	Pooled	Estimate confounding function
Hatt et al., 2022	IPD	RCT and OD	Non-parametric	Pooled	Estimate confounding function

AD = aggregate-level data, FL = federated learning, IPD = individual participant-level data, RCT = randomized controlled trial, OD = observational data

2.6.2 Parametric and Nonparametric Approaches

Meta-analyses have been in use for many years but are less often conceptualized in terms of identifying treatment effect moderation. This review and some other continuing work (i.e., Seo et al., 2021) have tied meta-analyses into this framework. Traditional methods for assessing moderation generally have involved parametric approaches that require pre-specification of the potential moderators. However, parametric regression models are limited by the need to pre-specify interaction terms, and complex non-linearities might be missed in the ultimate CATE function. Variable shrinkage techniques (including priors) could help to ensure that the most important interactions are included without overfitting the model (Seo et al., 2021).

Newer approaches listed in Section 2.5.2 include flexible machine learning methods that allow for complicated functional forms for the covariates in the CATE and do not require that moderators be pre-specified. The nonparametric side to estimation that is often employed when combining an RCT with observational data allows for the CATE function to be more complex, but there are some potential weaknesses of these methods compared with simpler parametric models. First, the resulting CATE estimates may be more difficult to interpret, particularly if the goal is to pick out individual effect moderators and assess their precise relationship with the treatment effect. Second, the desirable theoretical properties of these methods—consistency of the estimators, robustness against model misspecification, accuracy of the associated confidence intervals—are for the most part asymptotic, and so a priori one would expect that the nonparametric/machine learning methods are better

suited to situations with enough data. The point at which the robustness of the nonparametric approaches is to be preferred over the explicitness and simplicity of the parametric approaches is perhaps best assessed using a combination of contextual or scientific background knowledge, simulation studies, data splitting techniques like cross-validation and training/test/validation sets, and real-world experience with the methods.

In conclusion, parametric models may suffer from model misspecification but are easy to interpret and apply. Although machine learning methods are relatively untested, their statistical properties are mostly asymptotic, and their implementation can be more computationally intensive, they incorporate a large amount of flexibility and could be ideal when complex nonlinear associations are expected with a large number of variables.

2.6.3 Current Shortcomings and Future Directions

Because this field is growing rapidly and the methods discussed are somewhat new, many methods have not been thoroughly compared to one another in simulation studies or illustrated using real trials and/or observational datasets. There is therefore a broad opening for future research that assesses these approaches in comparison to one another through data applications. For meta-analysis, many real-world applications exist, but not all go in-depth into treatment effect heterogeneity. The remaining approaches discussed in this study are all very recent, and the new methods have not been tried out extensively in real data. Real-world applications will be important for understanding the practical implications and considerations such as differential

measurement across datasets, missing data, and more – such implications must be addressed for the methods to be fully useful in applications. Furthermore, any comparisons that have been done do not combine parametric and nonparametric approaches in this field of CATE estimation using multiple studies.

Another useful field of follow-up study is consolidating and evaluating assumptions. The assumptions of methods discussed here vary in whether they are required, relaxed, or unneeded. It would be helpful to be able to empirically evaluate the assumptions across datasets to examine their feasibility, although not all assumptions explored in this paper can be empirically assessed. Specific approaches for inference in the form of variance estimation and confidence intervals are also needed in many approaches. For parametric approaches discussed throughout the review, often standard methods such as Wald confidence intervals can be employed (Yang, Zeng, and Wang, 2022), or bootstrapping can be used to estimate intervals and standard errors as well. However, there is an opening for more work to determine the best inference approaches in the parametric and nonparametric cases, and how these approaches vary depending on the method.

More work could also be done when it comes to the type of data being combined. One might be interested in determining how to apply the meta-analytic framework to the combination of trial and observational data; this field has been called cross-design synthesis and has been debated in the literature (Debray et al., 2015). On the other hand, the methods geared towards combining an RCT with observational data could be tailored to combine

multiple RCTs, but this option was not discussed in the methods previously described aside from briefly in the federated learning setting (Tan, Chang, and Tang, 2021)

In terms of specific data availability settings, aggregate-level data consistently provides a challenge for estimating individual-level effect moderation, and there are only a couple of limited settings in which this goal can be achieved. Therefore, more IPD data access is the simplest solution to being able to derive an in-depth model to estimate the CATE. For the case when IPD is available but cannot be shared across studies (i.e., federated learning), the approaches discussed in this review could be tailored to deal with this. Very few methods exist in this field within federated learning; only one paper specifically discusses treatment effect heterogeneity when data is distributed privately across studies (Tan, Chang, and Tang, 2021). Thus, future work could be done to derive approaches to estimate the CATE in federated learning.

Data availability also can vary within a given set of studies, and researchers often run into the issue of systematically missing covariates – i.e., covariates available in some but not all data sources. Covariates also can be sporadically missing, where the covariate is present in all studies but missing for some individuals throughout the studies. Future development of the methods discussed previously should incorporate these considerations, as many of the new approaches leave this for future work. Some papers have looked into these types of missingness in a slightly separate context (Colnet et al., 2021a); for example, Audigier et al. (2018) investigated the performance of multiple imputation procedures for systematically and sporadically missing

data. Jolani et al. (2015) also describe a generalized imputation approach for IPD meta-analysis when covariates are systematically missing.

An appropriate follow-up question from this work is when to best implement each method. Because the machine learning methods have not been compared to one another in simulation studies, it is difficult to conclude which of the methods is optimal in which scenario. This review does attempt to clarify which type of data can be handled by each method, and whether the method works with RCT and observational data, or multiple RCTs. However, further study is needed to determine which approach will yield the most accurate predictions depending on the types of heterogeneity present in the study (i.e., heterogeneity across studies, heterogeneity within studies).

For those working in this field or those who want to learn more, it is important to continue to look out for new research that comes out, since this field is changing and growing rapidly. At the time of this review, many future directions of work are open for pursuit. The new methods mentioned throughout this review increase the feasibility of reproducible conclusions regarding individualized treatment decisions. Because we can employ data from multiple sources, we are developing a deeper understanding and can more effectively estimate individual treatment effects that are reliable and generalizable.

2.7 Acknowledgements

Research reported in this publication was partially funded through a Patient-Centered Outcomes Research Institute (PCORI) Award (ME-2020C3-21145;

PI: Stuart) and by the National Institute of Mental Health (R01MH126856; PI: Stuart). Ms. Brantner also received financial support in the form of a training grant through the National Institutes of Health (T32AG000247). The statements in this work are solely the responsibility of the authors and do not necessarily represent the views of the Patient-Centered Outcomes Research Institute (PCORI), its Board of Governors or Methodology Committee, or of the National Institute of Mental Health.

References

Trivedi, Madhukar H., A. John Rush, Stephen R. Wisniewski, Andrew A. Nierenberg, Diane Warden, Louise Ritz, Grayson Norquist, Robert H. Howland, Barry Lebowitz, Patrick J. McGrath, Kathy Shores-Wilson, Melanie M. Biggs, G. K. Balasubramani, Maurizio Fava, and STAR*D Study Team (2006). "Evaluation of outcomes with citalopram for depression using measurement-based care in STAR*D: implications for clinical practice". eng. In: *The American Journal of Psychiatry* 163.1, pp. 28–40. ISSN: 0002-953X. DOI: 10.1176/appi.ajp.163.1.28.

Samara, Myrto T., Adriani Nikolakopoulou, Georgia Salanti, and Stefan Leucht (2019). "How Many Patients With Schizophrenia Do Not Respond to Antipsychotic Drugs in the Short Term? An Analysis Based on Individual Patient Data From Randomized Controlled Trials". eng. In: *Schizophrenia Bulletin* 45.3, pp. 639–646. ISSN: 1745-1701. DOI: 10.1093/schbul/sby095.

Xie, Fangying, Juliana Cn Chan, and Ronald Cw Ma (2018). "Precision medicine in diabetes prevention, classification and management". eng. In: *Journal of Diabetes Investigation* 9.5, pp. 998–1015. ISSN: 2040-1124. DOI: 10.1111/jdi.12830.

Baron, Reuben M. and David A. Kenny (1986). "The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations". In: *Journal of Personality and Social Psychology* 51.6, pp. 1173–1182. ISSN: 1939-1315. DOI: 10.1037/0022-3514.51.6.1173.

Enderlein, G. (1988). "Fleiss, J. L.: The Design and Analysis of Clinical Experiments." en. In: *Biometrical Journal* 30.3, pp. 304–304. ISSN: 1521-4036. DOI: 10.1002/bimj.4710300308. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/bimj.4710300308 (visited on 03/30/2022).

Green, Angela K., Niti Trivedi, Jennifer J. Hsu, Nancy L. Yu, Peter B. Bach, and Susan Chimonas (2022). "Despite The FDA's Five-Year Plan, Black Patients Remain Inadequately Represented In Clinical Trials For Drugs: Study examines FDA's five-year action plan aimed at improving diversity

in and transparency of pivotal clinical trials for newly-approved drugs." en. In: *Health Affairs* 41.3, pp. 368–374. ISSN: 0278-2715, 1544-5208. DOI: 10.1377/hlthaff.2021.01432. URL: http://www.healthaffairs.org/doi/10.1377/hlthaff.2021.01432 (visited on 03/15/2022).

Berlin, Jesse A., Jill Santanna, Christopher H. Schmid, Lynda A. Szczech, Harold I. Feldman, and Anti-Lymphocyte Antibody Induction Therapy Study Group (2002). "Individual patient- versus group-level data meta-regressions for the investigation of treatment effect modifiers: ecological bias rears its ugly head". eng. In: *Statistics in Medicine* 21.3, pp. 371–387. ISSN: 0277-6715. DOI: 10.1002/sim.1023.

Brown, C. Hendricks, Zili Sloboda, Fabrizio Faggiano, Brent Teasdale, Ferdinand Keller, Gregor Burkhart, Federica Vigna-Taglianti, George Howe, Katherine Masyn, Wei Wang, Bengt Muthén, Peggy Stephens, Scott Grey, Tatiana Perrino, and Prevention Science and Methodology Group (2013). "Methods for synthesizing findings on moderation effects across multiple randomized trials". eng. In: *Prevention Science: The Official Journal of the Society for Prevention Research* 14.2, pp. 144–156. ISSN: 1573-6695. DOI: 10.1007/s11121-011-0207-8.

Kent, David M., Peter M. Rothwell, John PA Ioannidis, Doug G. Altman, and Rodney A. Hayward (2010). "Assessing and reporting heterogeneity in treatment effects in clinical trials: a proposal". In: *Trials* 11.1, p. 85. ISSN: 1745-6215. DOI: 10.1186/1745-6215-11-85. URL: https://doi.org/10.1186/1745-6215-11-85 (visited on 03/30/2022).

Kent, David M, Jessica K Paulus, David Van Klaveren, Ralph D'Agostino, Steve Goodman, Rodney Hayward, John PA Ioannidis, Bray Patrick-Lake, Sally Morton, Michael Pencina, et al. (2020). "The predictive approaches to treatment effect heterogeneity (PATH) statement". In: *Annals of internal medicine* 172.1, pp. 35–45.

Vo, Thanh Vinh, Trong Nghia Hoang, Young Lee, and Tze-Yun Leong (2021). "Federated Estimation of Causal Effects from Observational Data". en. In: *arXiv:2106.00456 [cs, stat]*. URL: http://arxiv.org/abs/2106.00456 (visited on 02/02/2022).

Debray, Thomas P. A., Karel G. M. Moons, Gert Valkenhoef, Orestis Efthimiou, Noemi Hummel, Rolf H. H. Groenwold, and Johannes B. Reitsma (2015). "Get real in individual participant data (IPD) meta-analysis: a review of the methodology". en. In: *Research Synthesis Methods* 6.4, pp. 293–309. ISSN: 1759-2879, 1759-2887. DOI: 10.1002/jrsm.1160. URL: https://onlinelibrary.wiley.com/doi/10.1002/jrsm.1160 (visited on 02/02/2022).

Burke, Danielle L., Joie Ensor, and Richard D. Riley (2017). "Meta-analysis using individual participant data: one-stage and two-stage approaches, and why they may differ". en. In: *Statistics in Medicine* 36.5, pp. 855–875. ISSN: 02776715. DOI: 10.1002/sim.7141. URL: https://onlinelibrary.wiley.com/doi/10.1002/sim.7141 (visited on 02/02/2022).

Seo, Michael, Ian R. White, Toshi A. Furukawa, Hissei Imai, Marco Valgimigli, Matthias Egger, Marcel Zwahlen, and Orestis Efthimiou (2021). "Comparing methods for estimating patient-specific treatment effects in individual patient data meta-analysis". en. In: *Statistics in Medicine* 40.6, pp. 1553–1573. ISSN: 0277-6715, 1097-0258. DOI: 10.1002/sim.8859. URL: https://onlinelibrary.wiley.com/doi/10.1002/sim.8859 (visited on 02/02/2022).

Rosenman, Evan TR, Art B Owen, Mike Baiocchi, and Hailey R Banack (2022). "Propensity score methods for merging observational and experimental datasets". In: *Statistics in Medicine* 41.1, pp. 65–86.

Rosenman, Evan, Guillaume Basse, Art Owen, and Michael Baiocchi (2020). "Combining Observational and Experimental Datasets Using Shrinkage Estimators". en. In: *arXiv:2002.06708 [math, stat]*. URL: http://arxiv.org/abs/2002.06708 (visited on 02/02/2022).

Cheng, David and Tianxi Cai (2021). "Adaptive Combination of Randomized and Observational Data". en. In: *arXiv:2111.15012 [stat]*. URL: http://arxiv.org/abs/2111.15012 (visited on 02/02/2022).

Yang, Shu, Donglin Zeng, and Xiaofei Wang (2020). "Elastic Integrative Analysis of Randomized Trial and Real-World Data for Treatment Heterogeneity Estimation". en. In: *arXiv:2005.10579 [stat]*. URL: http://arxiv.org/abs/2005.10579 (visited on 03/17/2022).

Kallus, Nathan, Aahlad Manas Puli, and Uri Shalit (2018). "Removing Hidden Confounding by Experimental Grounding". en. In: *arXiv:1810.11646 [cs, stat]*. URL: http://arxiv.org/abs/1810.11646 (visited on 02/02/2022).

Wu, Lili and Shu Yang (2021). "Integrative *R*-learner of heterogeneous treatment effects combining experimental and observational studies". In: *First Conference on Causal Learning and Reasoning*.

Hatt, Tobias, Jeroen Berrevoets, Alicia Curth, Stefan Feuerriegel, and Mihaela van der Schaar (2022). "Combining observational and randomized data for estimating heterogeneous treatment effects". In: *arXiv preprint arXiv:2202.12891*.

Colnet, Bénédicte, Imke Mayer, Guanhua Chen, Awa Dieng, Ruohong Li, Gaël Varoquaux, Jean-Philippe Vert, Julie Josse, and Shu Yang (2021b). "Causal

inference methods for combining randomized trials and observational studies: a review". en. In: *arXiv:2011.08047 [stat]*. URL: http://arxiv.org/abs/2011.08047 (visited on 02/02/2022).

Petrelli, Fausto and Sandro Barni (2012). "Surgery of primary tumors in stage IV breast cancer: an updated meta-analysis of published studies with meta-regression". In: *Medical oncology* 29, pp. 3282–3290.

Han, Larry, Jue Hou, Kelly Cho, Rui Duan, and Tianxi Cai (2021). "Federated Adaptive Causal Estimation (FACE) of Target Treatment Effects". en. In: *arXiv:2112.09313 [math, stat]*. URL: http://arxiv.org/abs/2112.09313 (visited on 02/02/2022).

Hong, Hwanhee, Haoda Fu, Karen L Price, and Bradley P Carlin (2015). "Incorporation of individual-patient data in network meta-analysis for multiple continuous endpoints, with application to diabetes treatment". In: *Statistics in Medicine* 34.20, pp. 2794–2819.

Brantner, Carly Lupton, Trang Quynh Nguyen, Tengjie Tang, Congwen Zhao, Hwanhee Hong, and Elizabeth A Stuart (2024). "Comparison of methods that combine multiple randomized trials to estimate heterogeneous treatment effects". In: *Statistics in Medicine*.

Rubin, Donald B. (1974). "Estimating causal effects of treatments in randomized and nonrandomized studies". In: *Journal of Educational Psychology* 66.5, pp. 688–701. ISSN: 1939-2176. DOI: 10.1037/h0037350.

Abrevaya, Jason, Yu-Chin Hsu, and Robert P. Lieli (2015). "Estimating Conditional Average Treatment Effects". en. In: *Journal of Business & Economic Statistics* 33.4, pp. 485–505. ISSN: 0735-0015, 1537-2707. DOI: 10.1080/07350015.2014.975555. URL: http://www.tandfonline.com/doi/full/10.1080/07350015.2014.975555 (visited on 03/30/2022).

Künzel, Sören R, Jasjeet S Sekhon, Peter J Bickel, and Bin Yu (2019). "Metalearners for estimating heterogeneous treatment effects using machine learning". In: *Proceedings of the national academy of sciences* 116.10, pp. 4156–4165.

Efthimiou, Orestis, Thomas P. A. Debray, Gert van Valkenhoef, Sven Trelle, Klea Panayidou, Karel G. M. Moons, Johannes B. Reitsma, Aijing Shang, and Georgia Salanti (2016). "GetReal in network meta-analysis: a review of the methodology". en. In: *Research Synthesis Methods* 7.3, pp. 236–263. ISSN: 17592879. DOI: 10.1002/jrsm.1195. URL: https://onlinelibrary.wiley.com/doi/10.1002/jrsm.1195 (visited on 02/02/2022).

Debray, Thomas PA, Ewoud Schuit, Orestis Efthimiou, Johannes B Reitsma, John PA Ioannidis, Georgia Salanti, and Karel GM Moons (2018). "An

overview of methods for network meta-analysis using individual participant data: when do benefits arise?" en. In: *Statistical Methods in Medical Research* 27.5, pp. 1351–1364. ISSN: 0962-2802, 1477-0334. DOI: 10.1177/0962280216660741. URL: http://journals.sagepub.com/doi/10.1177/0962280216660741 (visited on 02/02/2022).

Dahabreh, Issa J., Lucia C. Petito, Sarah E. Robertson, Miguel A. Hernán, and Jon A. Steingrimsson (2020). "Towards causally interpretable meta-analysis: transporting inferences from multiple studies to a target population". en. In: *arXiv:1903.11455 [stat]*. URL: http://arxiv.org/abs/1903.11455 (visited on 03/18/2022).

Yang, Shu, Donglin Zeng, and Xiaofei Wang (2022). "Improved Inference for Heterogeneous Treatment Effects Using Real-World Data Subject to Hidden Confounding". en. In: *arXiv:2007.12922 [stat]*. URL: http://arxiv.org/abs/2007.12922 (visited on 02/02/2022).

Teramukai, Satoshi, Yutaka Matsuyama, Sachiko Mizuno, and Junichi Sakamoto (2004). "Individual patient-level and study-level meta-analysis for investigating modifiers of treatment effect". In: *Japanese journal of clinical oncology* 34.12, pp. 717–721.

Lambert, P.C., A.J. Sutton, K.R. Abrams, and D.R. Jones (2002). "A comparison of summary patient-level covariates in meta-regression with individual patient data meta-analysis". en. In: *Journal of Clinical Epidemiology* 55.1, pp. 86–94. ISSN: 08954356. DOI: 10.1016/S0895-4356(01)00414-0. URL: https://linkinghub.elsevier.com/retrieve/pii/S0895435601004140 (visited on 03/17/2022).

Simmonds, M. C. and J. P. T. Higgins (2007). "Covariate heterogeneity in meta-analysis: Criteria for deciding between meta-regression and individual patient data". en. In: *Statistics in Medicine* 26.15, pp. 2982–2999. ISSN: 1097-0258. DOI: 10.1002/sim.2768. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.2768 (visited on 03/30/2022).

Kovalchik, Stephanie A (2013). "Aggregate-data estimation of an individual patient data linear random effects meta-analysis with a patient covariate-treatment interaction term". In: *Biostatistics* 14.2, pp. 273–283.

Godolphin, Peter J, Ian R White, Jayne F Tierney, and David J Fisher (2022). "Estimating interactions and subgroup-specific treatment effects in meta-analysis without aggregation bias: A within-trial framework". In: *Research Synthesis Methods*.

Yang, Qiang, Yang Liu, Yong Cheng, Yan Kang, Tianjian Chen, and Han Yu (2019). "Federated learning". In: *Synthesis Lectures on Artificial Intelligence and Machine Learning* 13.3, pp. 1–207.

Silva, Santiago, Boris A Gutman, Eduardo Romero, Paul M Thompson, Andre Altmann, and Marco Lorenzi (2019). "Federated learning in distributed medical databases: Meta-analysis of large-scale subcortical brain data". In: *2019 IEEE 16th international symposium on biomedical imaging (ISBI 2019)*. IEEE, pp. 270–274.

Tan, Xiaoqing, Chung-Chou H. Chang, and Lu Tang (2021). "A Tree-based Federated Learning Approach for Personalized Treatment Effect Estimation from Heterogeneous Data Sources". en. In: *arXiv:2103.06261 [cs, stat]*. URL: http://arxiv.org/abs/2103.06261 (visited on 02/02/2022).

Athey, Susan, Julie Tibshirani, and Stefan Wager (2019). "Generalized random forests". In: *The Annals of Statistics* 47.2, pp. 1148–1178.

Riley, Richard D, Lesley A Stewart, and Jayne F Tierney (2021). "Individual Participant Data Meta-Analysis for Healthcare Research". In: *Individual Participant Data Meta-Analysis: A Handbook for Healthcare Research*, pp. 1–6.

Tierney, Jayne F., Claire Vale, Richard Riley, Catrin Tudur Smith, Lesley Stewart, Mike Clarke, and Maroeska Rovers (2015). "Individual Participant Data (IPD) Meta-analyses of Randomised Controlled Trials: Guidance on Their Use". eng. In: *PLoS medicine* 12.7, e1001855. ISSN: 1549-1676. DOI: 10.1371/journal.pmed.1001855.

Dagne, Getachew A., C. Hendricks Brown, George Howe, Sheppard G. Kellam, and Lei Liu (2016). "Testing moderation in network meta-analysis with individual participant data: Testing moderation in network meta-analysis with individual participant data". en. In: *Statistics in Medicine* 35.15, pp. 2485–2502. ISSN: 02776715. DOI: 10.1002/sim.6883. URL: https://onlinelibrary.wiley.com/doi/10.1002/sim.6883 (visited on 02/02/2022).

Gelman, Andrew, Jennifer Hill, and Aki Vehtari (2020). *Regression and other stories*. Cambridge University Press.

Thomas, Doneal, Sanyath Radji, and Andrea Benedetti (2014). "Systematic review of methods for individual patient data meta- analysis with binary outcomes". en. In: *BMC Medical Research Methodology* 14.1, p. 79. ISSN: 1471-2288. DOI: 10.1186/1471-2288-14-79. URL: https://bmcmedresmethodol.biomedcentral.com/articles/10.1186/1471-2288-14-79 (visited on 02/02/2022).

McCandless, Lawrence (2009). *Bayesian methods for data analysis (3rd edn). Bradley P. Carlin and Thomas A. Louis, Chapman & Hall/CRC, Boca Raton, 2008. ISBN 9781584886976*.

Hua, Hairui, Danielle L. Burke, Michael J. Crowther, Joie Ensor, Catrin Tudur Smith, and Richard D. Riley (2017). "One-stage individual participant data meta-analysis models: estimation of treatment-covariate interactions must avoid ecological bias by separating out within-trial and across-trial information". eng. In: *Statistics in Medicine* 36.5, pp. 772–789. ISSN: 1097-0258. DOI: 10.1002/sim.7171.

Donegan, Sarah, Paula Williamson, Umberto D'Alessandro, and Catrin Tudur Smith (2012). "Assessing the consistency assumption by exploring treatment by covariate interactions in mixed treatment comparison meta-analysis: Individual patient-level covariates versus aggregate trial-level covariates". In: *Statistics in medicine* 31. DOI: 10.1002/sim.5470.

Hayward, Rodney A, Joel J Gagnier, Michael Borenstein, Geert JMG VanDerHeijden, Issa J Dahabreh, Xin Sun, Willi Sauerbrei, Michael Walsh, John PA Ioannidis, Lehana Thabane, et al. (2020). "Instrument for the Credibility of Effect Modification Analyses (ICEMAN) in randomized controlled trials and meta-analyses: manual Version 1.0". In.

Riley, Richard D, Paul C Lambert, Jan A Staessen, Jiguang Wang, Francois Gueyffier, Lutgarde Thijs, and Florent Boutitie (2008). "Meta-analysis of continuous outcomes combining individual patient data and aggregate data". In: *Statistics in medicine* 27.11, pp. 1870–1893.

Hong, Hwanhee, Haoda Fu, and Bradley P Carlin (2018). "Power and commensurate priors for synthesizing aggregate and individual patient level data in network meta-analysis". In: *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 67.4, pp. 1047–1069.

Saramago, Pedro, Alex J Sutton, Nicola J Cooper, and Andrea Manca (2012). "Mixed treatment comparisons using aggregate and individual participant level data". In: *Statistics in medicine* 31.28, pp. 3516–3536.

Kennedy, Edward H. (2020). "Optimal doubly robust estimation of heterogeneous causal effects". In: *arXiv:2004.14497 [math, stat]*. URL: http://arxiv.org/abs/2004.14497 (visited on 03/30/2022).

Nie, X and S Wager (2021). "Quasi-oracle estimation of heterogeneous treatment effects". en. In: *Biometrika* 108.2, pp. 299–319. ISSN: 0006-3444, 1464-3510. DOI: 10.1093/biomet/asaa076. URL: https://academic.oup.com/biomet/article/108/2/299/5911092 (visited on 03/30/2022).

Colnet, Bénédicte, Julie Josse, Erwan Scornet, and Gaël Varoquaux (2021a). "Causal effect on a target population: a sensitivity analysis to handle missing covariates". URL: https://hal.archives-ouvertes.fr/hal-03473691.

Audigier, Vincent, Ian R White, Shahab Jolani, Thomas PA Debray, Matteo Quartagno, James Carpenter, Stef Van Buuren, and Matthieu Resche-Rigon (2018). "Multiple imputation for multilevel data with continuous and binary variables". In: *Statistical Science* 33.2, pp. 160–183.

Jolani, Shahab, Thomas PA Debray, Hendrik Koffijberg, Stef van Buuren, and Karel GM Moons (2015). "Imputation of systematically missing predictors in an individual participant data meta-analysis: a generalized approach using MICE". In: *Statistics in medicine* 34.11, pp. 1841–1863.

Chapter 3

Comparison of Methods that Combine Multiple Randomized Trials to Estimate Heterogeneous Treatment Effects

Abstract:[1] Individualized treatment decisions can improve health outcomes, but using data to make these decisions in a reliable, precise, and generalizable way is challenging with a single dataset. Leveraging multiple randomized controlled trials allows for the combination of datasets with unconfounded treatment assignment to better estimate heterogeneous treatment effects. This paper discusses several non-parametric approaches for estimating heterogeneous treatment effects using data from multiple trials. We extend single-study methods to a scenario with multiple trials and explore their performance through a simulation study, with data generation scenarios that have differing levels of cross-trial heterogeneity. The simulations demonstrate that

[1]This chapter has undergone peer review and is published in *Statistics in Medicine*: Brantner, C. L., Nguyen, T. Q., Tang, T., Zhao, C., Hong, H., and Stuart, E. A. (2024). Comparison of methods that combine multiple randomized trials to estimate heterogeneous treatment effects. *Statistics in Medicine*.

methods that directly allow for heterogeneity of the treatment effect across trials perform better than methods that do not, and that the choice of single-study method matters based on the functional form of the treatment effect. Finally, we discuss which methods perform well in each setting and then apply them to four randomized controlled trials to examine effect heterogeneity of treatments for major depressive disorder.

3.1 Introduction

When tailoring treatment regimens to individual patients, one must strive to understand how different treatment options might affect the specific patient based on their characteristics or context. Rather than using a one-size-fits-all approach, clinicians and researchers are turning more towards personalized medicine with the goal of improving clinical outcomes. In this setting, the focus of estimation becomes conditional average treatment effects, i.e., how well the treatment is expected to work conditional on the person's known characteristics.

The benchmark for estimating treatment effects in an unbiased manner is most often a randomized controlled trial (RCT). In an RCT, participants are randomly assigned to treatment or control, therefore ensuring unconfounded treatment assignment and unbiased treatment effect estimates in the given sample. However, these trials often have sample sizes that are large enough to detect main effects but lack power to estimate heterogeneous treatment effects (Fleiss, 2011) and might not be representative of a broader population. To overcome these specific issues, researchers have started combining

information from multiple studies to improve treatment effect estimation. Multiple studies allow for larger sample sizes and at times a more representative sample of the target population. In the setting with multiple RCTs, meta-analysis or hierarchical models are common techniques to combine studies and estimate treatment effects (Debray et al., 2015; Seo et al., 2021). These approaches often do not explicitly target conditional average treatment effects though, and often only use aggregate-level data which makes it challenging to estimate treatment effects conditional on individual-level characteristics. Furthermore, meta-analysis is commonly applied within a parametric framework, which is highly interpretable but requires prespecification of effect moderators and distributional assumptions for parameters. Non-parametric approaches are worth exploring in this setting because they allow for high levels of flexibility in outcome and treatment effect functions. Relationships between covariates and treatment effect can be complex and non-linear in reality, and non-parametric machine learning methods can better handle those scenarios.

Many non-parametric approaches exist to estimate heterogeneous treatment effects (Künzel et al., 2019; Athey, Tibshirani, and Wager, 2019; Green and Kern, 2012; Kennedy, 2020; Nie and Wager, 2021; Dandl et al., 2022); however, these approaches have generally been developed only for the single-study setting. Several of the common approaches are discussed in the section to follow (3.3.1), and we subsequently extend these methods for use in multiple studies. Recent research has investigated a few non-parametric approaches for the multiple study setting, mostly geared towards combining data from one

RCT with a large observational dataset (Yang, Zeng, and Wang, 2020; Yang, Zeng, and Wang, 2022; Kallus, Puli, and Shalit, 2018; Rosenman et al., 2023). In that work, the focus is often on estimating the bias present in the observational data to determine the level at which the observational study estimates can be combined with the RCT estimates. These methods are therefore not as straightforward to use in the multiple RCT setting. With multiple RCTs, each individual trial has the benefit of unconfounded treatment assignment, but significant cross-trial heterogeneity could still exist due to both observed and unobserved factors. The focus in this case is no longer de-biasing one of the datasets, but instead determining the amount of heterogeneity present and how to account for it.

Brantner and colleagues (2023) wrote a comprehensive review of methods geared towards combining datasets to estimate treatment effect heterogeneity. That review included approaches for multiple RCTs; the most common were individual participant-level data one-stage meta-analyses (Debray et al., 2015). One alternative approach focuses on combining RCTs to estimate conditional average treatment effects in a non-parametric framework (Tan et al., 2022). However, that work by Tan and colleagues was done in the federated learning setting, in which individual-level data could not be shared across study sites and instead only aggregate results or models could be shared. In the sections to follow, we tailor Tan et al.'s method to when individual-level data can be shared across trials, and we add other new options for combining trials.

