Statistical Science
2024, Vol. 39, No. 1, 165–191
https://doi.org/10.1214/23-STS889

Causal Inference Methods for Combining Randomized Trials and Observational Studies: A Review

Bénédicte Colnet, Imke Mayer, Guanhua Chen, Awa Dieng, Ruohong Li, Gaël Varoquaux, Jean-Philippe Vert, Julie Josse and Shu Yang

Abstract. With increasing data availability, causal effects can be evaluated across different data sets, both randomized controlled trials (RCTs) and observational studies. RCTs isolate the effect of the treatment from that of unwanted (confounding) co-occurring effects but they may suffer from unrepresentativeness, and thus lack external validity. On the other hand, large observational samples are often more representative of the target population but can conflate confounding effects with the treatment of interest. In this paper, we review the growing literature on methods for causal inference on combined RCTs and observational studies, striving for the best of both worlds. We first discuss identification and estimation methods that improve generalizability of RCTs using the representativeness of observational data. Classical estimators include weighting, difference between conditional outcome models and doubly robust estimators. We then discuss methods that combine RCTs and observational data to either ensure unconfoundedness of the observational analysis or to improve (conditional) average treatment effect estimation. We also connect and contrast works developed in both the potential outcomes literature and the structural causal model literature. Finally, we compare the main methods using a simulation study and real world data to analyze the effect of tranexamic acid on the mortality rate in major trauma patients. A review of available codes and new implementations is also provided.

Key words and phrases: Causal effect generalization, transportability, double robustness, data integration, heterogeneous data, S-admissibility.

1. INTRODUCTION

Experimental data, collected through carefully designed and randomized protocols, are usually considered the gold standard approach for assessing the causal effect of an intervention or a treatment on an outcome of interest. In particular, the intensive use of randomized controlled trials (RCTs) grounds the so-called "evidence-based medicine," a keystone of modern medicine. In an RCT, the treatment allocation is under control, ensuring a *balanced* distribution of treated and control individuals; as a consequence, simple estimators can be used to measure the treatment effect, for example, with the difference in mean effect between the treated and control individuals (Imbens and Rubin, 2015). Still, RCTs

Bénédicte Colnet is Ph.D. candidate, Soda project-team, INRIA Saclay, Palaiseau, France (e-mail: benedicte.colnet@inria.fr). Imke Mayer is Research Scientist, Owkin, London, UK (e-mail: Imke.mayer@owkin.com). Guanhua Chen is Associate Professor, Department of Biostatistics and Medical Informatics, University of Wisconsin-Madison, Madison, WI 53726, USA (e-mail: gchen25@wisc.edu). Awa Dieng is Research Associate, Google DeepMind, Montreal, Canada (e-mail: awadieng@google.com). Ruohong Li is Senior Machine Learning Scientist, Microsoft, Kirkland, WA 98033 , USA (e-mail: hannahli@microsoft.com). Gaël Varoquaux is Research Director, Soda project-team, INRIA Saclay, Paries, France (e-mail: gael.varoquaux@inria.fr). Jean-Philippe Vert is Chief R&D Officer, Owkin, Paris, France (e-mail: jean-philippe.vert@owkin.com). Julie Josse is Head, Premedical project Inria team, University of Montpellier, France (e-mail: julie.josse@inria.fr). Shu Yang is Associate Professor, Department of Statistics, North Carolina State University, 2311 Stinson Drive, Raleigh, NC 27695, USA (e-mail: syang24@ncsu.edu).

come with practical drawbacks such as cost and time, but also with methodological issues such as restrictive inclusion/exclusion criteria, which can lead to a trial sample that differs markedly from the population potentially eligible for the treatment. Therefore, the findings from RCTs can lack generalizability to a target population of interest. This concern is related to the aim of *external validity*, central in medical research (Concato, Shah and Horwitz, 2000, Rothwell, 2005, Green and Glasgow, 2006, Frieden, 2017) policy research (Martel Garcia and Wantchekon, 2010, Deaton and Cartwright, 2018, Deaton et al., 2019, Jeong and Namkoong, 2022), psychology (Kennedy and Gelman, 2021) and other fields such as advertising (Gordon et al., 2019).

In contrast, *observational data*—collected without systematically designed interventions, such as disease registries, cohorts, biobanks, epidemiological studies or electronic health records—are promising as they are readily available, include large and representative samples and are less cost-intensive than RCTs. However, there are often concerns about the quality of these "big data," given that the lack of a controlled experimental intervention opens the door to *confounding bias*. This concern is referred to as a lack of *internal validity*. Under assumptions such as unconfoundedness, it is possible to estimate a causal treatment effect from observational data. In practice, methods such as matching, inverse propensity weighting (IPW) or augmented IPW (AIPW) are used (Imbens and Rubin, 2015). Even when a confounder is unobserved, solutions exist at the price of additional assumptions, for example, the *front-door criterion* (Pearl, 1993), instrumental variables (Angrist, Imbens and Rubin, 1996, Hernán and Robins, 2006, Imbens, 2014) and sensitivity analysis (Cornfield et al., 1959, Rosenbaum and Rubin, 1983, Imbens, 2003).

Combining information gathered from experimental and observational data opens the door to new tools for:

- (a) accounting for the lack of representativeness of RCT, as observational data can constitute an external representative sample of a target population of interest;
- (b) making observational evidence more credible using RCT to ground observational analysis, such as detecting a confounding bias;
- (c) improving statistical efficiency, for example, to better estimate heterogeneous treatment effects as RCTs are often underpowered in such settings.

As of today, there is an abundant literature about the different ways and purposes of combining both sources of information. Terms used to refer to similar problems are *generalizability* (Cole and Stuart, 2010, Stuart et al., 2011, Hernán and Van der Weele, 2011, Tipton, 2013, O'Muircheartaigh and Hedges, 2014, Stuart, Bradshaw and Leaf, 2015, Keiding and Louis, 2016, Dahabreh and Hernán, 2019, Dahabreh et al., 2019, Buchanan et al., 2018, Cinelli and Pearl, 2021), *representativeness* (Campbell, 1957), *external validity* (Rothwell, 2005, Stuart, Ackerman and Westreich, 2018, Westreich et al., 2018), *transportability* (Pearl and Bareinboim, 2011, Rudolph and van der Laan, 2017, Westreich et al., 2017), *recoverability* (Bareinboim and Pearl, 2012a, Bareinboim, Tian and Pearl, 2014) and finally *data fusion* (Bareinboim and Pearl, 2016); this review will explain the commonalities or differences between the terminologies. They have connections to inference from nonprobability samples in survey sampling (Yang, Kim and Song, 2020, Yang and Kim, 2020) and to the covariate shift problem in machine learning (Sugiyama and Kawanabe, 2012). This problem of data integration for causal inference is tackled by two main bodies of literature, namely the potential outcomes (PO) framework (Neyman, 1923, Rubin, 1974), and the work on structural causal models (SCM) using directed acyclic graphs (DAGs), pioneered by Pearl (1995) and his collaborators.

The present paper reviews this literature on combining experimental and observational data. Section 2 introduces the notation from the PO literature, as well as the common designs. Section 3 details how an observational sample can be used to generalize RCT findings to another population (point (a)). We detail the corresponding identifiability assumptions and present the main estimation methods that have been suggested to account for distributional shifts. In this section, only baseline covariates are required in the observational data. In Section 4, we consider the case where observational data also contain treatment and outcome data. This setting in particular provides the opportunity to tackle different scientific questions such as hidden confounding or statistical efficiency (points (b) and (c)). In Section 5, we present the SCM literature, using different notation and ways to formulate assumptions, thus capturing richer and more diverse identifiability scenarios. In Section 6, we first present existing implementations and software and then we illustrate the properties of the generalization estimators on simulated data with new implementations. In Section 7, we apply the various methods presented in Section 3 on a medical application involving major trauma patients. The aim of this study is to assess the effect of the drug tranexamic acid on mortality in head trauma patients. Both an RCT (the CRASH-3 trial) and an observational database (the Traumabase registry) are available. In this section, we also review methods for addressing data quality issues such as missing values.

2. PROBLEM SETTING

2.1 Notation in the PO Framework

Each individual in the RCT or observational population is described by a random tuple $(X, Y(0), Y(1), A, S)$,

with distribution $\mathbb{P}$, where X is a p-dimensional vector of covariates, A the binary treatment assignment (with $A = 0$ for the control and $A = 1$ for the treated individuals), $Y(a)$ is the binary or continuous outcome had the subject been given treatment a (for $a \in \{0, 1\}$) and S a binary variable indicating trial eligibility and willingness to participate.[1] We model the individuals belonging to an RCT sample of size n and to an observational data sample of size m by $n + m$ independent random tuples: $\{X_i, Y_i(0), Y_i(1), A_i, S_i\}_{i=1}^{n+m}$, where the RCT samples $i = 1, \ldots, n$ are identically distributed according to $\mathbb{P}(X, Y(0), Y(1), A, S \mid S = 1)$, and the observational data samples $i = n + 1, \ldots, n + m$ are identically distributed according to $\mathbb{P}(X, Y(0), Y(1), A, S)$. The sampling mechanisms of the RCT and observational samples are assumed to be independent, which corresponds to a so-called nonnested design as explained in Section 2.2.1. We also denote $\mathcal{R} = \{1, \ldots, n\}$ the set of indices of units observed in the RCT study, and $\mathcal{O} = \{n + 1, \ldots, n + m\}$ the set of indices of units observed in the observational study. For each RCT sample $i \in \mathcal{R}$, we observe $(X_i, A_i, Y_i, S_i = 1)$, while for observational data $i \in \mathcal{O}$, we consider two settings: *(i)* we only observe the covariates X_i (Section 3) and *(ii)* we also observe the treatment and outcome (X_i, A_i, Y_i) (Section 4).

In this review, we consider the absolute difference, and do not consider other contrast measures.[2] Doing so, we denote respectively by $\tau(x)$ and $\tau_1(x)$ the conditional average treatment effect (CATE) in the observational population:

$$\forall x \in \mathbb{R}^p, \quad \tau(x) = \mathbb{E}\big[Y(1) - Y(0) \mid X = x\big],$$

and the RCT population:

$$\forall x \in \mathbb{R}^p, \quad \tau_1(x) = \mathbb{E}\big[Y(1) - Y(0) \mid X = x, S = 1\big].$$

We also denote τ and τ_1 the population average treatment effect (ATE) in the observational population:

$$\tau = \mathbb{E}\big[Y(1) - Y(0)\big] = \mathbb{E}\big[\tau(X)\big],$$

and the RCT population:

$$\tau_1 = \mathbb{E}\big[Y(1) - Y(0) \mid S = 1\big],$$

where the population ATE can be different from the RCT ATE, that is, $\tau \neq \tau_1$ in general. We denote respectively by $e(x)$ and $e_1(x)$ the propensity score in the observational population:

$$e(x) = \mathbb{P}(A = 1 \mid X = x),$$

and in the RCT population:

$$e_1(x) = \mathbb{P}(A = 1 \mid X = x, S = 1),$$

where $e_1(x)$ is usually known in an RCT. We also denote by $\mu_a(x)$ and $\mu_{a,1}(x)$ the conditional mean outcome under treatment $a \in \{0, 1\}$ in the observational population:

$$\mu_a(x) = \mathbb{E}\big[Y(a) \mid X = x\big],$$

and in the RCT population:

$$\mu_{a,1}(x) = \mathbb{E}\big[Y(a) \mid X = x, S = 1\big].$$

Finally, we denote by $\alpha(x)$ the conditional odds that an individual with covariates x is in the RCT or in the observational sample:

$$\begin{aligned}\alpha(x) &= \frac{\mathbb{P}(i \in \mathcal{R} \mid \exists i \in \mathcal{R} \cup \mathcal{O}, X_i = x)}{\mathbb{P}(i \in \mathcal{O} \mid \exists i \in \mathcal{R} \cup \mathcal{O}, X_i = x)} \\ &= \frac{\pi_{\mathcal{R}}(x)}{\pi_{\mathcal{O}}(x)} \\ &= \frac{\pi_{\mathcal{R}}(x)}{1 - \pi_{\mathcal{R}}(x)},\end{aligned}$$

where $\pi_{\mathcal{R}}(x)$ (resp., $\pi_{\mathcal{O}}(x)$) is the probability that an individual with covariates x known to be in the concatenated data (RCT sample and observational sample) is in the RCT (resp., in the observational sample). In the literature, another widely used quantity is the selection score—or sampling propensity score (in particular this name was proposed by Tipton (2013))—denoted $\pi_S(x)$ and defined as

$$\pi_S(x) = \mathbb{P}(S = 1 \mid X = x).$$

Because $\pi_S(x)$ is the probability of being *sampled* in the trial given covariate values x, it is different from $\pi_{\mathcal{R}}(x)$. $\pi_S(x)$ is often used with a nested design (see Section 2.2.1 for a definition), but is not of interest in our setup (nonnested design) because it cannot be identified. Indeed,

$$\begin{aligned}\pi_S(x) &= \underbrace{\mathbb{P}(S = 1)}_{\text{Not known}} \underbrace{\frac{\mathbb{P}(X = x \mid S = 1)}{\mathbb{P}(X = x)}}_{\propto \pi_{\mathcal{R}}(x)/\pi_{\mathcal{O}}(x)} \\ &= \mathbb{P}(S = 1) \times \frac{\mathbb{P}(X = x \mid S = 1)}{\mathbb{P}(X = x)} \\ &= \mathbb{P}(S = 1) \times \frac{\mathbb{P}(X_i = x \mid i \in \mathcal{R})}{\mathbb{P}(X_i = x \mid i \in \mathcal{O})} \\ &= \underbrace{\mathbb{P}(S = 1)}_{\text{Not known}} \times \frac{n}{m} \underbrace{\frac{\pi_{\mathcal{R}}(x)}{\pi_{\mathcal{O}}(x)}}_{= \alpha(x)}.\end{aligned}$$

[1] Note that in the literature, S can have a slightly different meaning, for example, other works use two separate indicators, one for participation and one for eligibility (Nguyen et al., 2018, Dahabreh et al., 2019).

[2] Considering other measures such as the ratio or odds ratio can have an impact on the assumptions considered, for example, in generalization (Huitfeldt et al., 2019). As the large majority of the literature focuses on the absolute difference, this review reflects the practices and, therefore, considers the absolute difference.

TABLE 1

Illustration of data structure of RCT data (Set $\mathcal{R}$) and observational data (Set $\mathcal{O}$) with covariates X, trial eligibility S, binary treatment A and outcome Y. Left: with observed outcomes, right: with potential outcomes. Note that the S covariate can be either 0 or 1 in the observational data set (it is unknown in the nonnested design, hence the NA for not available), and is always equal to 1 for observations in the RCT. In the nested design (cf. Section E of the Supplementary Material of Colnet et al., 2024), $S = 0$ for all individuals in the observational data set

			Covariates			Treatment	Outcome
	S	Set	X_1	X_2	X_3	A	Y
1	1	$\mathcal{R}$	1.1	20	F	1	1
	1	$\mathcal{R}$	−6	45	F	0	1
n	1	$\mathcal{R}$	0	15	M	1	0
$n+1$	NA	$\mathcal{O}$		...		...	...
	NA	$\mathcal{O}$	−2	52	M	0	1
	NA	$\mathcal{O}$	−1	35	M	1	1
$n+m$	NA	$\mathcal{O}$	−2	22	M	0	0

		Covariates			Treatment	Outcome(s)	
S	Set	X_1	X_2	X_3	A	$Y(0)$	$Y(1)$
1	$\mathcal{R}$	1.1	20	F	1	NA	1
1	$\mathcal{R}$	−6	45	F	0	1	NA
1	$\mathcal{R}$	0	15	M	1	NA	1
NA	$\mathcal{O}$		...		...	...	...
NA	$\mathcal{O}$	−2	52	M	0	1	NA
NA	$\mathcal{O}$	−1	35	M	1	NA	1
NA	$\mathcal{O}$	−2	22	M	0	0	NA

Detailed derivations can be found in Section C of the Supplementary Material of Colnet et al. (2024). The quantity $\mathbb{P}(S = 1)$ is unknown because individuals in the target population could have participated in the RCT or not; S can be equal to 1 and 0 in the observational sample but this information is not known. Table 1 illustrates the considered type of data, and Table 2 summarizes the notation.

2.2 Study Designs and Goals

2.2.1 *Nested and nonnested study designs.* Following Dahabreh et al. (2023a) and Dahabreh and Hernán (2019), the study designs to obtain the trial and observational samples can be categorized into two types: *nested* study designs and *nonnested* study designs as illustrated on Figure 1. Designs imply different identifiability conditions and, therefore, estimators. This review focuses on what is called the nonnested design, as the trial sample and the observational sample are obtained separately from the target population(s). On the contrary, the nested design involves a two-stage nested sampling. For example, it can correspond to an embedded trial in a broader health system. As a concrete example, one can mention the Women Health Initiative, or the recent study on Medicaid where parts of

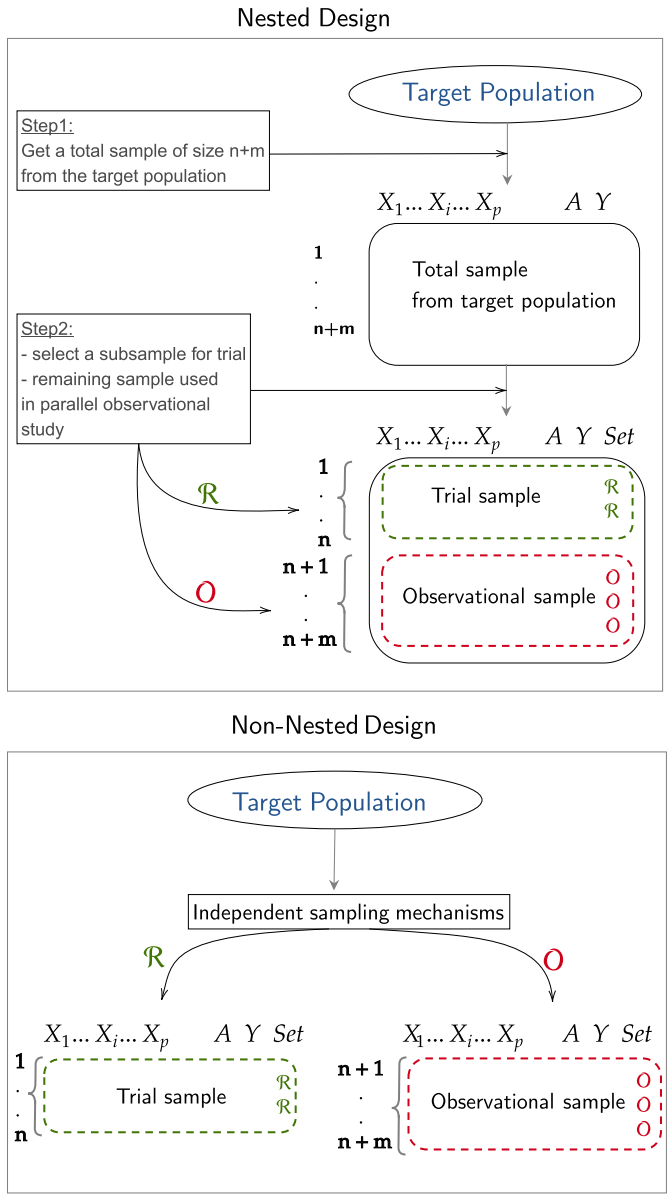

FIG. 1. *Schematics of the nested (left) and nonnested (right) designs, a similar schematic can be found in Josey et al. (2021).*

the participants are randomized (Degtiar et al., 2021). In this situation, data are not really combined as the overall data comes from one initial sampling in which two treatment assignment regimes (randomized or not) coexist. The nested design estimators are detailed in Section E of the Supplementary Material of Colnet et al. (2024).

