Published in final edited form as:

Stat Sci. 2023 November ; 38(4): 640–654. doi:10.1214/23-sts890.

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

Carly Lupton Brantner [PhD Candidate],
Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, Maryland 21205, USA

Ting-Hsuan Chang [PhD Student],
Department of Biostatistics, Columbia Mailman School of Public Health, New York, New York 10032, USA

Trang Quynh Nguyen [Associate Scientist],
Department of Mental Health, Johns Hopkins Bloomberg School of Public Health, Baltimore, Maryland 21205, USA

Hwanhee Hong [Associate Professor],
Department of Biostatistics and Bioinformatics, Duke University, Durham, North Carolina 27710, USA

Leon Di Stefano [PhD Candidate],
Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, Maryland 21205, USA

Elizabeth A. Stuart [Professor]
Departments of Biostatistics, Mental Health, and Health Policy and Management, Johns Hopkins Bloomberg School of Public Health, Baltimore, Maryland 21205, USA

Abstract

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 participant-level data. We define the conditional average treatment effect and discuss differences between parametric and

---

clupton1@jhu.edu .

SUPPLEMENTARY MATERIAL

Single-Study CATE Estimation Methods (DOI: 10.1214/23-STS890SUPP;.pdf). This supplement provides an overview of approaches that estimate the conditional average treatment effect (CATE) in a single randomized controlled trial or observational dataset. Both parametric and nonparametric methods are included, and the nonparametric methods are grouped into classes to help differentiate the approaches.

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.

Keywords

Treatment effect heterogeneity; combining data; generalizability and reproducibility

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 under-represented 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, 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 prespecified, 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 (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, 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. (2021a) 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. (2021a); 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., 2023a). 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), 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 3 discusses aggregate-level data; Section 4, federated learning; and Section 5, individual participant-level data (IPD). Finally, Section 6 compares methods and provides an overview of potential future areas for research.

2. NOTATION

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 $\boldsymbol{X}$:

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

(Abrevaya, Hsu and Lieli, 2015, Künzel et al., 2019), where $E[\cdot \mid \cdot]$ denotes conditional expectation in the target population of interest and $g(\cdot)$ 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

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

This $\tau(\cdot)$ 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 (1) defines a general CATE that is not study-specific. When estimating study-specific CATEs, equation (1) can be rewritten as

$$\tau_s(\boldsymbol{X}) = E[Y(1) - Y(0) \mid \boldsymbol{X}, S = s]. \tag{2}$$

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

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

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

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, 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 $h(\cdot)$ 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 prespecified, and $\tau(\boldsymbol{Z})$ is the the CATE function:

$$\tau(\boldsymbol{Z}) = \delta + \boldsymbol{\theta}^T \boldsymbol{Z}\,. \tag{3}$$

In this expression for $\tau(\boldsymbol{Z}), \delta$ 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 (1), this parametric form of $\tau(\boldsymbol{Z})$ can be expressed as multiple study-specific functions:

$$\tau_s(\boldsymbol{Z}) = \delta_s + \boldsymbol{\theta}_s^T \boldsymbol{Z}\,. \tag{4}$$

When combining an RCT with an observational dataset, there are a few within-study assumptions, including unconfoundedness (Assumption 1), positivity (Assumption 2), and consistency (Assumption 3) (Colnet et al., 2021a, Cheng and Cai, 2021).

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

Assumption 2. For almost all $\boldsymbol{X}$ with $\pi(\boldsymbol{X}) = P(A = 1 \mid \boldsymbol{X})$ (the propensity score), there exists a constant $c > 0$ such that $c < \pi(\boldsymbol{X}) < 1 - c$ within each study.

Assumption 3. $Y = AY(1) + (1 - A)Y(0)$ almost surely.

The unconfoundedness assumption (1) is satisfied by design in an RCT. Assumption 2 also holds by design in an RCT since the probability of treatment is independent of observed covariates and is prespecified.

When combining datasets, we expand upon the previous assumptions. In the setting where observational data is being combined with an RCT, the unconfoundedness assumption (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, 2022). Assumption 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 4) (Colnet et al., 2021a, Cheng and Cai, 2021) and unconfounded study membership (Assumption 5) (Hatt et al., 2022, Cheng and Cai, 2021, Kallus, Puli and Shalit, 2018).