To our knowledge, this paper is the first to describe and compare machine learning options for estimating heterogeneous treatment effects using data

from multiple RCTs, in the setting in which all data can be shared across trials. Because not many methods exist to do this, we demonstrate several options for extending current methods for single studies to the multiple-study setting. We also build off of the approach in Tan et al. (2022) by adapting it to the case when individual-level data can be shared across trials. Our goals are to assess estimation accuracy of the various methods within a given sample of trials and to determine whether and when pooling data is useful, or if it might ever worsen accuracy in the presence of high heterogeneity across trials. We conduct extensive simulations with varying data generation parameters to determine which of the single-study and aggregation methods perform best depending on different amounts of cross-trial heterogeneity in the effects. We then apply the approaches to a set of four RCTs of depression treatments and discuss the variability in estimates across the approaches and potential substantive conclusions that can be made.

3.2 Notation

The estimand considered in this paper is the conditional average treatment effect (CATE), defined under Rubin's potential outcomes framework (Rubin, 1974). Let A denote a binary treatment indicator (often treatment versus control), $\boldsymbol{X}$ represent covariates, and Y represent a continuous outcome. Under Rubin's framework, $Y(0)$ and $Y(1)$ denote the potential outcomes under control and treatment, respectively. In other words, $Y(0)$ is the value of Y that an individual would have if they are in the control group, while $Y(1)$ is the value of Y that they would have if they received treatment. The fundamental

problem of causal inference is that we cannot ever observe both $Y(0)$ and $Y(1)$ simultaneously for the same person; therefore, we must use design and analysis approaches to estimate the unobserved outcomes. Next, let S be a categorical variable representing the trial in which the individual participated and ranging from 1 to K, where K is the total number of RCTs. Finally, represent the probability of receiving treatment given covariates and trial membership (propensity score) as $\pi_s(\boldsymbol{X}) = P(A = 1|\boldsymbol{X}, S = s)$.

With a continuous outcome, the CATE is

$$\tau(\boldsymbol{X}) = E(Y(1)|\boldsymbol{X}) - E(Y(0)|\boldsymbol{X}). \tag{3.1}$$

In this paper, we note that the goal estimand is this "universal" CATE (3.1) built off of potential outcomes that are not dependent upon study membership. However, many methods in the following sections target a study-specific CATE:

$$\tau_s(\boldsymbol{X}) = E(Y(1)|\boldsymbol{X}, S = s) - E(Y(0)|\boldsymbol{X}, S = s). \tag{3.2}$$

To identify the estimand when combining data across RCTs, many of the standard causal inference assumptions are required, including the Stable Unit Treatment Value Assumption (SUTVA) within each RCT. Other standard assumptions include: unconfoundedness (Assumption 3.1), consistency (Assumption 3.2) and positivity (Assumptions 3.3 and 3.4) (Tan et al., 2022). Assumption 3.2 varies slightly depending on the estimand; under the universal CATE estimand (Equation 3.1), we assume overall consistency, while under the study-specific estimand (Equation 3.2), we assume consistency within each

study. Assumption 3.4, which requires that any $\boldsymbol{X}$ is possible to be observed in all studies, can be relaxed depending on the method.

Assumption 3.1 $\{Y(0), Y(1)\} \perp\!\!\!\perp A \mid \boldsymbol{X}, S = s$ *for all studies s.*

Assumption 3.2 $Y = AY(1) + (1 - A)Y(0)$ *almost surely (in each study).*

Assumption 3.3 *There exists a constant* $c > 0$ *such that* $c < \pi_s(\boldsymbol{x}) < 1 - c$ *for all studies s and for all* $\boldsymbol{x}$ *values in each study.*

Assumption 3.4 *(Can be relaxed) There exists a constant* $d > 0$ *such that* $d < P(S = s|\boldsymbol{X} = \boldsymbol{x}) < 1 - d$ *for all* $\boldsymbol{x}$ *and s.*

3.3 Methods

This paper includes methods developed for treatment effect estimation in a single study and aggregation approaches that apply these methods to multiple studies. This section discusses three single-study methods and several aggregation options that apply the single-study methods to the multi-study setting.

3.3.1 Single-Study Methods

For a given RCT, many machine learning methods have been developed for CATE estimation. The single-study methods that exist can be grouped into multiple categories, as delineated by Brantner et al (2023). For ease of comparison, three approaches are included that are user-friendly and have been shown to be effective in previous literature: the S-learner, X-learner

(Künzel et al., 2019), and causal forest (Athey, Tibshirani, and Wager, 2019). We ultimately selected these three approaches because they represent two distinct classes of methods for estimating heterogeneous treatment effects (see Brantner et al., 2023) and seem to be used in practice, especially the causal forest (Athey and Wager, 2019; Jawadekar et al., 2023). Specifically, the first two approaches are multi-step procedures that involve first estimating the conditional outcome mean under treatment or control and then combining the two into one CATE function, while the causal forest involves tree-based partitioning of the covariate space by treatment effect. In this paper, we use random forests as the base learners for both the S-learner and the X-learner to best compare with the causal forest, which is inherently forest-based. These single-study methods are different from those explored by Tan and colleagues (2022); we chose to focus on the causal forest over a causal tree because the causal forest is an aggregation of multiple trees, and we added in the X-learner and S-learner to provide a different type of method to compare with.

3.3.1.1 S-Learner

The first single-study machine learning method used in this paper is called the "S-learner" (Künzel et al., 2019). This method is classified as a "meta-learner" in that it combines base learners (i.e., regression models) of any form in a specific way (Künzel et al., 2019). The S-learner uses a base learner (i.e., a random forest) to estimate a conditional outcome mean function given observed covariates and assigned treatment:

$$\mu(\boldsymbol{X}, A) = E(Y|\boldsymbol{X}, A).$$

The conditional outcome mean function in this approach is not specific to treatment group, but instead treatment is included together with the covariates as features to be used by the random forest. The CATE can then be directly estimated by plugging in 0 and 1 for the treatment indicator to obtain predicted outcomes under treatment and control for each individual and calculate

$$\hat{\tau}(\boldsymbol{X}) = \hat{\mu}(\boldsymbol{X}, 1) - \hat{\mu}(\boldsymbol{X}, 0).$$

3.3.1.2 X-Learner

The second approach considered here is another meta-learner called the "X-learner" (Künzel et al., 2019). The X-learner takes a similar approach as the S-learner by modeling the conditional outcome mean functions before estimating the CATE directly. However, rather than estimating one outcome mean function for $Y(1)$ and $Y(0)$ simultaneously, the X-learner estimates two functions separately and then imputes treatment effects for each treatment group.

Specifically, the X-learner involves three steps. First, the conditional outcome mean functions are estimated using base learners (in this case, random forests) like in the S-learner, but separately by treatment group:

$$\mu_0(\boldsymbol{X}) = E(Y(0)|\boldsymbol{X}) \quad \text{and} \quad \mu_1(\boldsymbol{X}) = E(Y(1)|\boldsymbol{X}).$$

Next, the unobserved potential outcomes for individuals in the treatment and control groups are predicted using those models to get $\hat{\mu}_0(\boldsymbol{X}_{i:A=1})$ (estimate of the potential outcome under control for an individual who received treatment) and $\hat{\mu}_1(\boldsymbol{X}_{i:A=0})$ (estimate of the potential outcome under treatment for an

individual who received control). We then input these predictions along with the observed outcomes to impute individual treatment effects:

$$\tilde{D}_{i:A=1} = Y_{i:A=1} - \hat{\mu}_0(\boldsymbol{X}_{i:A=1}) \quad \text{and} \quad \tilde{D}_{i:A=0} = \hat{\mu}_1(\boldsymbol{X}_{i:A=0}) - Y_{i:A=0}.$$

Then $\tilde{D}$ is regressed on $\boldsymbol{X}$ to estimate $\tau(\boldsymbol{X})$. This is done within each treatment group separately, resulting in two estimates, labeled $\hat{\tau}_1(\boldsymbol{X})$ and $\hat{\tau}_0(\boldsymbol{X})$. Finally, these are combined to obtain one estimate of the CATE function:

$$\hat{\tau}(\boldsymbol{X}) = g(\boldsymbol{X})\hat{\tau}_1(\boldsymbol{X}) + (1 - g(\boldsymbol{X}))\hat{\tau}_0(\boldsymbol{X}),$$

where the weight $g(\boldsymbol{X})$ is often an estimate of the propensity score (the case in this paper) or can be chosen otherwise (Künzel et al., 2019).

3.3.1.3 Causal Forest

The third single-study approach is the causal forest (Athey, Tibshirani, and Wager, 2019). The causal forest is similar to a random forest, but the focal estimand is the treatment effect itself, rather than the outcome for a given individual. The causal forest is based off of a causal tree, which involves recursive partitioning of the covariates to best split based on treatment effect heterogeneity. Here, the treatment effect is estimated as the difference in average outcomes between the treatment and control group individuals within leaves. From there, the causal forest is the weighted aggregation of many causal trees.

One potential challenge with causal forests is that bias could occur when there is overlap between the data used to form the trees and data used to

estimate the treatment effects within leaves. A solution to that problem, called "honesty", has been proposed (Wager and Athey, 2018). This concept ensures that for every individual involved in creating the tree, their outcome is used either for splitting the tree or estimating the treatment effect within a leaf, but not both. Honesty has been used some in the literature, but there is not a widespread conclusion as to whether trees should be fit with or without honesty depending on the scenario. Dandl and colleagues compared honesty versus adaptive (not honest) forests in their simulations including causal forests and found that in their setting that was meant to represent an RCT, the adaptive forests performed better (Dandl et al., 2022). Additionally, honesty requires large sample sizes. Thus, we do not include honesty in the causal forests in the primary simulations but do investigate it in a second round of method comparisons.

3.3.2 Aggregation Methods

In many contexts, there are multiple RCTs available that compare the same two treatments. It is then worth considering methods that allow combining across trials. When aggregating to the multi-study level, the question becomes: how much does the treatment effect vary based on study membership? This variability can range along a continuum, where on one end is the possibility that the trials are all very homogeneous in terms of the CATE, meaning that participants in trial j and in trial k who have the same covariate values would have the same treatment effect. At the other extreme, individuals with the same covariates but in different trials could have completely different

treatment effects. These differences can be due to heterogeneity in the sites in which the trials were conducted, heterogeneity in trial procedures (including the treatment or control conditions themselves), heterogeneity in trial samples, or other reasons. The aggregation methods to follow take different approaches to incorporating trial membership into the treatment effect estimation, ranging from assuming trial membership does not matter at all, to allowing it to matter just as much as any other characteristic.

3.3.2.1 Complete Pooling

A complete pooling approach is very straightforward: the researcher simply takes all data from each of the K RCTs, creates a single dataset, and then fits one of the three previously described methods (S-learner, X-learner, or causal forest) to the pooled dataset. This approach is quick and easy to do, but requires many assumptions. Namely, this approach assumes a high level of homogeneity across trials and specifically that the CATE function is shared across studies. This method is included because it represents a naive comparison point and because it provides universal CATE estimates (i.e., not study-specific).

3.3.2.2 Pooling with Trial Indicator

An alternative pooling approach is to incorporate trial membership in the models but essentially still perform the pooling as before. Here, all of the individual data from each RCT is combined into one comprehensive dataset, but a categorical variable is included that represents the trial in which the individual participated. Then, the researcher can apply one of the single-study

approaches to this full dataset, allowing for all of the covariates, including trial membership, to be involved in the treatment effect function. In this way, if trial membership is important for estimating effects, estimates should be somewhat informed by trial membership; otherwise, the treatment effect estimates should be similar across trials. While the previous complete pooling approach gives estimates that were not trial-specific, this approach yields trial-specific CATE estimates.

3.3.2.3 Ensemble Approach

The next approach is based off of Tan and colleagues' (2022) methods for federated learning, originally developed for scenarios in which individual data cannot be shared across trial sites. Their original approach fits trial-specific models and then applies those models to data from a single coordinating site to derive an ensemble. We propose an adaptation of Tan's approach for settings where individual-level data from all trials are available to the analyst.

This adaptation of Tan et al.'s approach involves three steps.

1. First, the researcher builds localized models for the CATE within each trial, using one of the three single-study methods previously discussed (S-learner, X-learner, or causal forest).

2. Next, they apply these localized models to each individual across all of the RCTs to get for each individual their *trial-specific CATE estimates*, i.e., the estimated effects had the individual been part of study 1, study 2 and so on. For K studies with a total of N individuals in all studies combined, there will be K trial-specific CATE models. Once each of

these models are applied to all N data points, every individual will have K different estimates of their CATE. So there will ultimately be $N \cdot K$ CATE estimates in what Tan et al. define as an "augmented" dataset. The difference between the second step here and what Tan et al. did is that we apply the study-specific models to all data points in all trials, rather than having to restrict to a single coordinating site.

3. The third and final step is to fit an ensemble model to the augmented dataset that has CATE estimates for every individual crossed with every trial. In this ensemble model, the response variable is the CATE estimate, and the predictors are the individual covariates and a categorical variable indicating the local model that had been used to compute the CATE estimate. We use three different options for this final ensemble model fit to the augmented dataset: a regression tree, a random forest, and a lasso regression. The regression tree and random forest were explored in Tan et al.'s paper (2022), while we added lasso regression to provide a parametric comparison point.

The resulting functions from these ensemble approaches are trial-specific estimates of the CATE; however, they have been adapted based on the CATEs from the other trials. Therefore, this method allows for trial heterogeneity but incorporates information across trials to hopefully improve the model from each trial.

3.3.2.4 IPD Meta-Analysis

As a comparison point in the simulations to follow, we also include an individual patient-level data (IPD) meta-analysis with a random intercept for trial membership. This method is a standard approach taken by researchers when combining multiple RCTs and assessing treatment effects (Debray et al., 2015; Seo et al., 2021; Burke, Ensor, and Riley, 2017), and it also serves here as a parametric comparison to the primarily non-parametric approaches outlined above. A meta-analysis does not employ a single-study method like the S-learner, X-learner, or causal forest; instead, all of the data is pooled together and trial-level relationships can be included as fixed or random effects. The decision of how to parametrize a given meta-analysis is very important and can have major implications as to the assumptions of how the true data is distributed and the subsequent fit of the model. While the previous non-parametric approaches implicitly allow for any important moderating relationships and interactions to be picked up based on the modeling procedure, meta-analysis requires that we pre-specify moderation according to a priori hypotheses. In this paper, we set up the meta-analysis to mimic the setup of the first scenario in the simulation to follow except for the exact form of the moderator, so that we can see how well meta-analysis performs when it is mostly correctly specified versus when it is incorrectly specified (for the second and third CATE scenarios described in the simulations below). The model is as follows:

$$Y = (\alpha_0 + a_s) + (\alpha_1 + b_s)X_1 + \alpha_2 X_2 + \alpha_3 X_3 + \alpha_4 X_4 + (\zeta + z_s)A + (\theta + t_s)X_1 A + \epsilon.$$

In this model, we allow the intercept to include a fixed component (α_0) and a random component by study ($a_s \sim N(0, \sigma_a^2)$), and our residual error is $\epsilon \sim N(0, \sigma^2)$. The fixed effects are $\boldsymbol{\alpha} = \{\alpha_1, \alpha_2, \alpha_3, \alpha_4\}$, the coefficients relating the covariates to the outcome; ζ, the coefficient for treatment; and θ, the coefficient of the interaction between treatment and a moderator X_1 (Seo et al., 2021). The random effects by study are $b_s \sim N(0, \sigma_b^2)$, the random slope for the covariate X_1; $z_s \sim N(0, \sigma_z^2)$, the random slope for treatment; and $t_s \sim N(0, \sigma_t^2)$, the random slope for the treatment-X_1 interaction term. From here, the estimate of the conditional average treatment effect can be calculated as $\hat{\tau}_s(\boldsymbol{X}) = (\hat{\zeta} + \hat{z}_s) + (\hat{\theta} + \hat{t}_s)X_1$.

The meta-analysis framework assumes that the CATE function is shared across studies, but that the mean potential outcome under control can differ across studies. Notably, this functional form of the CATE assumes linear relationships, and one must prespecify all variables that might be relevant to the main effect of the covariates and to the treatment effect.

3.3.2.5 No Pooling

Finally, we also consider that there might be instances where trials are too heterogeneous to reliably combine information across trials. When this is the case, fitting models within each study would be the best approach; therefore, we include this option in our simulations as well. For this "no pooling" approach, one can fit a single model within every trial separately using a single-study method previously introduced, and CATEs can be estimated within each study using the given study's model. We provide results from this

method in the simulations to investigate if there are settings when pooling worsens estimation accuracy. However, it is important to mention that this approach is not technically an "aggregation approach" because it analyzes each study independently from the others and does not use data from multiple studies together. Particularly in the simulations to follow, the no pooling approach will find the best fit within each study and should therefore yield consistently high estimation accuracy. There will also be some differences in terms of variance; we assume that there would be higher variance when using only one study, but we do not explore this explicitly here. Note, though, that the current setup does not examine how well this approach will predict CATEs for individuals outside of the specific trials; we elaborate on this more in the sections to follow.

3.4 Simulation Setup

To compare both the single-study and aggregation methods, we performed a simulation study, simulating data from multiple randomized controlled trials and changing parameter values to compare which methods achieve the lowest mean squared error (MSE) between the estimated and true individual CATEs. Because there were three single-study methods (S-learner, X-learner, and causal forest) and six aggregation methods (complete pooling, pooling with trial indicator, ensemble tree, ensemble forest, ensemble lasso, and no pooling) being compared along with meta-analysis, there were $3 \cdot 6 + 1 = 19$ total combinations of methods applied to each simulated dataset.

3.4.1 Data Generating Mechanism

In the simulations to follow, the potential outcomes are generated using the following model (Tan et al., 2022):

$$Y_i(a) = m(\boldsymbol{x}_i, s_i) + \frac{2a-1}{2} \cdot \tau(\boldsymbol{x}_i, s_i) + \epsilon_i \tag{3.3}$$

where $m(\boldsymbol{x}_i, s_i)$ represents the outcome mean conditional on covariates and trial, and $\tau(\boldsymbol{x}_i, s_i)$ is the CATE. In the main setting for the data generation, we employed two options for m and τ. The first setup (1a) involves a linear m and piecewise linear τ, based on a similar setup by Tan et al. (Tan et al., 2022):

$$m(\boldsymbol{x}, s) = x_1/2 + \sum_{j=2}^{4} x_j + \beta_s + \delta_s \cdot x_1 \text{ and } \tau(\boldsymbol{x}, s) = x_1 \cdot I(x_1 > 0) + \beta_s + \delta_s \cdot x_1.$$

The second setup (1b) involves a more complicated non-linear function for τ, derived partially from a simulation setting by Kunzel et al. (2019):

$$m(\boldsymbol{x}, s) = 0 \text{ and } \tau(\boldsymbol{x}, s) = g(x_1)g(x_2) + \beta_s + \delta_s \cdot x_1$$

where $g(x) = \frac{2}{1+\exp(-12(x-1/2))}$. In both of these, the coefficients β_s represent trial-specific main effect coefficients, and δ_s represent trial-specific interaction effect coefficients (interaction between trial and the moderator x_1). In both setups, x_1 is an effect moderator, and in the second setup, x_2 is as well. If the coefficients β_s and δ_s differ across s (i.e., trial membership), then trial is making an impact in the moderation.

From this information, the components simulated are listed as follows:

1. For each simulation, the number of trials was $K = 10$.

2. Each trial had a sample size of 500 individuals.

3. Within each trial, we simulated five continuous covariates per person X_i, $i \in \{1, 2, 3, 4, 5\}$, where $E(X_i) = 0$, $Var(X_i) = 1$, and $Cov(X_i, X_j) = 0.2$ for all $i \neq j$.

4. Each person was then assigned a treatment status, 0 or 1, according to a propensity score of $\pi_i = 0.5$ within each trial.

5. Each person was also assigned an error term for their outcome function, so $\epsilon_i \sim N(0, 0.01)$.

6. We then sampled trial-specific main effect and interaction effect terms. Each of the $K = 10$ studies was assigned a main effect term according to $\beta_s \sim N(0, \sigma_\beta^2)$ and an interaction effect term according to $\delta_s \sim N(0, \sigma_\delta^2)$. The values of the standard deviations were:

$$(\sigma_\beta, \sigma_\delta) \in \{(0.5, 0), (1, 0), (1, 0.5), (1, 1), (3, 1)\}.$$

7. From this information, m, τ, and Y were calculated under either of the two setups described above (1a and 1b).

We then included some variations of the above setup to assess method performance under different adjustments. The first was including one other scenario (2) to see how the methods would perform when the functional form of the CATE itself differed across trials – a particularly challenging situation for pooling. For this scenario, we used the same form for Y_i as in Equation (3.3), and now we set m and τ to be such that m is linear and τ depends on

study:

$$m(\boldsymbol{x}, s) = x_1/2 + \sum_{j=2}^{4} x_j,$$

$$\tau(\boldsymbol{x}, s) = I(s \in \{1,2,3,4\}) \cdot g(x_1)g(x_2) + I(s \in \{5,6,7,8\}) \cdot x_1 \cdot I(x_1 > 0)$$

$$+ I(s \in \{9,10\}) \cdot 0$$

where $g(x)$ is as previously defined.

We also added settings with variation in the trial sample sizes. One new option involved one large trial (n=1000) and the rest smaller (n=200). The second new setting had half of the trials with n=500 and the other half with n=200. We assessed performance for these sample size adjustments under scenarios 1a, 1b, and 2 with trial main and interaction coefficient standard deviations of 1 and 0.5, respectively.

We then investigated the impact of covariate shift on method performance. In particular, we generated the data such that all even numbered studies had X_1 with mean 0 as above, but in odd numbered studies, the mean of X_1 was set to be 2. We assessed this setting under scenarios 1a and 1b with standard deviations of 1 and 0.5 of study main and interaction effect terms, and we allowed trial sample sizes to either all be the same or for one trial to be large and the rest smaller.

Finally, we added some simulations with $K = 30$ trials in two of the settings (scenario 1a and 1b with trial main and interaction coefficient standard deviations of 0.5 and 0, respectively) to determine if there were differences in performance based on number of trials (the remainder of the simulations had

$K = 10$).

For each simulation setup, we generated 1,000 simulated datasets. Necessary packages included causalToolbox for the S-learner and X-learner (Künzel et al., 2019), grf for the causal forest (Athey, Tibshirani, and Wager, 2019), rpart for the ensemble tree (Therneau et al., 2015), ranger for the ensemble forest (Wright and Ziegler, 2017), glmnet for the ensemble lasso (Friedman et al., 2017), and lme4 for the mixed effects meta-analysis (Bates, 2010). Ensembling functions were based off of those in the ifedtree package (Tan et al., 2022) but were adapted to the setting in which data could be shared across trials. In all non-parametric approaches, hyperparameters were set to be the defaults, except that the causal forest was set to use 1,000 trees instead of the default of 2,000 for computational ease, and honesty was set to false for the preliminary simulations. For each method and each iteration, performance of the different approaches was assessed based on the mean squared error (MSE) between the true individual CATEs and the estimated individual CATEs, and these MSEs were ultimately averaged across the 1,000 repetitions. Code containing all adapted methods and implementation of the simulations can be found at the github repo: `https://github.com/carlyls/CATE_multiRCT`.

3.5 Simulation Results

The following tables and figures display the performance results across 1,000 iterations of each parameter combination/scenario. Figure 3.1 displays the distribution of MSE for every approach for the two main scenarios (piecewise linear and non-linear CATE), broken down by the standard deviations of

the trial main and interaction effects. In the piecewise linear and non-linear CATE scenarios, as the trial coefficients (both main and interaction effects) increase in variability, the MSE increases, meaning the methods estimate individual CATEs more poorly. This is consistent with the idea that when trial membership is involved in the treatment effect function, the CATEs vary across trials and therefore are harder to estimate when data is pooled across studies. Notably, this increase in MSE happens much more quickly for the complete pooling approaches.

In the piecewise linear scenario (1a), the most consistently effective approaches in terms of MSE are when the causal forest is used as the single-study method and when the aggregation approach is either pooling with trial indicator or ensemble forest. The X-learner also performs relatively well in terms of MSE. Meta-analysis performs well, which is expected because the model was set up to mostly match the true functional form of the CATE in this scenario. For the non-linear scenario (1b), the ensemble lasso and meta-analysis perform notably worse, which makes sense due to the complexity of the functional form of the CATE, as it includes the product of two expit functions, and the lasso and meta-analysis assume a parametric linear relationship between covariates and outcome. The ensemble forest and pooling with trial indicator again estimate the CATEs well, with all single-study methods performing more similarly. While the S-learner was not very effective with the piecewise linear CATE (1a), it was more effective with the non-linear CATE (1b). In all main settings, the no pooling approach performs similarly well to pooling with trial indicator and ensemble forest (Figure B.1); we discuss more about

Figure 3.1: Distribution of MSE for main parameter combinations across all single-study and aggregation approaches.

Columns are broken down by simulation scenarios (piecewise linear versus non-linear CATE), and rows are by standard deviation of study main and study interaction coefficients.

this in the Discussion section.

Several boxplots in the Appendix display the results of the many variations upon the original simulation settings included. To assess the performance of methods with different trial sizes, Figure B.2 demonstrates that there do

not seem to be notable differences in patterns across methods depending on whether all trials have the same sample size, one trial is much larger, or half are larger while half are smaller. The MSE seems to be slightly higher overall when trial sizes are different, but not substantially different. Furthermore, Figure B.3 displays the results for the variable CATE scenario (2). Here, the causal forest is clearly performing the best of the three single-study methods, while the S-learner is not performing as well. The most effective aggregation methods are again pooling with trial indicator and ensemble forests, and meta-analysis performs relatively poorly.

When we introduced a shift in the covariate distributions between even versus odd numbered studies (Figure B.4), there again does not seem to be a difference in the patterns of results. The MSE generally is slightly higher across all methods compared to when the covariates all came from the same distributions across trials; however, methods like the causal forest with pooling with trial indicator and ensemble forest still perform consistently well. In the piecewise linear CATE with a shift in covariate distributions, meta-analysis performs very well and the best of all aggregation approaches, but it does not perform well when the CATE is non-linear.

Finally, for the two scenarios with 30 trials instead of 10, Figure B.5 demonstrates that the results and patterns are all similar to the results for K=10, except for the causal forest with pooling with trial indicator. Interestingly, this approach, which performed very well with 10 trials, has high MSE when there are 30. To understand this more fully we did further investigations, including some iterations with 15, 20, and 25 trials to see how the pattern

changes. Overall, the results of these investigations indicate that when there are more trials, the causal forest with pooling with trial indicators has more difficulty identifying the heterogeneity that exists across trials. In particular, the method rarely "picks up" the trial indicators of trials that do have different patterns in effects when $K > 20$, as indicated by the variable importance measures (weighted sum of the number of times the variable was used in a split at each level of the forest) (Athey, Tibshirani, and Wager, 2019). Table B.1 shows average variable importance values under the piecewise linear CATE scenario for different values of K. Based on the simulation setup, the causal forests should split often on moderating variables, which in this case are X_1 and study membership. The variable importance measures demonstrate that for all values of K, X_1 is involved in a high proportion of splits, as it should be as a moderator. For lower numbers of trials ($K = 10$ through around 20), the most heterogeneous studies (defined based on main coefficients) had notable variable importance, meaning they were involved in some of the splits in the causal forest. However, for higher values of K (more trials), the variable importance for these most heterogeneous studies approached zero, meaning study membership was no longer picked up much in the causal forest even though there was notable heterogeneity of the treatment effect based on study membership. In addition, for high values of K, the causal forest split more often on the non-moderating covariates, $X_2 - X_5$. These issues that arose with large numbers of trials likely contributed to the high MSE of the causal forest with pooling with trial indicator for large values of K. We reflect more on these results in the Discussion section.

To more formally examine the results of the main settings in our simulation, we regressed the average MSE across iterations on the methods and parameter combinations, just within the piecewise linear and non-linear CATE scenarios, excluding meta-analysis and no pooling, and excluding the settings with $K = 30$ and with covariate shift. Specifically, the regression is such that:

$$\begin{aligned} MSE =& \beta_0 + \beta_1 \cdot singlestudy + \beta_2 \cdot aggregation + \beta_3 \cdot singlestudy \cdot aggregation \\ &+ \beta_4 \cdot main_{sd} + \beta_5 \cdot interaction_{sd} + \beta_6 \cdot scenario + \beta_7 \cdot trialsizes + \epsilon. \end{aligned}$$

From this regression, there were no significant differences in performance across single-study methods, but all aggregation methods performed significantly better than complete pooling. The ensemble forest had the best average MSE for the S-learner and X-learner, and pooling with trial indicator had the best average MSE for the causal forest.

Finally, we also performed 500 more iterations using the same methods previously described, but with honest causal forests instead of traditional "adaptive" causal forests. These iterations were performed using the main data generation setups as above, except that covariates were not correlated. The resulting average MSEs are presented in the Appendix (Figure B.6). We found very similar results to the original 1,000 repetitions with adaptive causal forests, but the honest causal forests had slightly higher MSEs on average, indicating worse estimation accuracy than the adaptive causal forests. For the ensemble tree, forest, and lasso, the honest causal forests had slightly higher average MSE compared to the X-learner (Figure B.6), while the adaptive causal forests had slightly lower average MSE compared to the X-learner for

these same aggregation approaches in the original simulations. However, these differences are very small, so we can broadly make similar conclusions whether we use adaptive or honest causal forests in these scenarios.

3.6 Application to Real Dataset

After the simulations demonstrated differences across methods in several data generation setups, we applied the various methods to an existing dataset containing multiple randomized controlled trials that compared the same two medications.

3.6.1 Treatments for Major Depressive Disorder

The applied dataset used in the current paper consists of four randomized controlled trials (Mahableshwarkar, Jacobsen, and Chen, 2013; Mahableshwarkar et al., 2015; Boulenger, Loft, and Olsen, 2014; Baldwin, Loft, and Dragheim, 2012), each of which included three treatments: duloxetine, vortioxetine, and placebo, where duloxetine and vortioxetine are both treatments for major depressive disorder (MDD). At the time of the trials, duloxetine had been more commonly used to treat MDD so was primarily included in the trials as an active reference, while vortioxetine was a newer treatment not yet marketed (Schatzberg et al., 2014). Each of the four trials compared at least two different dosages of vortioxetine and therefore had more participants taking vortioxetine as opposed to duloxetine or placebo. For the purposes of the current application, we removed placebo participants and lumped all dosages of vortioxetine together to investigate the potential differences between the

efficacy of the active medications (duloxetine and vortioxetine), as well as identify features that might be moderating this difference.