2.2.2 *Transportability, generalizability, and recoverability.* Several terms are currently present in the literature to describe the process of predicting the effect of the treatment from an RCT to another population: generalization (Stuart et al., 2011, Buchanan et al., 2018, Dahabreh et al., 2019), transportability (Hernán and Van der Weele, 2011, Bareinboim and Pearl, 2016, Westreich et al., 2017), or recoverability (Bareinboim, Tian and Pearl, 2014). Differences in the definitions can be found

TABLE 2
List of notation

Symbol	Description
X	Covariates (also known as baseline covariates when measured at inclusion of the patient)
A	Treatment indicator ($A = 1$ for treatment, $A = 0$ for control)
Y	Outcome of interest
S	Trial eligibility and willingness to participate if invited to ($S = 1$ for eligibility, $S = 0$ for noneligibility)
n	Size of the RCT study
m	Size of the observational study
$\mathcal{R}$	Index set of units observed in the RCT study; $\mathcal{R}$={1,...,n}
$\mathcal{O}$	Index set of units observed in the observational study; $\mathcal{O}$={n+1,...,n+m}
$\pi_{\mathcal{R}}(x)$	Probability that a unit in $\mathcal{R} \cup \mathcal{O}$ with covariate x is in $\mathcal{R}$, defined as $\pi_{\mathcal{R}}(x) = \mathbb{P}(i \in \mathcal{R} \mid \exists i \in \mathcal{R} \cup \mathcal{O}, X_i = x)$
$\pi_{\mathcal{O}}(x)$	Probability that a unit in $\mathcal{R} \cup \mathcal{O}$ with covariate x is in $\mathcal{O}$, defined as $\pi_{\mathcal{O}}(x) = 1 - \pi_{\mathcal{R}}(x)$
$\alpha(x)$	Conditional odds $\alpha(x) = \pi_{\mathcal{R}}(x)/\pi_{\mathcal{O}}(x)$
τ	Population average treatment effect (ATE) defined as $\tau = \mathbb{E}[Y(1) - Y(0)]$
τ_1	Trial (or sample) average treatment effect defined as $\tau_1 = \mathbb{E}[Y(1) - Y(0) \mid S = 1]$
$\tau(x)$	Conditional average treatment effect (CATE) defined as $\tau(x) = \mathbb{E}[Y(1) - Y(0) \mid X = x]$
$\tau_1(x)$	Trial conditional average treatment effect defined as $\tau_1(x) = \mathbb{E}[Y(1) - Y(0) \mid X = x, S = 1]$
$e(x)$	Propensity score defined as $e(x) = \mathbb{P}(A = 1 \mid X = x)$
$e_1(x)$	Propensity score in the trial defined as $e_1(x) = P(A = 1 \mid X = x, S = 1)$, known by design
$\mu_a(x)$	Outcome mean defined as $\mu_a(x) = \mathbb{E}[Y(a) \mid X = x]$ for $a = 0, 1$
$\mu_{a,1}(x)$	Outcome mean in the trial defined as $\mu_{a,1}(x) = \mathbb{E}[Y(a) \mid X = x, S = 1]$ for $a = 0, 1$
$\pi_S(x)$	Selection score defined as $\pi_S(x) = \mathbb{P}(S = 1 \mid X = x)$
$f(X)$	Covariate distribution in the target population
$f(X\|S = 1)$	Covariate distribution conditional to trial-eligible individuals ($S = 1$)

in the literature, underlying a specific design such as the existence of a common superpopulation or assumptions such as the support overlap between different populations. For example, Dahabreh et al. (2020) highlight that several definitions are given:

> *We use the term generalizability when the target population coincides or is a subset of the trial-eligible population and transportability when the target population includes at least some individuals who are not trial-eligible (and who, by definition, cannot be trial participants) (others have proposed different definitions).*

Due to different definitions in the literature, several terms can be found to describe the same scientific goal. In this review, we call *generalization* the task that extends the RCT result to *its* larger population, where it was sampled with a bias (detailed in Section 3). The SCM literature also uses different terminologies corresponding to different assumptions—and corresponding diagrams—as detailed in Section 5. For example, what is called transportability refers to two distinct populations, and not necessarily to different covariate supports as suggested by Dahabreh et al. (2020). In particular, in this literature the task that we study in Section 3 is termed recoverability from a sampling bias, rather than generalization. This terminology has the merit of indicating that generalization can have a much broader coverage, including other types of problems. Note that granting some assumptions about a common support or nonzero probability to be sampled, then the two problems—namely recovering from a sampling bias and transportability—rely on the same estimators and procedure, as highlighted in Section 3.1.3 and in Pearl (2015).

3. WHEN OBSERVATIONAL DATA HAVE NO TREATMENT AND OUTCOME INFORMATION

We start by considering the case where only the covariates from the observational study are available or used. We consider the observational data as a random sample from the target population. Considering this setup, the question tackled in this section is how to generalize or transport the trial findings toward a target population of interest. Applied examples can be found in Lee et al. (2023), Lesko et al. (2016), Tipton et al. (2017), Li, Buchanan and Cole (2021), Yang and Wang (2022). In particular, He et al. (2020) review current practice, revealing that generalization implementation is still at the stage of prototyping without real usage for clinical and public health decisions yet.

3.1 Assumptions Needed to Identify the ATE on the Target Population

A fundamental problem in causal inference is that we can observe at most one of the potential outcomes for an individual subject. In order to nonetheless identify the

ATE from RCT and observational covariate data, we require some of the following assumptions.

3.1.1 *Internal validity of the RCT.*

ASSUMPTION 1 (Consistency).

$$Y = AY(1) + (1 - A)Y(0).$$

Assumption 1 implies that the observed outcome is the potential outcome under the actual assigned treatment.

ASSUMPTION 2 (Randomization).

$$\{Y(0), Y(1)\} \perp\!\!\!\perp A \mid S = 1, X.$$

Assumption 2 corresponds to internal validity. It holds by design in a completely randomized experiment, where the treatment is independent of all the potential outcomes and covariates. The more general case of conditional randomization is assumed throughout this review.

If Assumptions 1 and 2 hold, then the RCT is said to be compliant. In addition, in an RCT, it is common that the probability of treatment assignment, $e_1(x)$, is known. In a completely randomized trial, the propensity score is fixed as a constant, and usually $e_1(x) = 0.5$ for all x.

3.1.2 *Assumptions ensuring generalizability of the RCT to the target population*. The literature proposes different assumptions to generalize trial findings to a target population.

ASSUMPTION 3 (Ignorability assumption on trial participation—Hotz, Imbens and Mortimer, 2005, Stuart et al., 2011, Tipton, 2013, Hartman et al., 2015, Buchanan et al., 2018, Degtiar and Rose, 2023, Egami and Hartman, 2021).

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

A parallel can be made with the *strong ignorability condition* in causal inference with observational data (see Section B of the Supplementary Material of Colnet et al., 2024), but applied to the sample selection rather than treatment assignment. In other words, these assumptions require to control for all covariates being shifted and predictive of Y. We call shifted covariates, all the baseline covariates along which the two populations—trial and target—do not follow the same distribution. A weaker version of Assumption 3 can be found in Dahabreh et al. (2019), Dahabreh et al. (2020).

ASSUMPTION 4 (Mean exchangeability). For all x and for all $a \in \{0, 1\}$,

$$\mathbb{E}\big[Y(a) \mid X = x, S = 1\big] = \mathbb{E}\big[Y(a) \mid X = x\big]$$

Another assumption can be found, relying on the transportability of treatment effect rather than the potential outcomes.

ASSUMPTION 5 (Sample ignorability for treatment effects—Kern et al., 2016, Nguyen et al., 2018).

$$Y(1) - Y(0) \perp\!\!\!\perp S \mid X.$$

A weaker version can be found as well.

ASSUMPTION 6 (Transportability of the CATE).

$$\tau_1(x) = \tau(x) \quad \text{for all} x.$$

To meet these last two assumptions, one requires variables that are both *treatment effects modifiers* and *shifted*. Epidemiologists often use the term "effect modification" to indicate that the treatment effect varies across strata of baseline covariates, such baseline covariates being treatment effect modifiers. These assumptions are implied by Assumption 3, but this is not reciprocal as not all covariates predictive of the outcome are necessarily treatment effect modifiers. Note that a treatment effect modifier depends on the chosen scale. Here, we focus on the absolute difference, but if we had considered a risk ratio, the variables being treatment effects modifiers would not be the same. Mathematical definitions of a treatment effect modifier are hard to find, but we quote one from Van der Weele and Robins (2007) for the absolute scale.

DEFINITION 1 (Treatment effect modifier). We say that a variable X is a treatment effect modifier for the causal risk difference of A on Y if X is not affected by A and if there exist two levels of A, a_0 and a_1, such that $\mathbb{E}[Y^{(a_1)} \mid X = x] - \mathbb{E}[Y^{(a_0)} \mid X = x]$ is not constant in x.

In this work, we only rely on Assumption 5 for the identification formula. Finally, a last assumption is needed—the *positivity of trial participation* assumption.

ASSUMPTION 7 (Positivity of trial participation, also called overlap). There exists a constant $c > 0$ such that, almost surely, $\mathbb{P}(S = 1 \mid X) \geq c$.

Assumption 7 requires adequate overlap of the covariate distribution between the trial sample and the target population (in other words, all members of the target population have nonzero probability of being selected into the trial). Another formulation of this assumption can be found under the assumption of the target population's support included in the trial sample support (Nie, Imbens and Wager, 2021, Colnet et al., 2022b)

3.1.3 *Identifications formulas*. Under Assumptions 1, 2, 6 and 7, the ATE can be identified based on the following formulas (derivations in Section C of the Supplementary Material of Colnet et al., 2024):

1. Reweighting formulation:

$$\tau = \mathbb{E}\left[\frac{n}{m\alpha(X)}\tau_1(X) \mid S = 1\right], \tag{1}$$

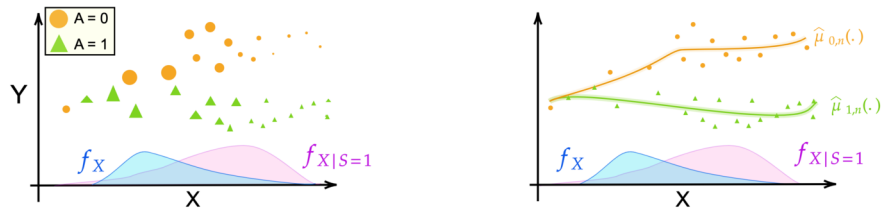

FIG. 2. *Illustrative schematics for the estimation strategies*: *In this example, the trial findings* $\hat{\tau}_{1,n}$ *would overestimate the target treatment effect* τ (*on an absolute scale*). *On the* left, *the IPSW* (*Definition* 2) *strategy, relying on weighting the RCT observations*; *on the* right, *the plug-in g-formula* (*Definition* 4) *strategy, relying on modeling the response using the RCT observations. Notation are the same as introduced in Table* 2, *that is,* f_X ($f_{X|S=1}$) *denotes the density of the target* (*resp.*, *trial*) *population, and* $\hat{\mu}_{a,n}(\cdot)$ *denotes the fitted response surface using the n trial observations.*

which can also be written as

$$\tau = \mathbb{E}\left[\frac{n}{m\alpha(X)}\left(\frac{A}{e_1(X)} - \frac{1-A}{1-e_1(X)}\right)Y \mid S=1\right].$$

Note that (1) can be understood as a transportability problem considering two distributions $\mathbb{P}_1$ and $\mathbb{P}$, and transporting evidence from population $\mathbb{P}_1$ to population $\mathbb{P}$,

$$\tau = \mathbb{E}_{\mathbb{P}}[\tau(X)] = \underbrace{\int_{\mathcal{X}} \tau(x) f(x)\, dx}_{\text{Integral on } \mathbb{P}}$$

$$= \underbrace{\int_{\mathcal{X}} \tau_1(x) \frac{f(x)}{f_1(x)} f_1(x)\, dx}_{\text{Integral on } \mathbb{P}_1}$$

$$= \int_{\mathcal{X}} \tau_1(x) \frac{n}{m} \frac{1}{\alpha(x)} f_1(x)\, dx,$$

noting that $\alpha(x) = \frac{\mathbb{P}(i\in\mathcal{R}|\exists i\in\mathcal{R}\cup\mathcal{O}, X_i=x)}{\mathbb{P}(i\in\mathcal{O}|\exists i\in\mathcal{R}\cup\mathcal{O}, X_i=x)} = \frac{\mathbb{P}(i\in\mathcal{R})}{\mathbb{P}(i\in\mathcal{O})} \times \frac{\mathbb{P}(X_i=x|i\in\mathcal{R})}{\mathbb{P}(X_i=x|i\in\mathcal{O})} = \frac{n}{m} \times \frac{f_1(x)}{f(x)}$, and using the transportability assumption (see Assumption 6) stating that $\tau(x) = \tau_1(x)$.

2. Regression formulation:

$$\tau = \mathbb{E}[\mu_{1,1}(X) - \mu_{0,1}(X)] = \mathbb{E}[\tau_1(X)]. \tag{2}$$

Different identification formulas motivate different estimation strategies as discussed next. These strategies are illustrated in Figure 2.

3.2 Estimation Methods to Generalize Trial Findings to a Target Population of Interest

All along this review, estimators are indexed with the number of observations used for estimation. For example, $\hat{\tau}_n$ indicates that the finite sample estimator only relies on the RCT individuals, or $\hat{\tau}_{n,m}$ if it depends on both data sets.

3.2.1 *IPSW and stratification*: *Modeling the probability of trial participation*. To overcome the bias due to covariate shift between populations, most existing methods rely on direct modeling of the selection score previously introduced. The selection score adjustment methods include IPSW (Cole and Stuart, 2010, Stuart et al., 2011, Lesko et al., 2017, Buchanan et al., 2018, Colnet et al., 2022b) and stratification (Stuart et al., 2011, Tipton, 2013, O'Muircheartaigh and Hedges, 2014).

3.2.1.1 *Inverse probability of sampling weighting* (*IPSW*). The IPSW approach can be seen as the counterpart of IPW methods for estimating the ATE from observational studies by controlling for confounding (see Section B of the Supplementary Material of Colnet et al., 2024 for details on IPW). Based on the identification formula (1), the IPSW estimator of the ATE is defined as the weighted difference of average outcomes between the treated and control group in the trial. The observations are weighted by the inverse odds $1/\alpha(x) = \pi_{\mathcal{O}}(x)/\pi_{\mathcal{R}}(x)$ to account for the shift of the covariate distribution from the RCT sample to the target population. The larger $\alpha(X_i)$, the smaller the weight of the observation i (as illustrated in Figure 2). The shape of the IPSW estimator is slightly different from the shape of the IPW estimator. In the latter, each observation is weighted by the inverse of the probability to be treated, whereas in the former, it is weighted by the inverse of the odds of the probability to be in the trial sample. This is due to the nonnested sampling design (see the IPSW estimator for the nested design (S5)), as highlighted by Kern et al. (2016) and Nguyen et al. (2018).

DEFINITION 2 (Inverse probability of sampling weighting—IPSW). The IPSW estimator is defined as follows:

$$\hat{\tau}_{\text{IPSW},n,m} = \frac{1}{n}\sum_{i=1}^{n} \frac{n}{m} \frac{Y_i}{\hat{\alpha}_{n,m}(X_i)}\left(\frac{A_i}{e_1(X_i)} - \frac{1-A_i}{1-e_1(X_i)}\right),$$

where $\hat{\alpha}_{n,m}$ is an estimate of the odds of the indicatrix of being in the RCT.

The IPSW estimator is consistent when the quantity α is consistently estimated by $\hat{\alpha}_{n,m}$ (Buchanan et al., 2018, Colnet et al., 2022a). In practice, various methods are used to estimate α, for example, by logistic regression (Stuart, 2010), while recent works rely on non-parametric methods such as random forest and gradient boosting (Kern et al., 2016) or the Hájek-style estimator to target the density ratio (Huang et al., 2023, Nie, Imbens and

Wager, 2021). Similar to IPW estimators, IPSW estimators are known to be highly unstable, especially when the weights are extreme. This can occur if the observational study contains units with very small probabilities of being in the trial. Normalized weights can be used to overcome this issue (Dahabreh and Hernán, 2019). Still, the major challenge remains that IPSW estimators require a correct model specification of the weights. Avoiding this problem requires either very strong domain expertise or turning to doubly robust methods (Section 3.2.4). Current theoretical guarantees and theorems are detailed in Section D of the Supplementary Material of Colnet et al. (2024). For example, Buchanan et al. (2018) propose a derivation of the asymptotic variance under parametric assumptions in the nested case, while Zivich et al. (2022) extend this to a nonnested design. Dahabreh et al. (2019) propose the use of sandwich-type variance estimators (for both nested and nonnested design) or nonparametric bootstrap approaches, and note that the latter may be preferred in practice. Colnet et al. (2022a) has formalized consistency results for any consistent estimator of α, including non-parametric estimators.

ASSUMPTION 8 (Consistency assumptions for α). Denoting by $\frac{n}{m\hat{\alpha}_{n,m}(x)}$ the estimated weights on the set X, the following conditions hold:

- $\sup_{x\in\mathcal{X}} |\frac{n}{m\hat{\alpha}_{n,m}(x)} - \frac{f_X(x)}{f_{X|S=1}(x)}| = \epsilon_{n,m} \xrightarrow{a.s.} 0$, when $n, m \to \infty$;
- for all n, m large enough $\mathbb{E}[\varepsilon_{n,m}^2]$ exists and $\mathbb{E}[\varepsilon_{n,m}^2] \xrightarrow{a.s.} 0$, when $n, m \to \infty$;
- Y is square integrable.