Assumption 4. For almost all $\boldsymbol{X}$, there exists a constant $d > 0$ such that $d < P(S = s \mid \boldsymbol{X} = \boldsymbol{x}) < 1 - d$.

Assumption 5. $\{Y(0), Y(1)\} \perp\!\!\!\perp S \mid \boldsymbol{X}$.

The following sections break down methods based on available data.

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.

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 (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., 2023).

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 15 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 (3) but as $\tau(\overline{\boldsymbol{Z}})$ where $\overline{\boldsymbol{Z}}$ 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.

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., 2022). 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.

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 3.1. A key difference here is that federated learning models apply a predetermined 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 (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.

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, Brantner et al., 2023b). 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)) 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.

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 Material (Brantner et al., 2023b).

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.

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).

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

$$g(E(Y_{\text{is}})) = \alpha_s + \delta_s A_{\text{is}} + \boldsymbol{\beta}_s^T \boldsymbol{X}_{\text{is}} + \boldsymbol{\theta}_s^T A_{\text{is}} \boldsymbol{Z}_{\text{is}}, \tag{5}$$

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, $\sigma_\alpha^2, \sigma_\delta^2$ and the diagonal elements of $\boldsymbol{\Sigma}_{\boldsymbol{\beta}}$ and $\boldsymbol{\Sigma}_{\boldsymbol{\theta}}$ measure the between-study variability of the effects. $\boldsymbol{\beta}_s$ and $\boldsymbol{\theta}_s$ are often assumed to be uncorrelated in the literature; however, we can extend this model to allow for correlation between $\boldsymbol{\beta}_s$ and $\boldsymbol{\theta}_s$.

If the outcome is continuous (as assumed in this paper), $g(\cdot)$ is often set to be the identity function; if the outcome is binary, $g(\cdot)$ 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 (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 (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 noninformative 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 (5):

$$
\begin{aligned}
& g(E(Y_{\text{is}})) = \\
& \alpha_s + \delta_s A_{\text{is}} + \boldsymbol{\beta}_{s,\ \text{within}}^T \left(\boldsymbol{X}_{\text{is}} - \overline{\boldsymbol{X}}_s \right) \\
& + \boldsymbol{\beta}_{\text{across}}^T \overline{\boldsymbol{X}}_s + \boldsymbol{\theta}_{s,\ \text{within}}^T A_{\text{is}} \left(\boldsymbol{Z}_{\text{is}} - \overline{\boldsymbol{Z}}_s \right) \\
& + \boldsymbol{\theta}_{\text{across}}^T A_{\text{is}} \overline{\boldsymbol{Z}}_s \, .
\end{aligned}
$$

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 (5) is a special case of this model when $\boldsymbol{\beta}_{\text{across}}$ and $\boldsymbol{\theta}_{\text{across}}$ are equal to the average of the $\boldsymbol{\beta}_{s,\text{within}}$'s and the $\boldsymbol{\theta}_{s,\text{within}}$'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.

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 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.

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., 2023a). 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. (2021a) 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, 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.

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:

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

(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 (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

$$\hat{\tau}_k(\lambda) = \hat{\tau}_k^r - \lambda(\hat{\tau}_k^r - \hat{\tau}_k^o), \tag{7}$$

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 (6). They also discuss an estimator that is the same as equation (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, $\hat{\tau}_s^o(\boldsymbol{X})$:

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

where $s = 0{,}1$ denotes RCT and observational data, respectively and $\upsilon_{\boldsymbol{X}}$ 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, Brantner et al., 2023b). 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 $\upsilon_{\boldsymbol{X}}$ 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 $\upsilon_{\boldsymbol{X}}$ 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 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 1 in the observational data) is sufficiently met or not (Yang, Zeng and Wang, 2020). To construct this test, Yang et al. (2020) introduce

$$H_{\psi 0}(\boldsymbol{X}) = Y - \tau_{\psi 0}(\boldsymbol{X})A \tag{8}$$

such that $E(H_{\psi_0} \mid A, \boldsymbol{X}, S) = E(Y(0) \mid A, \boldsymbol{X}, S)$. From here, they introduce a semiparametric efficient score of the parameters ψ_0 which we will call $\mathrm{SES}_{\psi 0}$. This semiparametric efficient score is used in their hypothesis test with a null hypothesis of $E(\mathrm{SES}_{\psi 0}^o) = 0$ where $\mathrm{SES}_{\psi 0}^o$ 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