Participants in each of the four trials shared similar eligibility criteria. All four trials required patients to be between the ages of 18 to 75, to have a Major Depressive Episode (MDE) as a primary diagnosis according to the DSM-IV-TR criteria over at least three months, and to have a Montgomery-Asberg Depression Rating Scale (MADRS) (Montgomery and Åsberg, 1979) score of at least 22 (one trial) or 26 (three trials) at both screening and baseline (Mahableshwarkar, Jacobsen, and Chen, 2013; Mahableshwarkar et al., 2015; Boulenger, Loft, and Olsen, 2014; Baldwin, Loft, and Dragheim, 2012). A primary outcome in the trials is the change in MADRS (Montgomery-Asberg Depression Rating Scale) score from baseline to the last observed follow-up in the study. Participants were meant to stay in the study for 8 weeks, at which point their final MADRS score was collected. For those who did not remain in the trial for 8 weeks, a last observation carried forward imputation approach was used for simplicity. This imputation approach is not the best way to account for missing data and many other options exist (Little et al., 2012), but it is used here for simplicity because this example is primarily illustrative. Predictors/effect modifiers used in the models were age, sex (female or male), smoking status (ever smoked or never smoked), weight, baseline MADRS score, baseline HAM-A (Hamilton Anxiety Rating) score (Hamilton, 1959), co-morbidity indicators (if ever had diabetes mellitus, hypothyroidism, anxiety), and medication indicators (if they are concomittantly taking an antidepressant, antipsychotic, thyroid medication). Since the outcome is the difference

in MADRS score (MADRS at follow up minus MADRS at baseline), a more negative outcome indicates a better result. We removed individuals who were either in the placebo group or who had missing treatment assignment, along with individuals with missing baseline MADRS or no post-randomization MADRS. After this, sample sizes were 575, 436, 418, and 418 for each of the trials. Further descriptive information about the samples in the four RCTs is reported in the Appendix (Table B.2). Little missing covariate data was present in the sample; however, conditional mean imputation was performed for missing values of weight (n=1) and baseline HAM-A score (n=2).

Following data preparation, we used each of the aforementioned method combinations (i.e., causal forest, S-learner, and X-learner as single-study methods paired with complete pooling, pooling with trial indicator, ensemble tree, ensemble forest, and ensemble lasso) to estimate the CATEs for every individual across the four trials. We then compared the CATE estimates across methods to see their concordance levels. Notably, it is not possible to compare the method performances with the truth, as the true CATEs are unknown in this real dataset.

3.6.2 Results

All methods broadly led to the conclusion of a positive average CATE. This indicates that vortioxetine is estimated to have less of a beneficial effect on the MADRS score on average. In each of the four RCTs, both treatments were associated with a reduction in depressive symptom severity over time (shown through a reduction in MADRS score), but this reduction was smaller

for the vortioxetine group than the duloxetine group. Table 3.1 contains the mean and standard deviation of the CATEs according to each method. Broadly, the S-learner approaches estimated lower CATEs on average than the other approaches, and there is some consistency between the aggregation approaches within each single-study method (S-learner, X-learner, and causal forest). There were especially high levels of similarity in the average CATE estimates across the causal forest methods, shown in the last column of Table 3.1. The variability of the CATE estimates differs depending on the approach as well; causal forest approaches had higher standard deviations than approaches that used the S-learner and X-learner. Complete pooling also yielded the highest standard deviations for CATE estimates out of all of the aggregation approaches. As a comparison point, we used a multiple linear regression with a random effect for trial to estimate an average treatment effect of 2.49 (SE = 0.49), which is similar to the averages of the CATEs according to the X-learner and causal forest approaches.

	S-Learner	**X-Learner**	**Causal Forest**
Complete Pooling	1.38 (1.6)	2.57 (1.4)	2.37 (2.8)
Pooling with Trial Indicator	0.91 (1.3)	2.52 (1.3)	2.37 (2.7)
Ensemble Tree	0.89 (1.3)	2.35 (1.5)	2.23 (2.5)
Ensemble Forest	0.89 (1.1)	2.36 (1.4)	2.30 (2.2)
Ensemble Lasso	0.89 (1.2)	2.32 (1.4)	2.23 (2.1)

Table 3.1: Mean (SD) of CATEs from all individuals in sample according to different single-study and aggregation method combinations.
The CATEs are individual-level estimates that indicate the difference in the estimated effect of vortioxetine versus duloxetine on the difference in MADRS score for a given patient. A positive CATE indicates that vortioxetine is estimated to have a smaller reduction of the MADRS score.

We then focused in on results from the causal forest with pooling with trial indicator approach, since that approach performed the best on average in

the simulations when there were not a large number of trials being combined. The CATE estimates and their 95% confidence intervals from this approach are displayed in Figure 3.2. These confidence intervals were calculated based on variance estimates provided through the grf package, where variance is calculated based on comparison of individual CATE predictions within and across small groups of fitted causal trees (Athey, Tibshirani, and Wager, 2019). These estimates support that the majority of individuals have a positive CATE estimate, but they also display very high levels of uncertainty, with all confidence intervals including zero.

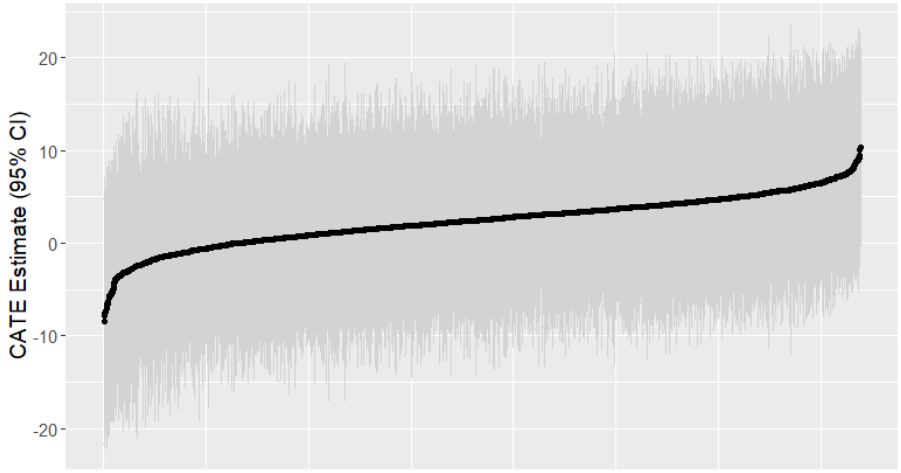

Figure 3.2: Point estimates and 95% confidence intervals for CATEs according to causal forest with pooling with trial indicator.

To learn more about the moderation within the CATE model, we can explore variable importance measures. As previously mentioned, variable importance from the grf package (Athey, Tibshirani, and Wager, 2019) is a weighted sum of the number of times the variable was used in a split at each level of the forest. We can investigate the variable importance measures

according to the grf package for all covariates, first in separate causal forests for each study (Figure 3.3), and second according to the causal forest with pooling with trial indicator (Figure 3.4). From Figure 3.3, there are a few variables that are consistently identified as effect moderators across studies (age, weight, baseline MADRS score, and baseline HAM-A score), and there are several that are not found to be major moderators (the comorbidity and medication indicators). However, notably there are some differences according to the separate models, indicating that the treatment effect functions are slightly different within each study. Figure 3.4 then displays the resulting importance measures from one aggregation model fit to all studies. Here, we can see that the same four variables (age, weight, baseline MADRS, and baseline HAM-A) are involved in a high proportion of the splits in the causal forest, and study membership is involved in some splits as well. The fact that these study indicators are not more highly involved in the partitioning of the treatment effect is a good sign, though, that there is not a very high level of heterogeneity in CATE estimates across studies.

The variable importance plots do not demonstrate the direction of the moderating effect, however. We briefly investigate these directional effects through an interpretation tree (Figure 3.5) and through exploratory plots such as Figure B.7. This interpretation tree was formed by fitting a regression tree, where the CATE estimates according to the causal forest with pooling with trial indicator were the outcomes, and the features (predictors) were every covariate in the original CATE model. The tree confirms what was shown in Figure 3.4 – that age, weight, baseline MADRS, and baseline HAM-A score

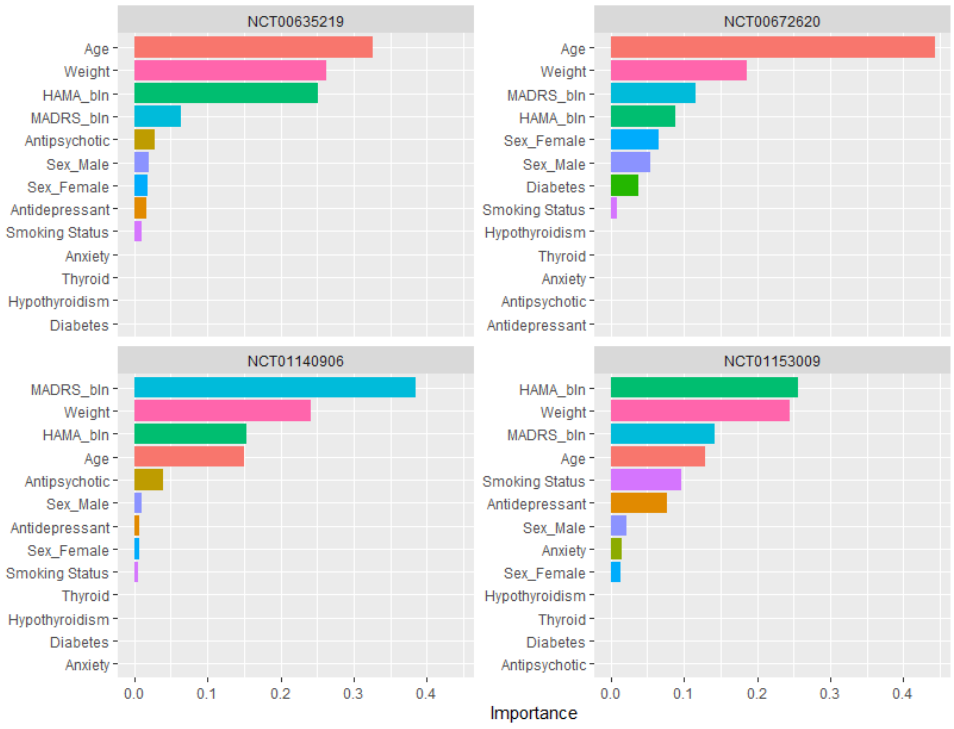

Figure 3.3: Variable importance for study-specific causal forest models.

are the strongest effect moderators. Study membership does not show up in this interpretation tree, supporting that there is low heterogeneity across trials. This is a helpful visual to see the direction of the relationships aggregated across trials, but it is exploratory and should not be interpreted in great detail. Another similar approach for investigating the CATE function in terms of individual moderators is to fit the best linear projection of the CATE estimates using a function in the grf package (Athey, Tibshirani, and Wager, 2019); the resulting coefficients from this regression using doubly-robust estimates of the CATE are reported in Table B.3.

Broadly, these interpretations of the CATE function derived by the causal

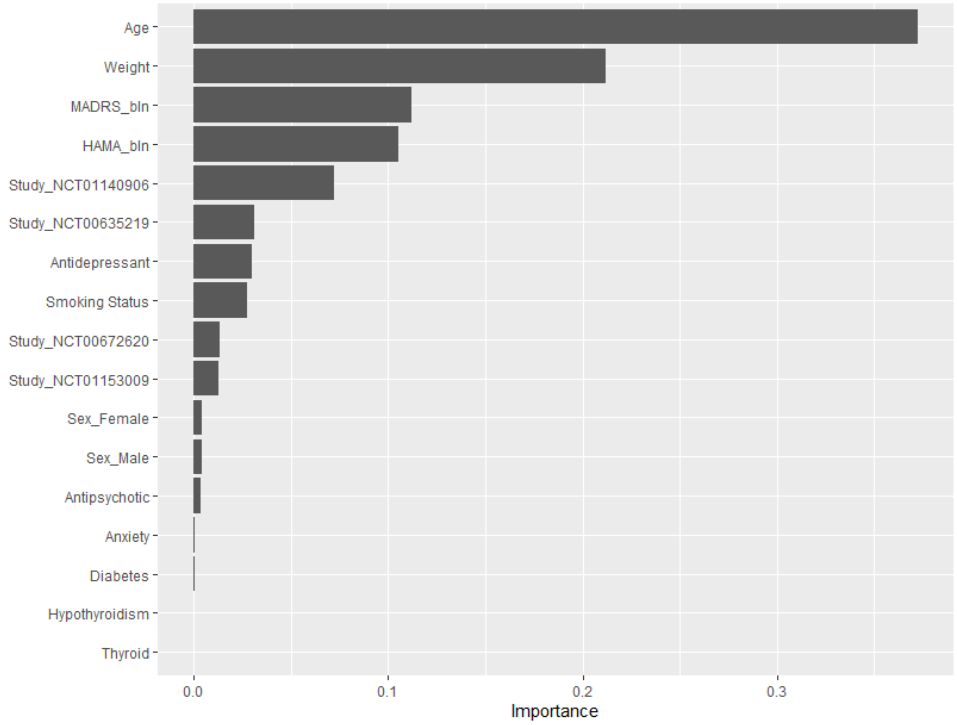

Figure 3.4: Variable importance for causal forest with pooling with trial indicator.

forest with pooling with trial indicator do not display high levels of heterogeneity, with the exception of potential heterogeneity by age. The scatterplot of CATE estimates by age in Figure B.7 and the best linear projection summarized in Table B.3 indicate somewhat higher CATE estimates for older individuals; however, there are very high levels of uncertainty in the confidence intervals (Figure 3.2). Other than this potential moderation by age, there does not appear to be heterogeneity across other variables, and in general we suggest further study, perhaps using more trials or observational data, to assess whether this age relationship is truly strong.

We also can compare the results of these pooled non-parametric methods with a more standard approach – IPD meta-analysis. In particular, we fit

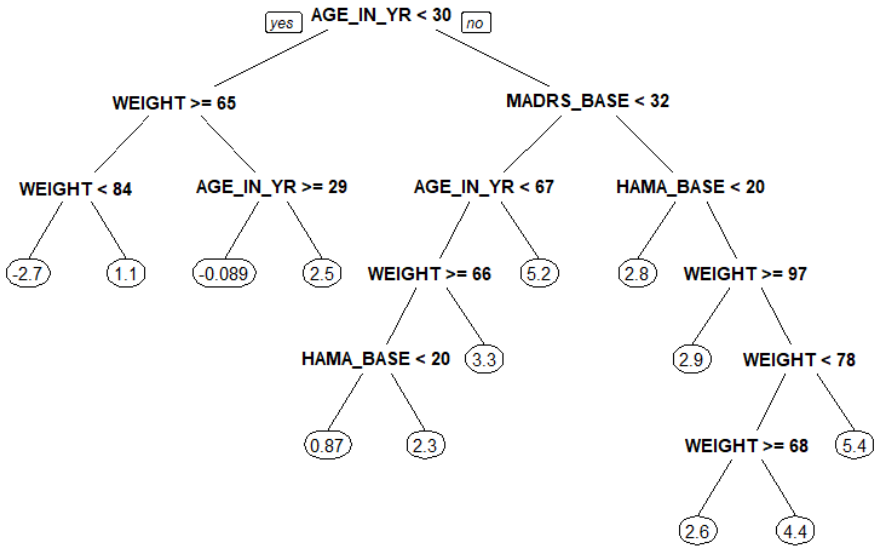

Figure 3.5: Interpretation tree for causal forest with pooling with trial indicator. Circled numbers represent the average CATE estimate for individuals in that leaf.

a linear regression with random effects for trial membership and included interaction terms to investigate potential moderation and compare results to the causal forest with pooling with trial indicator. As previously mentioned, the IPD meta-analysis yielded an average treatment effect estimate of 2.49 (SE = 0.49). To go a step further, we added interaction terms between treatment and each covariate in separate models to determine whether any interaction terms were significant. None were, although the interaction for age was close to significant (95% CI: (-0.01, 0.14)), which is consistent with our findings in the non-parametric approaches. We finally performed a subgroup analysis where we divided the sample into four groups based on age (18-34, 35-44, 45-54, and 55-75 years old) and fit mixed effects regression models with random effects

for trial membership to each subgroup separately. The resulting average treatment effect estimates for each subgroup are presented in Figure B.8, and they lead to a similar conclusion – that older individuals may have a higher treatment effect, but the moderation does not appear to be statistically significant.

This data application shows how to effectively apply the methods compared in simulations to a real dataset and assess potential moderation. The methods all agree broadly on the direction of the average treatment effect but imply somewhat different conclusions with respect to the individual CATE estimates. In comparing the causal forest with pooling with trial indicator versus the IPD meta-analysis with trial random effects, we reach similar conclusions. We expand upon the benefits and drawbacks of these approaches in the following section.

3.7 Discussion

In this paper, we compared methods to estimate the conditional average treatment effect in a single trial and methods to extend the single-trial approaches to multiple trials. In the absence of notable cross-trial heterogeneity of treatment effects, the methods examined all performed well, but when trial membership was involved in the treatment effect function, some methods performed worse than others. Specifically, and not surprisingly, methods that ignore trial membership (complete pooling) do not effectively estimate the CATE when there is cross-trial heterogeneity. On the other hand, some methods performed well no matter the level of heterogeneity: pooling with

trial indicator and ensemble forests had consistently low mean squared error despite increasing the variability of the trial membership coefficients in the treatment effect. This was especially true when the single-study method used was the causal forest (Figure 3.1). These patterns held across various data generation setups, including introducing different sample sizes across trials and a covariate shift. The patterns persisted for the most part with 30 trials as opposed to 10; however, the causal forest with pooling with trial indicator performed much worse with 30 trials. Therefore, this approach could be highly effective with a smaller number of trials but might miss key study-level differences with a large number of trials. Having 30 trials to combine is unlikely in practice, though, in our experience. Otherwise, the two best performing methods – causal forest with pooling with trial indicator and causal forest with ensemble forest – showed high accuracy across all other scenarios and could be good first choices for combining trials to estimate heterogeneous treatment effects.

When considering the three single-study approaches, the most consistently favorable method in the simulations was the causal forest, followed by the X-learner. The S-learner performed well in certain scenarios, such as scenario 1b, where the treatment effect function involved a bounded, non-linear expit function. The performances of the S-learner and X-learner in our simulations and applied example were consistent with results found previously (Künzel et al., 2019), in that the S-learner seemed to be somewhat biased towards 0 in the applied example (Table 3.1) and performed worse in the simulations when the treatment effect function was complicated (variable CATE scenario

and the piecewise linear and non-linear CATE with high variability). The X-learner performed well in the simulations with complex CATEs and with structural forms of the CATE, again consistent with previous work (Künzel et al., 2019). The causal forest performed well across all scenarios. These simulation results and the results from the applied data example of MDD medications demonstrate that it is important to carefully select the single-study method for a given question, as each of the three options can provide different estimates. A good starting point would be to consider expert knowledge of how heterogeneous across studies and complicated the outcomes or treatment effect might be. These results also indicate the need for more diagnostics to help researchers determine which approach to choose. In general though, the causal forest performed consistently well when combining 10 studies, so use of this method is supported by the simulations.

The simulations also incorporated some comparisons between the non-parametric and parametric approaches. Specifically, the use of a lasso regression as an ensemble showed how a parametric ensemble could perform compared to the ensemble tree and forest. The lasso performed very well when the treatment effect function was piecewise linear (scenario 1a) but quickly suffered in performance when the function was more non-linear (scenarios 1b and 2). Furthermore, the inclusion of a mixed effects meta-analysis demonstrated a common parametric technique used in the multiple-study setting. This model was set up to perform well when the CATE function was piecewise linear (scenario 1a), but it yielded high MSE in the non-linear and complex scenarios that it was not correctly parametrized for (scenario 1b and

2). The particular specification of a meta-analysis is therefore very important, and incorrect hypotheses of key interactions and moderating relationships have major implications for model fit and accuracy of estimates. These comparisons demonstrate that non-parametric machine learning approaches are very beneficial when the treatment effect function is complicated and non-linear, as the non-parametric methods do not require correct specification of any parameters. Although interpretability becomes more of a challenge, the non-parametric methods allow for flexible relationships and hopefully high levels of accuracy in CATE estimation.

In this work, we did not explore an exhaustive list of potential single-study and aggregation methods, and we also investigated a few data generation setups that do not cover every possible scenario of real data. We attempted to select single-study methods that were common, user-friendly, and shown to be effective or potentially effective in previous literature. However, as this is an ever-growing field, future work could include other single-study methods (Wendling et al., 2018; Powers et al., 2018) to see how they compare to the ones used in this study. For example, it would be interesting to investigate the performance of the X-learner with a different base learner, such as Bayesian additive regression trees (BART) (Chipman, George, and McCulloch, 2010; Künzel et al., 2019). In general, non-parametric methods for CATE estimation are notably flexible and effective in estimating complex functional forms of the CATE; however, reliable variance estimation for these approaches is somewhat lacking. Without the distributional assumptions present in parametric methods, the non-parametric approaches often require resampling procedures

to effectively estimate variance in predictions. Furthermore, with ensemble approaches such as those used in this paper, there are multiple sources of variance coming from both the original predictions and the predictions from the ensemble model. Therefore, variance estimation is an important area of future work for many of the methods discussed in this paper.

Another important point related to the non-parametric approaches used in this work is that they primarily serve to accurately estimate the true CATE function. They are not as straightforward to use when the goal is identification of key moderators; although we can use tools like variable importance, there are not statistical tests of moderation as there are in parametric approaches like meta-analysis. In the simulations, we were thus not able to efficiently evaluate the methods' ability to identify effect moderators and instead prioritized minimizing error in CATE estimation. If a research goal is to identify moderators, some of the more exploratory work in the applied example (plotting CATE estimates, best linear projections, etc.) could be a helpful starting point, and potential moderators could then also be included in a parametric model to more formally test for moderation.

The approaches discussed in this paper implicitly rely on the assumption that all of the trials being combined have observed the same covariates X necessary to estimate the CATE. We did not discuss cases where the trials contain different measures of a similar construct or cases of systematic missingness, meaning where certain covariates are not at all available in some trials. Approaches for dealing with systematic missingness have been discussed in the literature (Audigier et al., 2018; Jolani et al., 2015) but not in this specific

context, so future work should explore methods for addressing missingness and discordant measures of similar constructs.

It is important to note that with the exception of complete pooling, the resulting CATE estimates are trial-specific. Unless trial was not picked up in the aggregation methods, the majority of the methods discussed will produce trial-specific estimates of the CATE. This allows for improved accuracy of estimates but might be less helpful in real world applications. We are interested in continuing to identify ways in which researchers could aggregate across trials to develop estimates that are accurate but not trial-specific – this could be crucial for use of the resulting methods and models in practice, on data not coming from the specific trials used in the model formulation. However, the trial-specific estimates can still be useful; for example, if trials were done in separate hospitals, CATEs of future patients could be predicted using the hospital that they are being treated in, and the model that estimates their treatment effect should be more accurate after taking into consideration the data from the other hospitals. Similarly, the focal point of this paper and the simulations described above were the performance of models in the given sample. We thus assessed the performance in the simulations based on MSE across the trials used to fit the model, and we discuss accuracy in terms of the trials themselves. Future work will be focused on assessing how these methods perform when estimating CATEs in a target population, outside the specific trials used to estimate the CATE. This is where we might see even more of the benefits of pooling/ensembling approaches over methods like the no pooling approach, because we would be gaining information by combining

trials.

In the MDD trials, duloxetine was included as a reference medication because it was already marketed at the time of the trials, and patients were excluded from the study if they had previously not responded to duloxetine. On the other hand, vortioxetine was not yet marketed and was the more experimental medication; therefore, some bias could arise due to participants being excluded if they had previously not responded to duloxetine. Acknowledging this, we were able to estimate treatment effects according to each method combination, and we used variable importance and interpretation trees to investigate which variables might be important moderators of the treatment effect. Variable importance is a limited measure and can often be biased towards continuous variables with more possible split points (Strobl et al., 2007), so we encourage caution when interpreting those results. This example dataset shows how to combine multiple RCTs to get an improved assessment of treatment effect heterogeneity and better determine which treatment would be best suited to a given individual, based on their features and their estimated CATE. Notably, the four trials used in this dataset were run by the same organizations and had very similar protocols; this helps ensure that we can confidently combine datasets but also might limit the potential heterogeneity across trials that might exist in other applications. We also did not see high levels of heterogeneity in the treatment effects based on other covariates in these trials. A general idea is that studies need to be four times larger to identify effect moderators compared to an average treatment effect (Fleiss, 2011), and this study included precisely four trials. Therefore, our

findings would become more robust and we could more confidently assess heterogeneity or lack thereof with the inclusion of more studies.

There are many openings for future work, some of which have been mentioned. Broadly, it is important to further refine these methods and identify which are most helpful in specific data scenarios. It will also be helpful to determine when it is appropriate to develop universal CATE estimates, versus when the CATE estimates should be trial-specific. This paper demonstrated several approaches that take data from multiple studies and estimate heterogeneous treatment effects, using flexible models that allow for complex relationships – which is often the case in the real world.

3.8 Acknowledgments

The study was funded by the Patient-Centered Outcomes Research Institute (PCORI) through PCORI Award ME-2020C3-21145 (PI: Stuart) and the National Institute of Mental Health (NIMH) through Award R01MH126856 (PI: Stuart). Ms. Brantner also received financial support in the form of a training grant through the National Institutes of Health (T32AG000247). Disclaimer: Opinions and information in this content are those of the study authors and do not necessarily represent the views of PCORI or NIMH. Accordingly, PCORI and NIMH cannot make any guarantees with respect to the accuracy or reliability of the information and data.

Furthermore, this paper is based on research using data from data contributors, Takeda and Lundbeck, that has been made available through Vivli, Inc. Vivli has not contributed to or approved, and is not in any way responsible for,

the contents of this publication. This study, carried out under YODA Project 2022-4854, used data obtained from the Yale University Open Data Access Project, which has an agreement with Janssen Research & Development, L.L.C. The interpretation and reporting of research using this data are solely the responsibility of the authors and does not necessarily represent the official views of the Yale University Open Data Access Project or Janssen Research & Development, L.L.C.

References

Fleiss, Joseph L (2011). *Design and analysis of clinical experiments*. John Wiley & Sons.

Debray, Thomas P. A., Karel G. M. Moons, Gert Valkenhoef, Orestis Efthimiou, Noemi Hummel, Rolf H. H. Groenwold, Johannes B. Reitsma, and on behalf of the GetReal methods review group (2015). "Get real in individual participant data (IPD) meta-analysis: a review of the methodology". en. In: *Research Synthesis Methods* 6.4, pp. 293–309. ISSN: 1759-2879, 1759-2887. DOI: 10.1002/jrsm.1160. URL: https://onlinelibrary.wiley.com/doi/10.1002/jrsm.1160 (visited on 02/02/2022).

Seo, Michael, Ian R. White, Toshi A. Furukawa, Hissei Imai, Marco Valgimigli, Matthias Egger, Marcel Zwahlen, and Orestis Efthimiou (2021). "Comparing methods for estimating patient-specific treatment effects in individual patient data meta-analysis". en. In: *Statistics in Medicine* 40.6, pp. 1553–1573. ISSN: 0277-6715, 1097-0258. DOI: 10.1002/sim.8859. URL: https://onlinelibrary.wiley.com/doi/10.1002/sim.8859 (visited on 02/02/2022).

Künzel, Sören R, Jasjeet S Sekhon, Peter J Bickel, and Bin Yu (2019). "Metalearners for estimating heterogeneous treatment effects using machine learning". In: *Proceedings of the national academy of sciences* 116.10, pp. 4156–4165.

Athey, Susan, Julie Tibshirani, and Stefan Wager (2019). "Generalized random forests". en. In: *The Annals of Statistics* 47.2. ISSN: 0090-5364. DOI: 10.1214/18-AOS1709. URL: https://projecteuclid.org/journals/annals-of-statistics/volume-47/issue-2/Generalized-random-forests/10.1214/18-AOS1709.full (visited on 07/08/2022).

Green, D. P. and H. L. Kern (2012). "Modeling Heterogeneous Treatment Effects in Survey Experiments with Bayesian Additive Regression Trees".

en. In: *Public Opinion Quarterly* 76.3, pp. 491–511. ISSN: 0033-362X, 1537-5331. DOI: 10.1093/poq/nfs036. URL: https://academic.oup.com/poq/article-lookup/doi/10.1093/poq/nfs036 (visited on 07/08/2022).

Kennedy, Edward H (2020). "Towards optimal doubly robust estimation of heterogeneous causal effects". In.

Nie, X and S Wager (2021). "Quasi-oracle estimation of heterogeneous treatment effects". en. In: *Biometrika* 108.2, pp. 299–319. ISSN: 0006-3444, 1464-3510. DOI: 10.1093/biomet/asaa076. URL: https://academic.oup.com/biomet/article/108/2/299/5911092 (visited on 03/30/2022).

Dandl, Susanne, Torsten Hothorn, Heidi Seibold, Erik Sverdrup, Stefan Wager, and Achim Zeileis (2022). "What Makes Forest-Based Heterogeneous Treatment Effect Estimators Work?" In.

Yang, Shu, Donglin Zeng, and Xiaofei Wang (2020). "Elastic Integrative Analysis of Randomized Trial and Real-World Data for Treatment Heterogeneity Estimation". en. In: URL: http://arxiv.org/abs/2005.10579 (visited on 03/17/2022).

Yang, Shu, Donglin Zeng, and Xiaofei Wang (2022). "Improved Inference for Heterogeneous Treatment Effects Using Real-World Data Subject to Hidden Confounding". en. In: URL: http://arxiv.org/abs/2007.12922 (visited on 02/02/2022).

Kallus, Nathan, Aahlad Manas Puli, and Uri Shalit (2018). "Removing hidden confounding by experimental grounding". In: *Advances in neural information processing systems* 31.

Rosenman, Evan TR, Guillaume Basse, Art B Owen, and Mike Baiocchi (2023). "Combining observational and experimental datasets using shrinkage estimators". In: *Biometrics*. DOI: 10.1111/biom.13827.

Brantner, Carly Lupton, Ting-Hsuan Chang, Trang Quynh Nguyen, Hwanhee Hong, Leon Di Stefano, and Elizabeth A. Stuart (2023). "Methods for integrating trials and non-experimental data to examine treatment effect heterogeneity". In: *Statistical Science* 38.4, pp. 640–654.

Tan, Xiaoqing, Chung-Chou H Chang, Ling Zhou, and Lu Tang (2022). "A tree-based model averaging approach for personalized treatment effect estimation from heterogeneous data sources". In: *International Conference on Machine Learning*. PMLR, pp. 21013–21036.

Rubin, Donald B. (1974). "Estimating causal effects of treatments in randomized and nonrandomized studies". In: *Journal of Educational Psychology* 66.5, pp. 688–701. ISSN: 1939-2176. DOI: 10.1037/h0037350.

Athey, Susan and Stefan Wager (2019). "Estimating Treatment Effects with Causal Forests: An Application". en. In: *arXiv:1902.07409 [stat]*. URL: http://arxiv.org/abs/1902.07409 (visited on 02/02/2022).

Jawadekar, Neal, Katrina Kezios, Michelle C Odden, Jeanette A Stingone, Sebastian Calonico, Kara Rudolph, and Adina Zeki Al Hazzouri (2023). "Practical Guide to Honest Causal Forests for Identifying Heterogeneous Treatment Effects". In: *American Journal of Epidemiology*, kwad043.

Wager, Stefan and Susan Athey (2018). "Estimation and Inference of Heterogeneous Treatment Effects using Random Forests". en. In: *Journal of the American Statistical Association* 113.523, pp. 1228–1242. ISSN: 0162-1459, 1537-274X. DOI: 10.1080/01621459.2017.1319839. URL: https://www.tandfonline.com/doi/full/10.1080/01621459.2017.1319839 (visited on 05/09/2022).