THEOREM 3.1 (IPSW consistency—Colnet et al., 2022a). *Under causal assumptions* (*Assumptions* 1, 2, 6, 7), (*identifiability*), *and Assumption* 8 (*consistency*), *then* $\hat{\tau}_{IPSW,n,m}$ *converges toward* τ *in* L^1 *norm*,

$$\hat{\tau}_{IPSW,n,m} \xrightarrow[n,m\to\infty]{L^1} \tau.$$

More recently, Colnet et al. (2022b) have proposed a finite sample characterization of IPSW when X only contains categorical covariates.

3.2.1.2 *Stratification*. The stratification approach—or subclassification—is introduced by Cochran (1968) for a single observational data set, and has been further extended by Stuart et al. (2011), Tipton (2013) and O'Muircheartaigh and Hedges (2014) for the generalization's context. It is proposed as a solution to mitigate the risks of extreme weights in the IPSW formula. First, one has to estimate the conditional odds $\hat{\alpha}_{n,m}$ in the same manner as for the IPSW detailed above. Then, based on the values of the conditional odds obtained, L strata are defined (usually 5 as reported in (O'Muircheartaigh and Hedges, 2014), following the empirical seminal work of (Cochran, 1968)). In the trial, for each stratum l, one has to compute the average effect on this strata defined as $\overline{Y(1)}_l - \overline{Y(0)}_l$, where $\overline{Y(a)}_l$ denotes the average value of the outcome for units with treatment a in stratum l in the RCT. The generalized ATE is defined by the aggregation of the treatment effect estimates on each stratum l weighted by the proportion of the strata in the target population $\frac{m_l}{m}$, where m_l is the number of individuals in strata l in the target sample.

DEFINITION 3 (Stratification). The stratification estimator denoted $\hat{\tau}_{\text{strat},n,m}$ is defined as

$$\hat{\tau}_{\text{strat},n,m} = \sum_{l=1}^{L} \frac{m_l}{m} \underbrace{(\overline{Y(1)}_l - \overline{Y(0)}_l)}_{\text{from RCT}}.$$

Buchanan et al. (2018) proposed an asymptotic normality result for this estimator. Theoretical results for the stratification estimator are detailed in Section D of the Supplementary Material of Colnet et al. (2024).

3.2.2 *Plug-in g-formula estimators*: *Modeling the conditional outcome in the trial*. Other estimators to generalize RCT findings to a target population leverage the regression formulation (2), in the inspiration of (Robins, 1986). Known as plug-in g-formula estimators, they fit a model of the conditional outcome mean among trial participants, rather than modeling the probability of trial participation (as illustrated on Figure 2). Then a marginalization is done over the empirical covariate distribution of the target population.

DEFINITION 4 (Plug-in g-formula). The plug-in g-formula (or outcome model-based) estimator is then defined as

$$\hat{\tau}_{\text{G},n,m} = \frac{1}{m} \sum_{i=n+1}^{n+m} \big(\hat{\mu}_{1,1,n}(X_i) - \hat{\mu}_{0,1,n}(X_i)\big),$$

where $\hat{\mu}_{a,1,n}(X_i)$ is an estimator of $\mu_{a,1}(X_i)$ fitted using the RCT data.

In practice, any model can be use to fit $\mu_{a,1}(X_i)$, for example, standard ordinary least squares (OLS). Dahabreh et al. (2020) announce[3] consistency of the plug-in g-formula for parametric estimator of the response model $\mu_a(X)$. Note that derivations are made in the context of a nested design but said to extend to a nonnested design. They also recommend the use of sandwich-type variance estimators for computing confidence intervals when correctly specified parametric models are used. Machine-learning algorithms such as random forests can also be used to estimate $\mu_{a,1}(X_i)$ (Kern et al., 2016). As shown by Colnet et al. (2022a), if the model is correctly specified (see Assumption 9 below), the estimator is consistent.

[3] See their Appendix, Section A, pages 6–7.

ASSUMPTION 9 (Consistency of surface response estimators). Denote $\hat{\mu}_{0,n}$ (resp., $\hat{\mu}_{1,n}$) an estimator of μ_0 (resp., μ_1). Let $\mathcal{D}_n$ the RCT sample, so that:

- for $a \in \{0,1\}$, $\mathbb{E}[|\hat{\mu}_{a,n}(X) - \mu_a(X)| \mid \mathcal{D}_n] \xrightarrow{p} 0$, when $n \to \infty$;
- for $a \in \{0,1\}$, there exist C_1, N_1 so that for all $n \geq N_1$, a.s., $\mathbb{E}[\hat{\mu}^2_{a,n}(X) \mid \mathcal{D}_n] \leq C_1$.

THEOREM 3.2 (Consistency of the plug-in g-formula—Colnet et al. (2022a)). *Under causal assumptions (Assumptions 1, 2, 6, 7) and Assumption 9, the plug-in g-formula converges toward τ in L^1 norm,*

$$\hat{\tau}_{G,n,m} \xrightarrow[n,m\to\infty]{L^1} \tau.$$

3.2.3 *Calibration weighting*: *Balancing covariates*. Beyond propensity scores, other schemes use sample reweighting. Lee et al. (2023) propose a calibration weighting approach, similar to the idea of entropy balancing weights introduced by Hainmueller (2012). They calibrate subjects in the RCT sample in such a way that after calibration, the covariate distribution of the RCT sample empirically matches the target population.

DEFINITION 5 (Calibration weighting (CW)). Let $\boldsymbol{g}(X)$ be a vector of functions of X to be calibrated, for example, the moments, interactions and nonlinear transformations of components of X. Then assign a weight ω_i to each subject i in the RCT sample by solving the following optimization problem:

$$\min_{\omega_1,\ldots,\omega_n} \sum_{i=1}^{n} \omega_i \log \omega_i,$$

$$\text{subject to} \quad \omega_i \geq 0, \text{ for all } i,$$

$$\sum_{i=1}^{n} \omega_i = 1, \sum_{i=1}^{n} \omega_i \boldsymbol{g}(X_i) = \widetilde{\boldsymbol{g}}, \text{ (the balancing constraint)}$$

where $\widetilde{\boldsymbol{g}} = m^{-1} \sum_{i=n+1}^{m+n} \boldsymbol{g}(X_i)$ is a consistent estimator of $\mathbb{E}[\boldsymbol{g}(X)]$ from the observational sample. Based on the calibration weights, the CW estimator is then

$$\hat{\tau}_{\mathrm{CW},n,m} = \sum_{i=1}^{n} \hat{\omega}_{n,m}(X_i) Y_i \left(\frac{A_i}{e_1(X_i)} - \frac{1 - A_i}{1 - e_1(X_i)} \right),$$

where $\hat{\omega}_{n,m}(\cdot)$ is the estimated $\omega(\cdot)$ using the RCT and observational data.

The optimization problem in Definition 5 corresponds to the negative entropy of the calibration weights; thus, minimizing this criterion ensures that the empirical distribution of calibration weights is not too far away from the uniform distribution. This aims at minimizing the variability due to heterogeneous weights. This optimization problem can be solved using convex optimization with Lagrange multipliers. For an intuitive understanding of the calibration weighting framework, consider $\boldsymbol{g}(X) = X$. In such a setting, the balancing constraint is forcing the means of the observational data and of the RCT to be equal after reweighting. More complex constraints can enforce balance on higher-order moments. The calibration algorithm is inherently imposing a log-linear model on the sampling propensity score and solving the corresponding parameters by a set of estimating equations induced by covariate balance. Other objective functions of the weights correspond to different models for the sampling propensity score (Chu, Lu and Yang, 2023). Wu and Yang (2023) propose a cross-validation procedure to select the calibration weights that target the smallest mean squared error of the resulting estimator. The CW estimator $\hat{\tau}_{\mathrm{CW},n,m}$ is doubly robust in that it is a consistent estimator for τ if the selection score of RCT participation follows a log-linear model, that is, $\pi_S(X) = \exp\{\boldsymbol{\eta}_0^\top \boldsymbol{g}(X)\}$ for some $\boldsymbol{\eta}_0$, or if the CATE is linear in $\boldsymbol{g}(X)$, that is, $\tau(X) = \boldsymbol{\gamma}_0^\top \boldsymbol{g}(X)$, though not necessarily both. The authors suggest a bootstrap approach to estimate its variance.

3.2.4 *Doubly-robust estimators*. The model for the expectation of the outcomes among randomized individuals (used for the plug-in g-formula estimator in Definition 4) and the model for the probability of trial participation (used in the IPSW estimator in Definition 2) can be combined to form an Augmented IPSW estimator (AIPSW). It can be shown that this estimator is doubly robust, that is, consistent when either one of the two models for $\hat{\alpha}_{n,m}(\cdot)$ and $\widehat{\mu}_{a,1}(\cdot)$ $(a = 0, 1)$ is correctly specified. Dahabreh et al. (2020) has proposed a proof in the nested case (see their Appendix, Section A) said to follow the same principle in the nonnested design (Section B, p. 25). In the plain text, we recall the results from Colnet et al. (2022a).

DEFINITION 6 (Augmented IPSW (AIPSW)). The augmented IPSW estimator, denoted $\hat{\tau}_{\mathrm{AIPSW},n,m}$, is defined as

$$\begin{aligned}
&\hat{\tau}_{\mathrm{AIPSW},n,m} \\
&\quad = \frac{1}{n} \sum_{i=1}^{n} \frac{n}{m \hat{\alpha}_{n,m}(X_i)} \left(\frac{A_i (Y_i - \hat{\mu}_{1,1,n}(X_i))}{e_1(X_i)} \right) \\
&\qquad - \frac{1}{n} \sum_{i=1}^{n} \frac{n}{m \hat{\alpha}_{n,m}(X_i)} \left(\frac{(1 - A_i)(Y_i - \hat{\mu}_{0,1,n}(X_i))}{1 - e_1(X_i)} \right) \\
&\qquad + \frac{1}{m} \sum_{i=n+1}^{m+n} \big(\hat{\mu}_{1,1,n}(X_i) - \hat{\mu}_{0,1,n}(X_i) \big),
\end{aligned}$$

where $\hat{\mu}_{a,1,}$ are estimated on the RCT sample (see Definition 4), and $\hat{\alpha}_{n,m}$ (see Definition 2) on the concatenated RCT and observational samples.

ASSUMPTION 10 (Consistency assumptions—AIPSW). The nuisance parameters are bounded, and more particularly:

- There exists a function α_0 bounded from above and below (from zero), satisfying

$$\lim_{m,n\to\infty} \sup_{x\in\mathcal{X}} \left| \frac{n}{m\hat{\alpha}_{n,m}(x)} - \frac{1}{\alpha_0(x)} \right| = 0;$$

- There exist two bounded functions $\xi_1, \xi_0 : \mathcal{X} \to \mathbb{R}$, such that $\forall a \in \{0, 1\}$,

$$\lim_{n\to+\infty} \sup_{x\in\mathcal{X}} |\xi_{a,1}(x) - \hat{\mu}_{a,1,n}(x)| = 0.$$

THEOREM 3.3 (AIPSW consistency—Colnet et al., 2022a). *Assuming causal assumptions (Assumptions 1, 2, 6, 7) and Assumption 10 (consistency) and considering that estimated surface responses $\hat{\mu}_{a,1,n}(\cdot)$ where $a \in \{0, 1\}$ are obtained following a cross-fitting estimation, then if Assumption 9 **or** Assumption 8 also holds, then $\hat{\tau}_{AIPSW,n,m}$ converges toward τ in L^1 norm,*

$$\hat{\tau}_{AIPSW,n,m} \xrightarrow[n,m\to\infty]{L^1} \tau.$$

This estimator is also shown to be asymptotically normal when both the outcome mean and conditional odds models are consistently estimated at least at rate $n^{1/4}$ in Dahabreh and Hernán (2019) and Li, Hong and Stuart (2023). Note that machine-learning tools are tempting to avoid model misspecification when estimating nuisance parameters. Still this practice requires specific caution, such as using cross-fitting, due to overfitting and regularization. These issues are well described in the situation of a single observational data set. We refer to Chernozhukov et al. (2018) for a detailed explanation, and to Zhong et al. (2021), Bach et al. (2021), Bach et al. (2022) for implementations.

More recently, Lee et al. (2023) have proposed an augmented calibration weighting (ACW) estimator.

DEFINITION 7 (Augmented CW (ACW)). The ACW estimator, denoted $\hat{\tau}_{\mathrm{ACW},n,m}$, is defined as

$$\begin{aligned}\hat{\tau}_{\mathrm{ACW},n,m} &= \sum_{i=1}^{n} \hat{\omega}_{n,m}(X_i) \left(\frac{A_i(Y_i - \hat{\mu}_{1,1,n}(X_i))}{e_1(X_i)} \right) \\ &\quad - \sum_{i=1}^{n} \hat{\omega}_{n,m}(X_i) \left(\frac{(1 - A_i)(Y_i - \hat{\mu}_{0,1,n}(X_i))}{1 - e_1(X_i)} \right) \\ &\quad + \frac{1}{m} \sum_{i=n+1}^{m+n} \left(\hat{\mu}_{1,1,n}(X_i) - \hat{\mu}_{0,1,n}(X_i) \right),\end{aligned}$$

where the estimation of $\hat{\omega}_{n,m}(\cdot)$ is detailed in Definition 5, and where $\hat{\mu}_{a,1,n}$ are estimated on the RCT sample (see Definition 4).

They show that $\hat{\tau}_{\mathrm{ACW},n,m}$ achieves double robustness and local efficiency, that is, its asymptotic variance achieves the semiparametric efficiency bound when both the calibration weights and the outcome mean model are correctly specified. Moreover, the convergence rate of the ACW estimator corresponds to the product of the convergence rates of the nuisance estimators, enabling the use of machine-learning estimation of nuisance functions while preserving the $\sqrt{n}$-consistency of the ACW estimator, when both the outcome mean and calibration weights model are consistently estimated at rate $n^{1/4}$ (Lee et al., 2023). Furthermore, Lee, Yang and Wang (2022) and Lee, Ghosh and Yang (2022) extend the framework for handling survival outcomes.

3.2.5 *Practical issues: Nonparametric estimation, overlap and unobserved covariates.*

3.2.5.1 *Lack of overlap.* The overlap assumption (see Assumption 7) is restrictive because RCT inclusion and exclusion criteria can be strict as the goal of RCTs (at least in early stages) is to show a clear effect even on a restricted population. Whenever Assumption 7 does not hold, it is still possible to generalize on a different target population, such as the subset of the target population for which eligibility criteria of the trial are ensured. This has also been suggested before, for example, by Tipton (2013), page 245. The question asked would rather be "What would have been the estimated treatment effect in a situation where the trial had sampled individuals from the target population who met the trial eligibility criteria?" Another approach has been proposed by Chen, Chen and Yu (2023). Similar to the idea of trimming propensity scores for dealing with limited overlap between treated and control groups, they propose a generalizability score: a function of participation probability and propensity score, to select subpopulations from the observational data for causal generalization when overlap is limited.

3.2.5.2 *Unobserved treatment effect modifiers.* Finally, we point out the important caveat that all methods assume the ignorability conditions (see Assumptions 3, 4, 5 or 6): given the covariates X, the conditional treatment effect must be the same in the observational data and the RCT. In particular, this assumption could be violated if some shifted treatment effect modifiers were not captured in the concatenated data, which is a plausible scenario given that data are seldom collected jointly, and thus typically measure different covariates.

In case of a richer set of covariates in the RCT than in the observational study (which does not necessarily mean that a sufficient set of pretreatment covariates can be chosen; see, e.g., M-bias in Pearl (2000), p. 186), Egami and Hartman (2021) propose a method to select a sufficient set of covariates. But in the case of a low number of common covariates, standard practice is to consider the subset of covariates present in both data sets, but this violates the identifiability condition. Recently, sensitivity analyses

have been proposed to mitigate the consequences of missing covariates in the RCT, or in the observational sample or even in both data sets (Nguyen et al., 2017, Andrews and Oster, 2019, Nguyen et al., 2018, Dahabreh et al., 2023b, Colnet et al., 2022a, Nie, Imbens and Wager, 2021, Huang, 2022).

4. WHEN OBSERVATIONAL DATA CONTAIN TREATMENT AND OUTCOME INFORMATION

Section 3 studied how to correct RCT selection bias (with respect to the target population) while leveraging covariate distribution of an observational sample. When the observational sample also contains treatment and outcome information (Y, A), efficiency improvements can be obtained (Huang et al., 2023). But beyond the generalization question, such additional covariates enable different questions of interest. These questions are the purpose of Section 4. Indeed, RCTs can make causal conclusions from the observational sample more trustworthy, either by removing confounding bias (detailed in Section 4.1) or via more efficient estimation (detailed in Section 4.2). For completeness, we recall in Section B of the Supplementary Material of Colnet et al., 2024 how to perform causal inference from purely observational data.

4.1 Dealing with Unmeasured Confounders in Observational Data

4.1.1 *Motivation*. Unmeasured confounding implies that $\{Y(1), Y(0)\} \not\perp\!\!\!\perp A \mid X$, where X are the observed covariates. In such situations, standard causal inference estimators $\hat{\tau}_m^{\mathcal{O}}(x)$ (resp., $\hat{\tau}_m^{\mathcal{O}}$) of the CATE $\tau(X)$ (resp., ATE τ), that are designed for purely observational data of size m, face a so-called hidden confounding bias for these quantities, that is,

$$\lim_{m\to+\infty} \hat{\tau}_m^{\mathcal{O}}(x) \neq \tau(x), \quad \text{and} \quad \lim_{m\to+\infty} \hat{\tau}_m^{\mathcal{O}} \neq \tau.$$

In practice, former RCTs can be used as *negative controls*,[4] to ensure the observational study does not suffer from confounding. For example, in a recent observational study on a COVID-19 vaccine, Dagan et al. (2021) use such an approach to ensure that the previous trial results conclusion could be retrieved. When confounding remains, solutions such as sensitivity analysis have been developed to handle such situations (Rosenbaum, 2002, Imbens, 2003), but they typically rely on sensitivity parameters, which are difficult to set. Including additional experimental data brings interesting promises to handle such identification bias. Recent works described below propose to use an RCT to *ground* the observational analysis and debias the estimator that would be obtained on purely confounded observational data.