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

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

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 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., 2021a). 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, Brantner et al., 2023b), and the confounding effect is estimated using the following equation. For the propensity score in the RCT, $\pi^r(\boldsymbol{X}) = P(A = 1 \mid \boldsymbol{X}, S = 0)$, 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:

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \sum_{i=1}^{n^r} \left(q^r(\boldsymbol{X}_i)Y_i - \hat{\tau}^o(\boldsymbol{X}_i) - \boldsymbol{\theta}^T \boldsymbol{X}_i\right)^2$$

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

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

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), $\tau(\boldsymbol{X}) = E[Y(1) - Y(0) \mid \boldsymbol{X}, A = 1]$, and define a confounding function to estimate the effect of unobserved confounding in the observational data. They assume unconfoundedness in the RCT (Assumption 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:

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

When all confounders are measured, $\zeta(\boldsymbol{X}) = 0$, but in reality, unobserved confounders will lead the function to be nonzero. 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 (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 $S = 1$, that is, 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, Brantner et al., 2023b) 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:

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

where $c(X) = 0$ when there is no unobserved confounding in the observational dataset (Assumption 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 3), common support across the RCT and observational data (Assumption 4), and unconfoundedness in the RCT (Assumption 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,

$$E\left(Y^r \mid A^r = a, \boldsymbol{X}^r = \boldsymbol{x}\right) - E\left(Y^o \mid A^o = a, \boldsymbol{X}^o = \boldsymbol{x}\right) = h_a^r(\phi^*(\boldsymbol{x})) - h_a^o(\phi^*(\boldsymbol{x})).$$

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

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

6. DISCUSSION

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 1 summarizes the aforementioned approaches, with a focus on their data setting, modeling approach, and motivation.

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 prespecification of the potential moderators. However, parametric regression models are limited by the need to prespecify interaction terms, and complex nonlinearities 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 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 prespecified. 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.

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—that is, 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., 2022); 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.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGMENTS

The authors would like to thank the anonymous referees and the special issue Guest Editors for their constructive comments that improved the quality of this paper.

T.-H. Chang completed the work for this paper while employed as a Biostatistician at the Johns Hopkins Bloomberg School of Public Health.

FUNDING

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

Abrevaya J, Hsu Y-C and Lieli RP (2015). Estimating conditional average treatment effects. J. Bus. Econom. Statist 33 485–505. MR3416596 10.1080/07350015.2014.975555

Athey S, Tibshirani J and Wager S (2019). Generalized random forests. Ann. Statist 47 1148–1178. MR3909963 10.1214/18-AOS1709

Audigier V, White IR, Jolani S, Debray TPA, Quartagno M, Carpenter J, van Buuren S and Resche-Rigon M (2018). Multiple imputation for multilevel data with continuous and binary variables. Statist. Sci 33 160–183. MR3797708 10.1214/18-STS646

Baron RM and Kenny DA (1986). The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. J. Pers. Soc. Psychol 51 1173–1182. 10.1037/0022-3514.51.6.1173 [PubMed: 3806354]

Berlin JA, Santanna J, Schmid CH, Szczech LA, Feldman HI 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. Stat. Med 21 371–387. 10.1002/sim.1023 [PubMed: 11813224]

Brantner CL, Nguyen TQ, Tang T, Zhao C, Hong H and Stuart EA (2023a). Comparing machine learning methods for estimating heterogeneous treatment effects by combining data from multiple randomized controlled trials. arXiv preprint. Available at arXiv:2303.16299.