Burke, Danielle L., Joie Ensor, and Richard D. Riley (2017). "Meta-analysis using individual participant data: one-stage and two-stage approaches, and why they may differ". en. In: *Statistics in Medicine* 36.5, pp. 855–875. ISSN: 02776715. DOI: 10.1002/sim.7141. URL: https://onlinelibrary.wiley.com/doi/10.1002/sim.7141 (visited on 02/02/2022).

Therneau, Terry, Beth Atkinson, Brian Ripley, and Maintainer Brian Ripley (2015). "Package 'rpart'". In: *Available online: cran. ma. ic. ac. uk/web/packages/rpart/rpart. pdf (accessed on 20 April 2016)*.

Wright, Marvin N. and Andreas Ziegler (2017). "ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R". In: *Journal of Statistical Software* 77.1, pp. 1–17. DOI: 10.18637/jss.v077.i01. URL: https://www.jstatsoft.org/index.php/jss/article/view/v077i01.

Friedman, Jerome, Trevor Hastie, Noah Simon, Rob Tibshirani, Maintainer Trevor Hastie, and Depends Matrix (2017). "Package 'glmnet.'" In: *Journal of statistical software* 33.1, pp. 1–22.

Bates, Douglas M (2010). *lme4: Mixed-effects modeling with R*.

Mahableshwarkar, Atul R., Paula L. Jacobsen, and Yinzhong Chen (2013). "A randomized, double-blind trial of 2.5 mg and 5 mg vortioxetine (Lu AA21004) versus placebo for 8 weeks in adults with major depressive disorder". en. In: *Current Medical Research and Opinion* 29.3, pp. 217–226. ISSN: 0300-7995, 1473-4877. DOI: 10.1185/03007995.2012.761600. URL: http://www.tandfonline.com/doi/full/10.1185/03007995.2012.761600 (visited on 09/22/2022).

Mahableshwarkar, Atul R., Paula L. Jacobsen, Yinzhong Chen, Michael Serenko, and Madhukar H. Trivedi (2015). "A randomized, double-blind, duloxetine-referenced study comparing efficacy and tolerability of 2 fixed doses of vortioxetine in the acute treatment of adults with MDD". en. In: *Psychopharmacology* 232.12, pp. 2061–2070. ISSN: 0033-3158, 1432-2072. DOI: 10.1007/s00213-014-3839-0. URL: http://link.springer.com/10.1007/s00213-014-3839-0 (visited on 09/22/2022).

Boulenger, Jean-Philippe, Henrik Loft, and Christina Kurre Olsen (2014). "Efficacy and safety of vortioxetine (Lu AA21004), 15 and 20 mg/day: a randomized, double-blind, placebo-controlled, duloxetine-referenced study in the acute treatment of adult patients with major depressive disorder". en. In: *International Clinical Psychopharmacology* 29.3, pp. 138–149. ISSN: 0268-1315. DOI: 10.1097/YIC.0000000000000018. URL: http://journals.lww.com/00004850-201405000-00002 (visited on 09/22/2022).

Baldwin, David S, Henrik Loft, and Marianne Dragheim (2012). "A randomised, double-blind, placebo controlled, duloxetine-referenced, fixed-dose study of three dosages of Lu AA21004 in acute treatment of major depressive disorder (MDD)". In: *European Neuropsychopharmacology* 22.7, pp. 482–491.

Schatzberg, Alan F, Pierre Blier, Larry Culpepper, Rakesh Jain, George I Papakostas, and Michael E Thase (2014). "An overview of vortioxetine". In: *The Journal of Clinical Psychiatry* 75.12, p. 13677.

Montgomery, Stuart A and MARIE Åsberg (1979). "A new depression scale designed to be sensitive to change". In: *The British journal of psychiatry* 134.4, pp. 382–389.

Little, Roderick J, Ralph D'Agostino, Michael L Cohen, Kay Dickersin, Scott S Emerson, John T Farrar, Constantine Frangakis, Joseph W Hogan, Geert Molenberghs, Susan A Murphy, et al. (2012). "The prevention and treatment of missing data in clinical trials". In: *New England Journal of Medicine* 367.14, pp. 1355–1360.

Hamilton, MAX (1959). "The assessment of anxiety states by rating." In: *British journal of medical psychology*.

Wendling, T., K. Jung, A. Callahan, A. Schuler, N. H. Shah, and B. Gallego (2018). "Comparing methods for estimation of heterogeneous treatment effects using observational data from health care databases". en. In: *Statistics in Medicine* 37.23, pp. 3309–3324. ISSN: 02776715. DOI: 10.1002/sim.7820. URL: https://onlinelibrary.wiley.com/doi/10.1002/sim.7820 (visited on 03/30/2022).

Powers, Scott, Junyang Qian, Kenneth Jung, Alejandro Schuler, Nigam H. Shah, Trevor Hastie, and Robert Tibshirani (2018). "Some methods for heterogeneous treatment effect estimation in high dimensions: Some methods for heterogeneous treatment effect estimation in high dimensions". en. In: *Statistics in Medicine* 37.11, pp. 1767–1787. ISSN: 02776715. DOI: 10.1002/sim.7623. URL: https://onlinelibrary.wiley.com/doi/10.1002/sim.7623 (visited on 07/08/2022).

Chipman, Hugh A, Edward I George, and Robert E McCulloch (2010). "BART: Bayesian additive regression trees". In: *The Annals of Applied Statistics* 4.1, pp. 266–298.

Audigier, Vincent, Ian R White, Shahab Jolani, Thomas PA Debray, Matteo Quartagno, James Carpenter, Stef Van Buuren, and Matthieu Resche-Rigon (2018). "Multiple imputation for multilevel data with continuous and binary variables". In: *Statistical Science* 33.2, pp. 160–183.

Jolani, Shahab, Thomas PA Debray, Hendrik Koffijberg, Stef van Buuren, and Karel GM Moons (2015). "Imputation of systematically missing predictors in an individual participant data meta-analysis: a generalized approach using MICE". In: *Statistics in Medicine* 34.11, pp. 1841–1863.

Strobl, Carolin, Anne-Laure Boulesteix, Achim Zeileis, and Torsten Hothorn (2007). "Bias in random forest variable importance measures: Illustrations, sources and a solution". In: *BMC bioinformatics* 8.1, pp. 1–21.

Chapter 4

Combining Trials to Estimate Heterogeneous Treatment Effects in a Target Sample

Abstract: Estimating heterogeneous treatment effects can aid practitioners in determining which treatment would work best for a given individual based on their observed characteristics. However, estimating these effects can be challenging with small sample sizes. To better understand treatment effect heterogeneity, researchers can combine data from multiple randomized controlled trials (RCTs). However, combining RCTs requires taking into account that the data comes from different trials, and the treatment effect estimates are therefore conditional on trial membership. A key interest is in applying these models to make predictions for units who do not come from a particular trial, though, which is not straightforward. This paper introduces a method that examines how to best incorporate the resulting uncertainty from applying a model using multiple trials to a set of units in a new setting. The approach draws from meta-analytic prediction intervals to create 95% intervals for the

conditional average treatment effects in the target population. We employ a motivating example containing multiple trials that compare two treatments for major depression, duloxetine and vortioxetine, and we conduct simulations based on this real data that assess coverage of prediction intervals created using meta-analysis, causal forests, and Bayesian additive regression trees (BART). The non-parametric methods achieve high coverage of the true effects in the target population in simulations across data generating scenarios with varying forms, including heterogeneity of the treatment effect function. Finally, we form treatment effect prediction intervals for a representative set of patient profiles in a target population of patients with depression in a health care system. These approaches allow researchers to effectively leverage multiple RCTs to estimate treatment effects in a target population of interest and assess treatment effect heterogeneity across both trial and target data.

4.1 Introduction

In a clinical setting, as well as in decision-making in education, public policy, and more, practitioners are often interested in "what works for whom," (Roth and Fonagy, 2006) meaning identifying subgroups for whom certain interventions might work best. Understanding this can aid in allocation of resources and efficiency, as well as improve outcomes in the target group of interest. To make these decisions, researchers and practitioners often rely on estimation of heterogeneous treatment effects – the effects of treatment conditional on observed characteristics of the patients or units of interest.

Estimation of heterogeneous treatment effects is a common goal when

assessing treatment efficacy; however, conclusions are often limited by small sample sizes and by the fact that treatment effect heterogeneity can be due to unknown and complex interactions of characteristics (Yusuf et al., 1991). For example, when evaluating treatment effects from randomized controlled trials (RCTs), a commonly encountered issue is that the trials were powered to estimate the average effect rather than the conditional average treatment effect (CATE) (Mills et al., 2021), where the treatment effect is conditional on observed characteristics. Researchers often must choose a few pre-specified potential effect moderators to investigate for heterogeneity and might miss unknown subgroup differences. On the other hand, testing all variables as potential effect moderators can be problematic due to multiple testing concerns and a lack of pre-specification.

To learn about effect heterogeneity and estimate the CATE function, researchers could potentially address the above issues by combining individual, participant-level data (IPD) from multiple sources, e.g., multiple RCTs. A growing body of literature has focused on data integration methods to leverage the benefits of combining data sources and account for the limitations of single sources on their own (Brantner et al., 2023; Colnet et al., 2021). Many methods exist to combine RCTs, with meta-analysis being a common and standard approach to do so. Notably, meta-analysis is not commonly parameterized to identify heterogeneous treatment effects and instead generally focuses on estimating the average effect across studies. Some work has investigated meta-analysis in a causal inference setting; Sobel et al. (2017) defined a

causal framework in meta-analysis to estimate study-specific potential outcomes and assess reasons for heterogeneity across trials, and Dahabreh et al. (2020) focused on transporting causal estimation from a meta-analysis to a target population. In a recent paper by Brantner and colleagues (Brantner et al., 2024), several non-parametric methods were explored and extended to account for multiple trials when estimating the CATE as well. In most parametric and non-parametric approaches for estimating the CATE using multiple trials, a key consideration is accounting for the fact that the data came from different trials, which often yields CATE functions that depend not only on characteristics of patients, but also on trial.

The primary goal of this paper is guiding treatment decisions for a target set of patients who do not come from a randomized trial and who meet criteria to receive one of multiple potential treatments. In the current work, this set of potential treatments is restricted to two (i.e., treatment or control), and we assume that treatments have not yet been assigned in this target population. Importantly, this target population does not come from any of the trials used to fit the original CATE model; therefore, the method must account for the uncertainty of the CATE in this new population.

This paper proposes an approach to estimate the uncertainty of the CATE for new populations, to help ensure that treatment guidelines account for effect heterogeneity when needed, and that there is an understanding of when there is not strong evidence for variation in effects. This, ultimately, is often what proponents of personalized medicine are aiming for – use of existing

study results to predict outcomes and guide treatment decisions for individuals outside the original study samples. Dahabreh et al. (2020) and other approaches have a similar goal in their work on extending causal effects from one or more trials to a target population; however, they focus on transporting an *average* effect estimate to a setting where the target population may have different covariate distributions than the trial sample(s). The approach proposed in this paper builds off of meta-analytic prediction intervals (Riley, Higgins, and Deeks, 2011), which are commonly used to estimate a range of potential values for an average effect in a new study. We apply this prediction interval approach to the CATE, and we extend the method to the case when non-parametric methods (i.e., causal forest (Athey, Tibshirani, and Wager, 2019) and Bayesian additive regression trees (Hill, 2011)) are used to estimate the CATE from multiple trials.

As a case study, we include a comparison of two treatments for major depression, duloxetine and vortioxetine, using data from three randomized controlled trials (Mahableshwarkar, Jacobsen, and Chen, 2013; Mahableshwarkar et al., 2015; Boulenger, Loft, and Olsen, 2014) described in detail in Brantner et al. (2024). We estimate the conditional average treatment effect (CATE) across these trials, where the primary outcome is change in a depressive symptoms score from baseline to last observed follow-up. We also bring in electronic health record (EHR) data from Duke Health Care System and define a target group of patient profiles representing patients for whom clincians might be interested in understanding the CATE to aid in treatment decision-making. The methods described in the following sections are used to

form prediction intervals for the CATE in this new setting.

In the sections to follow, we introduce notation and the key estimands and assumptions required (Section 4.2). We then explain the approach for estimating the conditional average treatment effect (CATE) function in multiple trials and subsequently forming prediction intervals for the CATE in the target population, considering both parametric meta-analysis (Section 4.3.1) and non-parametric approaches (Section 4.3.2). We then investigate performance in simulations based on real data (Section 4.4) and apply the methods to estimate the conditional average treatment effects of vortioxetine versus duloxetine for treatment of patient profiles of individuals with major depressive disorder (Section 4.5). Finally, we discuss conclusions, limitations, and future directions in Section 4.6.

4.2 Notation

Let A represent treatment assignment, where $A \in \{0,1\}$ is binary. Let $\boldsymbol{X}$ represent individual-level covariates and Y represent a continuous outcome. One can then define the potential outcomes $Y(0)$ and $Y(1)$ under Rubin's framework (1974) as the outcomes that would have been observed if the individual had received control or treatment, respectively. The estimand of interest is the conditional average treatment effect (CATE):

$$\tau(\boldsymbol{X}) = E(Y(1)|\boldsymbol{X}) - E(Y(0)|\boldsymbol{X}). \tag{4.1}$$

This estimand – the true CATE function – is universal and not dependent upon study membership. In other words, for a given covariate profile $\boldsymbol{X}^*$,

we assume that there is a true, universal $\tau(\boldsymbol{X}^*)$ as in Equation 4.1. However, when combining multiple trials, the methods explored in this paper ultimately estimate study-specific CATEs, where study is represented by a categorical variable $S \in \{1, 2, ..., K\}$. These study-specific estimates of the estimand introduced in Equation 4.1 can be expressed as:

$$\hat{\tau}_s(\boldsymbol{X}) = (\hat{Y}(1)|\boldsymbol{X}, S = s) - (\hat{Y}(0)|\boldsymbol{X}, S = s). \tag{4.2}$$

4.2.1 Assumptions

To combine data from multiple trials to estimate the CATE, we employ standard causal inference assumptions, including the Stable Unit Treatment Value Assumption (SUTVA) within each RCT. We also assume unconfoundedness (Assumption 4.1), consistency (Assumption 4.2), and positivity of treatment assignment (Assumption 4.3) within each trial. These assumptions are generally satisfied in RCTs and have been described in detail elsewhere (Brantner et al., 2023). In this particular setting where we estimate the CATE using multiple trials and subsequently predict in a target population, we also require the assumption that every covariate profile in the target population has positive probability of being found in the trials (Assumption 4.4).

Assumption 4.1 $\{Y(0), Y(1)\} \perp\!\!\!\perp A \mid \boldsymbol{X}, S = s$ *for all studies s.*

Assumption 4.2 $Y = AY(1) + (1 - A)Y(0)$ *almost surely in each study.*

Assumption 4.3 *There exists a constant* $b > 0$ *such that* $b < P(A = 1|\boldsymbol{X} = \boldsymbol{x}, S = s) < 1 - b$ *for all studies s and for all* $\boldsymbol{x}$ *values in each study.*

Assumption 4.4 *There exists a constant* $c > 0$ *such that* $c < P(S \in \{1, 2, ..., K\} | \boldsymbol{X} = \boldsymbol{x}) < 1 - c$ *for all* $\boldsymbol{x}$ *in the target population.*

Finally, depending on the method used to combine trials and estimate the CATE, we sometimes need to assume positivity of study membership (Assumption 4.5).

Assumption 4.5 *There exists a constant* $d > 0$ *such that* $d < P(S = s | \boldsymbol{X} = \boldsymbol{x}) < 1 - d$ *for all* $\boldsymbol{x}$ *and* s.

4.3 Methods

In order to ultimately estimate heterogeneous treatment effects in a target population, we start by estimating effects using data from a set of randomized controlled trials comparing the same two treatments. Methods for estimating the CATE across multiple trials were investigated in depth in a previous paper (Brantner et al., 2024), where traditional parametric meta-analysis was compared with non-parametric methods for aggregating information and accounting for heterogeneity across trials. Some of these approaches are outlined in the following sections.

Once the CATE function is estimated from the multiple RCTs, the key goal of this work is predicting the CATE in an external target population. The models produced from the multiple trials would ideally be helpful for this target population to determine who should receive which treatment; however, the multi-trial CATE models are study-specific as in Equation 4.2. For the target population, we are interested in the universal CATE estimand instead

(Equation 4.1). Specifically, the target population does not come from a trial, so the variable S representing study membership is missing in this new group.

To address this issue of missing study membership in the target population, we provide an approach that essentially integrates out study membership from the original CATE function estimated using the multiple trials. This approach draws from the meta-analytic prediction interval literature and focuses on variance estimation. We now outline this approach using parametric and non-parametric models, describing both the CATE estimation using multiple trials and CATE prediction in the target population.

4.3.1 Meta-Analysis

Before discussing the non-parametric approaches that can be used to estimate the CATE in both the trials and the target population, we focus on a parametric and standard approach – meta-analysis – to introduce a form of CATE prediction intervals. Individual participant-level data (IPD) meta-analysis is a common modeling approach when combining trials; however, heterogeneous treatment effect estimation is not generally a focus of meta-analysis. This approach can have some limitations in that the hypothesized treatment effect heterogeneity and effect moderators have to be defined in the model in advance, and complex non-linearities and covariate interactions might be missed by the parametric specification. However, this model is common in practice, highly interpretable, and beneficial for introducing prediction intervals that can be extended to non-parametric methods as well.

4.3.1.1 Estimating CATE in Multiple Trials

We define the meta-analysis model as follows:

$$Y_{si} = (\beta_0 + a_s) + \boldsymbol{\beta}_1 \boldsymbol{X}_{si} + (\beta_2 + b_s) A_{si} + (\boldsymbol{\beta}_3 + \boldsymbol{c}_s) \boldsymbol{X}_{si}^{mod} A_{si} + \epsilon_{si}$$

where $s = 1, ..., K$ represents study membership, $i = 1, ..., n_s$ represents individual i in study s, and $a_s \sim N(0, \sigma_a^2)$, $b_s \sim N(0, \sigma_b^2)$, and $\boldsymbol{c}_s \sim N(\boldsymbol{0}, \Sigma_c)$ represent random study effects for the intercept, treatment main effect, and treatment-moderator interaction effects, respectively. Here, $\boldsymbol{X}_{si}^{mod}$ represents a usually low-dimensional subset of $\boldsymbol{X}_{si}$ that consists of hypothesized effect moderators.

From this model, we are interested in the conditional average treatment effect for a particular covariate profile $\boldsymbol{X}^*$. We can estimate the CATE for this covariate profile in study s (as defined broadly in Equation 4.2) as

$$\begin{aligned} \hat{\tau}_s(\boldsymbol{X}^*) &= (\hat{Y}(1) | \boldsymbol{X}^*, S = s) - (\hat{Y}(0) | \boldsymbol{X}^*, S = s) \\ &= (\hat{\beta}_2 + \hat{b}_s) + (\hat{\boldsymbol{\beta}}_3 + \hat{\boldsymbol{c}}_s) \boldsymbol{X}^{*,mod}. \end{aligned}$$

4.3.1.2 Estimating CATE in Target Population

For simplicity of notation, let us consider $\hat{\tau}_s = \hat{\tau}_s(\boldsymbol{X}^*)$. This mixed effects model parameterization assumes the following for $\hat{\tau}_s$:

$$\begin{aligned} \hat{\tau}_s &\sim N(\tau_s, SE(\hat{\tau}_s)^2) \\ \tau_s &\sim N(\tau, \theta^2) \end{aligned} \tag{4.3}$$

(Riley et al., 2021), where $\tau = \tau(\boldsymbol{X}^*)$ represents the average parameter across studies, $SE(\hat{\tau}_s)^2$ represents the variance of the fixed effects, and θ^2 represents the variance of the random effects in the model.

Notably, we are interested in the treatment effect in the absence of study membership. Prediction intervals in random effects meta-analysis estimate a range of potential parameter values in a new study (Riley, Higgins, and Deeks, 2011), so they can be implemented here to determine what the effects might be in the target population. The general form for a prediction interval for a parameter $\tau = \tau(\boldsymbol{X}^*)$ is based off of the above assumptions and models and can be expressed as follows:

$$\tau \in \left\{ \hat{\tau} \pm t_{K-2} \sqrt{SE(\hat{\tau})^2 + \hat{\theta}^2} \right\} \tag{4.4}$$

We can calculate this prediction interval for $\tau(\boldsymbol{X}^*)$ by leveraging the model assumptions that $b_s \sim N(0, \sigma_b^2)$ and $\boldsymbol{c}_s \sim N(\boldsymbol{0}, \Sigma_c)$. From here, the average CATE estimate for $\boldsymbol{X}^*$ across studies is

$$\hat{\tau} = \hat{\beta}_2 + \hat{\beta}_3 \boldsymbol{X}^{*,mod}.$$

We can then estimate the variance terms by determining the variance of the fixed effects (within-study variance) and the variance of the random effects (between-study variance):

$$SE(\hat{\tau})^2 = \text{Var}(\hat{\beta}_2 + \hat{\boldsymbol{\beta}}_3 \boldsymbol{X}^{*,mod}) \quad \text{and} \quad \hat{\theta}^2 = \text{Var}(b_s + \boldsymbol{c}_s \boldsymbol{X}^{*,mod}).$$

For more details of these variance calculations in matrix form, see the Appendix (C.1).

4.3.2 Non-Parametric Approaches

In non-parametric CATE estimation, there are no longer distributional assumptions on parameters, nor is there a need to prespecify a functional form of the CATE. However, we can still utilize the prediction interval approach similarly to Equation 4.4. The key difference here is that instead of calculating the variance of the fixed and random effects as done in meta-analysis, we utilize the study-specific estimates of the treatment effect and calculate variance by combining the within- and between-study variability of those study-specific estimates. We start with estimation of the CATE using non-parametric approaches applied to multiple RCTs.

4.3.2.1 Estimating CATE in Multiple Trials

Several non-parametric approaches for estimating the CATE using multiple trials are described in Brantner et al. (2024). The flexible modeling approaches used in Brantner et al. (2024) allow for trial membership to interact with other potential effect moderators. Several aggregation approaches were explored, with pooling with trial indicator and ensemble forests showing particularly good performance in simulations; for this paper, we focus on the approach labeled "pooling with trial indicator" due to its high estimation accuracy, existing methods for assessing uncertainty, and computational efficiency.

The pooling with trial indicator approach involves first pooling all individual-level data from all trials into one large dataset. This pooled dataset then includes a categorical variable that indicates the trial that the individual participated in. A non-parametric method like a causal forest (Athey, Tibshirani,

and Wager, 2019) (see Section 4.3.2.3) can then be fit to the pooled dataset to estimate the conditional average treatment effect as in Equation 4.2, conditioning on individual-level covariates as well as their trial membership.

4.3.2.2 Estimating CATE in Target Population

Once the CATE function has been estimated, the next step is predicting the CATE in a target population. We again focus on a given covariate profile $\boldsymbol{X}^*$ and estimate the treatment effect $\tau = \tau(\boldsymbol{X}^*)$ in the target population by utilizing the study-specific treatment effect estimates $\{\hat{\tau}_1, \hat{\tau}_2, ..., \hat{\tau}_K\}$ (again, removing the $\boldsymbol{X}^*$ for simplicity of notation). We assume that these treatment effects represent random draws from some distribution of the true treatment effect for covariate profile $\boldsymbol{X}^*$, where each estimated $\hat{\tau}_i$ has its own uncertainty as well. There is therefore some within-study uncertainty represented by the variance of $\hat{\tau}_i$, $i = 1, ..., K$, as well as between-study variability represented by the variance of the vector of study-specific estimates, $\{\hat{\tau}_1, \hat{\tau}_2, ..., \hat{\tau}_K\}$.

Thus, a prediction interval for the treatment effect conditional on covariate profile $\boldsymbol{X}^*$ can be constructed as follows:

$$\tau \in \{\hat{\tau} \pm t_{K-2}\sqrt{\text{var}_{\text{within}} + \text{var}_{\text{between}}}\} \tag{4.5}$$

where

$$\hat{\tau} = \frac{1}{K}\sum_{i=1}^{K}\hat{\tau}_i,$$

t_{K-2} is the critical value of the t-distribution at $K-2$ degrees of freedom (Riley,

Higgins, and Deeks, 2011),

$$\text{var}_{\text{within}} = \frac{1}{K}\sum_{i=1}^{K} \text{Var}(\hat{\tau}_i), \text{ and}$$

$$\text{var}_{\text{between}} = \text{Var}\{\hat{\tau}_1, \hat{\tau}_2, ..., \hat{\tau}_K\}.$$

This general form can be applied to any non-parametric approach that yields study-specific estimates of the CATE. In the following subsections, we introduce two potential non-parametric methods for use in this setting.

4.3.2.3 Causal Forest

The causal forest (Athey, Tibshirani, and Wager, 2019) is a non-parametric method that involves a weighted aggregation of causal trees. Each causal tree is formed by recursively partitioning covariates, where splits are chosen to maximize treatment effect heterogeneity. In each leaf, the treatment effect is estimated as the difference in average outcomes between the treatment and control group individuals within the subgroup that falls in that particular leaf. The causal forest does not rely on estimating the outcomes conditional on covariates and instead directly proceeds by estimating the treatment effects conditional on covariates.

Wager and Athey (2018) also introduce a concept called "honesty" to their causal forest implementation, which ensures that within each tree, every individual's outcome is used only for defining tree splits or estimating the treatment effect within a leaf, but not both. This concept is discussed in more depth in Wager and Athey's paper (2018), and honest versus adaptive (not honest) causal forests are compared briefly in Brantner et al. (2024). This paper

provides results from both honest and adaptive causal forests.

Based on the work done previously (Brantner et al., 2024) to extend single-study non-parametric methods to the multi-study setting, we focus on the causal forest pooling with trial indicator approach introduced in Section 4.3.2.1. In this approach, a single causal forest is fit to the pooled dataset, and the model covariates include patient-level characteristics as well as a categorical variable representing trial membership. Once the causal forest with pooling with trial indicator is fit to the multiple RCTs, this yields a CATE function that is study-specific (Equation 4.2). Furthermore, the causal forest is non-parametric so does not provide model coefficients like in meta-analysis. Instead, the model provides CATE estimates and variance estimates per covariate profile and study. Therefore, for a given covariate profile $\boldsymbol{X}^*$ in the target population, we can calculate estimated means and variances of their treatment effect if they were in each trial:

$$\big\{\{\hat{\tau}_1, \text{Var}(\hat{\tau}_1)\}, \{\hat{\tau}_2, \text{Var}(\hat{\tau}_2)\}, ..., \{\hat{\tau}_K, \text{Var}(\hat{\tau}_K)\}\big\}.$$

From this information, we can calculate the necessary information for Equation 4.5 as described above: $\hat{\tau}(\boldsymbol{X}^*)$ is the average of the estimates across studies; $\text{var}_{\text{within}}$ is the average of the variances within each study; $\text{var}_{\text{between}}$ is the variance of the estimates across studies.

4.3.2.4 Bayesian Additive Regression Trees

Another non-parametric approach for CATE estimation in a single-study is Bayesian Additive Regression Trees (BART) (Hill, 2011; Carnegie, Dorie, and

Hill, 2019). BART is a sum-of-trees model that uses regularization priors to restrict the amount of relationships that each tree can explain. BART is similar to the causal forest in that both are tree-based, but it is different in that it is a Bayesian implementation and focuses on first estimating the outcome conditional on covariates rather than the treatment effect. To estimate the CATE using BART, one can estimate the conditional mean outcome under treatment and control, and then directly calculate their difference (Hill, 2011; Carnegie, Dorie, and Hill, 2019). BART also provides draws from the posterior distributions for outcomes conditional on covariates, so intervals can be created either using the mean and variance of those draws and assuming a normal distribution, or using quantiles of the posterior distribution (Dorie et al., 2022). In this paper, we use a normal distribution assumption to stay consistent with approaches like the causal forest but discuss the alternative of posterior quantiles in the Appendix (C.2).

Here, we also apply the pooling with trial indicator aggregation method to fit BART to multiple trials at one time. Specifically, we fit a single BART model (sometimes called an S-learner (Künzel et al., 2019)) to the pooled dataset, where covariates include patient-level characteristics, trial membership, and treatment. We estimate each individual's counterfactual according to the fitted BART model by including as "test" data the same dataset but with opposite treatment assignment. Finally, we estimate the CATE by subtracting the estimated outcome under treatment minus the estimated outcome under control (averaged across posterior draws), and we estimate the variance of the CATE by adding together the variance of the outcome under treatment

across posterior draws with the variance of the outcome under control across posterior draws. Then, just like the causal forest, for a given covariate profile X^* we have:

$$\{\{\hat{\tau}_1, \text{Var}(\hat{\tau}_1)\}, \{\hat{\tau}_2, \text{Var}(\hat{\tau}_2)\}, ..., \{\hat{\tau}_K, \text{Var}(\hat{\tau}_K)\}\}.$$

We can then follow the same procedure as the causal forest to estimate a prediction interval according to Equation 4.5 as described in Section 4.3.2.2.

4.4 Simulations

4.4.1 Setup

We now present a simulation study to assess performance of the methods described above for estimating CATE prediction intervals. We set up the simulation to closely represent the real RCTs discussed in the following section (4.5); specifically, we estimated means and covariances of variables in the real data to guide covariate distributions in the simulated data, and we fit models to the real data to estimate reasonable treatment effect functions for the simulated data.

In primary simulations, we simulated $K = 10$ studies, each with $n = 500$ individuals and with the same covariate distributions across all trials. Each individual had probability 0.5 of receiving the treatment, and individuals had five observed covariates: sex (defined as binary, female or male), smoking status (defined as binary, have smoked or never smoked), weight in kilograms, age in years, and baseline Montgomery-Asberg Depression Rating

Scale (MADRS) score (Montgomery and Åsberg, 1979). These covariates were simulated using a multivariate normal distribution within each study with continuous variables simulated on a standardized (mean 0, standard deviation 1) scale.

While the trials were randomly sampled according to the above information for every iteration of the simulation, we also created a *single* target sample with $n = 100$ individuals representing a range of covariate profiles, where baseline covariates were sampled from the same multivariate normal distribution that the trials were sampled from. We saved this set of covariate profiles to be used across all simulation iterations, where outcomes and treatment effects for the target sample were defined for each iteration depending on the data generation setup.