4.1.2 *Using an assumption on secondary outcomes or surrogates*. The use of surrogate outcomes arises in different contexts, for example, in clinical studies (Prentice, 1989, Begg and Leung, 2000), where it may be difficult to observe long-term outcomes, for example, the effect of early childhood medical or economic interventions. Athey, Chetty and Imbens (2020), Athey et al. (2020) observe that the effect of class size reduction leads to a decrease in children 3rd grades in the observational data, while a famous RCT, the Tennessee Student/Teacher Achievement Ratio (STAR) study (Krueger, 1999), concludes on a positive effect. This difference could come from the fact that the two populations are different, but they assume the apparent difference can be entirely explained by confounding.[5] In their setup, they consider two outcomes, a primary long-term outcome $Y^{1^{st}}$ (8th grades) and a secondary short-term outcome $Y^{2^{nd}}$ (3rd grades). The RCT contains information on the surrogate but not the long-term outcome while this is the opposite for the observational sample. Their central assumption to recover identifiability is called *latent unconfoundedness*, that is,

$$A \perp\!\!\!\perp Y^{1^{st}}(a) \mid Y^{2^{nd}}(a), \quad i \in \mathcal{R}, \text{ for } a = 0, 1,$$

which corresponds to the assumption that hidden confounders violating identification of the effect on $Y^{1^{st}}$ are the same than for $Y^{2^{nd}}$. In other words, their method consists in adjusting the estimates of the treatment effects on the primary outcome using the differences observed on the secondary outcome. Their assumptions can be understood as a missing data problem, that is, the missing data in the primary outcomes are missing at random in the concatenated data (Rubin, 1976). For estimation, they suggest three methods, namely *(i)* imputing the missing primary outcome in the RCT, *(ii)* weighting the units in the observational sample and *(iii)* using control function methods.

4.1.3 *Deconfounding using the bias/confounding function*. Kallus, Puli and Shalit (2018) propose to use an RCT sample to deconfound the CATE estimated on a single observational data set, denoted $\hat{\tau}_m^{\mathcal{O}}(x)$. Due to possible unmeasured confounding, $\hat{\tau}_m^{\mathcal{O}}(x)$ may be biased for $\tau(x)$, that is, $\eta(x) \neq 0$ where $\eta(x) := \tau(x) - \hat{\tau}_m^{\mathcal{O}}(x)$ is the bias function. To correct for this bias, they assume they have at hand a narrow RCT (as it is usually the case with

[4] The term negative controls comes from usual routine precaution in biological laboratory experiments, where such controls are used to—at least partially—check that the experiment is not undermined. For example, it can test the absence of reagents or components that are necessary for a detection of something particular. For example, one of the two bars of the covid antigenic test is one of these controls. The analogy of this principle in causal inference is detailed in (Lipsitch, Tchetgen and Cohen, 2010).

[5] Assuming the bias comes from an unobserved confounder and not from inherent differences between populations can be stated as $S \perp\!\!\!\perp \{Y(1), Y(0)\}$, which means that the two samples come from comparable populations (see Section 3).

strict eligibility criteria in trial) with high internal validity, and with covariate support included in the observational sample support. Given that $\hat{\tau}_m^{\mathcal{O}}(x)$ is obtained from the observational data, one can estimate $\eta(\cdot)$ on the common support between the RCT and the observational data using the (unconfounded) RCT data. Another assumption is required, being that the bias can be well approximated by a function with low complexity, for example, a linear function of the covariates x: $\eta(x) = \theta^T x$. Kallus, Puli and Shalit (2018) then propose to estimate the bias as $\hat{\eta}_{m,n}(x) = \hat{\theta}_{m,n}^T x$ by solving the following minimization problem:

$$\begin{aligned}\hat{\theta}_{m,n} &= \operatorname*{argmin}_{\eta} \sum_{i=1}^{n} \big(\mathrm{Y}_i^* - \hat{\tau}_m^{\mathcal{O}}(X_i) - \eta(X_i)\big)^2 \\ &= \operatorname*{argmin}_{\theta} \sum_{i=1}^{n} \big(\mathrm{Y}_i^* - \hat{\tau}_m^{\mathcal{O}}(X_i) - \theta^T X_i\big)^2,\end{aligned}$$

where $Y_i^* = (e(X_i)^{-1}A_i - \{1 - e(X_i)\}^{-1}(1 - A_i))Y_i$, which satisfies $\mathbb{E}[Y_i^* \mid X_i] = \tau(X_i)$.

Note that the linear assumption guarantees the validity of the framework even if the observational data do not fully overlap with the experimental data as the bias, that i,, the confounding error, is assumed to be extrapolable.

Finally, $\hat{\tau}_{m,n}(x) = \hat{\tau}_m^{\mathcal{O}}(x) + \hat{\eta}_{m,n}(x)$ is the estimated conditional average treatment effect. They prove that under conditions of parametric identification of η, $\hat{\tau}_{m,n}(x)$ is a consistent estimate of $\tau(x)$, which converges at a rate governed by the rate of estimating $\mathbb{E}[\hat{\tau}_m^{\mathcal{O}}(x)]$ by $\hat{\tau}_m^{\mathcal{O}}(x)$.

More recently, Yang, Zeng and Wang (2020) proposed another approach. Rather than $\eta(x)$, they consider what they call the *confounding function* $\lambda(x)$,

$$\begin{aligned}\lambda(x) = \mathbb{E}\big[Y(0) \mid A = 1, X = x\big] \\ - \mathbb{E}\big[Y(0) \mid A = 0, X = x\big],\end{aligned}$$

summarizing the impact of unmeasured confounders on the potential outcome distribution between the treated and untreated patients. In the absence of unmeasured confounding, $\lambda(x)$ is zero for any $x \in \mathcal{X}$, while if there is unmeasured confounding, $\lambda(x) \neq 0$ for some x. Assuming a parametric model assumption for the CATE $\tau(x) := \tau_{\varphi_0}(x)$ with $\varphi_0 \in \mathbb{R}^{p_1}$, and for $\lambda(x) := \lambda_{\phi_0}(x)$ with $\phi_0 \in \mathbb{R}^{p_2}$, the coupling of RCT and observational data allows identifiability of $\tau(x)$ and $\lambda(x)$. The key insight is to introduce the following random variable:

$$H_{\psi_0} = Y - \tau_{\varphi_0}(X)A - (1 - S)\lambda_{\phi_0}(X)\big\{A - e(X)\big\},$$

where $\psi_0 = (\varphi_0^{\mathrm{T}}, \phi_0^{\mathrm{T}})^{\mathrm{T}}$ is the full vector of model parameters in the CATE and confounding function, and where here $S = 1$ (resp., $S = 0$) denotes trial participation (resp., observational study participation). By separating the treatment effect $\tau_{\varphi_0}(X)A$ and $(1 - S)\lambda_{\phi_0}(X)\{A - e(X)\}$ from the observed Y, H_{ψ_0} mimics the potential outcome $Y(0)$. They then derive the semiparametric efficient score of ψ_0:

$$\begin{aligned}S_{\psi_0}(V) = &-\mathbf{K}\big(\sigma_S^2(X)\big)^{-1}\mathbb{E}[H_{\psi_0} \mid X, S]\big(A - e(X)\big) \\ &+ \mathbf{K}\big(\sigma_S^2(X)\big)^{-1} H_{\psi_0},\end{aligned} \tag{3}$$

where

$$\mathbf{K} := \begin{pmatrix} \dfrac{\partial \tau_{\varphi_0}(X)}{\partial \varphi_0} \\ \dfrac{\partial \lambda_{\phi_0}(X)}{\partial \phi_0}(1 - S) \end{pmatrix},$$

and $\sigma_S^2(X) = \mathbb{V}[Y(0) \mid X, S]$. A semiparametric efficient estimator of ψ_0 can be obtained by solving the estimating equation based on (3). If the predictors in $\tau_{\varphi_0}(X)$ and $\lambda_{\phi_0}(X)$ are not linearly dependent, they show that the integrative estimator of the CATE is strictly more efficient than the RCT estimator. As a by-product, this framework can be used to generalize the ATE from the RCT to a target population without requiring an overlap covariate distribution assumption between the RCT and observational data. Wu and Yang (2022) propose an integrative R-learner that extends the framework of Yang, Zeng and Wang (2020) to allow flexible machine-learning methods for approximating CATE, confounding function and nuisance functions.

4.2 Toward More Efficient Estimation

Under Assumptions 1, 2 and 6, the CATE can be estimated based on the RCT, while under the classical unconfoundedness assumption (see Section B of the Supplementary Material of Colnet et al., 2024), the CATE can be estimated using the observational sample. Therefore, when both sets of assumptions are met, the two data sources can be pooled to improve estimation efficiency. Toward this end, Yang, Wang and Zeng (2023) use the semiparametric efficiency theory to derive the semiparametrically efficient integrative estimator of φ_0 for the CATE $\tau_{\varphi_0}(X)$. However, if the unconfoundedness assumption is violated, integrating the observational sample would bias the CATE estimation. Leveraging the design advantage of RCTs, Yang, Wang and Zeng (2023) derive a preliminary test statistic for the comparability and reliability assessment of the observational data and decide whether to use it in an integrative analysis. Denote the efficient score based solely on the RCT and observational data as $S_{\mathrm{rct},\varphi_0}(V)$ and $S_{\mathrm{os},\varphi_0}(V)$, respectively, where V is a full vector of variables. Their basic idea is to derive an RCT estimator $\widehat{\varphi}_{\mathrm{rct}}$ for φ_0 and construct the preliminary test statistics based on $S_{\mathrm{os},\widehat{\varphi}_{\mathrm{rct}}}(V)$. The rationale is that if the observational sample is comparable to the RCT sample for estimating φ_0, $S_{\mathrm{os},\widehat{\varphi}_{\mathrm{rct}}}(V)$ is expected to be close to zero; otherwise, $S_{\mathrm{os},\widehat{\varphi}_{\mathrm{rct}}}(V)$ is expected to deviate from

zero. This thought process leads to the test statistics

$$T = \left\{ n^{-1/2} \sum_{i=n+1}^{n+m} S_{\mathrm{os},\widehat{\varphi}_{\mathrm{rct}}}(V_i) \right\}^{\mathrm{T}} \cdot \widehat{\Sigma}_{SS}^{-1} \cdot \left\{ n^{-1/2} \sum_{i=n+1}^{n+m} S_{\mathrm{os},\widehat{\varphi}_{\mathrm{rct}}}(V_i) \right\}, \tag{4}$$

where $\widehat{\Sigma}_{SS}$ is a consistent estimator for the asymptotic variance of $n^{-1/2} \sum_{i=n+1}^{n+m} S_{\mathrm{os},\widehat{\varphi}_{\mathrm{rct}}}(V_i)$. Under H_0 that the observational sample is comparable to the RCT sample, $T \to \chi^2_p$, a chi-square distribution with degrees of freedom $\dim(\varphi_0)$, as $n \to \infty$. This result serves to detect the violation of the assumption required for the observational data.

Yang, Wang and Zeng (2023) propose the elastic integrative estimator by solving

$$\sum_{i=1}^{n} \widehat{S}_{\mathrm{rct},\varphi}(V_i) + \mathbb{I}(T < c_\gamma) \sum_{i=n+1}^{n+m} \widehat{S}_{\mathrm{os},\varphi}(V_i) = 0, \tag{5}$$

where c_γ is the $100(1-\gamma)$th percentile of χ^2_p, serving as a switch to decide on combining or not. The methodological contribution of Yang, Wang and Zeng (2023) is to derive a data-adaptive selection of c_γ such that the resulting estimator has the smallest mean squared error, and thus performs at least similar to the RCT-only estimator, if not better. Moreover, the elastic integrative estimator is non-regular and belongs to pretest estimation by construction. The theoretical contributions of Yang, Wang and Zeng (2023) include characterizing the distribution of the elastic integrative estimator under local alternatives, which better approximates the finite-sample behaviors, and provides data-adaptive confidence intervals that are uniformly valid.

4.3 Other Use Cases

Beyond generalizability or overcoming confounding, there are other purposes motivating the combination of experimental and observational data. We provide a brief list of these purposes and methodologies. A detailed or exhaustive survey is beyond the scope of this review.

4.3.1 *Using hybrid controls.* A hybrid control arm is a control arm constructed from a combination of randomized patients and patients receiving usual care in standard clinical practice, as introduced by Pocock (1976) and pursued by Hobbs, Sargent and Carlin (2012), Schmidli et al. (2014). Recently, the FDA has detailed their usage in the regulatory purposes (FDA, 2018). Using hybrid controls has the potential to decrease the cost of randomized trials, and to reduce ethic constraints on control groups.

4.3.2 *Case-control studies.* In certain applications, for example, in epidemiology, the observational data at hand comes from a case-control study where the selection of observations is driven by the outcome of interest Y. Thus, the RCT and observational data differ in terms of the outcome distribution, typically a preferential selection on the outcome for the observational data set. Several solutions have been proposed to handle this type of selection bias. Robins (2000) and Hernán et al. (2005) propose marginal structural model approaches to eliminate this bias given sufficient knowledge of the selection model given treatment. Guo et al. (2021) propose a control variates technique (Tan, 2006, Yang and Ding, 2020) identifying and estimating an estimand that is sufficiently correlated with the target estimand of interest for the observational cohort.

4.3.3 *Encouragement design intervention.* An encouragement design intervention is a design in which some individuals or groups are randomly assigned to receive encouragement to take up the program. (Rudolph and van der Laan, 2017) provide a semiparametric efficiency score for transporting the ATE from one study following an encouragement design, to another population. Due to the design, their setup is a variant of the generalization work from Section 3, but with treatment allocation information in the target population.

5. STRUCTURAL CAUSAL MODELS (SCM) AND TRANSPORTABILITY

Within the SCM framework (Pearl, 1995, 2009b), Bareinboim and Pearl (2016) have proposed answers for transportability and combination of different datasources, also called *data fusion*. This section is split off from the previous section as it builds on additional concepts.

Let us first briefly introduce the SCM framework, using as much as possible the notation of Section 2.1 that we introduced for the PO framework (Section F of the Supplementary Material of Colnet et al., 2024 gives a more general primer on the SCM framework, and in particular the *do*-operator). The covariates X, treatment A and response Y are modeled in the SCM framework as random variables with joint distribution $P(X, A, Y)$. Each intervention, such as setting A to $a = 0$ or $a = 1$, defines an alternative distribution over (X, A, Y) that can be systematically deduced from the no-intervention (or observational) distribution P using the SCM model, and which is written $P(X, A, Y|do(A = a))$. In this framework, the CATE is written:

$$\tau(x) = \mathbb{E}\big[Y|do(A=1), X=x\big] - \mathbb{E}\big[Y|do(A=0), X=x\big];$$

and the ATE:

$$\tau = \mathbb{E}\big[Y|do(A=1)\big] - \mathbb{E}\big[Y|do(A=0)\big].$$

These expressions mirror the corresponding expressions in the PO framework (Table 2) when one identifies the variable $Y(a)$ in the PO framework to the variable Y under the intervention $do(A=a)$ in the SCM framework, namely when we set $P(Y(a), X) = P(Y, X|do(A=a))$. In fact, this analogy is valid in the sense that any theorem that holds for SCM counterfactuals holds in the PO framework, and vice versa (Pearl, 2009b, Chapter 7; Pearl, 2009a, Chapter 4). In spite of this formal equivalence, the two frameworks differ in how they allow practitioners to express causal assumptions, and to derive corresponding estimands of causal effects. The SCM framework provides a convenient graphical representation known as causal diagrams to encode potentially complex causal assumptions between variables, and provides a complete language known as do-calculus to express causal effects (i.e., some expectation under the $do(A=a)$ probability) as a function of observational data (i.e., some expectation under the no-intervention distribution) (Pearl (1995, 2009b)). When this reduction is possible, the causal effect is called *identifiable*. In addition, the do-calculus is complete in the sense that a causal effect is identifiable if and only if it can be reduced to a function of observational data using do-calculus (Huang and Valtorta, 2006, Shpitser and Pearl, 2006). Interestingly, this provides a variety of formulas to correctly infer causal effects even in the presence of unmeasured confounders, which cannot be handled by the PO framework (without additional structural and modeling assumptions), such as the front-door adjustment formula (Pearl, 1995).

5.1 Formulating Transportability in the SCM Framework

The SCM literature and do-calculus naturally cover the problem of generalizing an RCT experiment to a different target population. Following our notation in the PO setting (Section 2.1), we again denote by S a binary random variable that indicates which individuals can be in the RCT. The RCT population then follows the distribution $P(X, Y, A|S=1)$, and by design the RCT allows estimating the conditional distributions $\mathbb{P}(Y|do(A=a), X, S=1)$ for $a=0,1$. The problem of generalization to the target population in this setting is then to deduce the distributions of $\mathbb{P}(Y|do(A=a), X)$ for $a=0,1$ from these two distributions and the observed distribution of the covariates $P(X)$ in the target distribution (as in Section 3), or of the covariates, treatments and responses $\mathbb{P}(X, A, Y)$ in the target population (as in Section 4). If this deduction (using do-calculus) is possible, then the causal effect on the target population is identifiable, and the deduction provides a formula for the causal effect that can then be estimated from a finite population using some consistent estimator.

Interestingly, this formalism covers two important situations: (i) the *sample selection bias* problem, when the RCT population is a subset of the target population that fulfills some eligibility criterion,[6] and (ii) the *transportability* problem, where the RCT population differs more drastically from the target, for example, when one wants to generalize an RCT conducted in one country to a population in another country (Pearl, 2015). To model sample selection bias, on the one hand, one typically adds a node S with incoming edges to a causal graph in order to capture the eligibility conditions that may depend on pre- or post-treatment variables. It is then possible to derive conditions under which one can recover from selection bias when the probability of selection is available (Cooper, 1995, Lauritzen and Richardson, 2008, Geneletti, Richardson and Best, 2008) or when no quantitative knowledge is available about probability of selection (Didelez, Kreiner and Keiding, 2010, Bareinboim and Pearl, 2012a). We provide examples of such conditions in Section F.3 of the Supplementary Material of Colnet et al. (2024). To model transportability to a different population, on the other hand, the node S has typically no incoming edge, and instead points to variables that differ between the RCT and the target population, either in their functional dependency to their parents in the causal graph, or in the distribution of their exogenous variables. The resulting graph is called a *selection diagram* and allows to encode graphically detailed assumptions about the differences between populations (Pearl and Bareinboim, 2011, 2014, Bareinboim and Pearl, 2012b, 2013). Note that even if the two situations imply different causal diagrams, the problem of selection bias "*has some unique features, but can also be viewed as a nuance of the transportability problem, thus inheriting all the theoretical results of transportability*" (Pearl, 2015); this remark is connected to the discussion from Section 2.2.

The SCM approach thus provides powerful machinery to generalize causal effect across populations, and entails a detailed description of the causal assumptions between variables in the selection diagram, including the selection variable S. The two selection diagrams of Figure 3 represent, for example, transportability problems with a distributional change of covariates X between the RCT and target populations (with an arrow from S to X), and where the interventional nature of the RCT versus the target population is also represented with an arrow from S to A. In addition, in Figure 3(a) the arrow from S to Y indicates that the conditional distribution of Y given X and A differs between the two

[6] This setting has been termed as *generalizability* in the introduction of the different study designs in Section 2.2.