Brantner CL, Chang T-H, Nguyen TQ, Hong H, Di Stefano L and Stuart EA (2023b). Supplement to “Methods for integrating trials and non-experimental data to examine treatment effect heterogeneity.” 10.1214/23-STS890SUPP

Brown CH, Sloboda Z, Faggiano F, Teasdale B, Keller F, Burkhart G, Vigna-Taglianti F, Howe G, Masyn K et al. (2013). Methods for synthesizing findings on moderation effects across multiple randomized trials. Prev. Sci 14 144–156. 10.1007/s11121-011-0207-8 [PubMed: 21360061]

Burke DL, Ensor J and Riley RD (2017). Meta-analysis using individual participant data: One-stage and two-stage approaches, and why they may differ. Stat. Med 36 855–875. MR3597661 10.1002/sim.7141 [PubMed: 27747915]

Cheng D and Cai T (2021). Adaptive combination of randomized and observational data. Available at arXiv:2111.15012.

Colnet B, Josse J, Varoquaux G and Scornet E (2022). Causal effect on a target population: A sensitivity analysis to handle missing covariates. J. Causal Inference 10 372–414. MR4512969 10.1515/jci-2021-0059

Colnet B, Mayer I, Chen G, Dieng A, Li R, Varoquaux G, Vert J, Josse J and Yang S (2021a). Causal inference methods for combining randomized trials and observational studies: A review. Available at arXiv:2011.08047.

Dagne GA, Brown CH, Howe G, Kellam SG and Liu L (2016). Testing moderation in network meta-analysis with individual participant data. Stat. Med 35 2485–2502. MR3513700 10.1002/sim.6883 [PubMed: 26841367]

Dahabreh IJ, Petito LC, Robertson SE, Hernán MA and Steingrimsson JA (2020). Towards causally interpretable meta-analysis: Transporting inferences from multiple studies to a target population. Available at arXiv:1903.11455.

Debray TPA, Moons KGM, Valkenhoef G, Efthimiou O, Hummel N, Groenwold RHH and Reitsma JB (2015). Get real in individual participant data (IPD) meta-analysis: A review of the methodology. Res. Synth. Methods 6 293–309. 10.1002/jrsm.1160 [PubMed: 26287812]

Debray TPA, Schuit E, Efthimiou O, Reitsma JB, Ioannidis JPA, Salanti G, Moons KGM and Workpackage G (2018). An overview of methods for network meta-analysis using individual participant data: When do benefits arise? Stat. Methods Med. Res 27 1351–1364. MR3777761 10.1177/0962280216660741 [PubMed: 27487843]

Donegan S, Williamson P, D'Alessandro U and Tudur Smith C (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. Stat. Med 31 3840–3857. MR3041777 10.1002/sim.5470 [PubMed: 22786621]

Efthimiou O, Debray TPA, Van Valkenhoef G, Trelle S, Panayidou K, Moons K, Reitsma JB, Shang A and Salanti G (2016). GetReal in network meta-analysis: A review of the methodology. Res. Synth. Methods 7 236–263. 10.1002/jrsm.1195 [PubMed: 26754852]

Enderlein G (1988). Fleiss, J. L.: The design and analysis of clinical experiments. Biom. J 30 304–304. 10.1002/bimj.4710300308

Gelman A, Hill J and Vehtari A (2020). Regression and Other Stories. Cambridge Univ. Press, Cambridge.

Godolphin PJ, White IR, Tierney JF and Fisher DJ (2023). Estimating interactions and subgroup-specific treatment effects in meta-analysis without aggregation bias: A within-trial framework. Res. Synth. Methods 14 68–78. 10.1002/jrsm.1590 [PubMed: 35833636]

Green AK, Trivedi N, Hsu JJ, Yu NL, Bach PB and Chimonas S (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. Health Aff. 41 368–374. 10.1377/hlthaff.2021.01432

Han L, Hou J, Cho K, Duan R and Cai T (2021). Federated Adaptive Causal Estimation (FACE) of target treatment effects. Available at arXiv:2112.09313.

hatt T, Berrevoets J, Curth A, Feuerriegel S and van der Schaar M (2022). Combining observational and randomized data for estimating heterogeneous treatment effects. arXiv preprint. Available at arXiv:2202.12891.

Hayward RA, Gagnier JJ, Borenstein M, Vanderheijden GJMG, Dahabreh IJ, Sun X, Sauerbrei W, Walsh M, Ioannidis JPA et al. (2020). Instrument for the Credibility of Effect Modification Analyses (ICEMAN) in randomized controlled trials and meta-analyses: Manual version 1.0.