We defined average outcomes for each covariate profile in the training trials and target population according to the following model:

$$Y = m(\boldsymbol{X}) + A * \tau(\boldsymbol{X}) + \epsilon,$$

where Y represented the change in MADRS score from baseline to last observed follow-up, $m(\boldsymbol{X})$ represented a main effect function, A represented treatment (0 representing control and 1 treatment), $\tau(\boldsymbol{X})$ represented the CATE function, and $\epsilon \sim N(0, 0.05^2)$ was a random error term. We defined two settings for m and τ:

1. Age is the only moderator, CATE is linear

$$m(\boldsymbol{X}) = (-17.40 + a_s) - 0.13 * \text{Age} - 2.05 * \text{MADRS} - 0.11 * \text{Sex}$$

$$\tau(\boldsymbol{X}) = (2.505 + b_s) + (0.82 + c_s) * \text{Age}$$

2. Age is the only moderator, CATE is non-linear:

$$m(\boldsymbol{X}) = (-17.52 + a_s) - 0.08 * \text{Age}$$

$$\tau(\boldsymbol{X}) = (2.20 + b_s) * \exp[(0.35 + c_s) * \text{Age}]$$

In these setups, $a_s \sim N(0, \sigma_a^2)$, $b_s \sim N(0, \sigma_b^2)$, and $c_s \sim N(0, \sigma_c^2)$ represent heterogeneity due to study membership. In the target population, we set $a_s = b_s = c_s = 0$. In the trial data within each of the three setups, we included different values for the standard deviations of these study-level terms to allow for varying heterogeneity in the main and treatment effects across studies. Specifically, we use four sets of values:

1. Low heterogeneity: $\sigma_a = 0.05$, $\sigma_b = 0.05$, $\sigma_c = 0.05$
2. Heterogeneous intercept: $\sigma_a = 1$, $\sigma_b = 0.05$, $\sigma_c = 0.05$
3. Heterogeneous intercept and main treatment effect: $\sigma_a = 1$, $\sigma_b = 0.5$, $\sigma_c = 0.05$
4. High heterogeneity: $\sigma_a = 1$, $\sigma_b = 1$, $\sigma_c = 0.5$

Finally, we reran the above settings with a few other changes to parameters. First, we allowed for two different settings in terms of the covariate distributions; one setting varied the average baseline depression across trials, and another setting contained a single trial with notably older individuals compared to the other trials. We also reran all simulations with $K = 3$ RCTs, which is the number of trials used in the applied example to follow.

For each simulation setup, we ran 500 replications. We used several R packages to conduct the simulations, including lme4 for the mixed effects meta-analysis (Bates, 2010), grf for the causal forest (Athey, Tibshirani, and Wager, 2019), and dbarts for BART (Dorie et al., 2023). In the causal forest and BART, hyperparameters were set to be the defaults, except that the causal forest was set to use 1,000 trees instead of the default of 2,000 for computational ease. Performance of the methods were assessed based on prediction interval coverage, prediction interval length, and absolute bias. Specifically, for each covariate profile in the target population, we assessed whether their true treatment effect was contained within the prediction interval produced by the modeling procedure, and we calculated the percent of the 500 iterations for which this occurred to determine coverage. Code containing all methods and implementation of the simulations can be found at the repository: `https://github.com/carlyls/OOSE_multiRCT`.

4.4.2 Results

We first present results from the primary simulations, including $K = 10$ RCTs with $n = 500$ individuals in each with the same covariate distributions across trials, and where age is the only moderator involved in the treatment effect function. Results from both a linear and non-linear CATE function with varying levels of heterogeneity in the coefficients across trials are included.

Figure 4.1 displays boxplots of CATE prediction interval coverage for the 100 covariate profiles in the target population, where coverage for each person was calculated as the percentage of the 500 iterations for which the covariate

profile's true treatment effect was contained within their prediction interval. Overall, there are high levels of coverage for the target population profiles across the majority of methods and simulation scenarios; specifically, most profiles have close to 100% coverage. The exceptions to this are the honest causal forest, where the coverage is centered around 95% when heterogeneity across trials was low, and meta-analysis, where the coverage was variable and had some profiles with notably low coverage when the treatment effect function was non-linear and heterogeneity across trials was low. Interestingly, all models had high coverage for all covariate profiles when heterogeneity of the CATE across trials was high; we explore this more by investigating interval length and bias.

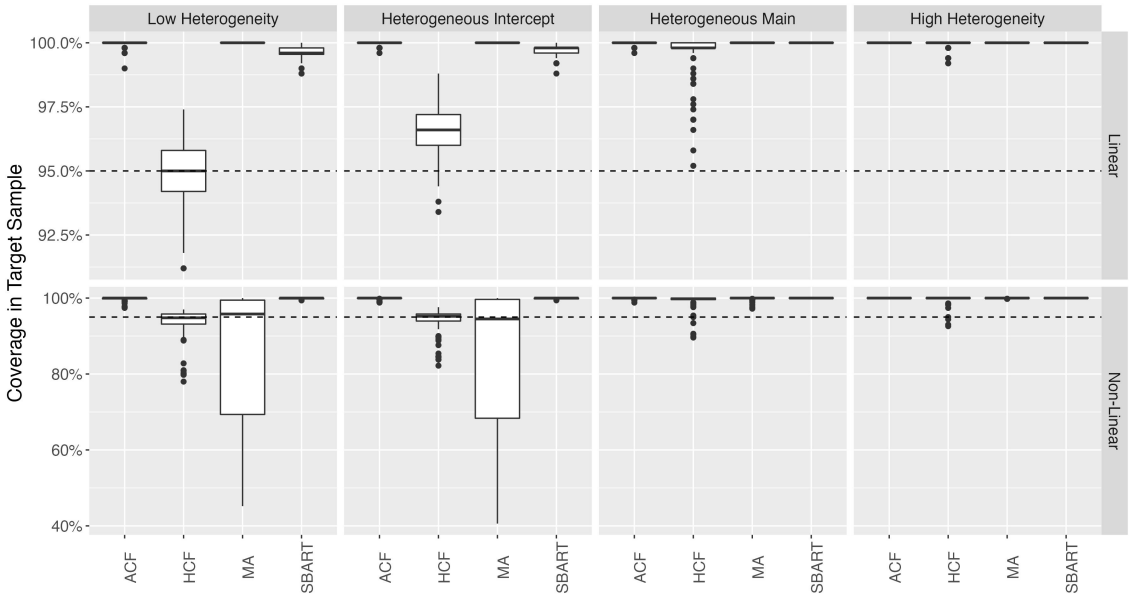

Figure 4.1: Distributions of coverage for each covariate profile in the target population across each method and data generation scenario.
Coverage was calculated as the percent of 500 iterations for which the profile's true treatment effect was contained within the estimated prediction interval. *Method abbreviations:* ACF = adaptive causal forest, HCF = honest causal forest, MA = meta-analysis, SBART = Bayesian Additive Regression Trees with S-learner.

Figure 4.2 presents boxplots of average prediction interval length across the 500 simulation iterations for each covariate profile in the target population. As heterogeneity of the CATE increases across trials (moving from the left-most column to the right-most column of plots), the intervals become wider, which is to be expected. The interval lengths are similar between the linear and non-linear CATE setups; however, there is a bit more variability in the distributions of interval lengths when the CATE is non-linear. Several profiles had very high average interval lengths in the non-linear CATE setting with high heterogeneity, which was omitted in the plot to be able to best visualize differences across methods and settings. In terms of methods, the honest causal forest and BART fit using an S-learner had lower interval length distributions across all settings displayed in Figure 4.2. Meta-analysis had the highest interval lengths in the non-linear setting, which aligns with the fact that meta-analysis assumes linearity and so is not tailored correctly to the non-linear CATE function.

Figure 4.3 presents average absolute bias for each covariate profile in the target population. Here, meta-analysis has considerably lower bias compared to the other methods when the CATE was linear, but higher bias when the CATE was non-linear. Again, the honest causal forest and BART with the S-learner performed similarly to one another. For all methods, higher bias occurred when heterogeneity across trials increased.

To further examine the variability in metrics like coverage, interval length, and absolute bias, we plotted these metrics versus values of the age covariate across the covariate profiles in the target population (Figures 4.4, 4.5, 4.6).

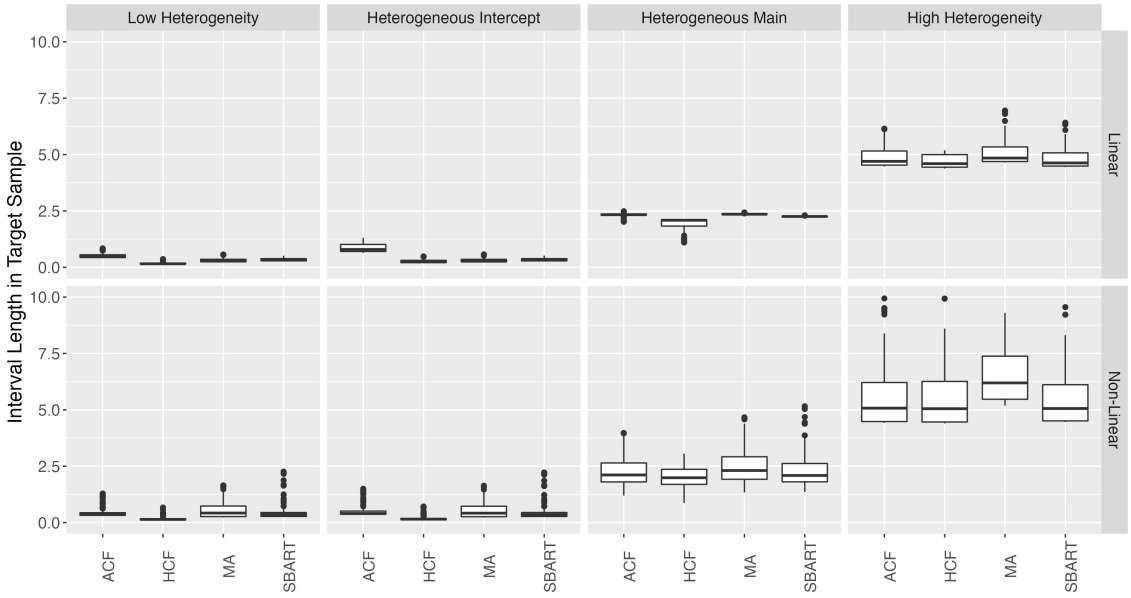

Figure 4.2: Distributions of average interval length for each covariate profile in the target population across each method and data generation scenario.
Length was calculated as the average length of the profile's prediction interval across 500 iterations. Y-axis was cut off at 10 for ease of visualization; some profiles in the non-linear, high heterogeneity CATE had higher average interval length. *Method abbreviations:* ACF = adaptive causal forest, HCF = honest causal forest, MA = meta-analysis, SBART = Bayesian Additive Regression Trees with S-learner.

The goal here was to determine whether the methods performed equally well across all values of the covariate or not. In terms of coverage, BART with the S-learner had high coverage for all ages across all scenarios. The honest causal forest had high coverage for profiles with age closer to the mean and lower coverage for profiles with age further from the mean. Meta-analysis had high coverage for all ages when the CATE was linear or non-linear and heterogeneous, but had much messier coverage results in the non-linear CATE setting with lower levels of heterogeneity. This messy coverage occurred in the non-linear and low heterogeneity CATE setting because the true CATE was non-linear and the estimated CATE was linear using meta-analysis, so

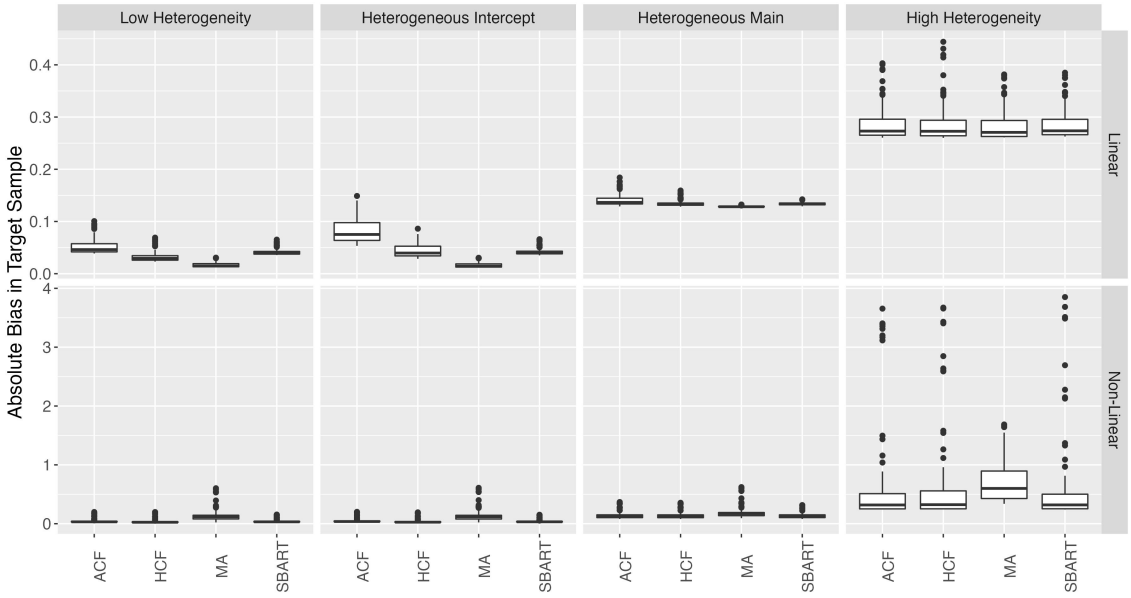

Figure 4.3: Distributions of absolute bias for each covariate profile in the target population across each method and data generation scenario.

Absolute bias was calculated as the average absolute difference between the covariate profile's true treatment effect and the estimated treatment effect across 500 iterations. *Method abbreviations:* ACF = adaptive causal forest, HCF = honest causal forest, MA = meta-analysis, SBART = Bayesian Additive Regression Trees with S-learner.

profiles towards the middle of the age range had small intervals but high bias and therefore low coverage.

Interval length also varied by age, data generation setup, and method; overall, interval length was higher for profiles with age further from the mean. In the high heterogeneity, non-linear CATE setting, some profiles had very wide prediction intervals. Absolute bias showed a similar pattern in that profiles who had age further from the mean had higher absolute bias. With the non-linear CATE, meta-analysis had high absolute bias for profiles with close to average age as well due to the fact that meta-analysis was constructed using a misspecified linear model.

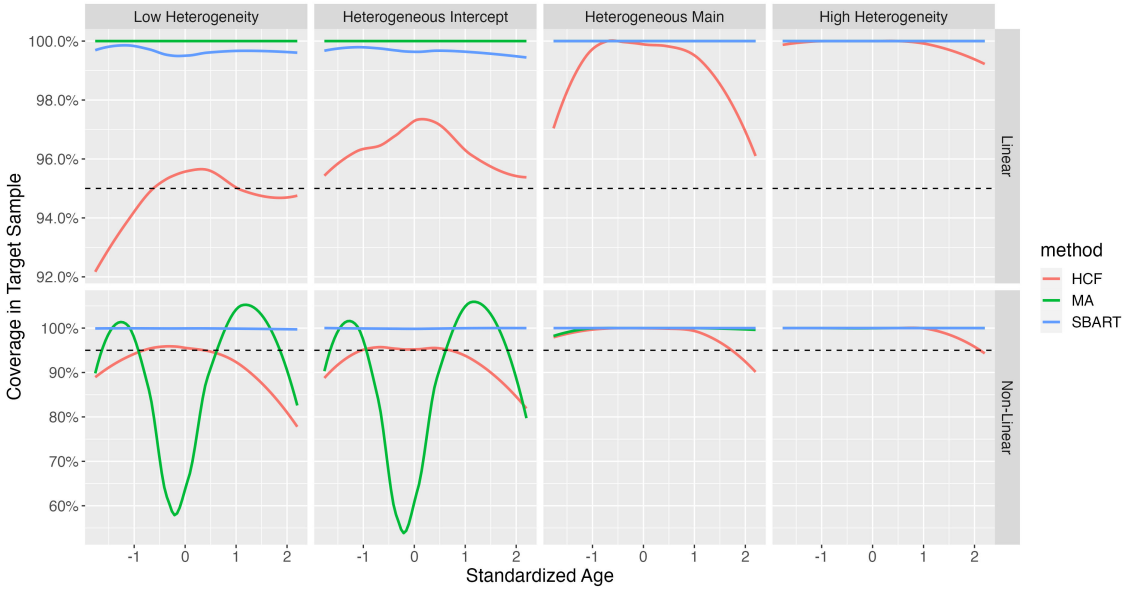

Figure 4.4: A LOESS plot of coverage for each covariate profile in the target population based on their standardized age across each method and data generation scenario. *Method abbreviations:* ACF = adaptive causal forest, HCF = honest causal forest, MA = meta-analysis, SBART = Bayesian Additive Regression Trees with S-learner.

4.5 Applied Example: Major Depression Treatments

4.5.1 Datasets

After comparing method performance in the simulation study described above, we applied the methods to data from three trials investigating two treatments for major depression: duloxetine and vortioxetine. Each trial included participants who were between 18 to 75 years old, had a Major Depressive Episode (MDE) as a primary diagnosis according to the DSM-IV-TR criteria over at least three months, and had a Montgomery-Asberg Depression Rating Scale (MADRS) (Montgomery and Åsberg, 1979) score of at least 22 (one trial) or 26 (three trials) at both screening and baseline (Mahableshwarkar, Jacobsen, and Chen, 2013; Mahableshwarkar et al., 2015; Boulenger, Loft, and

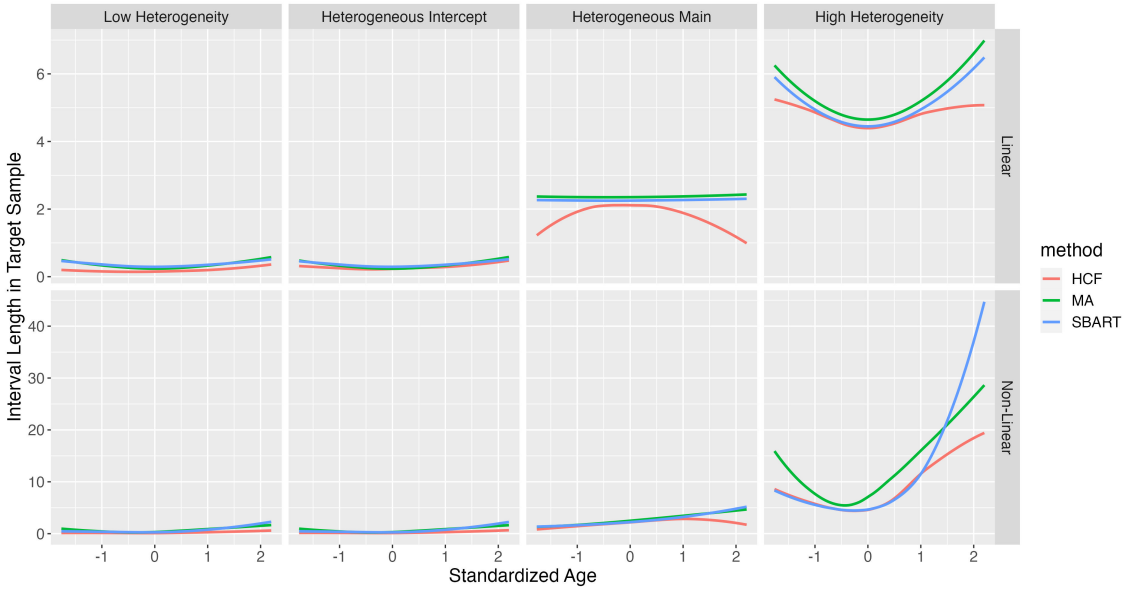

Figure 4.5: A LOESS plot of average interval length for each covariate profile in the target population based on their standardized age across each method and data generation scenario.

Method abbreviations: ACF = adaptive causal forest, HCF = honest causal forest, MA = meta-analysis, SBART = Bayesian Additive Regression Trees with S-learner.

Olsen, 2014). Participants were randomly assigned to receive duloxetine, vortioxetine, or placebo; we removed individuals who were randomly assigned to placebo for this analysis. We treat duloxetine as the control condition here because it was already in use at the time of the trials. The primary outcome was change in MADRS score from baseline to last observed follow-up, where the goal was to follow patients for 8 weeks. More information on these trials can be found in their original papers (Mahableshwarkar, Jacobsen, and Chen, 2013; Mahableshwarkar et al., 2015; Boulenger, Loft, and Olsen, 2014) or in Brantner et al. (2024); note that we removed one trial used in Brantner et al. because it did not collect BMI. Table 4.1 presents descriptive statistics for the three trials used in the current analysis.

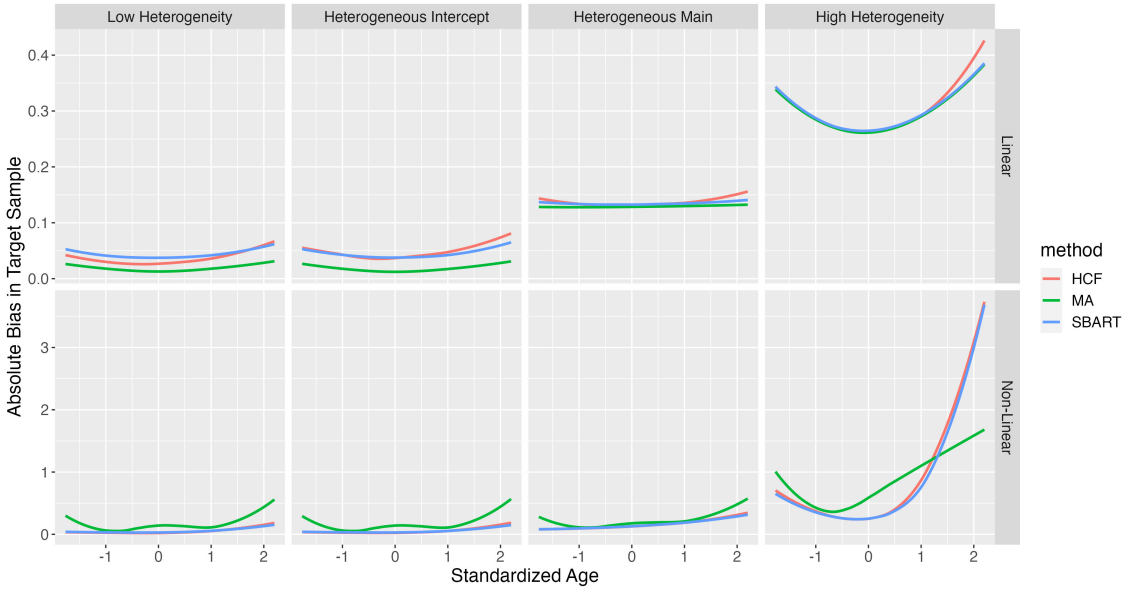

Figure 4.6: A LOESS plot of average absolute bias for each covariate profile in the target population based on their standardized age across each method and data generation scenario.

Method abbreviations: ACF = adaptive causal forest, HCF = honest causal forest, MA = meta-analysis, SBART = Bayesian Additive Regression Trees with S-learner.

We constructed our external target population using data from patients at the Duke Health Care System. The diagram explaining the construction of this patient sample can be found in the Appendix (C.4). We identified patients from Duke psychiatry or primary care with visits that were assigned diagnoses of major depression, bipolar disorder, and/or persistent mood disorder, and we filtered to patients who were prescribed either duloxetine or vortioxetine between January 1, 2014 to December 31, 2021. We then subset to patients who had at least one year of electronic health record (EHR) data before they had been prescribed either of the medications, and we finally included only patients aged 18 to 65 at the time of their prescription who had a non-missing PHQ-9 (Patient Health Questionnaire) score of at least 10 (indicating less than

moderate depression) (Kroenke, Spitzer, and Williams, 2001). The final sample size from the EHR data was 2,123 patients. These patients represent a set of covariate profiles that are possible in the target population, and the goal is to estimate the CATE for this set of profiles. For the purposes of this analysis, we ignore which treatment the patients actually received and use their baseline characteristics to estimate treatment effect prediction intervals based on the approaches previously described. Descriptive statistics for this patient sample can also be found in Table 4.1. Notably, the EHR sample is similar to the trials in terms of age, BMI, sex, and baseline depression; however, the EHR data has much higher prevalence of the comorbidities and medications measured.

Across both samples, conditional mean imputation was performed for missing values of BMI (n=2 in the trials and n=19 in the EHR data). Most variables were similarly defined across the trials and the EHR data; however, the RCTs used MADRS to measure depression, while the EHR data includes PHQ-9 to measure depression. In order to have a similar measure of baseline depression across all datasets, we created a binary indicator of moderate or severe depression based on criteria defined for both scales (Snaith et al., 1986; Herrmann et al., 1998; Kroenke, Spitzer, and Williams, 2001).

4.5.2 Results

After pre-processing the data, we applied the methods described previously to first fit a CATE model to the three RCTs and subsequently form treatment effect prediction intervals for the covariate profiles in the EHR data. Here, we focus on the results from the honest causal forest with pooling with trial

	NCT00672620 N=418	**NCT01140906** N=436	**NCT01153009** N=418	**Duke Patients** N=2,123
	Mean (SD)	*Mean (SD)*	*Mean (SD)*	*Mean (SD)*
Age	43.0 (13.8)	46.3 (13.9)	43.4 (12.2)	44.7 (12.7)
BMI	30.4 (7.9)	25.9 (4.9)	31.2 (7.9)	31.5 (8.7)
	%	%	%	%
Female	64.4	65.4	74.2	75.7
Diabetes Mellitus	4.8	1.4	2.4	20.2
Hypothyroidism	5.0	3.2	4.5	11.1
Anxiety	1.9	0.2	3.8	61.3
Antidepressant	1.7	33.5	19.4	58.7
Thyroid Medication	0.7	3.4	3.6	9.3
Severe (vs. Moderate) Baseline Depression	13.2	17.2	28.7	27.3

Table 4.1: Descriptive statistics for three randomized controlled trials and EHR data from patients at the Duke Health Care System.

indicator. Detailed results from a similar model fit to these trials can be found in previous work (Brantner et al., 2024). This paper uses three of the same trials analyzed in Brantner et al. (2024), where the fourth trial was removed in this analysis because it did not report body mass index (BMI). Most notably, there was not a high level of heterogeneity in the treatment effect, both in terms of potential covariates and across the studies being combined. There did seem to be some differences in the treatment effect across age, where older individuals had a more positive effect estimate, but the uncertainty intervals were wide. Overall, it seemed that duloxetine was the preferable treatment compared to vortioxetine across individual characteristics in the sample.

In these three trials, the main takeaways from Brantner et al. (2024) hold. From the honest causal forest with pooling with trial indicator, the average CATE estimate across the sample of 1,272 individuals was 2.68 (SD = 0.97). The average confidence interval length in the trials was 4.78, and 60.77% of the sample had a confidence interval that did not cross zero. Every set of covariates represented in this sample according to this model had a positive treatment effect estimate, thus in favor of duloxetine. There again is potential heterogeneity of the treatment effect by age, where older individuals had higher estimates of the effect.

Figure 4.7 displays the effect estimates and 95% prediction intervals for all of the covariate profiles in the target population of patients from the EHR data. All EHR patients have positive effect estimates, again indicating that duloxetine is estimated to be the better treatment for reduction of depressive symptoms in this patient population. Specifically, the average effect estimate in the target population is 2.79 (SD = 0.92). 66.8% of the covariate profiles in this target population have a prediction interval that does not cross zero, and the average interval length is 4.64. Because only three trials were being combined, the critical value used to calculate the prediction intervals was approximated as 1.96 rather than using the t-distribution with $K - 2 = 3 - 2 = 1$ degree of freedom. Therefore, these prediction intervals are likely a bit too narrow, but using 1.96 as an approximation of the critical value was shown to perform relatively well in simulated data with $K = 3$ trials (see Appendix C.3).

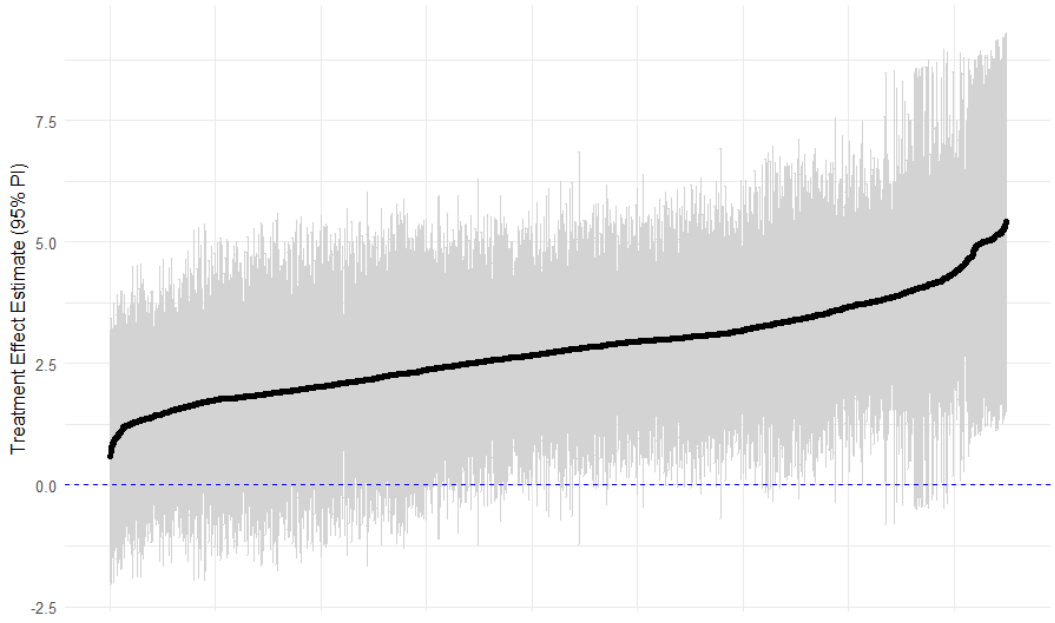

Figure 4.7: 95% prediction intervals for treatment effect estimates in target population. X axis is ordered by magnitude of treatment effect estimate.

4.6 Discussion

This paper introduced and assessed a prediction interval-based approach for predicting conditional average treatment effects in a target population after having estimated the CATE from multiple trials. We discussed this approach in conjunction with meta-analysis as well as with non-parametric methods, including the causal forest and BART. The non-parametric methods performed well in simulations in terms of prediction interval coverage of the true effects, across levels of heterogeneity in the effect across trials and both a linear and non-linear CATE function. Meta-analysis performed well when the true CATE was linear but poorly when the true CATE was non-linear.

The performance of all methods did vary across levels of age; specifically,

coverage, length, and absolute bias were worse for many of the approaches and data generation setups for values of age that were further from the mean age. This difference in performance for absolute bias and interval length was more drastic when the true CATE was very heterogeneous across studies and was non-linear. Therefore, in practice, it is important to explore the variability in the covariate distributions in the target population and in comparison to the trials to identify if there might be subgroups for whom the CATE prediction intervals might perform more poorly.

The methods also performed well even with a small number of studies (Appendix C.3) and when the critical value was set to be 1.96 rather than t_{K-2}. There were slight decreases in coverage compared to the settings with 10 studies, but coverage was still relatively high overall. In the other setups where covariate distributions varied across trials, the main results still held for the most part. Coverage did get lower for several covariate profiles in the setting where one trial had a very different mean age compared to the others (Appendix C.3).