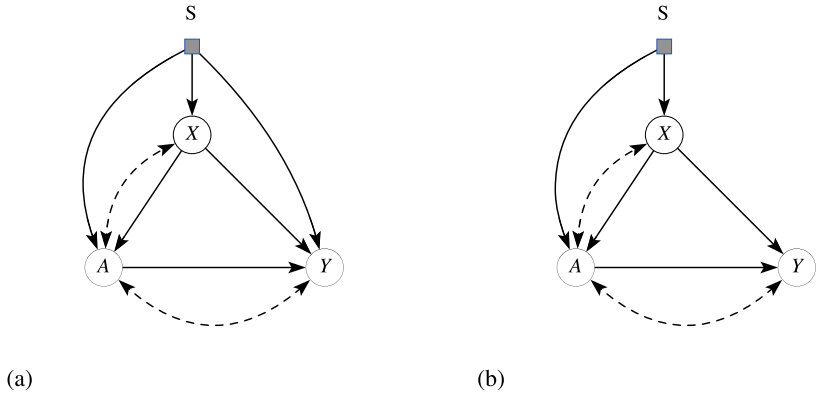

FIG. 3. *Illustration of selection diagrams depicting differences between source and target populations*: *In* (*a*) *and* (*b*), *the two populations differ by covariate distributions* (*indicated by S pointing to X*) *and the two populations differ in their interventional nature* (*S pointing to A*). *Assumption* 6 (*transportability assumption*) *is assumed on* (*b*), *but not on* (*a*) (*since S points to Y in* (*a*)). *These two examples are inspired by* Pearl and Bareinboim (2011).

populations, which in general prevents any transportability of causal effect, while the lack of arrow from S to A in Figure 3(b) encodes the independence assumption $\mathbb{P}(Y|X, A) = \mathbb{P}(Y|X, A, S = 1)$, which implies the transportability assumption $\mathbb{P}(Y|do(A = a), X, S = 1) = \mathbb{P}(Y|do(A = a), X)$ (which itself implies Assumption 6 in the PO framework). In that case, one easily deduces by simple conditioning on X that the distribution of Y under intervention on the whole population is given by $\mathbb{P}(Y \mid do(A = a)) =$

$$\sum_x \underbrace{\mathbb{P}\big(Y \mid do(A = a), X = x, S = 1\big)}_{\text{RCT}} \underbrace{\mathbb{P}(X = x)}_{\text{Obs.}}. \tag{6}$$

This transport formula, also known as *recalibration*, *reweighting* or *post-stratification* formula (Pearl, 2015), thus combines experimental results obtained in the RCT population and the observational description of the target population to estimate the causal effect in the target population. In particular, we easily deduce the ATE on the target population by integrating (6) in Y to get

$$\tau = \sum_x \underbrace{\tau_1(x)}_{\text{RCT}} \underbrace{\mathbb{P}(X = x)}_{\text{Obs.}}, \tag{7}$$

where $\tau_1(x)$ is by design identifiable by conditioning on treatment in the RCT population. This formula (7) is equivalent to the regression formula (2) in the PO framework, which is valid under Assumption 6. Interestingly, Pearl and Bareinboim (2011) show that the transport formula (6) holds more generally as soon as X is a set of pre-treatment variables, which is *S-admissible*, that is, if $S \perp\!\!\!\perp Y \mid X, do(A = a)$ for $a = 0, 1$. Graphically, S-admissibility holds whenever X blocks all paths from S to Y after deleting from the graph all incoming arrows into A. We note that S-admissibility implies the mean exchangeability assumption (Assumption 4) and is equivalent to the S-ignorability assumption $S \perp\!\!\!\perp Y(a) \mid X$ (Assumption 3) used in the PO literature when X and S are pre-treatment variables, and entails similar transport formula in that situation. However, the two notions differ for treatment-dependent selection and covariates, as discussed by Pearl (2015), where several examples illustrate how the S-admissibility assumption can lead to different transport formulas when both pre- and post-treatment variables are leveraged. Such an example is presented on Figure 4, where the covariate X is a post-treatment variable, for example, a biomarker, believed to mediate between treatment and outcome.

Here, we presented how Assumptions 2, 3 and 4 are translated in the SCM literature and how another scenario with post-treatment covariates can be identified. More identifiability scenarios have been discussed in the SCM literature (Huang and Valtorta, 2012, Bareinboim et al., 2013, Pearl, 2015, Lee, Correa and Bareinboim, 2020b), and to our knowledge we have found no similar identifiability scenario in the PO literature. It is worth mentioning that the transportation problem discussed so far, to export a causal effect estimated in an RCT to a general population, is only one specific instance of the more general problem of *data fusion* (Pearl and Bareinboim, 2011,

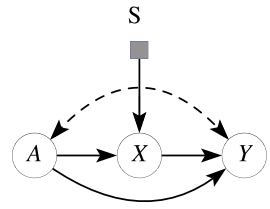

FIG. 4. *Post-treatment covariate adjustment*: *On this selection diagram, the arrow from S to X indicates the assumption of different effects of A on X in the two populations. Here, X is S-admissible but not S-ignorable, and the corresponding transport formula is* $\mathbb{P}(Y \mid do(A = a)) = \sum_x \mathbb{P}(Y \mid do(A = a), X = x, S = 1)\mathbb{P}(X = x \mid A = a)$, *where it invokes an unconventional average of the CATE weighted by a conditional probability in the target population. This example is taken from* Pearl (2015).

Bareinboim and Pearl, 2012b, 2016, Hünermund and Bareinboim, 2019, Lee, Correa and Bareinboim, 2020a), which simultaneously accounts for confounding issues of observational data, sample selection issues, as well as extrapolation of causal claims across heterogeneous environments. The SCM framework, with its elegant way of formalizing the problem, helps practitioners formulate and discuss causal assumptions across variables and environments. In particular, subject to a good knowledge of the graph, it helps selecting sets of variables that are sufficient to establish identifiability and exclude variables that would bias the analysis. As we will see in Section 7, already in the early phase of a study, the causal and selection diagrams offer a very convenient tool to discuss with clinicians and explicitly lay out conditional independence assumptions. Once a diagram encodes assumptions about a system, algorithmic solutions implementing the *do*-calculus are available to determine whether nonparametric identifiability holds, and to provide correct formula if it holds (Correa, Tian and Bareinboim, 2018, Tikka, Hyttinen and Karvanen, 2019).

While the SCM literature provides powerful and versatile sets of concepts and tools to *identify* causal effects, practical estimators with publicly available implementations and detailed consistency, convergence rates or robustness results are still scarce. Some recent work has proposed solutions for this estimation task in the context of either experimental or observational data by extending weighting-based methods developed for the back-door case to more general settings Jung, Tian and Bareinboim (2020a, 2020b), or extending the double/debiased machine learning (DML) approach proposed by Chernozhukov et al. (2018) under ignorability assumption to any identifiable causal effect (Jung, Tian and Bareinboim, 2021). In the same spirit, Karvanen, Tikka and Hyttinen (2020) propose a combination of data from a survey and a meta-analysis of 34 trials, where identifiability and transport formula are the output of the algorithm `do-search` (see Section 6), and estimation is performed with the real data at hand. Additionally, even if a causal effect is not identifiable, partial-identifiability techniques have been proposed for deriving bounds for the causal effect (Tian and Pearl, 2000, Dawid, Humphreys and Musio, 2019). Cinelli and Pearl (2021) give an example illustrating partial identifiability on real data, with experiments assessing the effect of the Vitamin A supplementation. In this setting, the existence of experimental data from one source population leads to identify bounds on the transported causal effect, but the availability of two trials instead of one leads to a point estimate. Finally, Dahabreh et al. (2019), Dahabreh, Robins and Hernán (2020) propose an alternative approach for generalizability and integrative analyses of trials and observational studies using structural equation models under weaker error assumptions and represented using single world intervention graphs (Richardson and Robins, 2013).

6. SOFTWARE FOR COMBINING RCT AND OBSERVATIONAL DATA

6.1 Review of Available Implementations

An important point to bridge the gap between theory and practice is the availability of software. In recent years, there have been more and more solutions for users interested in causal inference and causation; see Tikka and Karvanen (2017), Guo et al. (2018), Yao et al. (2020) for surveys and Mayer et al. (2022) for a task view of `R` implementations. Regarding the specific subject of this article, we present in Table 3 the implementations available for both identifiability and estimation.

The available implementations are often dedicated to specific sampling designs, and as mentioned, estimators are different from nested and nonnested framework. As a consequence, a new user has to pay attention to all of these practical—but fundamental—details.

6.2 Simulation Study of Generalization Estimators

This part presents simulation results to illustrate the different estimators introduced in Section 3 and their behavior under several misspecification patterns. The code to reproduce the results is available on Github.[7] We implement in R (R Core Team, 2021) our own version of the estimators to match exactly the formulas introduced in the review (`IPSW` and `IPSW.normd`; see Definition 2, `stratification`; Definition 3, `plug-in g-formula`; Definition 4, and `AIPSW`; Definition 6), except for the CW and ACW estimators (Definitions 5) and 7) for which we use the `genRCT` package.

6.2.1 *Scenario* 1: *Well-specified models*. Similar to Lee et al. (2023), we generate nonnested trial settings as follows. First, we draw a sample of size 50,000 from a covariate distribution with four covariates are generated independently as with $X_j \sim \mathcal{N}(1,1)$ for each $j = 1, \ldots, 4$. From this sample, we then select an RCT sample of size $n \sim 1000$ with trial selection scores defined using a logistic regression model:

$$\begin{aligned}\operatorname{logit}\{\pi_S(X)\} &= -2.5 - 0.5X_1 - 0.3X_2 \\ &\quad - 0.5X_3 - 0.4X_4.\end{aligned} \tag{8}$$

Then we generate the treatment according to a Bernoulli distribution with probability equals to 0.5, $e_1(x) = e_1 = 0.5$ and the outcome according to a linear model:

$$\begin{aligned}Y(a) &= -100 + 27.4aX_1 + 13.7X_2 + 13.7X_3 \\ &\quad + 13.7X_4 + \epsilon \quad \text{with } \epsilon \sim \mathcal{N}(0,1).\end{aligned} \tag{9}$$

[7] `BenedicteColnet/combine-rct-rwd-review`.

TABLE 3
Inventory of publicly available code for generalization (*top*: *software for identification*; *bottom*: *software for estimation*)

Name	Method—Setting	Source and Reference
Identification		
causaleffect	Identification and transportation of causal effects, for example, conditional causal effect identification algorithm	R package on CRAN, Tikka and Karvanen (2017)
dosearch	Identification of causal effects from arbitrary observational and experimental probability distributions via do-calculus	R package on CRAN, Tikka, Hyttinen and Karvanen (2019)
Causal Fusion	Identifiability in data fusion framework, (Section 5)	Browser beta version upon request Bareinboim and Pearl (2016)
Estimation		
ExtendingInferences	IPSW (Definition 2), plug-in g-formula equation S7—Nested AIPSW S9—Nested Continuous outcome	R code on GitHub, Dahabreh et al. (2020)
generalize	IPSW (Definition 2), TMLE (Section 3.2.4)	R package on GitHub Ackerman et al. (2021)
genRCT	IPSW (Definition 2), calibration weighting (Section 3.2.4) Continuous and binary outcome	R package Lee et al. (2023)
IntegrativeHTE	Integrative HTE (Section 4.1)	R package on GitHub, Yang, Wang and Zeng (2023)
IntegrativeHTEcf	Includes confounding functions (Section 4.1)	R package on GitHub, Yang, Wang and Zeng (2023)
generalizing	SCM with probabilistic graphical model for Bayesian inference Binary outcome	R package on GitHub, Cinelli and Pearl (2021)
RemovingHiddenConfounding	Unmeasured confounder (Section 4.1)	R package on GitHub, Kallus, Puli and Shalit (2018)
senseweight	Sensitivity analysis (IPSW Definition 2)	R package on Github Huang (2022)
transport	Targeted maximum likelihood estimators (TMLEs) Transport	R package on GitHub, Rudolph et al. (2018)
combine-rct-rwd-review	Generalization estimators of Section 3	R code on GitHub

This outcome model implies a target population ATE of $\tau = 27.4$ and $\mathbb{E}[X_1] = 27.4$. Finally, we generate an observational sample by drawing a new sample of size $m = 10{,}000$ from the distribution of the covariates.

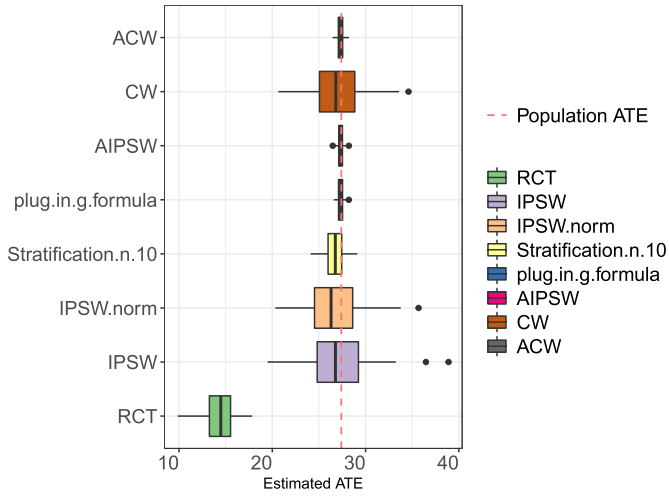

FIG. 5. *Well-specified model. Estimated ATE with the inverse propensity of sampling weighting with and without weights normalization* (*IPSW and IPSW.norm*; *Definition* 2), *stratification* (*with* 10 *strata*; *Definition* 3), *plug-in g-formula* (*Definition* 4), *calibration weighting* (*CW*; *Definition* 5), *augmented IPSW* (*AIPSW*; *Definition* 6) *and ACW* (*Definition* 7)) *over* 100 *simulations*.

Figure 5 presents estimated ATE over 100 simulations. The true ATE is represented with a dashed line. The ATE estimated only with the RCT sample is also displayed as a baseline. As expected, it is biased downward (its mean is equal to 14.24) as the distribution of the covariates and in particular the treatment effect modifiers such as X_1 is not the same in the trial sample and in the population (as illustrated in Table S5 in Section G of the Supplementary Material of Colnet et al., 2024). Note that in this simulation, all the estimators are unbiased. The variability of the two IPSW estimators are larger than the others. The number of strata in the stratification estimator plays an important role. As shown in Figure S7 in Section G of the Supplementary Material of Colnet et al. (2024), the results are biased when the number of strata is smaller than 10.

6.2.2 *Scenario* 2: *Misspecification of the sampling propensity score or outcome model*. To study the impact of misspecification of the sampling propensity score model, we generate the RCT selection according to the model

$$\operatorname{logit}\{\pi_S(X)\} = -2.5 - 0.5e^{X_1} - 0.3e^{X_2} - 0.5e^{X_3} - 0.4e^{X_4} + 3,$$

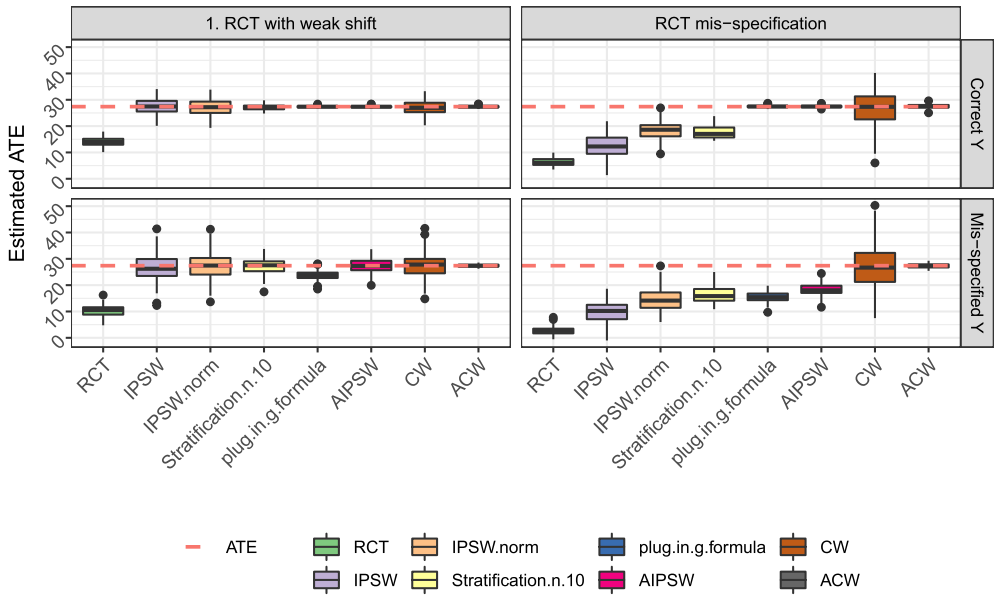

FIG. 6. *Misspecified models. Estimated ATE when selection in RCT and/or outcome models are misspecified. Estimators used being IPSW (IPSW and IPSW.norm; Definition 2), stratification (with 10 strata; Definition 3), plug-in g-formula (Definition 4), calibration weighting (CW; Definition 5), augmented IPSW (AIPSW; Definition 6) and ACW (Definition 7) over 100 simulations.*

and outcome according to the model

$$Y(a) = -100 + 27.4aX_1X_2 + 13.7X_2 + 13.7X_3 + 13.7X_4 + \epsilon.$$

The analysis is then performed using classical logistic and linear estimators on the four covariates. As shown in Figure 6, when the sampling propensity score model is misspecified, the IPSW estimators are biased; whereas when the outcome model is misspecified, the plug-in g-estimator is biased. In both settings, the double robust estimator (AIPSW) is unbiased and robust to misspecification. In the case where both models are misspecified, all estimators are biased except the CW and ACW estimators. This demonstrates some robust properties of calibration against slight model misspecification.

Section G of the Supplementary Material of Colnet et al. (2024) investigates the effect of a missing covariate, homogeneous treatment effect, and the impact of a stronger covariate shift, that is, poorly satisfied Assumption 7.