Hong H, Fu H and Carlin BP (2018). Power and commensurate priors for synthesizing aggregate and individual patient level data in network meta-analysis. J. R. Stat. Soc. Ser. C. Appl. Stat 67 1047–1069. MR3832263 10.1111/rssc.12275

Hong H, Fu H, Price KL and Carlin BP (2015). Incorporation of individual-patient data in network meta-analysis for multiple continuous endpoints, with application to diabetes treatment. Stat. Med 34 2794–2819. MR3375982 10.1002/sim.6519 [PubMed: 25924975]

Hua H, Burke DL, Crowther MJ, Ensor J, Tudur Smith C and Riley RD (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. Stat. Med 36 772–789. MR3597655 10.1002/sim.7171 [PubMed: 27910122]

Jolani S, Debray TPA, Koffijberg H, van Buuren S and Moons KGM (2015). Imputation of systematically missing predictors in an individual participant data meta-analysis: A generalized approach using MICE. Stat. Med 34 1841–1863. MR3334696 10.1002/sim.6451 [PubMed: 25663182]

Kallus N, Puli AM and Shalit U (2018). Removing hidden confounding by experimental grounding. Available at arXiv:1810.11646.

Kennedy EH (2020). Optimal doubly robust estimation of heterogeneous causal effects. Available at arXiv:2004.14497.

Kent DM, Paulus JK, Van Klaveren D, D'Agostino R, Goodman S, Hayward R, Ioannidis JPA, Patrick-Lake B, Morton S et al. (2020). The predictive approaches to treatment effect heterogeneity (PATH) statement. Ann. Intern. Med 172 35–45. [PubMed: 31711134]

Kent DM, Rothwell PM, Ionnnidis JPA, Altman DG and Hayward RA (2010). Assessing and reporting heterogeneity in treatment effects in clinical trials: A proposal. Trials 11 85. 10.1186/1745-6215-11-85 [PubMed: 20704705]

Kovalchik SA (2013). Aggregate-data estimation of an individual patient data linear random effects meta-analysis with a patient covariate-treatment interaction term. Biostatistics 14 273–283. 10.1093/biostatistics/kxs035 [PubMed: 23001065]

Künzel SR, Sekhon JS, Bickel PJ and Yu B (2019). Metalearners for estimating heterogeneous treatment effects using machine learning. In Proceedings of the National Academy of Sciences 116 4156–4165.

Lambert PC, Sutton AJ, Abrams KR and Jones DR (2002). A comparison of summary patient-level covariates in meta-regression with individual patient data meta-analysis. J. Clin. Epidemiol 55 86–94. 10.1016/S0895-4356(01)00414-0 [PubMed: 11781126]

McCandless L (2009). Bayesian Methods for Data Analysis, 3rd ed. Carlin Bradley P. and Louis Thomas A., Chapman & Hall/CRC, Boca Raton, 2008. ISBN 9781584886976.

Nie X and Wager S (2021). Quasi-oracle estimation of heterogeneous treatment effects. Biometrika 108 299–319. MR4259133 10.1093/biomet/asaa076

Petrelli F and Barni S (2012). Surgery of primary tumors in stage IV breast cancer: An updated meta-analysis of published studies with meta-regression. Med. Oncol 29 3282–3290. 10.1007/s12032-012-0310-0 [PubMed: 22843291]

Riley RD, Lambert PC, Staessen JA, Wang J, Gueyffier F, Thijs L and Boutitie F (2008). Meta-analysis of continuous outcomes combining individual patient data and aggregate data. Stat. Med 27 1870–1893. MR2420350 10.1002/sim.3165 [PubMed: 18069721]

Riley RD, Stewart LA and Tierney JF (2021). Individual participant data meta-analysis for healthcare research. Individual Participant Data Meta-Analysis: A Handbook for Healthcare Research 1–6.

Rosenman E, Basse G, Owen A and Baiocchi M (2020). Combining observational and experimental datasets using shrinkage estimators. Available at arXiv:2002.06708.