In the real data, three RCTs were combined to estimate the CATE function and subsequently predict CATEs for health records of patients in a health care system. This analysis revealed that there was potential heterogeneity of the treatment effect by age but that variability was high, and in general, all patient profiles were estimated to benefit more from duloxetine compared to vortioxetine. It is important to note that in the original trials, duloxetine was included as an active reference medication, and two of the trials had eligibility criteria where patients were excluded from the study if they had previously

not responded to duloxetine. Therefore, this comparison could be somewhat biased due to this exclusion criteria. The prediction intervals were calculated for this target population based on the causal forest with pooling with trial indicator; we estimated the critical value as 1.96 in this setting and so intervals might be slightly underestimating the true potential variability. This analysis could be helpful for clinicians looking to decide between treatment options for patients based on previous treatment effect heterogeneity estimated from randomized trials. These results again demonstrated that duloxetine would be preferable as a whole; however, other treatment comparisons might reveal more notable differences in treatment effectiveness across patient characteristics. In general, this approach allowed estimation of effects in the patient sample without needing to observe any outcomes or treatment assignment in this group beforehand.

Notably, throughout this paper we have emphasized the idea of the CATE for a set of covariate profiles. This emphasis is first due to the nature of non-parametric, machine learning methods. Specifically, while meta-analysis provides an interpretable functional form of the CATE, approaches like the causal forest and BART instead give predicted CATEs and intervals but not a functional form with parameter estimates. Therefore, we represent the CATE by estimating it across a set of covariate profiles. Another key point is the important distinction between conditional average treatment effects (CATEs) and individual treatment effects (ITEs). The two estimands are often treated as the same but are not, and it is important to draw distinctions between them. An ITE is the difference between the potential outcomes for an

individual and is very challenging (almost impossible) to estimate given that for each individual we only directly observe one of their potential outcomes. In contrast, the CATE represents average effects across individuals in the population that share the same observed covariates (Vegetabile, 2021), and it is more easily estimable. It is thus important to emphasize that the effects discussed in this paper are averages for a covariate profile rather than effect estimates for particular individuals.

The approaches discussed in this paper do have some limitations and opportunity for future refinement. Specifically, we relied on several assumptions to implement the methods, including an assumption of overlapping covariate distributions across the target and trial data. A key concern might be predicting effects for covariate profiles in the hospital data who were not represented in the trials, which would violate our assumption. In the real data analysis, we removed covariate profiles from our target population who had less than moderate depression, but there might be interest in understanding this group further. Future development could work to address this covariate overlap assumption; one approach could be augmenting the trial data with observational data that did have observed treatment and outcomes for this underrepresented group.

Use of EHR data can also come with other challenges, including measurement error. More specifically, the measures in the EHR data might be subjective, highly missing, or different measures of similar constructs compared to the trial data. Future work could explore imputation approaches to deal with missing data in this context or ways to leverage open text data to

fill in more information about patient care. Furthermore, the issue of differential measures that was encountered here – where the EHR data measured depression using a different scale than the trial data – is not an uncommon one. This was addressed in the present analysis by creating a binary variable from both scales to separate into moderate or severe depression; another approach could have been to map the scales to one another continuously, or create standardized measures. These discordant measures can also arise in combining the trials; the trials used here had the same outcome measures, but many other applications of this approach might involve different trial outcomes and predictors that would need to be harmonized.

Future work should examine in more detail the choice of critical value in the prediction interval construction. We chose the t-distribution with $K - 2$ degrees of freedom based on the literature, but the best choice for this critical value might be more complicated (Riley, Higgins, and Deeks, 2011). The true distribution of the treatment effects might not be approximately normal in reality, and future work could further refine the distributional assumptions and work to achieve close to 95% coverage across non-parametric and parametric approaches. Another potential approach could be to explore Bayesian meta-analysis (Higgins, Thompson, and Spiegelhalter, 2009) to leverage prior information about treatment effects when integrating information across trials.

Estimating heterogeneous treatment effects is of high interest in clinical practice but can be challenging statistically. Furthermore, practitioners are often interested in understanding the predicted effects before an individual has received any sort of treatment so that the treatment allocation can be optimal

for patient outcomes. This paper introduced an approach for predicting effects in a target population based on previously conducted trials. With further refinement of the variance and distribution estimation and approaches for effectively dealing with measurement error and non-overlap in the EHR data, we can move closer towards aiding clinical decision-making based on which treatment is predicted to lead to better outcomes for the patient.

4.7 Acknowledgments

The study was funded by the Patient-Centered Outcomes Research Institute (PCORI) through PCORI Award ME-2020C3-21145 (PI: Stuart) and the National Institute of Mental Health (NIMH) through Award R01MH126856 (PI: Stuart). Disclaimer: Opinions and information in this content are those of the study authors and do not necessarily represent the views of PCORI or NIMH. Accordingly, PCORI and NIMH cannot make any guarantees with respect to the accuracy or reliability of the information and data.

This paper is based on research using data from data contributors, Takeda and Lundbeck, made available through Vivli, Inc. Vivli has not contributed to or approved, and is not in any way responsible for, the contents of this publication. This study, carried out under YODA Project 2022-4854, used data obtained from the Yale University Open Data Access Project, in agreement with Janssen Research & Development, L.L.C. The interpretation and reporting of research using this data are solely the responsibility of the authors and do not necessarily represent the official views of the Yale University Open Data Access Project or Janssen Research & Development, L.L.C.

References

Roth, Anthony and Peter Fonagy (2006). "What works for whom?: a critical review of psychotherapy research". In.

Yusuf, Salim, Janet Wittes, Jeffrey Probstfield, and Herman A Tyroler (1991). "Analysis and interpretation of treatment effects in subgroups of patients in randomized clinical trials". In: *Jama* 266.1, pp. 93–98.

Mills, Harriet L, Julian PT Higgins, Richard W Morris, David Kessler, Jon Heron, Nicola Wiles, George Davey Smith, and Kate Tilling (2021). "Detecting heterogeneity of intervention effects using analysis and meta-analysis of differences in variance between trial arms". In: *Epidemiology (Cambridge, Mass.)* 32.6, p. 846.

Brantner, Carly Lupton, Ting-Hsuan Chang, Trang Quynh Nguyen, Hwanhee Hong, Leon Di Stefano, and Elizabeth A. Stuart (2023). "Methods for integrating trials and non-experimental data to examine treatment effect heterogeneity". In: *Statistical Science* 38.4, pp. 640–654.

Colnet, Bénédicte, Imke Mayer, Guanhua Chen, Awa Dieng, Ruohong Li, Gaël Varoquaux, Jean-Philippe Vert, Julie Josse, and Shu Yang (2021). "Causal inference methods for combining randomized trials and observational studies: a review". en. In: *arXiv:2011.08047 [stat]*. URL: http://arxiv.org/abs/2011.08047 (visited on 02/02/2022).

Sobel, Michael, David Madigan, and Wei Wang (2017). "Causal Inference for Meta-Analysis and Multi-Level Data Structures, with Application to Randomized Studies of Vioxx". en. In: *Psychometrika* 82.2, pp. 459–474. ISSN: 0033-3123, 1860-0980. DOI: 10.1007/s11336-016-9507-z. URL: http://link.springer.com/10.1007/s11336-016-9507-z (visited on 03/18/2022).

Dahabreh, Issa J., Lucia C. Petito, Sarah E. Robertson, Miguel A. Hernán, and Jon A. Steingrimsson (2020). "Towards causally interpretable meta-analysis: transporting inferences from multiple studies to a target population". en.

In: *arXiv:1903.11455 [stat]*. URL: http://arxiv.org/abs/1903.11455 (visited on 03/18/2022).

Brantner, Carly Lupton, Trang Quynh Nguyen, Tengjie Tang, Congwen Zhao, Hwanhee Hong, and Elizabeth A Stuart (2024). "Comparison of methods that combine multiple randomized trials to estimate heterogeneous treatment effects". In: *Statistics in Medicine*.

Riley, R. D., J. P. T. Higgins, and J. J. Deeks (2011). "Interpretation of random effects meta-analyses". en. In: *BMJ* 342.feb10 2, pp. d549–d549. ISSN: 0959-8138, 1468-5833. DOI: 10.1136/bmj.d549. URL: https://www.bmj.com/lookup/doi/10.1136/bmj.d549 (visited on 02/21/2023).

Athey, Susan, Julie Tibshirani, and Stefan Wager (2019). "Generalized random forests". In: *The Annals of Statistics* 47.2, pp. 1148–1178.

Hill, Jennifer L. (2011). "Bayesian Nonparametric Modeling for Causal Inference". en. In: *Journal of Computational and Graphical Statistics* 20.1, pp. 217–240. ISSN: 1061-8600, 1537-2715. DOI: 10.1198/jcgs.2010.08162. URL: http://www.tandfonline.com/doi/abs/10.1198/jcgs.2010.08162 (visited on 09/07/2023).

Mahableshwarkar, Atul R., Paula L. Jacobsen, and Yinzhong Chen (2013). "A randomized, double-blind trial of 2.5 mg and 5 mg vortioxetine (Lu AA21004) versus placebo for 8 weeks in adults with major depressive disorder". en. In: *Current Medical Research and Opinion* 29.3, pp. 217–226. ISSN: 0300-7995, 1473-4877. DOI: 10.1185/03007995.2012.761600. URL: http://www.tandfonline.com/doi/full/10.1185/03007995.2012.761600 (visited on 09/22/2022).

Mahableshwarkar, Atul R., Paula L. Jacobsen, Yinzhong Chen, Michael Serenko, and Madhukar H. Trivedi (2015). "A randomized, double-blind, duloxetine-referenced study comparing efficacy and tolerability of 2 fixed doses of vortioxetine in the acute treatment of adults with MDD". en. In: *Psychopharmacology* 232.12, pp. 2061–2070. ISSN: 0033-3158, 1432-2072. DOI: 10.1007/s00213-014-3839-0. URL: http://link.springer.com/10.1007/s00213-014-3839-0 (visited on 09/22/2022).

Boulenger, Jean-Philippe, Henrik Loft, and Christina Kurre Olsen (2014). "Efficacy and safety of vortioxetine (Lu AA21004), 15 and 20 mg/day: a randomized, double-blind, placebo-controlled, duloxetine-referenced study in the acute treatment of adult patients with major depressive disorder". en. In: *International Clinical Psychopharmacology* 29.3, pp. 138–149. ISSN: 0268-1315. DOI: 10.1097/YIC.0000000000000018. URL: http://journals.lww.com/00004850-201405000-00002 (visited on 09/22/2022).

Rubin, Donald B. (1974). "Estimating causal effects of treatments in randomized and nonrandomized studies". In: *Journal of Educational Psychology* 66.5, pp. 688–701. ISSN: 1939-2176. DOI: 10.1037/h0037350.

Riley, Richard D., Thomas P.A. Debray, Tim P. Morris, and Dan Jackson (2021). "The Two-stage Approach to IPD Meta-Analysis". en. In: *Individual Participant Data Meta-Analysis*. John Wiley & Sons, Ltd, pp. 87–125. ISBN: 978-1-119-33378-4. DOI: 10.1002/9781119333784.ch5. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/9781119333784.ch5 (visited on 02/28/2023).

Wager, Stefan and Susan Athey (2018). "Estimation and inference of heterogeneous treatment effects using random forests". In: *Journal of the American Statistical Association* 113.523, pp. 1228–1242.

Carnegie, Nicole, Vincent Dorie, and Jennifer L. Hill (2019). "Examining treatment effect heterogeneity using BART". en. In: *Observational Studies* 5.2, pp. 52–70. ISSN: 2767-3324. DOI: 10.1353/obs.2019.0002. URL: https://muse.jhu.edu/article/793357 (visited on 10/24/2023).

Dorie, Vincent, George Perrett, Jennifer L. Hill, and Benjamin Goodrich (2022). "Stan and BART for Causal Inference: Estimating Heterogeneous Treatment Effects Using the Power of Stan and the Flexibility of Machine Learning". In: *Entropy* 24.12, p. 1782. ISSN: 1099-4300. DOI: 10.3390/e24121782. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9778579/ (visited on 09/18/2023).

Künzel, Sören R, Jasjeet S Sekhon, Peter J Bickel, and Bin Yu (2019). "Metalearners for estimating heterogeneous treatment effects using machine learning". In: *Proceedings of the national academy of sciences* 116.10, pp. 4156–4165.

Montgomery, Stuart A and MARIE Åsberg (1979). "A new depression scale designed to be sensitive to change". In: *The British journal of psychiatry* 134.4, pp. 382–389.

Bates, Douglas M (2010). *lme4: Mixed-effects modeling with R*.

Dorie, Vincent, Hugh Chipman, Robert McCulloch, Armon Dadgar, R Core Team, Guido U Draheim, Maarten Bosmans, Christophe Tournayre, Michael Petch, Rafael de Lucena Valle, et al. (2023). "Package 'dbarts'". In.

Kroenke, Kurt, Robert L Spitzer, and Janet BW Williams (2001). "The PHQ-9: validity of a brief depression severity measure". In: *Journal of general internal medicine* 16.9, pp. 606–613.

Snaith, RP, FM Harrop, t DA Newby, and C Teale (1986). “Grade scores of the Montgomery—Åsberg depression and the clinical anxiety scales”. In: *The British journal of psychiatry* 148.5, pp. 599–601.

Herrmann, Nathan, SE Black, J Lawrence, C Szekely, and JP Szalai (1998). “The Sunnybrook Stroke Study: a prospective study of depressive symptoms and functional outcome”. In: *Stroke* 29.3, pp. 618–624.

Vegetabile, Brian G (2021). “On the distinction between" conditional average treatment effects"(cate) and" individual treatment effects"(ite) under ignorability assumptions”. In: *arXiv preprint arXiv:2108.04939*.

Higgins, Julian P. T., Simon G. Thompson, and David J. Spiegelhalter (2009). “A re-evaluation of random-effects meta-analysis”. en. In: *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 172.1, pp. 137–159. ISSN: 1467-985X. DOI: 10.1111/j.1467-985X.2008.00552.x. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-985X.2008.00552.x (visited on 03/02/2023).

Chapter 5

Discussion and Conclusion

This dissertation has outlined, developed, and implemented data integration methods for estimating treatment effects conditional on observed characteristics. Chapter 2 provided an in-depth review of methods, broken down by their data setting and type, to reveal the current state of the field and areas for future work. This review showed that while several methods exist to combine data and estimate heterogeneous treatment effects, many methods were created in isolation and without concrete code or data applications, therefore making it challenging to determine which methods would work well in practice based on the data and questions of interest. Furthermore, combining trials was mainly handled using meta-analysis, which is interpretable and straightforward to implement, but is not typically geared towards treatment effect heterogeneity and might miss key relationships that are non-linear or complex. Across all methods, more applications to real data were also needed to identify new challenges and methodological considerations that might not be picked up when applying the methods to simulated datasets (Brantner et al., 2023).

The next two chapters in this dissertation addressed some of the openings revealed by Chapter 2. Specifically, Chapter 3 developed aggregation methods to apply non-parametric approaches to estimate the conditional average treatment effect (CATE) function across multiple randomized controlled trials. This chapter revealed more about the benefits and drawbacks of parametric meta-analysis compared to non-parametric approaches like the causal forest (Athey, Tibshirani, and Wager, 2019). Chapter 3 compared several aggregation options and applied them to real data looking at treatment effects of major depression medications, duloxetine and vortioxetine (Brantner et al., 2024).

From Chapter 3, an important next question was determining how to apply the CATE function to a new target population that was distinct from the trial populations. Chapter 4 examined this question, which is of the utmost clinical interest and relevance. This chapter explored a prediction interval-based approach to implement with both meta-analysis and non-parametric methods. In this way, the work done in Chapter 3 could be applied to predict effects in a target population as in Chapter 4. CATE prediction intervals were constructed for sets of covariate profiles and assessed in extensive simulations. Furthermore, the CATE function was estimated by combining three of the RCTs introduced in Chapter 3, and then prediction intervals were formed for covariate profiles represented by real electronic health record data from patients in a health care system.

A primary goal of this dissertation is moving towards making treatment allocation decisions to optimize outcomes. The chapters in this dissertation progressively moved closer to this goal. Chapter 2 started by providing an

overview of the options and opportunities for combining data to improve heterogeneous treatment effect estimation; Chapter 3 estimated these effects by combining multiple trials; and Chapter 4 applied the information gained from multiple trials to a target population.

In Chapters 3 and 4, real data was utilized to identify new considerations when applying the methods outside of a simulated setting. In the real data, the methods demonstrated that duloxetine was estimated to be the more beneficial medication than vortioxetine for nearly everyone in the trials as well as in the target population. There seemed to be minimal heterogeneity if any, where older patients had a larger magnitude of the treatment effect, meaning that duloxetine would be even more beneficial for older individuals, but confidence and prediction intervals were wide. The applications of these methods showed how conclusions can be drawn based on trial data in a target population to help guide treatment decision-making. In this particular case, the lack of substantial heterogeneity could provide more confidence that all patients could receive the same medication, duloxetine, since duloxetine was not only estimated to be preferable on average, but also for almost all covariate profiles found in the trials and target population. This application also demonstrated that even with combination of multiple trials, there were still high levels of uncertainty of the CATE in the trials and wide prediction intervals in the target population. This further supports the fact that estimation of effect heterogeneity is challenging, and more can be done to continue to leverage multiple sources while effectively accounting for uncertainty.

There are several opportunities for future work to continue moving forward in this field of heterogeneous treatment effect estimation using multiple datasets. First, while Chapter 2 introduced methods for bringing together trial and observational data, the subsequent chapters focused primarily on trial data and then an external, target population. Future work could dive more into the methods of combining trial and observational data by applying some of the approaches discussed in Chapter 3 but accounting for confounded treatment assignment in the observational data, and comparing methods using simulated and real data. Bringing in the observational data would add new complications but would considerably increase the sample size and might help extend the covariate distributions of the trials with more strict eligibility criteria. More work could also be done to assess settings when assumptions are violated; for example, this dissertation often relied on the assumption that the trials and target population had similar covariate distributions, but this is often not fully the case in real applications.

A key focus of Chapter 4 was effectively estimating uncertainty of the CATE to produce prediction intervals that were wide enough to capture the truth but not overly conservative. This idea of variance estimation could be further explored, especially in conjunction with new non-parametric methods like the causal forest (Athey, Tibshirani, and Wager, 2019), S- and X-learner (Künzel et al., 2019), and ensemble forests (Tan et al., 2022; Brantner et al., 2024). Many of these newer non-parametric approaches still have more to be done to reliably capture uncertainty, especially when used in conjunction with aggregation methods described in Chapter 3 to combine multiple studies.

Finally, as mentioned previously, bringing in real data reveals important challenges that need to be addressed methodologically. In this dissertation, the key challenges that came up mostly relate to bringing in EHR data. First, the EHR data of patients with depression often had no measures of depression included, and when it did, the measure was a different depression scale (PHQ-9) than the scale used in the trials (MADRS). The EHR data also included some groups of patients who were not represented in the trials due to eligibility criteria. Each of these considerations were addressed briefly in Chapter 4; for example, individuals were only included in the target population if they had non-missing PHQ-9 score. This step excluded a large proportion of available individuals in the EHR data, so future work could investigate other proxies measuring depression severity, like medication changes or hospitalizations. Much more exploration can be done to understand how the methods discussed in this dissertation can be effectively tailored to account for these real data challenges.

With more open data access and careful, ethical sharing of individual participant-level data, questions about treatment effect heterogeneity seem more feasible to answer than before. Leveraging multiple data sources and the methods discussed in this dissertation can continue to examine heterogeneity of the treatment effect and move practice closer towards precise intervention decisions. The work done in the previous chapters did also reveal reasons for caution, in that heterogeneous treatment effects can still involve high levels of uncertainty, and sometimes heterogeneity might not actually be highly prevalent based on the characteristics that are observed. This dissertation aimed

to address the challenges that arise when bringing multiple data sources together, maintaining a focus on a target population of interest and optimizing outcomes for this population. Continual work in this field can harness the data available, account for heterogeneity and uncertainty effectively, and prioritize interpretable results to impact real groups of patients in their treatment journeys.

References

Brantner, Carly Lupton, Ting-Hsuan Chang, Trang Quynh Nguyen, Hwanhee Hong, Leon Di Stefano, and Elizabeth A. Stuart (2023). "Methods for integrating trials and non-experimental data to examine treatment effect heterogeneity". In: *Statistical Science* 38.4, pp. 640–654.

Athey, Susan, Julie Tibshirani, and Stefan Wager (2019). "Generalized random forests". In: *The Annals of Statistics* 47.2, pp. 1148–1178.

Brantner, Carly Lupton, Trang Quynh Nguyen, Tengjie Tang, Congwen Zhao, Hwanhee Hong, and Elizabeth A Stuart (2024). "Comparison of methods that combine multiple randomized trials to estimate heterogeneous treatment effects". In: *Statistics in Medicine*.

Künzel, Sören R, Jasjeet S Sekhon, Peter J Bickel, and Bin Yu (2019). "Metalearners for estimating heterogeneous treatment effects using machine learning". In: *Proceedings of the national academy of sciences* 116.10, pp. 4156–4165.

Tan, Xiaoqing, Chung-Chou H Chang, Ling Zhou, and Lu Tang (2022). "A tree-based model averaging approach for personalized treatment effect estimation from heterogeneous data sources". In: *International Conference on Machine Learning*. PMLR, pp. 21013–21036.

Appendix A

Supplemental Material for Chapter 2

A.1 Single-Study CATE Estimation Methods

In this section, we review several approaches geared towards CATE estimation in a single RCT or observational dataset when we have access to the individual participant-level data (IPD). One option is through a regression, using

$$g(E(Y_i)) = \beta_0 + \beta_a A_i + \boldsymbol{\beta}_x^T \boldsymbol{X}_i + \boldsymbol{\beta}_z^T A_i \boldsymbol{Z}_i,$$

where $\boldsymbol{Z}$ represents effect moderators and is a subset of $\boldsymbol{X}$. Traditionally, regressions like this are used to examine pre-determined subgroups; however, this model can also be used for CATE estimation. Specifically, we can define the CATE from this as

$$\tau(\boldsymbol{Z}_i) = \beta_a + \boldsymbol{\beta}_z^T \boldsymbol{Z}_i.$$

This approach is built upon in the IPD meta-analysis framework discussed in the main paper in the section entitled, "One-Stage IPD Meta-Analysis". In observational data, we can also incorporate propensity score methods before

modeling using the above regression to account for confounded treatment assignment. As another option, $\boldsymbol{Z}$ could be a risk score, rather than a set of effect moderators (Kent et al., 2010; Kent et al., 2020). This approach is the same but requires some preliminary modeling to derive a risk score.

There are also several machine learning methods for CATE estimation in the single study setting that are still relatively new. These methods can be grouped into the following four classes. The first relies mainly on modeling the conditional mean of the outcome given covariates under each intervention (treatment and control), with the CATE function taken as the difference between the two conditional mean outcome functions. This can be achieved via two different models fit to the two treatment groups separately, or via a single model fit to the full sample; these two strategies have been labeled "T-learner" and "S-learner" (with T standing for "two" models and S for "single" model), respectively (Künzel et al., 2019). Based off of these options, the "X-learner" has also been developed. In this approach, one first estimates the conditional mean outcomes in each treatment group, $\mu(\boldsymbol{x}, 0) = E(Y(0)|\boldsymbol{X} = \boldsymbol{x})$ and $\mu(\boldsymbol{x}, 1) = E(Y(1)|\boldsymbol{X} = \boldsymbol{x})$. Next, individual counterfactual outcomes for each treatment group are imputed by using outcome estimators fit to individuals from the *other* group Here, the subscript $i : A = 1$ refers to individual i from the treatment group, and $i : A = 0$ refers to individual i from the control group.

$$\tilde{D}_{i:A=1} = Y_{i:A=1} - \hat{\mu}(\boldsymbol{X}_{i:A=1}, 0)$$

$$\tilde{D}_{i:A=0} = \hat{\mu}(\boldsymbol{X}_{i:A=0}, 1) - Y_{i:A=0}.$$

Finally, a CATE estimator is calculated in each treatment group using a regression with $\tilde{D}_{i:A=1}$ and $\tilde{D}_{i:A=0}$ as outcomes, respectively; the ultimate CATE estimate is then a weighted average (where weights are often estimates of propensity scores) of the CATE functions from each group (Künzel et al., 2019). These methods can utilize approaches like random forests (Breiman, 2001; Athey, Tibshirani, and Wager, 2019) or Bayesian additive regression trees (BART) (Chipman, George, and McCulloch, 2010) to perform the first step of outcome function estimation. A helpful review of these and related methods are in Caron, Baio, and Manolopoulou (2020).

The second class of single-study methods for estimating treatment effect heterogeneity involves transformation of either the outcome or the covariates. An option for transforming the outcome can be written as

$$Y_i^* = Y_i \frac{A_i}{\pi(\boldsymbol{X}_i)} - Y_i \frac{1 - A_i}{1 - \pi(\boldsymbol{X}_i)}$$

where $\pi(\boldsymbol{X}_i)$ are the propensity scores (probability of treatment assignment given covariates) (Signorovitch, 2007; Powers et al., 2018). Since $E(Y_i^* | \boldsymbol{X}_i) = \tau(\boldsymbol{X}_i)$, a regression model can be fit to this transformed outcome to estimate $\tau(\boldsymbol{X})$, obviating the need to model outcome mean functions (treated as "nuisance" parameters). Transformation of the covariates is another option; this transformation is often via some sign flipping and scaling so that the systematic part of the model fit to the transformed variables represents treatment effect variation, while variation in the mean outcome that is unrelated to treatment effect is relegated to the error part of the model. The modified covariate method (MCM) (Tian et al., 2014) is an example of this approach.

The third class similarly includes transformation of the covariates but still estimates nuisance parameters in the process. The "R-learner" (Nie and Wager, 2021) is one example, which first estimates conditional mean outcomes and propensity scores and then uses those estimates in a "quasi-oracle" objective function that is optimized. This class also includes transformed outcome methods, where the first step assembles a function, $f(.)$, such that $E(f(.)|\boldsymbol{X}) = \tau(\boldsymbol{X})$, and the second step regresses $f(.)$ on $\boldsymbol{X}$. Kennedy (2020) uses this type of method through an influence function for the average treatment effect. The Bayesian causal forest (Hahn, Murray, and Carvalho, 2020) also exists within this class. The Bayesian causal forest parameterizes a function f such that

$$f(\boldsymbol{X}, \boldsymbol{A}) = \mu(\boldsymbol{X}, \boldsymbol{\pi}, 0) + \tau(\boldsymbol{X})\boldsymbol{A}$$

where $\mu(\boldsymbol{x}, \pi, 0) = E(Y(0)|\boldsymbol{X} = \boldsymbol{x}, \boldsymbol{\pi} = \pi)$, $\boldsymbol{\pi}$ represent the propensity scores, and $\mu(0)$ and τ have pre-specified prior distributions.

The final class of methods includes trees and forests that partition the covariate space to locally maximize the distance in $\tau(\boldsymbol{X})$ between the sides of each split. Causal inference trees were introduced by Su et al. (2012) who used recursive partitioning to split data into strata by propensity score and treatment effect. Moving forward from this, causal forests have been developed and then extended to "honest" causal forests (Athey, Tibshirani, and Wager, 2019), wherein for each tree, Y_i can only be used in one of the following: to determine the splitting in the tree, or to estimate the treatment effect within a given leaf (Wager and Athey, 2018). The R-learner approach previously mentioned (Nie and Wager, 2021) can also be considered within this class,

as it is a causal forest based on residualized exposure and outcome, and the Bayesian causal forest (Hahn, Murray, and Carvalho, 2020) is also a part of this class as well.

In all classes, many but not all methods also have built in propensity score-based adjustment for confounding for use in non-randomized studies. A review of several such methods in the observational data setting is by Wendling et al. (2018).

References

Kent, David M., Peter M. Rothwell, John PA Ioannidis, Doug G. Altman, and Rodney A. Hayward (2010). "Assessing and reporting heterogeneity in treatment effects in clinical trials: a proposal". In: *Trials* 11.1, p. 85. ISSN: 1745-6215. DOI: 10.1186/1745-6215-11-85. URL: https://doi.org/10.1186/1745-6215-11-85 (visited on 03/30/2022).

Kent, David M, Jessica K Paulus, David Van Klaveren, Ralph D'Agostino, Steve Goodman, Rodney Hayward, John PA Ioannidis, Bray Patrick-Lake, Sally Morton, Michael Pencina, et al. (2020). "The predictive approaches to treatment effect heterogeneity (PATH) statement". In: *Annals of internal medicine* 172.1, pp. 35–45.

Künzel, Sören R, Jasjeet S Sekhon, Peter J Bickel, and Bin Yu (2019). "Metalearners for estimating heterogeneous treatment effects using machine learning". In: *Proceedings of the national academy of sciences* 116.10, pp. 4156–4165.

Breiman, Leo (2001). "Random forests". In: *Machine learning* 45.1, pp. 5–32.

Athey, Susan, Julie Tibshirani, and Stefan Wager (2019). "Generalized random forests". In: *The Annals of Statistics* 47.2, pp. 1148–1178.

Chipman, Hugh A, Edward I George, and Robert E McCulloch (2010). "BART: Bayesian additive regression trees". In: *The Annals of Applied Statistics* 4.1, pp. 266–298.

Caron, Alberto, Gianluca Baio, and Ioanna Manolopoulou (2020). "Estimating individual treatment effects using non-parametric regression models: A review". In: *arXiv preprint arXiv:2009.06472*.

Signorovitch, James E. (2007). "Identifying informative biological markers in high-dimensional genomic data and clinical trials". PhD thesis. Department of Biostatistics, Harvard University School of Public Health.

Powers, Scott, Junyang Qian, Kenneth Jung, Alejandro Schuler, Nigam H Shah, Trevor Hastie, and Robert Tibshirani (2018). "Some methods for heterogeneous treatment effect estimation in high dimensions". In: *Statistics in medicine* 37.11, pp. 1767–1787.

Tian, Lu, Ash A Alizadeh, Andrew J Gentles, and Robert Tibshirani (2014). "A simple method for estimating interactions between a treatment and a large number of covariates". In: *Journal of the American Statistical Association* 109.508, pp. 1517–1532.

Nie, X and S Wager (2021). "Quasi-oracle estimation of heterogeneous treatment effects". en. In: *Biometrika* 108.2, pp. 299–319. ISSN: 0006-3444, 1464-3510. DOI: 10.1093/biomet/asaa076. URL: https://academic.oup.com/biomet/article/108/2/299/5911092 (visited on 03/30/2022).

Kennedy, Edward H. (2020). "Optimal doubly robust estimation of heterogeneous causal effects". In: *arXiv:2004.14497 [math, stat]*. URL: http://arxiv.org/abs/2004.14497 (visited on 03/30/2022).

Hahn, P Richard, Jared S Murray, and Carlos M Carvalho (2020). "Bayesian regression tree models for causal inference: Regularization, confounding, and heterogeneous effects (with discussion)". In: *Bayesian Analysis* 15.3, pp. 965–1056.

Su, Xiaogang, Joseph Kang, Juanjuan Fan, Richard A Levine, and Xin Yan (2012). "Facilitating score and causal inference trees for large observational studies". In: *Journal of Machine Learning Research* 13, p. 2955.

Wager, Stefan and Susan Athey (2018). "Estimation and inference of heterogeneous treatment effects using random forests". In: *Journal of the American Statistical Association* 113.523, pp. 1228–1242.