7. APPLICATION: EFFECT OF TRANEXAMIC ACID

To illustrate the methodological question of combining experimental and observational data and demonstrate some of the previously discussed methods, we consider an open medical question about major trauma patients. We focus on trauma patients suffering from a traumatic brain injury (TBI): brain damage caused by a blow or jolt to the head. Tranexamic acid (TXA) is an antifibrinolytic agent that limits excessive bleeding, commonly given to surgical patients. Previous clinical trials showed that TXA decreases mortality in patients with traumatic *extracranial* bleeding (Shakur-Still et al., 2009). Such prior result raises the possibility that it might also be effective in TBI, because *intracranial* hemorrhage is common in TBI patients, with risks of raised intracranial pressure, brain herniation and death. Therefore, the aim here is to assess the potential decrease of mortality in patients with intracranial bleeding when using TXA. To answer this question, we have at our disposal both an RCT, *CRASH-3* and an observational study, the *Traumabase*. Both data have previously been analyzed separately in CRASH-3 (2019), Cap (2019) (for the RCT) and in Mayer et al. (2020) (for the observational study) and the medical teams of both studies want to share their respective data to answer both medical and methodological questions. Such initiatives allow to: (1) collate the results from the observational study with the RCT findings; (2) assess the generalizability methods, considering the Traumabase as the target population and assess the estimators presented in this review in a real application. We first present the two data sources, treatment effect analyses and findings from these, before turning to the combined analysis in Section 7.3. The code to reproduce all these analyses is available on Github;[8] however, the medical data cannot be publicly shared for privacy concerns.

7.1 The Observational Data: Traumabase

7.1.1 *Context*. The Traumabase regroups 23 French Trauma centers that collect detailed clinical data from

[8] https://github.com/BenedicteColnet/combine-rct-rwd-review.

major trauma patients from the scene of the accident to hospital discharge in form of a registry. The data, currently counting over 30,000 patient records, are of unique granularity and size in Europe. However, they are highly heterogeneous, with both categorical—sex, type of illness, . . . — and quantitative—blood pressure, hemoglobin level, . . .—features, multiple sources and many missing data (98% of the records are incomplete). Here, we use 8270 patients suffering from TBI extracted from the Traumabase. Mayer et al. (2020) performed a first, purely observational, study to assess the effect of TXA on mortality for traumatic brain injury patients from this data: the treatment variable is the administration of TXA during pre-hospital care or on admission to a Trauma Center[9] within 3 hours of the initial trauma. The Traumabase analysis contains many missing values (see Section H of the Supplementary Material of Colnet et al., 2024), which implies additional assumptions to perform causal inference.

7.1.2 *Purely-observational results from two different estimation strategies*. The direct causal effect of TXA on 28-day intrahospital TBI-related mortality and on all cause intrahospital mortality among traumatic brain injury patients is estimated by adjusting for confounding using 17 confounding variables. In addition, 21 variables predictive of the outcome but not related to the treatment are included (see Mayer et al. (2020) for the detailed adjustment set). We recall the results from this study, which put a focus on how to estimate treatment effects in the presence of incomplete data. The presented methods rely either on logistic regressions or generalized random forests (Athey, Tibshirani and Wager, 2019) for the nuisance components, denoted respectively by *GLM* and *GRF* in Table 4. The doubly robust results (AIPW) in Table 4 show that from this study there is no evidence for an effect of TXA on mortality of TBI patients. These findings—obtained prior to the publication of CRASH-3—are consistent with the main conclusion of the CRASH-3 study. However, the results from IPW conclude on a possible deleterious effect. In such a situation, the possibility to generalize the treatment effect from the RCT is also a step to comfort the results. In Section H of the Supplementary Material of Colnet et al. (2024), we additionally recall results on sub-groups obtained by stratifying along trauma severity.

7.2 The RCT: CRASH-3

7.2.1 *Context*. CRASH-3 is a multicentric randomized and placebo-controlled trial launched over 175 hospitals in 29 different countries (Dewan et al., 2012). This trial recruited 9202 adults—unusually large for a medical

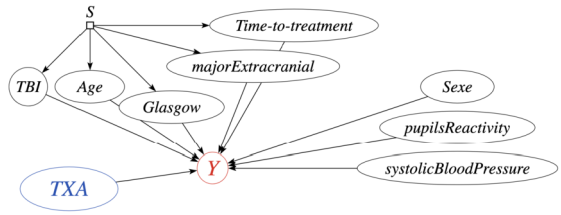

FIG. 7. *Structural causal diagram representing treatment, outcome, inclusion criteria with S and other predictors of outcome (Figure generated using the Causal Fusion software presented in Section 6 from* Bareinboim and Pearl (2016)).

RCT—all suffering from TBI with only intracranial bleeding, that is, without major extracranial bleeding. All participants were randomly administrated TXA (CRASH-3, 2019, Cap, 2019). The primary outcome studied is head-injury-related death in hospital within 28 days of injury in patients included and randomized within 3 hours of injury. The study concludes that the risk of head-injury-related death is 18.5% in the TXA group versus 19.8% in the placebo group. The causal effect, measured as a Risk Ratio (RR) was not significant (RR $= 0.94$ [95% CI 0.86 -1.02])). Note that CRASH-3 revealed a positive effect of TXA only when considering mild and moderate cases. In Section H.4.3 of the Supplementary Material of Colnet et al. (2024), we provide a complementary analysis to study this subgroup.

7.2.2 *RCT selection*. Six covariates are present at baseline, being age, sex, time since injury, systolic blood pressure, Glasgow Coma Scale score (GCS)[10] and pupil reaction. The inclusion criteria of the trial are patients with a GCS score of 12 or lower or any intracranial bleeding on CT scan (computed tomography), and no major extra cranial bleeding. We provide a DAG summarizing the trial selection and predictors of the outcome present in CRASH-3 in Figure 7.

7.3 Transporting the ATE on the Observational Data

With the two separate analyses in mind, we can now turn to the combined analysis, more specifically, the generalization from the RCT results to the target population defined by the observational Traumabase registry. Before any analysis aiming to compare and combine two data sets, an important step is to assess that baseline covariates, treatment and outcome are the same (for details, see Section H.2 of the Supplementary Material of Colnet et al., 2024).

7.3.1 *Descriptive analyses*.

[9]More precisely, to the resuscitation room of a hospital equipped to treat major trauma patients.

[10]The Glasgow Coma Scale (GCS) is a neurological scale, which aims to assess a person's consciousness. The lower the score, the higher the severity of the trauma.

TABLE 4

ATE estimations from the Traumabase for TBI-related 28-day mortality. Red cells conclude on deteriorating effect, white cells conclude on no effect. GLM stands for Generalized Linear Models and GRF for Generalized Random Forests. Additional results can be found in Table S1 in Section H of the Supplementary Material of Colnet et al. (2024)

	Multiple imputation (MICE)				GRF		Unadjusted ATE $\times 10^2$
	IPW (95% CI) $\times 10^2$		AIPW (95% CI) $\times 10^2$		IPW (95% CI) $\times 10^2$	AIPW (95% CI) $\times 10^2$	
	GLM	GRF	GLM	GRF			
Total ($n = 8248$)	15 (6.8, 23)	11 (6.0, 16)	3.4 (-9.0, 16)	-0.1 (-4.7, 4.4)	9.3 (4.0, 15)	-0.4 (-5.2, 4.4)	16

7.3.1.1 *Missing values*. The RCT contains almost no missing values, whereas the variables for determining eligibility in the observational data contain important fractions of missing values, ranging from 0.27 to 29%. Thus, the methods discussed in this review must be adapted to account for missing values.[11] In order to estimate the nuisance components, that is, the conditional odds and the outcome model(s), despite the missing data, we explore two alternative strategies: (1) logistic regression with incomplete covariates using an expectation maximization algorithm (Dempster, Laird and Rubin, 1977), a computationally efficient variant of this method using stochastic approximation is implemented in the R package `misaem` (Jiang et al., 2020); (2) generalized regression forest with missing incorporated in attributes (Twala, Jones and Hand, 2008, Josse et al., 2019), this method is implemented in the R package `grf` (Tibshirani, Athey and Wager, 2020).

7.3.1.2 *Distribution shift*. Simple comparisons of the means of the covariates between the treatment groups of the two studies—Figure 8—reveal the fundamental difference between the two studies, namely the treatment assignment bias in the observational study and the balanced treatment groups in the RCT. In Section H.3.1 of the Supplementary Material of Colnet et al. (2024), we further explore the distribution shift with univariate histograms (Figures S12–S16).

TABLE 5

Sample sizes for both studies

Traumabase			CRASH-3		
m	#treated	#death	n	#treated	#death
8248	683	1411	9168	4632	1745

[11] If we assumed the missing values being missing completely at random (MCAR), we could "throw away" the incomplete observations and perform the analyses on the complete observations, but this would reduce the total sample size to 917 observations. And as explained in Section 7.1, the MCAR assumption is not plausible for the present observational data, thus such a *complete case analysis* would be biased.

7.3.2 *Analyses*.

7.3.2.1 *Notation and estimator details*. We use two consistent ATE estimators from the CRASH-3 data, namely the difference in mean estimator (`Difference in means`; Section A of the Supplementary Material of Colnet et al., 2024) and the difference in conditional mean relying on OLS (`Difference in conditional means`). We also present the results from the purely observational study outlined earlier: AIPW coupled with multiple imputation (`MI AIPW`) and AIPW based on nuisance parameters estimated via generalized random forest (`GRF AIPW`) that can directly handle missing values when needed with missing incorporated in attribute strategy.

To generalize the ATE to the target population, we apply the estimators discussed in this review while implementing strategies to handle the missing values. The resulting estimators are presented in Table 6.

The confidence intervals of these estimators are computed with a stratified bootstrap in the RCT and the observational data set in order to maintain the ratio of relative size of the two studies (with 100 bootstrap samples). Note that the Calibration Weighting estimators (CW and

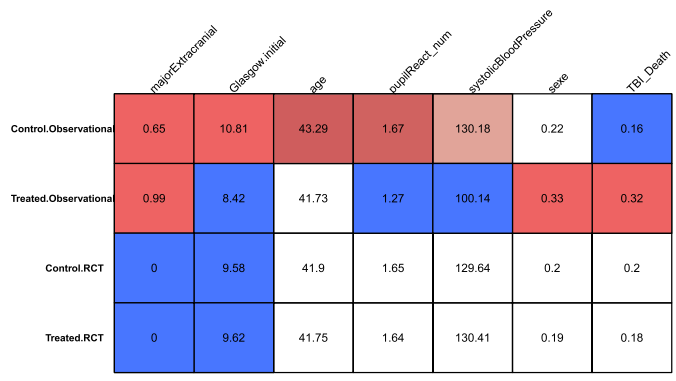

FIG. 8. *Distributional shift and difference in terms of univariate means of the trial inclusion criteria* (*red*: *group mean greater than overall mean*, *blue*: *group mean less than overall mean*, *white*: *no significant difference with overall mean*, *numeric values*: *group mean* (*resp.*, *proportion for binary variables*). *Graph obtained with the* `catdes` *function of the* `FactoMineR` *package* (Lê, Josse and Husson, 2008).

TABLE 6
Overview of generalization estimators based on different missing values handling strategies used in the data analysis

		Missing values strategy	
		Logistic regression with missing values	Generalized random forests (grf)—MIA
$\hat{\tau}_{n,m}$	IPSW	`EM IPSW`	`GRF IPSW`
	Plug-in g-formula	`EM Plug-in g-formula`	`GRF Plug-in g-formula`
	AIPSW	`EM AIPSW`	`GRF AIPSW`

ACW) are not used in this analysis as they would require a specific adaptation to the case of the missing values.

7.3.2.2 *Results of the combined analysis*. Figure 9 gives the generalization from the RCT to the target population using all the observations from both data sets, showing certain discrepancies with respect to the separate analysis results. On the one hand, one-half of the generalization estimators support the CRASH-3 conclusion about the treatment effect: no significant effect. On the other hand, some estimators point toward a deleterious treatment effect. Recall that the AIPW ATE estimations from the purely observational data study do not reject the null hypothesis of no treatment effect. Note that these results are to be interpreted carefully due to the potential impact of missing values on the performance of the chosen estimators. For example, the large confidence intervals for the `GRF` estimators when used to estimate weights are likely to be due to the imbalanced proportions of missing values in the RCT and the observational data. Indeed, the variance is much smaller using the plug-in g-formula with GRF. Dealing with missing values when generalizing a treatment effect remains an open research question.

Here, we present the results transported onto the total TBI Traumabase population, but the CRASH-3 study highlights a specific subgroup of patients (mild and moderate patients) for which a positive effect of the tranexamic acid is measured. The generalization of the CRASH-3 findings onto this subgroup in the Traumabase raises multiple methodological issues that still need to be addressed in future works (detailed in Section H.4.3 of the Supplementary Material of Colnet et al., 2024).

Overall this data analysis highlights the interest of combining two different data sets, but also some challenges: the need for a good understanding of the com-

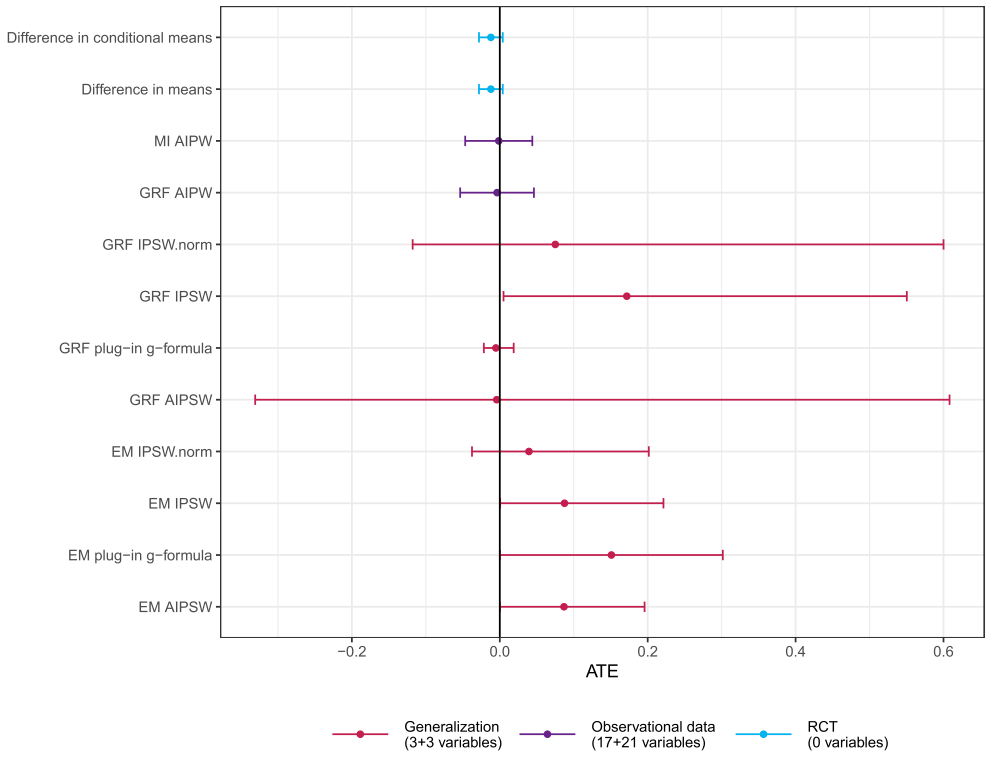

FIG. 9. *Juxtaposition of different estimation results with ATE estimators computed on the Traumabase (observational data set), on the CRASH-3 trial (RCT) and transported from CRASH-3 to the Traumabase target population. All the observations are used. Number of variables used in each context is given in the legend.*

mon covariates, exposure and outcome of interest before combining the data sets, different missing data patterns and poor overlap when considering specific target (sub)populations.

8. CONCLUSION

Combining observational data and RCTs can improve many aspects of causal inference, from increased statistical power to better external validity. A large part of this review is dedicated to generalizability and transportability of RCT from one population to another. The corresponding rich and prolific literature answers a real practical concern: external validity. Indeed, questions about external validity arise as soon as there are treatment effect heterogeneities in the populations under study. We find that, as any growing scientific field, the ideas are in flux: notation differ, implementations are scattered and the proposed methods still lack real-world benchmarks, generated hand-in-hand with practitioners. In addition, many open questions still remain as detailed below.

8.1 Discrepancies Between RCTs and Observational Data

The application on tranexamic acid effect hinted to moderate external validity of the RCT as the generalized ATE is concordant with the findings from the RCT, at least for half of the estimators. Additionally, the purely observational data study also supports the results from the RCT. Determining which analysis to trust depends on the assumptions we are willing to make—either related to transportability or unconfoundedness—as well as the suitability of the selected variables. Beyond these assumptions, caution is needed when interpreting the results, as observing the methods in action reveals threats to their validity. The target population of interest and overlap also raise concerns. Considering certain strata revealed violated positivity, which leads to a nontransportable treatment effect on the strata of interest: mild and moderate patients. Therefore, further discussions and analyses with the medical expert committee are necessary to redefine a target population of interest on which generalization is possible and medically relevant. As it is generally the case, beyond methodological and theoretical guarantees, a major step to be taken before applying a set of methods is to clearly state the causal question and estimand(s) and the associated identifiability requirements. This task is even more complex when combining data sets. A primary and fundamental concern is whether outcome, treatment and covariates are comparable in the two studies (Lodi et al., 2019).

8.2 Right Choice of Covariates to Answer the Question

Domain expertise can be used to postulate a causal graph: a directed acyclic graph representing the mechanisms (as Figure 7). The SCM framework is then convenient to assess whether the question of interest can be formulated in an identifiable way. This approach offers a principled way of selecting variables needed for identification of the causal effect and to avoid biased causal effect estimates. Without such an approach, identifiability claims are limited. A common practical recommendation is to include as many variables as possible to avoid violation of any assumption as proposed, for example, by Stuart and Rhodes (2017), Ling et al. (2022) and Dahabreh and Hernán (2019): "it is probably best to include as many outcome predictors as possible in regression models for the expectation of the outcome or the probability of trial participation." On the contrary, a recent work alerts about the bad consequences of adding covariates that are shifted between the two populations while not being treatment effect modifiers, resulting in variance inflation (Colnet et al., 2022b). In its current state, the field probably lacks work on covariate selection and its impact on bias and variance. Some recent works propose the use of causal graphs to select optimal adjustment sets that allow the reduction of the variance of the final estimation (Rotnitzky and Smucler, 2020, Witte et al., 2020, Guo and Perković, 2022), but such methods have not yet been developed for generalization or data fusion.

8.3 Challenges in Handling Missing Values

In our data analysis, we have seen the need to account for missing values, and in particular different missing value patterns between data sources. Missing values typically occur more often in observational data since in RCTs, investigators typically deploy significant efforts to avoid them. RCTs may however suffer from participants missing scheduled visits or completely dropping out from the study. The literature for RCT mainly focuses on missing outcome data and calls for sensitivity analysis given that available strategies to handle such missing data (weighting, multiple imputation) rely on untestable assumptions about the missing values mechanism (Carpenter and Kenward, 2007, National Research Council, 2012, Kenward, 2013, O'Kelly and Ratitch, 2014, Li and Stuart, 2019, Cro et al., 2020). Missing values may lead to subtle biases in the inferences, as they are seldom uniformly distributed across both data sets—occurring more in one than in the other. While a recent research work proposes an assessment of the effect of different missing data patterns (Mayer, Josse and Traumabase Group, 2023), further research is needed to clarify identifiability conditions and estimators in this setting in order to better understand the scope of each method.