Rosenman ETR, Owen AB, Baiocchi M and Banack HR (2022). Propensity score methods for merging observational and experimental datasets. Stat. Med 41 65–86. MR4376789 10.1002/sim.9223 [PubMed: 34671998]

Rubin DB (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. J. Educ. Psychol 66 688–701. 10.1037/h0037350

Samara MT, Nikolakopoulou A, Salanti G and Leucht S (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. Schizophr. Bull 45 639–646. 10.1093/schbul/sby095 [PubMed: 29982701]

Saramago P, Sutton AJ, Cooper NJ and Manca A (2012). Mixed treatment comparisons using aggregate and individual participant level data. Stat. Med 31 3516–3536. MR3041828 10.1002/sim.5442 [PubMed: 22764016]

Seo M, White IR, Furukawa TA, Imai H, Valgimigli M, Egger M, Zwahlen M and Efthimiou O (2021). Comparing methods for estimating patient-specific treatment effects in individual patient data meta-analysis. Stat. Med 40 1553–1573. MR4212329 10.1002/sim.8859 [PubMed: 33368415]

Silva S, Gutman BA, Romero E, Thompson PA, Altmann A and Lorenzi M (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) 270–274. IEEE, Los Alamitos, CA.

Simmonds MC and Higgins JPT (2007). Covariate heterogeneity in meta-analysis: Criteria for deciding between meta-regression and individual patient data. Stat. Med 26 2982–2999. MR2370988 10.1002/sim.2768 [PubMed: 17195960]

Tan X, Chang C-CH and Tang L (2021). A tree-based federated learning approach for personalized treatment effect estimation from heterogeneous data sources. Available at arXiv:2103.06261.

Teramukai S, Matsuyama Y, Mizuno S and Sakamoto J (2004). Individual patient-level and study-level meta-analysis for investigating modifiers of treatment effect. Jpn. J. Clin. Oncol 34 717–721. 10.1093/jjco/hyh138 [PubMed: 15640501]

Thomas D, Radji S and Benedetti A (2014). Systematic review of methods for individual patient data meta-analysis with binary outcomes. BMC Med. Res. Methodol 14. 10.1186/1471-2288-14-79

Tierney JF, Vale C, Riley R, Smith CT, Stewart L, Clarke M and Rovers M (2015). Individual participant data (IPD) meta-analyses of randomised controlled trials: Guidance on their use. PLoS Med. 12 e1001855. 10.1371/journal.pmed.1001855 [PubMed: 26196287]

Trivedi MH, Rush AJ, Wisniewski SR, Nierenberg AA, Warden D, Ritz L, Norquist G, Howland RH, Lebowitz B et al. (2006). Evaluation of outcomes with citalopram for depression using measurement-based care in STAR*D: Implications for clinical practice. Am. J. Psychiatr 163 28–40. 10.1176/appi.ajp.163.1.28 [PubMed: 16390886]

Vo TV, Hoang TN, Lee Y and Leong T-Y (2021). Federated estimation of causal effects from observational data. Available at arXiv:2106.00456.

Wu L and Yang S (2021). Integrative R-learner of heterogeneous treatment effects combining experimental and observational studies. In First Conference on Causal Learning and Reasoning.

Xie F, Chan JC and Ma RC (2018). Precision medicine in diabetes prevention, classification and management. J. Diabetes Investig 9 998–1015. 10.1111/jdi.12830

Yang Q, Liu Y, Cheng Y, Kang Y, Chen T and Yu H (2022). Federated Learning. Synthesis Lectures on Artificial Intelligence and Machine Learning 43. Springer, Cham. Reprint of the 2020 original MR4592510 10.1007/978-3-031-01585-4

Yang S, Zeng D and Wang X (2020). Elastic integrative analysis of randomized trial and real-world data for treatment heterogeneity estimation. Available at arXiv:2005.10579.

Yang S, Zeng D and Wang X (2022). Improved inference for heterogeneous treatment effects using real-world data subject to hidden confounding. Available at arXiv:2007.12922.

Table 1

Comparison of approaches to estimate 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	Nonparametric	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	Nonparametric	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	Nonparametric	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	Nonparametric	Pooled	Estimate confounding function
Hatt et al. (2022)	IPD	RCT and OD	Nonparametric	Pooled	Estimate confounding function

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