Wendling, T., K. Jung, A. Callahan, A. Schuler, N. H. Shah, and B. Gallego (2018). "Comparing methods for estimation of heterogeneous treatment effects using observational data from health care databases". en. In: *Statistics in Medicine* 37.23, pp. 3309–3324. ISSN: 02776715. DOI: 10.1002/sim.7820. URL: https://onlinelibrary.wiley.com/doi/10.1002/sim.7820 (visited on 03/30/2022).

Appendix B

Supplemental Material for Chapter 3

B.1 More Simulation Results

K	**Avg Importance for X_1** Mean (SD)	**Avg Importance for $X_2 - X_5$** Mean (SD)	**Avg Importance for 20% Most Heterogeneous Studies** Mean (SD)	**Largest Absolute Value Study Main Coefficient** Mean (SD)
10	0.757 (0.07)	0.008 (<0.01)	0.090 (0.04)	0.906 (0.24)
15	0.783 (0.02)	0.006 (<0.01)	0.059 (0.01)	1.002 (0.25)
20	0.819 (0.02)	0.007 (<0.01)	0.034 (0.01)	1.013 (0.21)
25	0.807 (0.01)	0.044 (0.02)	0.002 (<0.01)	1.105 (0.22)
30	0.722 (0.01)	0.069 (0.03)	<0.001 (<0.01)	1.178 (0.23)

Table B.1: Average variable importance measures across 50 iterations of causal forest with pooling with trial indicator for different values of K (the number of trials).

Data is generated under a setting of a piecewise linear CATE with all trials the same size (n=500), no covariate shift, a study main coefficient standard deviation of 0.5, and a study interaction coefficient standard deviation of 0. Numbers reported represent average and standard deviations of variable importance measures according to the causal forest. The top 20% of studies refer to the studies that had coefficients that were furthest in absolute value from the mean coefficient across all studies, meaning studies that had the most heterogeneity of the treatment effect.

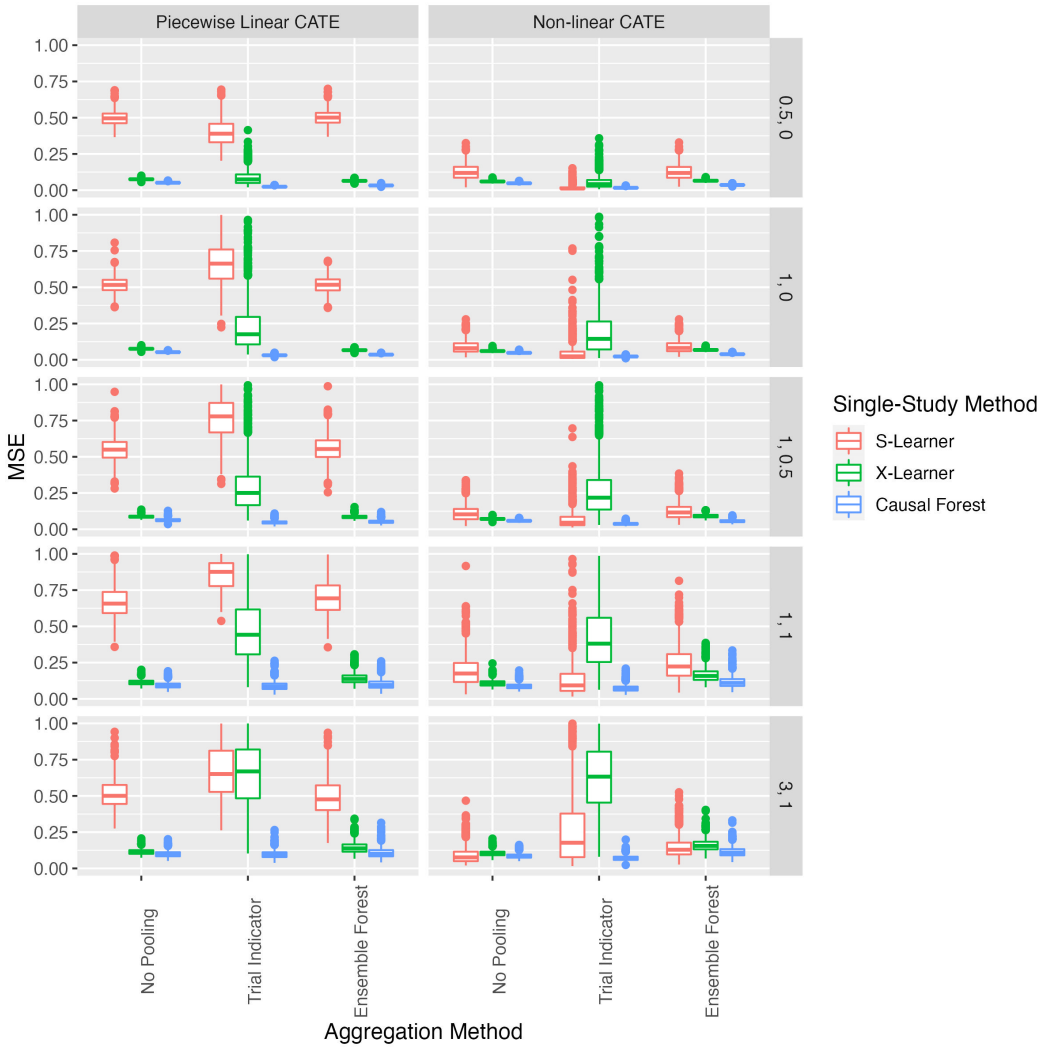

Figure B.1: Distribution of MSE for no pooling versus best performing pooling/ensembling methods

Columns are broken down by simulation scenarios (piecewise linear versus non-linear CATE), and rows are by standard deviation of study main and study interaction coefficients. Y-axis is cutoff for ease of visualization.

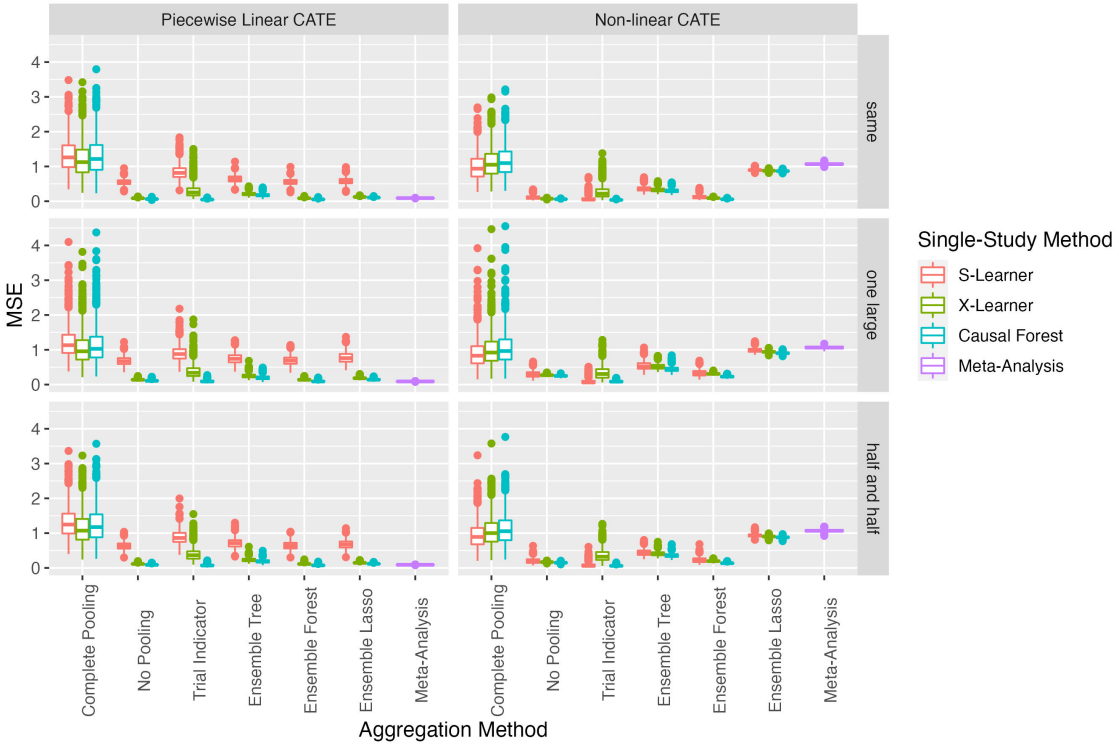

Figure B.2: Distribution of MSE for trials with different sample sizes
Columns are broken down by simulation scenarios (piecewise linear versus non-linear CATE), and rows are by trial sample sizes (same: all trials with n=500, one large: one trial with n=1,000 and the rest with n=200, half and half: five trials with n=500 and five with n=200). SD of study main and study interaction coefficients were 1 and 0.5, respectively for all iterations.

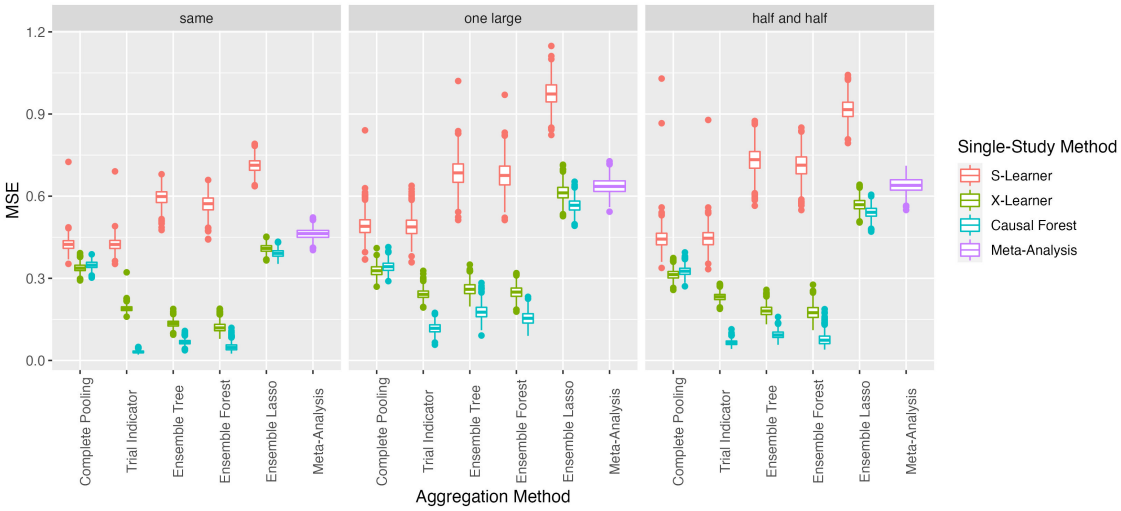

Figure B.3: Distribution of MSE for trials with variable CATE function
Columns are broken down by trial sample sizes (same: all trials with n=500, one large: one trial with n=1,000 and the rest with n=200, half and half: five trials with n=500 and five with n=200). SD of study main and study interaction coefficients were 1 and 0.5, respectively for all iterations.

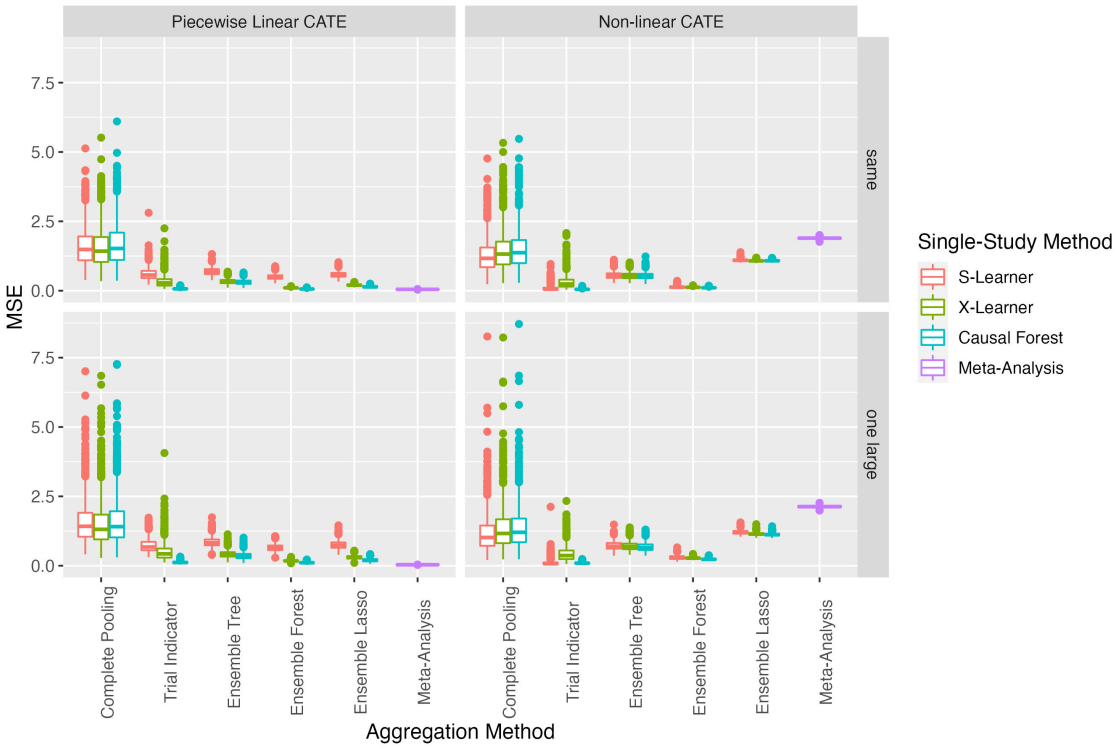

Figure B.4: Distribution of MSE for trials with covariate shift
Columns are broken down by simulation scenarios (piecewise linear versus non-linear CATE), and rows are by trial sample sizes (same: all trials with n=500, one large: one trial with n=1,000 and the rest with n=200). SD of study main and study interaction coefficients were 1 and 0.5, respectively for all iterations.

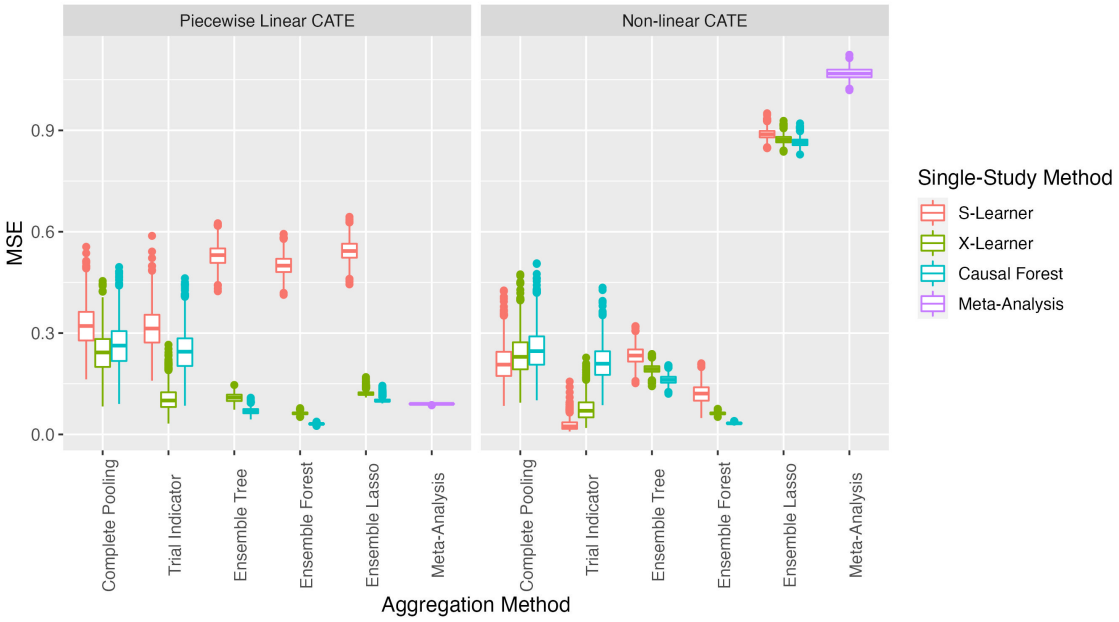

Figure B.5: Distribution of MSE for K=30 trials
Columns are broken down by simulation scenarios (piecewise linear versus non-linear CATE). SD of study main and study interaction coefficients were 0.5 and 0, respectively for all iterations.

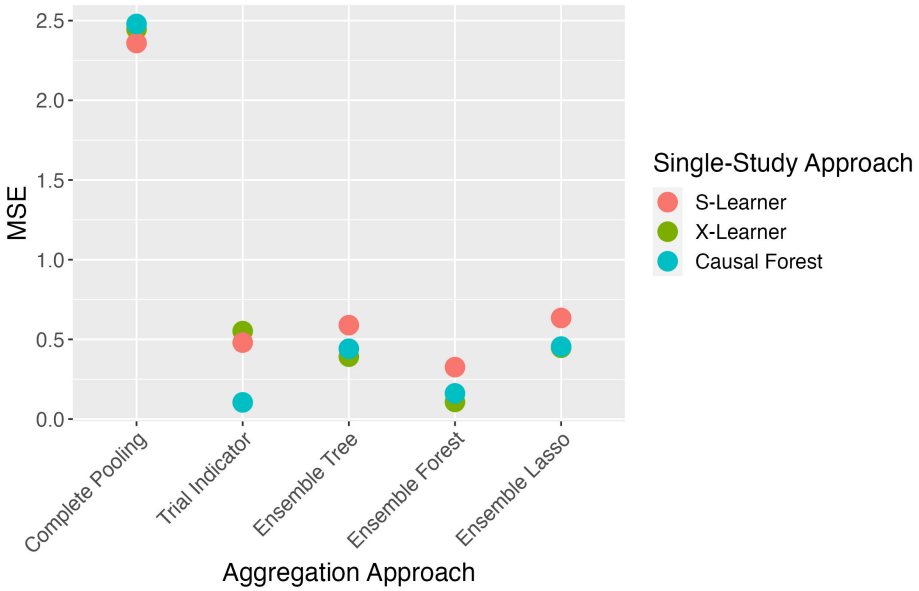

Figure B.6: Average MSE across all scenarios and iterations using honest causal forests.

	NCT00635219		NCT01140906		NCT00672620		NCT01153009	
	Duloxetine (N=134)	Vortioxetine (N=441)	Duloxetine (N=144)	Vortioxetine (N=292)	Duloxetine (N=134)	Vortioxetine (N=284)	Duloxetine (N=140)	Vortioxetine (N=278)
Age								
Mean (SD)	45.9 (11.2)	46.4 (11.6)	45.7 (13.5)	46.7 (14.1)	42.2 (14.8)	43.3 (13.3)	44.0 (12.1)	43.2 (12.2)
Sex								
Female	67.9%	67.6%	70.1%	63.0%	61.2%	65.8%	77.1%	72.7%
Male	32.1%	32.4%	29.9%	37.0%	38.8%	34.2%	22.9%	27.3%
Weight								
Mean (SD)	70.9 (16.1)	70.4 (16.0)	74.8 (17.5)	73.8 (14.8)	87.7 (21.7)	87.1 (25.4)	87.7 (24.6)	87.4 (22.7)
Missing	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	1 (0.4%)	0 (0%)	0 (0%)
Have ever smoked								
Yes	35.8%	31.1%	30.6%	38.4%	23.1%	29.2%	24.3%	26.3%
Have diabetes mellitus								
Yes	3.0%	1.8%	1.4%	1.4%	5.2%	4.6%	2.1%	2.5%
Have hypothyroidism								
Yes	2.2%	2.9%	2.8%	3.4%	6.0%	4.6%	3.6%	5.0%
Have anxiety								
Yes	4.5%	3.4%	0.7%	0%	2.2%	1.8%	6.4%	2.5%
On an antidepressant								
Yes	27.6%	23.4%	38.9%	30.8%	2.2%	1.4%	19.3%	19.4%
On an antipsychotic								
Yes	29.1%	22.0%	20.1%	12.3%	0%	0%	0%	0.4%
On thyroid medication								
Yes	2.2%	2.7%	2.8%	3.8%	1.5%	0.4%	2.1%	4.3%
Baseline MADRS								
Mean (SD)	31.8 (4.0)	32.0 (4.2)	31.1 (3.5)	31.5 (3.4)	29.4 (4.4)	30.0 (4.5)	32.8 (4.3)	32.0 (4.2)
Baseline HAM-A								
Mean (SD)	22.9 (6.3)	23.1 (6.4)	20.5 (6.7)	20.9 (6.8)	17.3 (5.5)	19.0 (5.6)	18.3 (5.6)	17.7 (5.4)
Missing	1 (0.7%)	1 (0.2%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
Difference in MADRS								
Mean (SD)	-17.7 (9.0)	-16.8 (10.3)	-20.4 (9.0)	-16.7 (9.8)	-15.2 (9.1)	-12.0 (9.4)	-16.2 (9.6)	-14.4 (9.8)

Table B.2: Descriptive statistics of participants of four randomized controlled trials, broken down by treatment group.

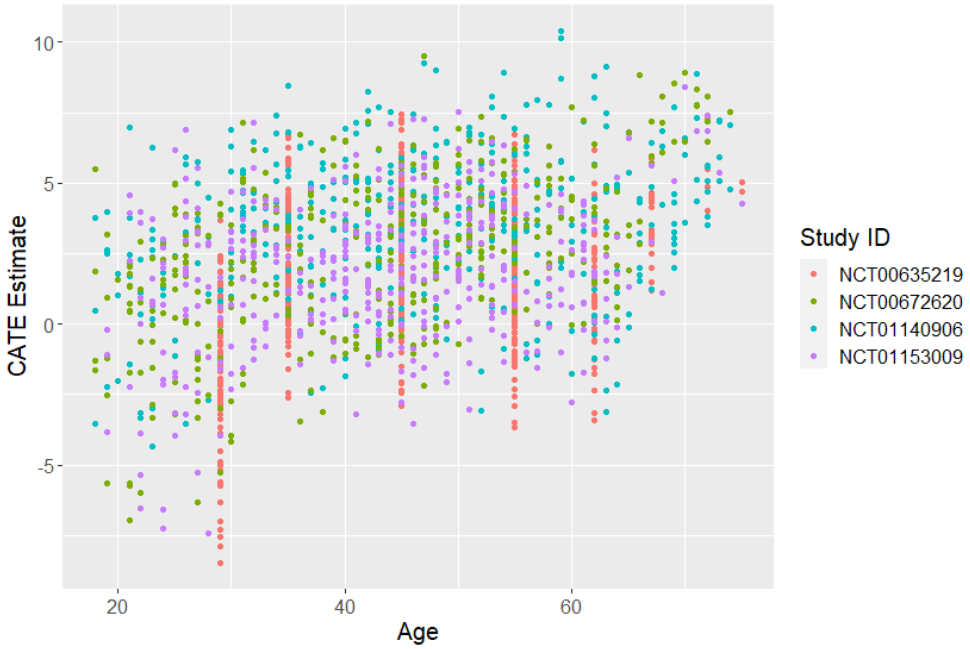

Figure B.7: CATE estimates by age of individual according to causal forest with pooling with trial indicator.

Note that uncertainty of the CATE estimates is not reflected in this plot.

	Estimate	Standard Error	P-Value
(Intercept)	-6.32	5.06	0.21
Age	0.09	0.04	0.03*
Female	0.45	1.07	0.67
Smoker	-1.47	1.10	0.18
Weight	-0.01	0.03	0.72
Baseline MADRS	0.09	0.13	0.49
Baseline HAM-A	0.08	0.09	0.38
Has Diabetes Mellitus	-3.97	3.36	0.24
Has Hypothyroidism	-1.81	3.56	0.61
Has Anxiety	3.58	3.63	0.32
Takes Antidepressant	1.55	1.32	0.24
Takes Antipsychotic	-0.21	1.93	0.91
Takes Thyroid Medication	2.23	4.22	0.60
Study NCT00635219	-1.09	1.63	0.50
Study NCT01140906	2.71	1.62	0.09
Study NCT00672620	2.93	1.58	0.06

Table B.3: Results of best linear projection of the CATE according to the causal forest with pooling with trial indicator.

*Indicates a p-value less than 0.05.

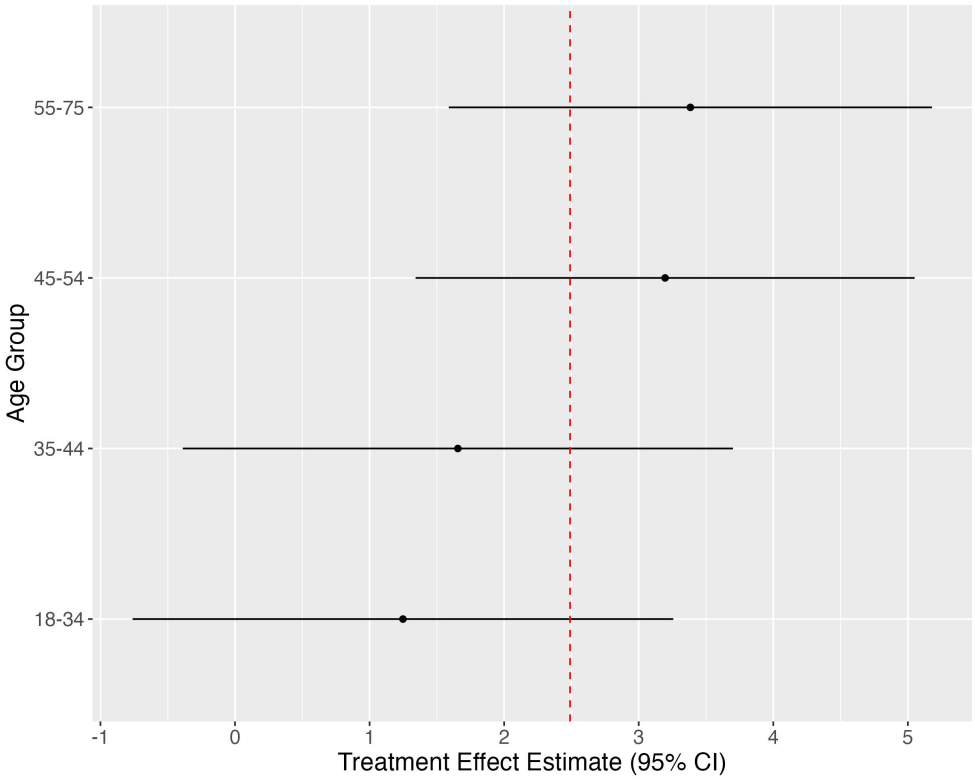

Figure B.8: Average treatment effect with 95% confidence interval by subgroup of age
Vertical red line represents the overall average treatment effect estimate.

Appendix C

Supplemental Material for Chapter 4

C.1 Meta-Analytic Prediction Intervals

This paper extends the prediction interval format towards different estimands, rather than just the overall average treatment effect. In particular, we allow for effect moderation in the form of treatment-covariate interaction terms, which yields treatment effect estimates that are dependent upon a covariate profile. This heterogeneity in the treatment effect is not commonly addressed in meta-analysis, nor is the CATE a common estimand of interest.

We define the meta-analysis model as follows:

$$Y_{si} = (\beta_0 + a_s) + \beta_1 \boldsymbol{X}_{si} + (\beta_2 + b_s) A_{si} + (\boldsymbol{\beta}_3 + \boldsymbol{c}_s) \boldsymbol{X}_{si}^{mod} A_{si} + \epsilon_{si}$$

where $a_s \sim N(0, \sigma_a^2)$, $b_s \sim N(0, \sigma_b^2)$, and $\boldsymbol{c}_s \sim N(\boldsymbol{0}, \Sigma_c)$.

Then we can define the CATE as

$$\tau(\boldsymbol{X}_{si}) = E(Y_{si}(1)|\boldsymbol{X}_{si}) - E(Y_{si}(0)|\boldsymbol{X}_{si})$$

$$= (\beta_2 + b_s) + (\boldsymbol{\beta}_3 + \boldsymbol{c}_s)\boldsymbol{X}_{si}^{mod}$$

We can convert this CATE to matrix notation to facilitate calculation of the variance.

$$\tau(\boldsymbol{X}) = \tilde{\boldsymbol{X}}\tilde{\boldsymbol{\beta}} + \tilde{\boldsymbol{Z}}\tilde{\boldsymbol{u}} = \begin{bmatrix} \tilde{\boldsymbol{X}}_1 \\ ... \\ \tilde{\boldsymbol{X}}_k \end{bmatrix} \tilde{\boldsymbol{\beta}} + \begin{bmatrix} \tilde{\boldsymbol{Z}}_1 & ... & 0 \\ 0 & ... & \tilde{\boldsymbol{Z}}_k \end{bmatrix} \tilde{\boldsymbol{u}}$$

where

$$\tilde{\boldsymbol{X}}_s = \tilde{\boldsymbol{Z}}_s = \begin{bmatrix} 1 & \boldsymbol{X}_{s1} \\ ... & ... \\ 1 & \boldsymbol{X}_{sn} \end{bmatrix} \tilde{\boldsymbol{\beta}} = \begin{bmatrix} \beta_2 \\ \boldsymbol{\beta}_3 \end{bmatrix} \tilde{\boldsymbol{u}}_s = \begin{bmatrix} b_s \\ \boldsymbol{c}_s \end{bmatrix}$$

Then, the variance of the CATE estimate can be calculated as

$$Var(\hat{\tau}(\tilde{\boldsymbol{X}})) = Var(\tilde{\boldsymbol{X}}\hat{\boldsymbol{\beta}} + \tilde{\boldsymbol{Z}}\hat{\boldsymbol{u}})$$

$$= \tilde{\boldsymbol{X}} Var(\hat{\boldsymbol{\beta}})\tilde{\boldsymbol{X}}^T + \tilde{\boldsymbol{Z}} Var(\hat{\boldsymbol{u}})\tilde{\boldsymbol{Z}}^T.$$

C.2 CATE Estimation Using Bayesian Additive Regression Trees

Bayesian additive regression trees (BART) is a sum-of-trees modeling procedure in conjunction with a regularization prior, where the prior restricts the amount that each tree can contribute to the overall model fit. An important distinction with BART is that it estimates the expected outcome conditional on covariates. To apply BART to estimate the CATE, several approaches exist, including the S-learner, T-learner (Künzel et al., 2019), and the Bayesian causal

forest (Hahn, Murray, and Carvalho, 2020). We focus on the S-learner in this paper, as it is relatively straightforward to implement and applies well to the setting with multiple trials.