ACKNOWLEDGMENTS

This work was initiated by a SAMSI working group jointly led by JJ and SY in the 2020 causal inference program. We would like to acknowledge the helpful discussions during the SAMSI working group meetings. We also

would like to acknowledge the discussions and insights from the Traumabase group and physicians, in particular, Drs. François-Xavier AGERON, Tobias GAUSS and Jean-Denis MOYER. In addition, none of the data analysis part could have been done without the help of Dr. Ian ROBERTS and the CRASH-3 group, who shared with us the clinical trial data. Part of this work was performed while JJ was a visiting researcher at Google Brain Paris. Finally, we would like to warmly thank Issa DAHABREH for his comments, suggestions of additional references and insightful discussions.

FUNDING

SY is partially supported by NSF SES 2242776, NIH 1R01AG066883 and FDA 1U01FD007934.

SUPPLEMENTARY MATERIAL

Additional Information (DOI: 10.1214/23-STS889SUPP; .pdf). The supplementary material contains details on treatment effect estimation performed separately on RCT data (Section A) and on observational data (Section B), derivations of the different identification formula for the generalization problem (Section C), a review of formal results for estimators discussed in the review (Section D), estimators in the case of nested study designs (Section E), additional information on the structural causal model framework (Section F), additional simulations results (Section G), as well as additional information on the clinical data from the application (Section H).

REFERENCES

ACKERMAN, B., LESKO, C. R., SIDDIQUE, J., SUSUKIDA, R. and STUART, E. A. (2021). Generalizing randomized trial findings to a target population using complex survey population data. *Stat. Med.* **40** 1101–1120. MR4384364 https://doi.org/10.1002/sim.8822

ANDREWS, I. and OSTER, E. (2019). A simple approximation for evaluating external validity bias. *Econom. Lett.* **178** 58–62. MR3921042 https://doi.org/10.1016/j.econlet.2019.02.020

ANGRIST, J. D., IMBENS, G. W. and RUBIN, D. B. (1996). Identification of causal effects using instrumental variables. *J. Amer. Statist. Assoc.* **91** 444–455.

ATHEY, S., CHETTY, R. and IMBENS, G. (2020). Combining experimental and observational data to estimate treatment effects on long term outcomes. Preprint. Available at arXiv:2006.09676.

ATHEY, S., CHETTY, R., IMBENS, G. and KANG, H. (2020). Estimating treatment effects using multiple surrogates: The role of the surrogate score and the surrogate index.

ATHEY, S., TIBSHIRANI, J. and WAGER, S. (2019). Generalized random forests. *Ann. Statist.* **47** 1148–1178. MR3909963 https://doi.org/10.1214/18-AOS1709

BACH, P., CHERNOZHUKOV, V., KURZ, M. S. and SPINDLER, M. (2021). DoubleML—An object-oriented implementation of double machine learning in R. Available at arXiv:2103.09603 [stat.ML].

BACH, P., CHERNOZHUKOV, V., KURZ, M. S. and SPINDLER, M. (2022). DoubleML—An object-oriented implementation of double machine learning in Python. *J. Mach. Learn. Res.* **23** 1–6.

BAREINBOIM, E., LEE, S., HONAVAR, V. and PEARL, J. (2013). Transportability from multiple environments with limited experiments.

BAREINBOIM, E. and PEARL, J. (2012a). Controlling selection bias in causal inference. In *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics* (N. D. Lawrence and M. Girolami, eds.). *Proceedings of Machine Learning Research* **22** 100–108. La Palma, Canary Islands. PMLR.

BAREINBOIM, E. and PEARL, J. (2012b). Transportability of causal effects: Completeness results. In *Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence*, *AAAI'*12 698–704. AAAI Press, Menlo Park.

BAREINBOIM, E. and PEARL, J. (2013). A general algorithm for deciding transportability of experimental results. *J. Causal Inference* **1** 107–133. MR4289403 https://doi.org/10.1515/jci-2012-0004

BAREINBOIM, E. and PEARL, J. (2016). Causal inference and the data-fusion problem. *Proc. Natl. Acad. Sci. USA* **113** 7345–7352.

BAREINBOIM, E., TIAN, J. and PEARL, J. (2014). Recovering from selection bias in causal and statistical inference. In *Proceedings of the AAAI Conference on Artificial Intelligence* **28**.

BEGG, C. B. and LEUNG, D. H. Y. (2000). On the use of surrogate end points in randomized trials. *J. Roy. Statist. Soc. Ser. A* **163** 15–28.

BUCHANAN, A. L., HUDGENS, M. G., COLE, S. R., MOLLAN, K. R., SAX, P. E., DAAR, E. S., ADIMORA, A. A., ERON, J. J. and MUGAVERO, M. J. (2018). Generalizing evidence from randomized trials using inverse probability of sampling weights. *J. Roy. Statist. Soc. Ser. A* **181** 1193–1209. MR3876388 https://doi.org/10.1111/rssa.12357

CAMPBELL, D. T. (1957). Factors relevant to the validity of experiments in social settings. *Psychol. Bull.* **54** 297–312.

CAP, A. P. (2019). Crash-3: A win for patients with traumatic brain injury. *Lancet* **394** 1687–1688.

CARPENTER, J. R. and KENWARD, M. G. (2007). Missing data in randomised controlled trials: a practical guide.

CHEN, R., CHEN, G. and YU, M. (2023). A generalizability score for aggregate causal effect. *Biostatistics* **24** 309–326. MR4597229 https://doi.org/10.1093/biostatistics/kxab029

CHERNOZHUKOV, V., CHETVERIKOV, D., DEMIRER, M., DUFLO, E., HANSEN, C., NEWEY, W. and ROBINS, J. (2018). Double/debiased machine learning for treatment and structural parameters. *Econom. J.* **21** C1–C68. MR3769544 https://doi.org/10.1111/ectj.12097

CHU, J., LU, W. and YANG, S. (2023). Targeted optimal treatment regime learning using summary statistics. *Biometrika*. https://doi.org/10.1093/biomet/asad020

CINELLI, C. and PEARL, J. (2021). Generalizing experimental results by leveraging knowledge of mechanisms. *Eur. J. Epidemiol.* **36** 149–164. https://doi.org/10.1007/s10654-020-00687-4

COCHRAN, W. G. (1968). The effectiveness of adjustment by subclassification in removing bias in observational studies. *Biometrics* **24** 295–313. MR0228136 https://doi.org/10.2307/2528036

COLE, S. R. and STUART, E. A. (2010). Generalizing evidence from randomized clinical trials to target populations: The ACTG 320 trial. *Amer. J. Epidemiol.* **172** 107–115. https://doi.org/10.1093/aje/kwq084

COLNET, B., JOSSE, J., VAROQUAUX, G. and SCORNET, E. (2022a). Causal effect on a target population: A sensitivity analysis to handle missing covariates. *J. Causal Inference* **10** 372–414. MR4512969 https://doi.org/10.1515/jci-2021-0059

COLNET, B., JOSSE, J., VAROQUAUX, G. and SCORNET, E. (2022b). Reweighting the RCT for generalization: finite sample analysis and variable selection.

COLNET, B., MAYER, I., CHEN, G., DIENG, A., LI, R., VAROQUAUX, G., VERT, J.-P, JOSSE, J. and YANG, S. (2024). Supplement to "Causal inference methods for combining randomized trials and observational studies: A review." https://doi.org/10.1214/23-STS889SUPP

CONCATO, J., SHAH, N. and HORWITZ, R. I. (2000). Randomized, controlled trials, observational studies, and the hierarchy of research designs. *N. Engl. J. Med.* **342** 1887–1892. https://doi.org/10.1056/NEJM200006223422507

COOPER, G. (1995). Causal discovery from data in the presence of selection bias. In *Proceedings of the Fifth International Workshop on Artificial Intelligence and Statistics* 140–150.

CORNFIELD, J., HAENSZEL, W., HAMMOND, E. C., LILIENFELD, A. M., SHIMKIN, M. B. and WYNDER, E. L. (1959). Smoking and lung cancer: Recent evidence and a discussion of some questions. *J. Natl. Cancer Inst.* **22** 173–203.

CORREA, J. D., TIAN, J. and BAREINBOIM, E. (2018). Generalized adjustment under confounding and selection biases. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

NATIONAL RESEARCH COUNCIL (2012). The prevention and treatment of missing data in clinical trials. *N. Engl. J. Med.* **367** 1355–1360.

CRASH-3 (2019). Effects of tranexamic acid on death, disability, vascular occlusive events and other morbidities in patients with acute traumatic brain injury (CRASH-3): A randomised, placebo-controlled trial. *Lancet* **394** 1713–1723.

CRO, S., MORRIS, T. P., KAHAN, B. C., CORNELIUS, V. R. and CARPENTER, J. R. (2020). A four-step strategy for handling missing outcome data in randomised trials affected by a pandemic. *BMC Med. Res. Methodol.* **20** 208. https://doi.org/10.1186/s12874-020-01089-6

DAGAN, N., BARDA, N., KEPTEN, E., MIRON, O., PERCHIK, S., KATZ, M. A., HERNÁN, M. A., LIPSITCH, M., REIS, B. et al. (2021). Bnt162b2 mrna Covid-19 vaccine in a nationwide mass vaccination setting. *N. Engl. J. Med.* **384** 1412–1423.

DAHABREH, I. J. and HERNÁN, M. A. (2019). Extending inferences from a randomized trial to a target population. *Eur. J. Epidemiol.* **34** 719–722.

DAHABREH, I. J., ROBERTSON, S. E., STEINGRIMSSON, J. A., STUART, E. A. and HERNÁN, M. A. (2020). Extending inferences from a randomized trial to a new target population. *Stat. Med.* **39** 1999–2014. MR4105273 https://doi.org/10.1002/sim.8426

DAHABREH, I. J., ROBERTSON, S. E., TCHETGEN, E. J., STUART, E. A. and HERNÁN, M. A. (2019). Generalizing causal inferences from individuals in randomized trials to all trial-eligible individuals. *Biometrics* **75** 685–694. MR3999190 https://doi.org/10.1111/biom.13009

DAHABREH, I. J., ROBINS, J. M., HANEUSE, S. J.-P. A., SAEED, I., ROBERTSON, S. E., STUART, E. A. and HERNÁN, M. A. (2023a). Sensitivity analysis using bias functions for studies extending inferences from a randomized trial to a target population. *Stat. Med.* **42** 2029–2043. MR4594699 https://doi.org/10.1002/sim.9550

DAHABREH, I. J., ROBINS, J. M., HANEUSE, S. J.-P. A., SAEED, I., ROBERTSON, S. E., STUART, E. A. and HERNÁN, M. A. (2023b). Sensitivity analysis using bias functions for studies extending inferences from a randomized trial to a target population. *Stat. Med.* **42** 2029–2043. MR4594699 https://doi.org/10.1002/sim.9550

DAHABREH, I. J., ROBINS, J. M. and HERNÁN, M. A. (2020). Benchmarking observational methods by comparing randomized trials and their emulations. *Epidemiology* **31** 614–619.

DAWID, P., HUMPHREYS, M. and MUSIO, M. (2019). Bounding causes of effects with mediators Technical Report. arXiv:1907.00399.

DEATON, A. and CARTWRIGHT, N. (2018). Understanding and misunderstanding randomized controlled trials. *Soc. Sci. Med.* **210** 2–21. https://doi.org/10.1016/j.socscimed.2017.12.005

DEATON, A., CASE, S. C., CÔTÉ, N., DRÈZE, J., EASTERLY, W., KHERA, R., PRITCHETT, L. and REDDY, C. R. (2019). *Randomization in the Tropics Revisited*: *A Theme and Eleven Variations*. *Randomized Controlled Trials in the Field of Development*: *A Critical Perspective*. Oxford Univ. Press, London. Forthcoming.

DEGTIAR, I., LAYTON, T., WALLACE, J. and ROSE, S. (2021). Conditional cross-design synthesis estimators for generalizability in medicaid.

DEGTIAR, I. and ROSE, S. (2023). A review of generalizability and transportability. *Annu. Rev. Stat. Appl.* **10** 501–524. MR4567803 https://doi.org/10.1146/annurev-statistics-042522-103837

DEMPSTER, A. P., LAIRD, N. M. and RUBIN, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc. Ser. B* **39** 1–38. MR0501537

DEWAN, Y., KOMOLAFE, E., MEJÌA-MANTILLA, J., PEREL, P., ROBERTS, I. and SHAKUR-STILL, H. (2012). CRASH-3: Tranexamic acid for the treatment of significant traumatic brain injury: Study protocol for an international randomized, double-blind, placebo-controlled trial. *Trials* **13** 87.

DIDELEZ, V., KREINER, S. and KEIDING, N. (2010). Graphical models for inference under outcome-dependent sampling. *Statist. Sci.* **25** 368–387. MR2791673 https://doi.org/10.1214/10-STS340

EGAMI, N. and HARTMAN, E. (2021). Covariate selection for generalizing experimental results: Application to a large-scale development program in Uganda. *J. Roy. Statist. Soc. Ser. A* **184** 1524–1548. MR4344647 https://doi.org/10.1111/rssa.12734

FDA (2018). Framework for fda's real-world evidence program.

FRIEDEN, T. R. (2017). Evidence for health decision making—beyond randomized, controlled trials. *N. Engl. J. Med.* **377** 465–475. https://doi.org/10.1056/NEJMra1614394

GENELETTI, S., RICHARDSON, S. and BEST, N. (2008). Adjusting for selection bias in retrospective, case–control studies. *Biostatistics* **10** 17–31.

GORDON, B. R., ZETTELMEYER, F., BHARGAVA, N. and CHAPSKY, D. (2019). A comparison of approaches to advertising measurement: Evidence from big field experiments at Facebook. *Mark. Sci.* **38** 193–225.

GREEN, L. W. and GLASGOW, R. E. (2006). Evaluating the relevance, generalization, and applicability of research: Issues in external validation and translation methodology. *Eval. Health Prof.* **29** 126–153. https://doi.org/10.1177/0163278705284445

GUO, F. R. and PERKOVIĆ, E. (2022). Efficient least squares for estimating total effects under linearity and causal sufficiency. *J. Mach. Learn. Res.* **23** Paper No. [104], 41. MR4576689

GUO, R., CHENG, L., LI, J., HAHN, P. R. and LIU, H. (2018). A survey of learning causality with data: Problems and methods. Preprint. Available at arXiv:1809.09337.

GUO, W., WANG, S., DING, P., WANG, Y. and JORDAN, M. I. (2021). Multi-source causal inference using control variates. Preprint. Available at arXiv:2103.16689.

HAINMUELLER, J. (2012). Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies. *Polit. Anal.* **20** 25–46.

HARTMAN, E., GRIEVE, R., RAMSAHAI, R. and SEKHON, J. S. (2015). From sample average treatment effect to population average treatment effect on the treated: Combining experimental with observational studies to estimate population treatment effects. *J. Roy. Statist. Soc. Ser. A* **178** 757–778. MR3348358 https://doi.org/10.1111/rssa.12094

HE, Z., TANG, X., YANG, X., GUO, Y., GEORGE, T., CHARNESS, N., HEM, K., HOGAN, W. and BIAN, J. (2020). Clinical

trial generalizability assessment in the big data era: A review. *Clin. Transl. Sci.* **13**.

HERNÁN, M. and ROBINS, J. (2006). Instruments for causal inference: An epidemiologist's dream? In *Epidemiology* **17** 360–72, Cambridge, Mass.

HERNÁN, M. A., COLE, S. R., MARGOLICK, J., COHEN, M. and ROBINS, J. M. (2005). Structural accelerated failure time models for survival analysis in studies with time-varying treatments. *Pharmacoepidemiol. Drug Saf.* **14** 477–491. https://doi.org/10.1002/pds.1064

HOBBS, B. P., SARGENT, D. J. and CARLIN, B. P. (2012). Commensurate priors for incorporating historical information in clinical trials using general and generalized linear models. *Bayesian Anal.* **7** 639–673. MR2981631 https://doi.org/10.1214/12-BA722

HOTZ, V. J., IMBENS, G. W. and MORTIMER, J. H. (2005). Predicting the efficacy of future training programs using past experiences at other locations. *J. Econometrics* **125** 241–270. MR2143377 https://doi.org/10.1016/j.jeconom.2004.04.009

HUANG, M. (2022). Sensitivity analysis in the generalization of experimental results.

HUANG, M., EGAMI, N., HARTMAN, E. and MIRATRIX, L. (2023). Leveraging population outcomes to improve the generalization of experimental results: Application to the JTPA study. *Ann. Appl. Stat.* **17** 2139–2164. MR4637661 https://doi.org/10.1214/22-aoas1712

HUANG, Y. and VALTORTA, M. (2006). Pearl's calculus of intervention is complete. In *Proceedings of the Twenty-Second Conference on Uncertainty in Artificial Intelligence*, *UAI'*06 217–224. AUAI Press, Arlington, VA, USA.

HUANG, Y. and VALTORTA, M. (2012). Pearl's calculus of intervention is complete.

HUITFELDT, A., SWANSON, S. A., STENSRUD, M. J. and SUZUKI, E. (2019). Effect heterogeneity and variable selection for standardizing causal effects to a target population. *Eur. J. Epidemiol.* **34** 1119–1129.

HÜNERMUND, P. and BAREINBOIM, E. (2019). Causal inference and data-fusion in econometrics. Preprint. Available at arXiv:1912.09104.

IMBENS, G. (2003). Sensitivity to exogeneity assumptions in program evaluation. *Amer. Econ. Rev.*

IMBENS, G. W. (2014). Instrumental variables: An econometrician's perspective. *Statist. Sci.* **29** 323–358. MR3264545 https://doi.org/10.1214/14-STS480

IMBENS, G. W. and RUBIN, D. B. (2015). *Causal Inference—for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge Univ. Press, New York. MR3309951 https://doi.org/10.1017/CBO9781139025751

JEONG, S. and NAMKOONG, H. (2022). Assessing external validity over worst-case subpopulations. arXiv:2007.02411 [cs, econ, stat].

JIANG, W., JOSSE, J., LAVIELLE, M. and TRAUMABASE GROUP (2020). Logistic regression with missing covariates—parameter estimation, model selection and prediction within a joint-modeling framework. *Comput. Statist. Data Anal.* **145** 106907, 20. MR4053706 https://doi.org/10.1016/j.csda.2019.106907

JOSEY, K. P., BERKOWITZ, S. A., GHOSH, D. and RAGHAVAN, S. (2021). Transporting experimental results with entropy balancing. *Stat. Med.* **40** 4310–4326. MR4300088 https://doi.org/10.1002/sim.9031

JOSSE, J., PROST, N., SCORNET, E. and VAROQUAUX, G. (2019). On the consistency of supervised learning with missing values. Preprint. Available at arXiv:1902.06931.

JUNG, Y., TIAN, J. and BAREINBOIM, E. (2020a). Estimating causal effects using weighting-based estimators. In *Proceedings of the AAAI Conference on Artificial Intelligence* **34** 10186–10193.

JUNG, Y., TIAN, J. and BAREINBOIM, E. (2020b). Learning causal effects via weighted empirical risk minimization. In *Advances in Neural Information Processing Systems* (H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan and H. Lin, eds.) **33** 12697–12709. Curran Associates, Red Hook.

JUNG, Y., TIAN, J. and BAREINBOIM, E. (2021). Estimating identifiable causal effects through double machine learning. *Proc. AAAI Conf. Artif. Intell.* **35** 12113–12122.

KALLUS, N., PULI, A. M. and SHALIT, U. (2018). Removing hidden confounding by experimental grounding. In *Advances in Neural Information Processing Systems* 10888–10897.

KARVANEN, J., TIKKA, S. and HYTTINEN, A. (2020). Do-search—a tool for causal inference and study design with multiple data sources.

KEIDING, N. and LOUIS, T. A. (2016). Perils and potentials of self-selected entry to epidemiological studies and surveys. *J. Roy. Statist. Soc. Ser. A* **179** 319–376. MR3461587 https://doi.org/10.1111/rssa.12136

KENNEDY, L. and GELMAN, A. (2021). Know your population and know your model: Using model-based regression and poststratification to generalize findings beyond the observed sample. *Psychol. Methods* **26** 547.

KENWARD, M. (2013). The handling of missing data in clinical trials. *Clin. Invest.* **3** 241–250.

KERN, H. L., STUART, E. A., HILL, J. and GREEN, D. P. (2016). Assessing methods for generalizing experimental impact estimates to target populations. *J. Res. Educ. Eff.* **9** 103–127.

KRUEGER, A. B. (1999). Experimental estimates of education production functions. *Q. J. Econ.* **114** 497–532.

LAURITZEN, S. L. and RICHARDSON, T. S. (2008). Discussion of mccullagh: Sampling bias and logistic models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **70** 671.

LÊ, S., JOSSE, J. and HUSSON, F. (2008). FactoMineR: A package for multivariate analysis. *J. Stat. Softw.* **25** 1–18.

LEE, D., GHOSH, S. and YANG, S. (2022). Transporting survival of an HIV clinical trial to the external target populations. Preprint. Available at arXiv:2210.02571.

LEE, D., YANG, S., DONG, L., WANG, X., ZENG, D. and CAI, J. (2023). Improving trial generalizability using observational studies. *Biometrics* **79** 1213–1225. MR4606348 https://doi.org/10.1111/biom.13609

LEE, D., YANG, S. and WANG, X. (2022). Doubly robust estimators for generalizing treatment effects on survival outcomes from randomized controlled trials to a target population. *J. Causal Inference* **10** 415–440. MR4519820 https://doi.org/10.1515/jci-2022-0004

LEE, S., CORREA, J. and BAREINBOIM, E. (2020a). General transportability—synthesizing observations and experiments from heterogeneous domains. *Proc. AAAI Conf. Artif. Intell.* **34** 10210–10217.

LEE, S., CORREA, J. D. and BAREINBOIM, E. (2020b). General identifiability with arbitrary surrogate experiments. In *Proceedings of Machine Learning Research* (R. P. Adams and V. Gogate, eds.) **115** 389–398. PMLR.

LESKO, C. R., BUCHANAN, A. L., WESTREICH, D., EDWARDS, J. K., HUDGENS, M. G. and COLE, S. R. (2017). Generalizing study results: A potential outcomes perspective. *Epidemiology* **28** 553–561. https://doi.org/10.1097/EDE.0000000000000664

LESKO, C. R., COLE, S. R., HALL, H. I., WESTREICH, D., MILLER, W. C., ERON, J. J., LI, J., MUGAVERO, M. J. and FOR THE CNICS INVESTIGATORS (2016). The effect of antiretroviral therapy on all-cause mortality, generalized to persons diagnosed with HIV in the USA, 2009–2011. *Int. J. Epidemiol.* **45** 140–150.

LI, F., BUCHANAN, A. L. and COLE, S. R. (2021). Generalizing trial evidence to target populations in non-nested designs: Applications to aids clinical trials.

LI, F., HONG, H. and STUART, E. A. (2023). A note on semiparametric efficient generalization of causal effects from randomized trials to target populations. *Comm. Statist. Theory Methods* **52** 5767–5798. MR4608916 https://doi.org/10.1080/03610926.2021.2020291

LI, P. and STUART, E. A. (2019). Best (but oft-forgotten) practices: Missing data methods in randomized controlled nutrition trials. *Am. J. Clin. Nutr.* **109** 504–508.

HERNÁN, M. A. and VAN DER WEELE, T. J. (2011). Compound treatments and transportability of causal inference. *Epidemiology* **22** 368–377.

LING, A. Y., MONTEZ-RATH, M. E., CARITA, P., CHANDROSS, K., LUCATS, L., MENG, Z., SEBASTIEN, B., KAPPHAHN, K. and DESAI, M. (2022). A critical review of methods for real-world applications to generalize or transport clinical trial findings to target populations of interest.

LIPSITCH, M., TCHETGEN, E. J. T. and COHEN, T. (2010). Negative controls: A tool for detecting confounding and bias in observational studies. *Epidemiology* **21** 383–388.

LODI, S., PHILLIPS, A., LUNDGREN, J., LOGAN, R., SHARMA, S., COLE, S., BABIKER, A., LAW, M., CHU, H. et al. (2019). Effect estimates in randomized trials and observational studies: Comparing apples with apples. *Amer. J. Epidemiol.* **188**.

MARTEL GARCIA, F. and WANTCHEKON, L. (2010). Theory, external validity, and experimental inference: Some conjectures. *Ann. Am. Acad. Polit. Soc. Sci.* **628** 132–147.

MAYER, I., JOSSE, J. and TRAUMABASE GROUP (2023). Generalizing treatment effects with incomplete covariates: Identifying assumptions and multiple imputation algorithms. *Biom. J.* **65** Paper No. 2100294, 30. MR4603527 https://doi.org/10.1002/bimj.202100294

MAYER, I., SVERDRUP, E., GAUSS, T., MOYER, J.-D., WAGER, S. and JOSSE, J. (2020). Doubly robust treatment effect estimation with missing attributes. *Ann. Appl. Stat.* **14** 1409–1431. MR4152139 https://doi.org/10.1214/20-AOAS1356

MAYER, I., ZHAO, P., GREIFER, N., HUNTINGTON-KLEIN, N. and JOSSE, J. (2022). Cran task view: Causal inference.

NEYMAN, J. (1923). Sur les applications de la thar des probabilities aux experiences Agaricales: Essay de principle. English translation of excerpts by Dabrowska, D. and Speed, T.. *Statist. Sci.* **5** 465–472.

NGUYEN, T., ACKERMAN, B., SCHMID, I., COLE, S. and STUART, E. (2018). Sensitivity analyses for effect modifiers not observed in the target population when generalizing treatment effects from a randomized controlled trial: Assumptions, models, effect scales, data scenarios, and implementation details. *PLoS ONE* **13** e0208795.

NGUYEN, T. Q., EBNESAJJAD, C., COLE, S. R. and STUART, E. A. (2017). Sensitivity analysis for an unobserved moderator in RCT-to-target-population generalization of treatment effects. *Ann. Appl. Stat.* **11** 225–247. MR3634322 https://doi.org/10.1214/16-AOAS1001

NIE, X., IMBENS, G. and WAGER, S. (2021). Covariate balancing sensitivity analysis for extrapolating randomized trials across locations.

O'KELLY, M. and RATITCH, B. (2014). *Clinical Trials with Missing Data*: *A Guide for Practitioners*. Wiley, New York.

O'MUIRCHEARTAIGH, C. and HEDGES, L. V. (2014). Generalizing from unrepresentative experiments: A stratified propensity score approach. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **63** 195–210. MR3234340 https://doi.org/10.1111/rssc.12037

PEARL, J. (1993). [Bayesian analysis in expert systems]: Comment: Graphical models, causality and intervention. *Statist. Sci.* **8** 266–269.

PEARL, J. (1995). Causal diagrams for empirical research. *Biometrika* **82** 669–710. MR1380809 https://doi.org/10.1093/biomet/82.4.669

PEARL, J. (2000). *Causality*: *Models*, *Reasoning*, *and Inference*. Cambridge Univ. Press, Cambridge. MR1744773

PEARL, J. (2009a). Causal inference in statistics: An overview. *Stat. Surv.* **3** 96–146. MR2545291 https://doi.org/10.1214/09-SS057

PEARL, J. (2009b). *Causality*: *Models*, *Reasoning*, *and Inference*, 2nd ed. Cambridge Univ. Press, Cambridge. MR2548166 https://doi.org/10.1017/CBO9780511803161

PEARL, J. (2015). Generalizing experimental findings. *J. Causal Inference* **3** 259–266. MR4289440 https://doi.org/10.1515/jci-2015-0025

PEARL, J. and BAREINBOIM, E. (2011). Transportability of causal and statistical relations: A formal approach. In 2011 *IEEE* 11*th International Conference on Data Mining Workshops* (*ICDMW*) 540–547. IEEE.

PEARL, J. and BAREINBOIM, E. (2014). External validity: From do-calculus to transportability across populations. *Statist. Sci.* **29** 579–595. MR3300360 https://doi.org/10.1214/14-STS486

POCOCK, S. J. (1976). The combination of randomized and historical controls in clinical trials. *J. Chronic Dis.* **29** 175–188.

PRENTICE, R. (1989). Surrogate endpoints in clinical trials: Definition and operational criteria. *Stat. Med.* **8** 431–440.

RICHARDSON, T. S. and ROBINS, J. M. (2013). Single world intervention graphs (swigs): A unification of the counterfactual and graphical approaches to causality. *Center for the Statistics and the Social Sciences, University of Washington Series. Working Paper*, 128:2013.

ROBINS, J. (1986). A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. *Math. Model.* **7** 1393–1512.

ROBINS, J. M. (2000). Marginal structural models versus structural nested models as tools for causal inference. In *Statistical Models in Epidemiology*, *the Environment*, *and Clinical Trials* (*Minneapolis*, *MN*, 1997). *IMA Vol. Math. Appl.* **116** 95–133. Springer, New York. MR1731682 https://doi.org/10.1007/978-1-4612-1284-3_2

ROSENBAUM, P. R. (2002). Sensitivity to hidden bias. In *Observational Studies* 105–170. Springer, Berlin.

ROSENBAUM, P. R. and RUBIN, D. B. (1983). Assessing the sensitivity to an unobserved binary covariate in an observational study with binary outcome. *J. Roy. Statist. Soc. Ser. B* **45** 212–218.

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

ROTNITZKY, A. and SMUCLER, E. (2020). Efficient adjustment sets for population average causal treatment effect estimation in graphical models. *J. Mach. Learn. Res.* **21** Paper No. 188, 86. MR4209474

RUBIN, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *J. Educ. Psychol.* **66** 688–701.

RUBIN, D. B. (1976). Inference and missing data. *Biometrika* **63** 581–592. MR0455196 https://doi.org/10.1093/biomet/63.3.581

RUDOLPH, K. E., SCHMIDT, N. M., GLYMOUR, M. M., CROWDER, R., GALIN, J., AHERN, J. and OSYPUK, T. L. (2018). Composition or context: Using transportability to understand drivers of site differences in a large-scale housing experiment. *Epidemiology* **29** 199–206.

RUDOLPH, K. E. and VAN DER LAAN, M. J. (2017). Robust estimation of encouragement design intervention effects transported across sites. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **79** 1509–1525. MR3731673 https://doi.org/10.1111/rssb.12213

SCHMIDLI, H., GSTEIGER, S., ROYCHOUDHURY, S., O'HAGAN, A., SPIEGELHALTER, D. and NEUENSCHWANDER, B. (2014). Robust meta-analytic-predictive priors in clinical trials with historical control information. *Biometrics* **70** 1023–1032. MR3295763 https://doi.org/10.1111/biom.12242

SHAKUR-STILL, H., ROBERTS, I., BAUTISTA, R., CABALLERO, J., COATS, T., DEWAN, Y., EL-SAYED, H., TAMAR, G., GUPTA, S. et al. (2009). Effects of tranexamic acid on death, vascular occlusive events, and blood transfusion in trauma patients with significant haemorrhage (CRASH-2): A randomised, placebo-controlled trial. *Lancet* **376** 23–32.

SHPITSER, I. and PEARL, J. (2006). Identification of joint interventional distributions in recursive semi-Markovian causal models. In *Proceedings of the* 21*st National Conference on Artificial Intelligence—Volume* 2, *AAAI'*06 1219–1226. AAAI Press, Menlo Park.

STUART, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statist. Sci.* **25** 1–21. MR2741812 https://doi.org/10.1214/09-STS313

STUART, E. A., ACKERMAN, B. and WESTREICH, D. (2018). Generalizability of randomized trial results to target populations: Design and analysis possibilities. *Res. Soc. Work Pract.* **28** 532–537.

STUART, E. A., BRADSHAW, C. P. and LEAF, P. J. (2015). Assessing the generalizability of randomized trial results to target populations. *Prev. Sci.* **16** 475–485. https://doi.org/10.1007/s11121-014-0513-z

STUART, E. A., COLE, S. R., BRADSHAW, C. P. and LEAF, P. J. (2011). The use of propensity scores to assess the generalizability of results from randomized trials. *J. Roy. Statist. Soc. Ser. A* **174** 369–386. MR2898850 https://doi.org/10.1111/j.1467-985X.2010.00673.x

STUART, E. A. and RHODES, A. (2017). Generalizing treatment effect estimates from sample to population: A case study in the difficulties of finding sufficient data. *Eval. Rev.* **41** 357–388. PMID: 27491758.

SUGIYAMA, M. and KAWANABE, M. (2012). *Machine Learning in Non-stationary Environments*: *Introduction to Covariate Shift Adaptation*. MIT Press, Cambridge.

TAN, Z. (2006). A distributional approach for causal inference using propensity scores. *J. Amer. Statist. Assoc.* **101** 1619–1637. MR2279484 https://doi.org/10.1198/016214506000000023

R CORE TEAM (2021). *R*: *A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

TIAN, J. and PEARL, J. (2000). Probabilities of causation: Bounds and identification. *Ann. Math. Artif. Intell.* **28** 287–313. MR1797625 https://doi.org/10.1023/A:1018912507879

TIBSHIRANI, J., ATHEY, S. and WAGER, S. (2020). *grf: Generalized random forests*. R package version 1.2.0.

TIKKA, S., HYTTINEN, A. and KARVANEN, J. (2019). Causal effect identification from multiple incomplete data sources: A general search-based approach. Preprint. Available at arXiv:1902.01073.

TIKKA, S. and KARVANEN, J. (2017). Identifying causal effects with the R package causaleffect. *J. Stat. Softw.* **76** 1–30.

TIPTON, E. (2013). Improving generalizations from experiments using propensity score subclassification: Assumptions, properties, and contexts. *J. Educ. Behav. Stat.* **38** 239–266.

TIPTON, E., HALLBERG, K., HEDGES, L. V. and CHAN, W. (2017). Implications of small samples for generalization: Adjustments and rules of thumb. *Eval. Rev.* **41** 472–505. https://doi.org/10.1177/0193841X16655665

TWALA, B., JONES, M. and HAND, D. J. (2008). Good methods for coping with missing data in decision trees. *Pattern Recogn. Lett.* **29** 950–956.

VAN DER WEELE, T. J. and ROBINS, J. M. (2007). Four types of effect modification: A classification based on directed acyclic graphs. *Epidemiology* **18** 561–568.

WESTREICH, D., EDWARDS, J. K., LESKO, C. R., COLE, S. R. and STUART, E. A. (2018). Target validity and the hierarchy of study designs. *Amer. J. Epidemiol.* **188** 438–443.

WESTREICH, D., EDWARDS, J. K., LESKO, C. R., STUART, E. and COLE, S. R. (2017). Transportability of trial results using inverse odds of sampling weights. *Amer. J. Epidemiol.* **186** 1010–1014.

WITTE, J., HENCKEL, L., MAATHUIS, M. H. and DIDELEZ, V. (2020). On efficient adjustment in causal graphs. *J. Mach. Learn. Res.* **21** Paper No. 246, 45. MR4213427

WU, L. and YANG, S. (2022). Integrative r-learner of heterogeneous treatment effects combining experimental and observational studies. In *Proceedings of the* 1*st Conference on Causal Learning and Reasoning*.

WU, L. and YANG, S. (2023). Transfer learning of individualized treatment rules from experimental to real-world data. *J. Comput. Graph. Statist.* **32** 1036–1045. MR4641479 https://doi.org/10.1080/10618600.2022.2141752

YANG, S. and DING, P. (2020). Combining multiple observational data sources to estimate causal effects. *J. Amer. Statist. Assoc.* **115** 1540–1554. MR4143484 https://doi.org/10.1080/01621459.2019.1609973

YANG, S. and KIM, J. K. (2020). Statistical data integration in survey sampling: A review. *Jpn. J. Stat. Data Sci.* **3** 625–650. MR4181993 https://doi.org/10.1007/s42081-020-00093-w

YANG, S., KIM, J. K. and SONG, R. (2020). Doubly robust inference when combining probability and non-probability samples with high dimensional data. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **82** 445–465. MR4084171

YANG, S. and WANG, X. (2022). RWD-integrated randomized clinical trial analysis. In *ASA Biopharmaceutical Report Real World Evidence* (H. Pang, L. Wang and K. L. Griffiths, eds.) **29** 15–21.

YANG, S., WANG, X. and ZENG, D. (2023). Elastic integrative analysis of randomized trial and real-world data for treatment heterogeneity estimation. *J. Roy. Statist. Soc. Ser. B* https://doi.org/10.1093/jrsssb/qkad017

YANG, S., ZENG, D. and WANG, X. (2020). Improved inference for heterogeneous treatment effects using real-world data subject to hidden confounding. Preprint. Available at arXiv:2007.12922.

YAO, L., CHU, Z., LI, S., LI, Y., GAO, J. and ZHANG, A. (2020). A survey on causal inference. Preprint. Available at arXiv:2002.02770.

ZHONG, Y., KENNEDY, E. H., BODNAR, L. M. and NAIMI, A. I. (2021). Aipw: An R package for augmented inverse probability weighted estimation of average causal effects. *Amer. J. Epidemiol.*

ZIVICH, P., KLOSE, M., COLE, S., EDWARDS, J. and SHOOK-SA, B. (2022). Delicatessen: M-estimation in Python.