When using BART as an S-learner and in the setting with multiple trials, we fit a single model to the entire training data to estimate $E(Y|\boldsymbol{X}, A, S)$ – the expected outcome given covariates, treatment, and study membership. As "testing" data, we replicate the training data but assign the opposite treatment to what the individual actually received, to estimate their counterfactual outcome. BART ultimately provides draws from the posterior distribution for the training outcomes under their true and counterfactual treatment assignment, and we can use these draws to estimate the treatment effects conditional on covariates and create credible intervals for these effect estimates. There are two options for how to estimate these credible intervals:

1. We can estimate $E(Y|\boldsymbol{X}, 1, S)$ and $E(Y|\boldsymbol{X}, 0, S)$ by taking the average of the posterior draws for each treatment condition, and we can estimate $\text{Var}(Y|\boldsymbol{X}, 1, S)$ and $\text{Var}(Y|\boldsymbol{X}, 0, S)$ by taking the variance of the posterior draws for each treatment condition. We can then estimate the CATE by defining $\tau_S(\boldsymbol{X}) = E(Y|\boldsymbol{X}, 1, S) - E(Y|\boldsymbol{X}, 0, S)$, and we can estimate the variance of that CATE estimate by adding $\text{Var}(Y|\boldsymbol{X}, 1, S) + \text{Var}(Y|\boldsymbol{X}, 0, S)$, under the conservative assumption that the potential outcomes under treatment and control are uncorrelated. We then create an interval under the assumption that

$$\hat{\tau}_S(\boldsymbol{X}) \sim N(\tau_S(\boldsymbol{X}), \text{Var}(Y|\boldsymbol{X}, 1, S) + \text{Var}(Y|\boldsymbol{X}, 0, S)).$$

2. Instead of aggregating across posterior draws first, we can start by subtracting $(\hat{Y}|\boldsymbol{X}, 1, S) - (\hat{Y}|\boldsymbol{X}, 0, S)$ within each posterior draw. We can then take the mean of these treatment effect estimates as well as the 2.5th and 97.5th percentiles of the CATE across these draws to construct a credible interval.

We use option 1 above in this paper, but did explore both options in preliminary simulations. In our setting, it seemed that the more conservative option 1 performed better, especially when focusing on interval coverage in the target population. Notably, the above options refer to constructing intervals in the original trials. In the target population, we rely on estimating the treatment effect for the given covariate profile within each study, and then accounting for the within- and between-study variance to produce prediction intervals for the new setting.

C.3 More Simulation Results

The following figures display results from secondary simulations with new data generation settings. Figures C.1 and C.2 display coverage results when study distributions varied in baseline MADRS and mean age, respectively.

Figures C.3, C.4, and C.5 display coverage, interval length, and absolute bias across 500 iterations of the same simulation settings as in the main results, but now with $K = 3$ RCTs instead of 10.

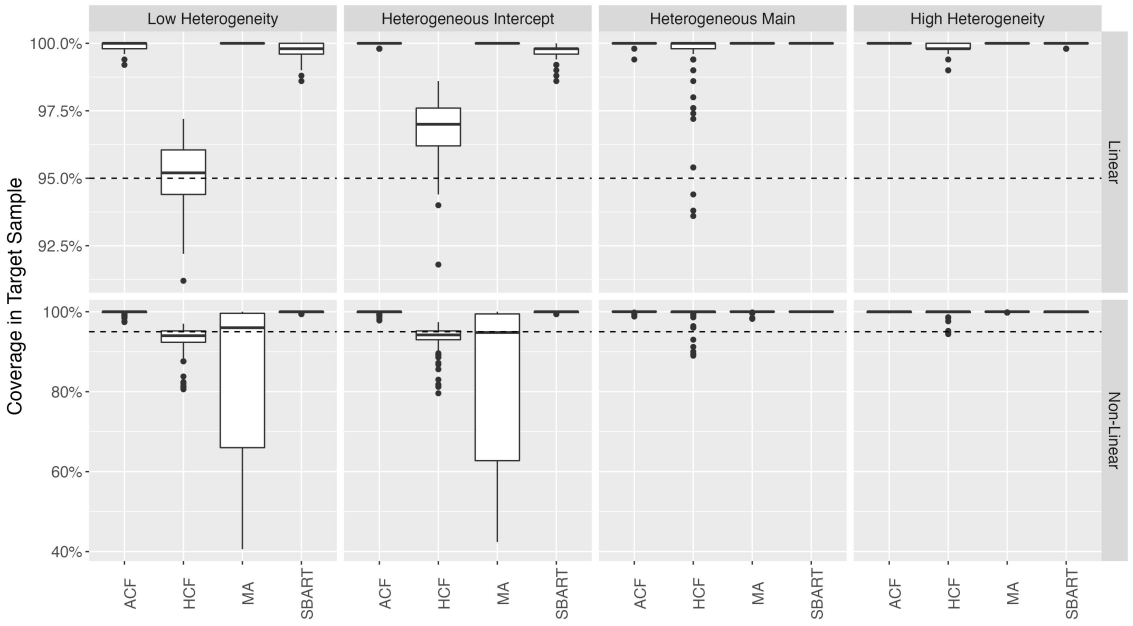

Figure C.1: Distributions of coverage for each covariate profile in the target population across each method and data generation scenario where studies had different mean MADRS score.

Coverage was calculated as the percent of 500 iterations for which the profile's true treatment effect was contained within the estimated prediction interval. *Method abbreviations:* ACF = adaptive causal forest, HCF = honest causal forest, MA = meta-analysis, SBART = Bayesian Additive Regression Trees with S-learner.

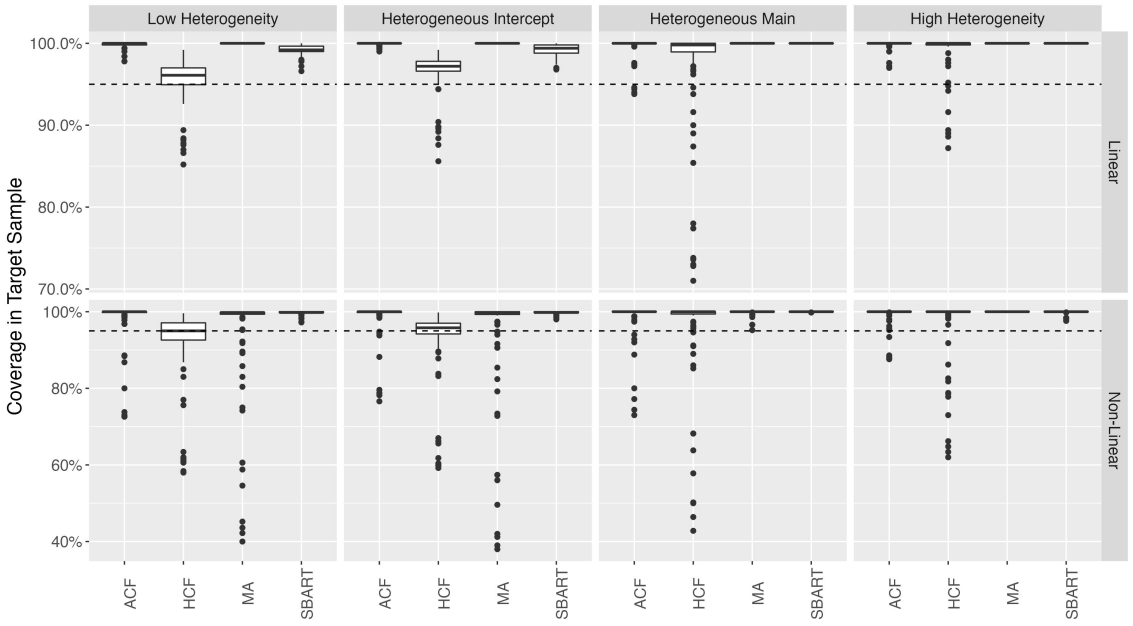

Figure C.2: Distributions of coverage for each covariate profile in the target population across each method and data generation scenario where one study had very different mean age.

Coverage was calculated as the percent of 500 iterations for which the profile's true treatment effect was contained within the estimated prediction interval. *Method abbreviations:* ACF = adaptive causal forest, HCF = honest causal forest, MA = meta-analysis, SBART = Bayesian Additive Regression Trees with S-learner.

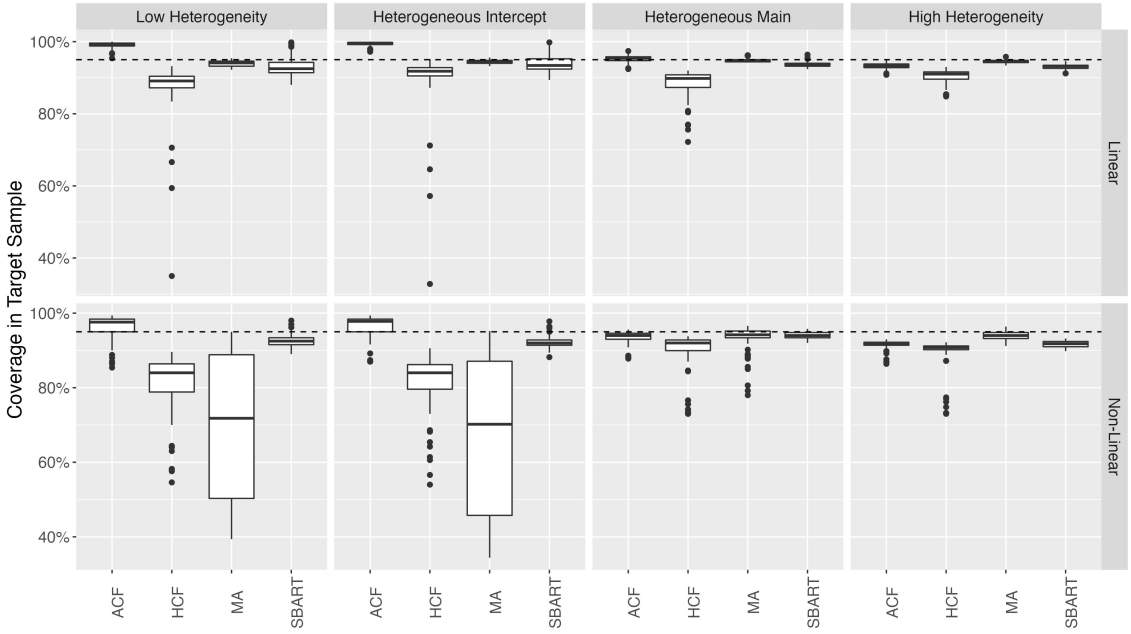

Figure C.3: Distributions of coverage for each covariate profile in the target population across each method and data generation scenario using $K = 3$ RCTs.
Coverage was calculated as the percent of 500 iterations for which the profile's true treatment effect was contained within the estimated prediction interval. *Method abbreviations:* ACF = adaptive causal forest, HCF = honest causal forest, MA = meta-analysis, SBART = Bayesian Additive Regression Trees with S-learner.

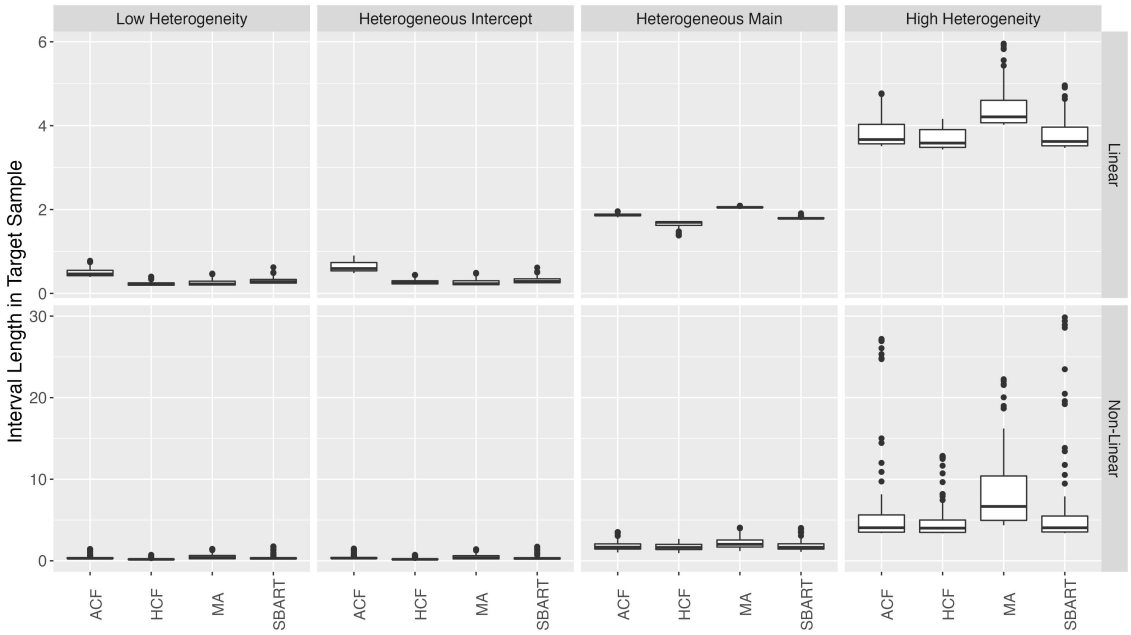

Figure C.4: Distributions of average interval length for each covariate profile in the target population across each method and data generation scenario using $K = 3$ RCTs. Length was calculated as the average length of the profile's prediction interval across 500 iterations. *Method abbreviations:* ACF = adaptive causal forest, HCF = honest causal forest, MA = meta-analysis, SBART = Bayesian Additive Regression Trees with S-learner.

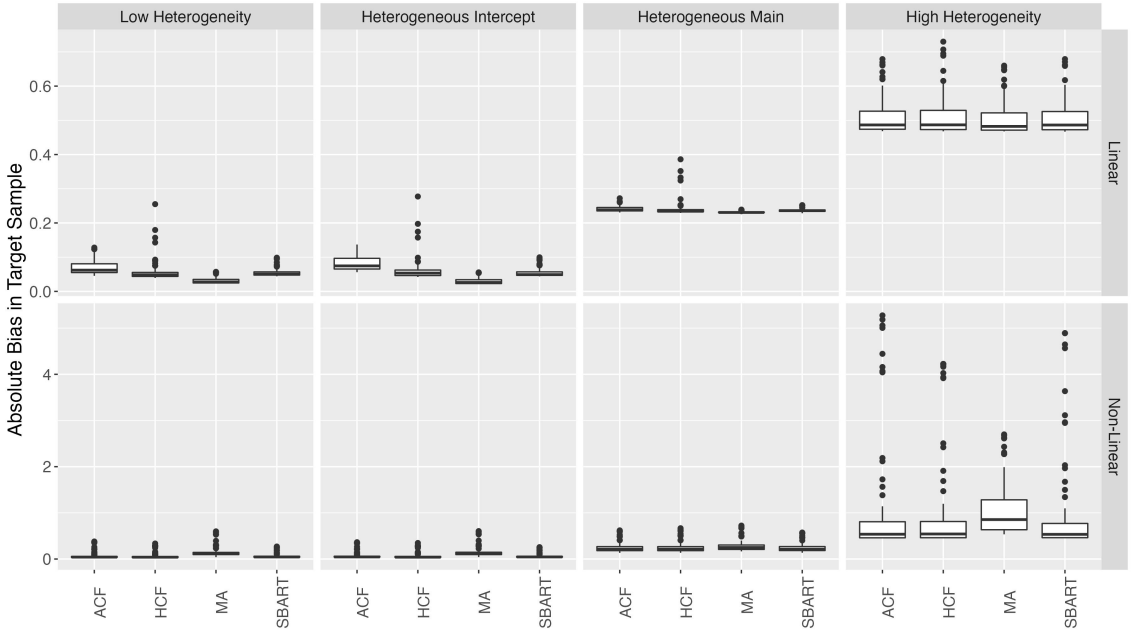

Figure C.5: Distributions of absolute bias for each covariate profile in the target population across each method and data generation scenario using $K = 3$ RCTs.
Absolute bias was calculated as the average absolute difference between the profile's true treatment effect and the estimated treatment effect across 500 iterations. *Method abbreviations:* ACF = adaptive causal forest, HCF = honest causal forest, MA = meta-analysis, SBART = Bayesian Additive Regression Trees with S-learner.

C.4 Duke EHR Cohort

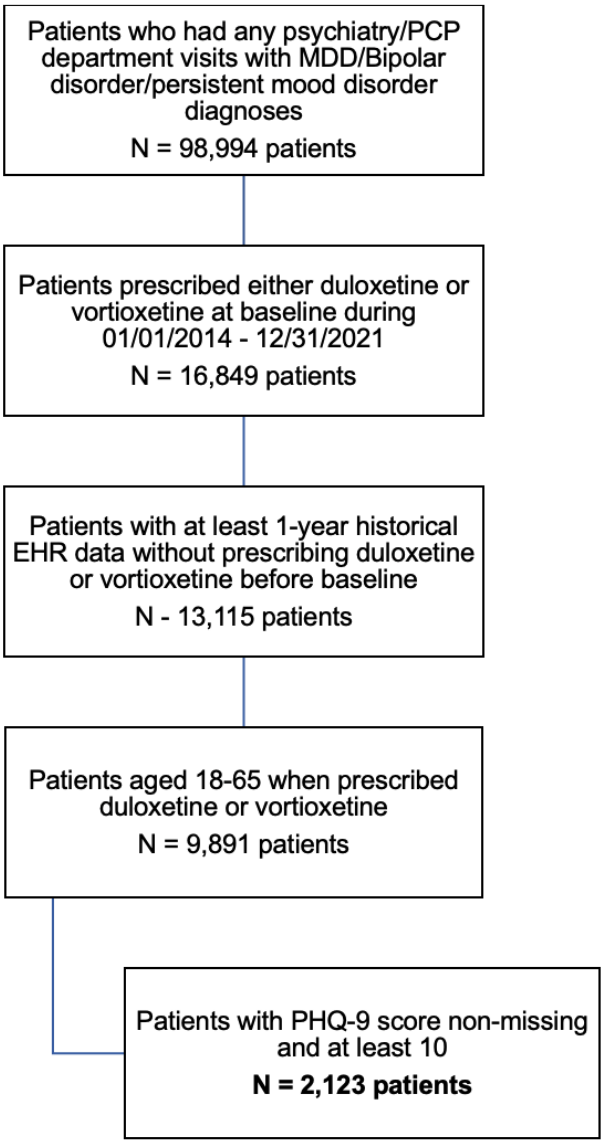

Figure C.6: CONSORT diagram for producing sample of EHR patients from Duke Health Care System.

References

Künzel, Sören R, Jasjeet S Sekhon, Peter J Bickel, and Bin Yu (2019). "Metalearners for estimating heterogeneous treatment effects using machine learning". In: *Proceedings of the national academy of sciences* 116.10, pp. 4156–4165.

Hahn, P. Richard, Jared S. Murray, and Carlos M. Carvalho (2020). "Bayesian Regression Tree Models for Causal Inference: Regularization, Confounding, and Heterogeneous Effects (with Discussion)". In: *Bayesian Analysis* 15.3, pp. 965–1056. ISSN: 1936-0975, 1931-6690. DOI: 10.1214/19-BA1195. URL: https://projecteuclid.org/journals/bayesian-analysis/volume-15/issue-3/Bayesian-Regression-Tree-Models-for-Causal-Inference--Regularization-Confounding/10.1214/19-BA1195.full (visited on 02/21/2023).

Carly Lupton Brantner

Email: clupton1@jhu.edu
Website: www.carlybrantner.com
GitHub: carlyls

Education

In progress	Ph.D.	**Biostatistics** Johns Hopkins Bloomberg School of Public Health Advisor: Dr. Elizabeth A. Stuart
2020	M.S.	**Applied Mathematics and Statistics** Johns Hopkins University
2020	B.S.	**Applied Mathematics and Statistics, Psychological and Brain Sciences** Johns Hopkins University

Professional Experience

2023	**Senior Data Science Intern**, Mathematica Policy Research
2019 – present	**Research Assistant**, Johns Hopkins Bloomberg School of Public Health Advisor: Dr. Elizabeth A. Stuart
2021 – present	**Teaching Assistant**, Johns Hopkins Bloomberg School of Public Health
2020 – 2023	**Epidemiology and Biostatistics of Aging Training Grant Trainee**, Johns Hopkins Bloomberg School of Public Health PI: Dr. Karen Bandeen-Roche
2019	**CIPhER Intern in Education Research and Assessment**, UNC Eshelman School of Pharmacy
2018 – 2020	**Course Assistant**, Center for Leadership Education at Johns Hopkins University
2018	**Student Researcher**, Summer Institute in Biostatistics (SIBS) at North Carolina State University
2017	**Applied Statistics Intern**, UNC Eshelman School of Pharmacy

Publications

Peer-Reviewed Articles

1. **Brantner, C. L.,** Nguyen, T. Q., Tang, T., Zhao, C., Hong, H., & Stuart, E. A. (2024). Comparison of methods that combine multiple randomized trials to estimate heterogeneous treatment effects. *Statistics in Medicine.*

2. Nguyen T. Q., Roberts Lavigne L. C., **Brantner C. L.,** Kirk G. D., Mehta S. H., Linton S. L. (2024). Estimation of place-based vulnerability scores for HIV viral non-suppression: an application leveraging data from a cohort of people with histories of using drugs. *BMC Medical Research Methodology*, *24*:21. doi: 10.1186/s12874-023-02133-x
3. **Brantner, C. L.,** Chang, T., Nguyen, T. Q., Hong, H., Di Stefano, L., & Stuart, E. A. (2023). Methods for integrating trials and non-experimental data to examine treatment effect heterogeneity. *Statistical Science, 38*(4), 640-654*.* doi:10.1214/23-STS890
4. **Brantner, C. L.,** Bentley, J. P., & Roth, D. L. (2023). Subtypes of transitions into a family caregiving role: A latent class analysis. *Journal of Applied Gerontology.* doi: 10.1177/07334648231210680
5. Ettman, C. K., **Brantner, C. L.,** Albert, M., Goes, F., Mojtabai, R., Spivak, S., Stuart, E. A., & Zandi, P. P. (2023). Trends in telehealth and in-person psychiatric care from 2017-2022 among patients with depression in a large US academic medical system. *Psychiatric Services.* doi:10.1176/appi.ps.20230064
6. Olsen, A. A., **Brantner, C. L.,** Dallaghan, G. L. B., & McLaughlin, J. E. (2023). A review of interprofessional education research: Disciplines, authorship practices, research design, and dissemination trends. *Journal of Interprofessional Education & Practice.* doi:10.1016/j.xjep.2023.100653
7. **Lupton-Smith, C.,** Badillo-Goicoechea, E., Collins, M., Lessler, J., Grabowski, M. K., & Stuart, E. A. (2022). Consistency between household and county measures of onsite schooling during the COVID-19 pandemic. *Journal of Research on Educational Effectiveness.* doi:10.1080/19345747.2022.2131660
8. McLaughlin, J. E., Lyons, K., **Lupton-Smith, C.,** & Fuller, K. (2022). An introduction to text analytics for educators. *Currents in Pharmacy Teaching and Learning.* doi:10.1016/j.cptl.2022.09.005
9. **Lupton-Smith, C.,** Badillo-Goicoechea, E., Chang, T. Maniates, H., Riehm, K. E., Schmid, I., & Stuart, E. A. (2022). Factors associated with county-level mental health during the COVID-19 pandemic. *Journal of Community Psychology, 50*(5), 2431-2442*.* doi:10.1002/jcop.22785
10. Riehm, K. E., Badillo-Goicoechea, E., Wang, F. M., Kim, E., Aldridge, L. R., **Lupton-Smith, C.**, Presskreischer, R., Chang, T., LaRocca, S., Kreuter, F., & Stuart, E. A. (2022). Association of Non-Pharmaceutical Interventions with Anxiety and Depressive Symptoms during the COVID-19 Pandemic: A Multi-National Study of 43 Countries. *International Journal of Public Health, 67,* 1604430. doi:10.3389/ijph.2022.1604430
11. Lessler, J., Grabowski, M. K., Grantz, K. H., Badillo-Goicoechea, E., Metcalf, C. J. E., **Lupton-Smith, C.,** Azman, A. S., Stuart, E. A. (2021). Household COVID-19 risk and in-person schooling. *Science, 372*(6546), 1092-1097. doi:10.1126/science.abh2939
12. **Lupton-Smith, C.,** Stuart, E., McGinty, B., Dalcin, A., Jerome, G. J., Wang, N. Y., & Daumit, G. (2021). Determining predictors of weight loss in a behavioral intervention:

A case study in the use of Lasso regression. *Frontiers in Psychiatry, 12*, 707707. doi:10.3389/fpsyt.2021.707707

13. Olsen, A., **Lupton-Smith, C.,** Rodgers, P., & McLaughlin, J. E. (2021). Characterizing research about interprofessional education within pharmacy. *American Journal of Pharmaceutical Education, 85*(8), 8541. doi:10.5688/ajpe8541.
14. Wolcott, M. D., **Lupton-Smith, C**., Cox, W. C., & McLaughlin, J. E. (2019). A 5-minute Situational Judgment Test to assess empathy in first-year student pharmacists*. American Journal of Pharmaceutical Education, 83*(6), Article 6960. doi:10.5688/ajpe6960
15. McLaughlin, J. E., **Lupton-Smith, C**., & Wolcott, M. D. (2018). Text mining as a method for examining the alignment between educational outcomes and the workforce needs. *Education in the Health Professions, 1*(2), 55-60. doi:10.4103/EHP.EHP_25_18

In Progress/Under Review

1. **Brantner, C. L.,** Nguyen, T. Q., & Stuart, E. A. Combining trials to estimate heterogeneous treatment effects in a target sample. In progress.
2. Ettman, C. E., **Brantner, C. L.,** Ringlein, G., Stuart, E. A., & Zandi, P. Trends in continuous treatment of mental health before and during the COVID-19 pandemic among patients with depression. In progress.

Presentations

Presentations at Scientific Meetings

1. **Brantner, C. L.** March 10, 2024 (accepted). Combining trials to estimate heterogeneous treatment effects in a target sample. Eastern North American Region (ENAR) International Biometrics Society. *Part of invited session (organizer: **Brantner, C. L.**): Integrating Data from Multiple Sources to Estimate Causal Effects.*
2. **Brantner, C. L.** March 22, 2023. Combining datasets to estimate heterogeneous treatment effects. Eastern North American Region (ENAR) International Biometrics Society.
3. **Brantner, C. L.** December 18, 2022. Combining datasets to estimate heterogeneous treatment effects. International Conference of the ERCIM WG on Computational and Methodological Statistics (CMStatistics 2022).
4. **Lupton-Smith, C.** January 24, 2022. Family caregiving subtypes and well-being in the Caregiving Transitions Study: A latent class analysis. Research in Progress presentation for Epidemiology and Biostatistics of Aging Training Group.
5. **Lupton-Smith, C**. October 8, 2021. Factors associated with county-level mental health during the COVID-19 pandemic. Invited presentation for Hopkins Biostatistics Virtual Prospective Student Event.
6. **Lupton-Smith, C.,** Badillo Goicoechea, E., Chang, T., & Stuart, E. A. October 7, 2021. Factors associated with county-level mental health during the COVID-19 pandemic. Women in Statistics and Data Science Conference.

Posters

1. **Brantner, C. L.** (2023). Combining datasets to estimate heterogeneous treatment effects. *American Causal Inference Conference.*
2. Ettman, C. K., **Brantner, C. L.,** Albert, M., et al. (2023). Trends in telehealth and in-person psychiatric care from 2017-2022 among patients with depression in a large US academic medical system. *Society for Epidemiologic Research.*
3. Ettman, C. K., **Brantner, C. L.,** Albert, M., et al. (2023). Trends in telehealth and in-person psychiatric care from 2017-2022 among patients with depression in a large US academic medical system. *AcademyHealth.*
4. Ettman, C. K., **Brantner, C. L.,** Albert, M., et al. (2023). Trends in telehealth and in-person psychiatric care from 2017-2022 among patients with depression in a large US academic medical system. *International Association for Population Health Sciences.*
5. Bentley, J., **Lupton-Smith, C.,** & Roth, D. (2021). Family caregiving subtypes in the Caregiving Transitions Study: A latent class analysis. *Gerontological Society of America.*
6. Chang, T., **Lupton-Smith, C.**, Badillo Goicoechea, E., & Stuart, E. A. (2021). Examining county-level mental health using Facebook COVID-19 Symptom Survey data. *American Psychopathological Association.*
7. McLaughlin, J. E., **Lupton-Smith, C**., Bell, E. L., Hubal, R., Persky, A. (2020). Automated analysis of course evaluation comments: The use of sentiment analysis to characterize classroom teaching. *American Association of Colleges of Pharmacy.* ***Recipient of AACP SAS Best Poster Award**

Podcasts

1. **Brantner, C. L.** May 30, 2022. School Reopening Analysis. Data Skeptic. https://dataskeptic.com/blog/episodes/2022/school-reopening-analysis.

Leadership Experience

2023 – present	**Council of Emerging and New Statisticians Member**, Eastern North American Region (ENAR) of International Biometrics Society
2022 – present	**Co-President of Biostatistics Student Organization**, Johns Hopkins Bloomberg School of Public Health
2021 – present	**Mental Health Graduate Network Representative**, Johns Hopkins University
2020 – present	**Biostatistics Student Organization Mentoring Committee and Curriculum Committee Member**, Johns Hopkins Bloomberg School of Public Health
2019	**Varsity Women's Soccer Captain**, Johns Hopkins University

2019 **Assistant New Member Director**, Phi Mu Gamma Tau Chapter

2019 – 2020 **Blue Jays LEAD Captains Cohort Member**, Johns Hopkins University

Honors and Awards

1. **Gertrude M. Cox Scholarship Honorable Mention,** American Statistical Association, 2023
2. **June B. Culley Award** for the best preliminary examination paper, Johns Hopkins Department of Biostatistics, 2023
3. **Jane and Steve Dykacz Award** for an outstanding paper in medical statistics, Johns Hopkins Department of Biostatistics, 2023
4. **Epidemiology and Biostatistics of Aging Training Grant Recipient,** Johns Hopkins University, 2020-2023
5. **Dean's List**, Johns Hopkins University, 2016-2020
6. **Phi Beta Kappa,** Johns Hopkins University, 2020
7. **Applied Mathematics and Statistics Achievement Award,** Johns Hopkins University, 2020
8. **Applied Mathematics and Statistics Prize for Outstanding Master's Research,** Johns Hopkins University, 2020
9. **Applied Mathematics and Statistics Mathematical Modeling Competition Prize,** Johns Hopkins University, 2020
10. **Naddor Prize,** Johns Hopkins University Department of Applied Mathematics and Statistics, 2019
11. **Centennial Conference Women's Soccer Scholar-Athlete of the Year,** Johns Hopkins Women's Soccer, 2019
12. **CoSIDA Third Team Academic All-American,** Johns Hopkins Women's Soccer, 2019
13. **CoSIDA Academic All-District Team,** Johns Hopkins Women's Soccer, 2019
14. **Chi Alpha Sigma,** Scholar Athlete Honor Society, Johns Hopkins University, 2019

Teaching Experience

Lead Teaching Assistant, Johns Hopkins Bloomberg School of Public Health

1. Statistical Methods in Public Health I, II — Fall 2022 & 2023
   Instructors: Dr. Marie Diener-West and Dr. Karen Bandeen-Roche

Teaching Assistant, Johns Hopkins Bloomberg School of Public Health

1. Causal Inference in Medicine and Public Health I — Spring 2022 & 2023
   Instructor: Dr. Elizabeth Stuart
2. Statistical Methods in Public Health I, II — Fall 2021
   Instructors: Dr. Marie Diener-West and Dr. Karen Bandeen-Roche

Course Assistant, Johns Hopkins University

1. Leading Teams — Fall 2018 – Spring 2020
   Instructor: Dr. William Smedick

Skills

Programming: R, SAS, MPlus, Python, MATLAB, Stata, SQL, Git
Research: Machine learning, data integration, causal inference, multiple imputation, nonparametric statistics, propensity score estimation, variable selection, regression, latent class analysis, electronic health records, text mining, sentiment analysis, social network analysis