HHS Public Access

Efficient and robust methods for causally interpretable meta-analysis: transporting inferences from multiple randomized trials to a target population

Issa J. Dahabreh[1,2,3,*], **Sarah E. Robertson**[1,2], **Lucia C. Petito**[4], **Miguel A. Hernán**[1,2,5], **Jon A. Steingrimsson**[6]

[1]CAUSALab, Harvard T.H. Chan School of Public Health, Boston, MA

[2]Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, MA

[3]Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA

[4]Department of Preventative Medicine, Feinberg School of Medicine, Northwestern University, Chicago, IL

[5]Harvard-MIT Division of Health Sciences and Technology, Boston, MA

[6]Department of Biostatistics, School of Public Health, Brown University, Providence, RI

Summary:

We present methods for causally interpretable meta-analyses that combine information from multiple randomized trials to estimate potential (counterfactual) outcome means and average treatment effects in a target population. We consider identifiability conditions, derive implications of the conditions for the law of the observed data, and obtain identification results for transporting causal inferences from a collection of independent randomized trials to a new target population in which experimental data may not be available. We propose an estimator for the potential (counterfactual) outcome mean in the target population under each treatment studied in the trials. The estimator uses covariate, treatment, and outcome data from the collection of trials, but only covariate data from the target population sample. We show that it is doubly robust, in the sense that it is consistent and asymptotically normal when at least one of the models it relies on is correctly specified. We study the finite sample properties of the estimator in simulation studies and demonstrate its implementation using data from a multi-center randomized trial.

Keywords

meta-analysis; transportability; causal inference

---

* idahabreh@hsph.harvard.edu .

1. Introduction

When examining a body of evidence that consists of multiple trials, decision makers are typically interested in learning about the effects of interventions in some well-defined *target population*. In other words, they are interested in *synthesizing the evidence* across trials and *transporting causal inferences* from the collection of trials to a target population in which further experimentation may not be possible. Typically, each trial samples participants from a different underlying population, by recruiting participants from centers with different referral patterns or in different geographic locations. The goal of evidence synthesis in this context is to use the information from these diverse trials to draw causal inferences about the target population, accounting for any differences between the target population and the populations underlying the trials.

"Meta-analysis," an umbrella term for statistical methods for synthesizing evidence across multiple trials (Cooper et al., 2009), traditionally focuses on modeling the distribution of treatment effects (effect sizes) across studies or on obtaining unbiased and minimum variance summaries of data from multiple trials (Higgins et al., 2009; Rice et al., 2018). Standard meta-analysis methods produce estimates that do not have a clear causal interpretation outside of the sample of participants enrolled in the trials because the estimates do not pertain to any well-defined target population (Dahabreh et al., 2020). Recent work on "generalizability" and "transportability" has considered methods for extending causal inferences to a target population, when data are available from a single randomized trial (Westreich et al., 2017; Rudolph and van der Laan, 2017; Dahabreh et al., 2020). When multiple trials are available, methods have been proposed for assessing case-mix heterogeneity in individual patient data meta-analyses without specifying or using data from a target population of substantive interest (Vo et al., 2019, 2021).

Here, we propose methods for *causally interpretable meta-analysis* that can be used to extend inferences about potential (counterfactual) outcome means and average treatment effects from a collection of randomized trials to a target population in which experimental data may not be available (Dahabreh et al., 2020). We consider identifiability conditions, derive implications of the conditions for the law of the observed data, and obtain identification results. We then propose novel estimators for potential outcome means and average treatment effects in the target population. The estimators use data on baseline covariates, treatments, and outcomes from the collection of trials, but only data on baseline covariates from the sample of the target population. We show that the estimators are doubly robust, in the sense that they remain consistent and asymptotically normal provided at least one of the two working models on which they rely is correctly specified. Last, we study the finite sample properties of the estimators in simulation studies and demonstrate the implementation of the methods using data from a multi-center trial of treatments for hepatitis C infection.

2. Data, sampling scheme, and causal estimands

Suppose we have data from a collection of randomized trials $\mathcal{S}$, indexed by $s = 1, \ldots, m$. For each trial participant we have information on the trial S in which they participated,

treatment assignment A, baseline covariates X, and outcomes Y. We assume that the same finite set of treatments, $\mathscr{A}$, has been compared in each trial of the collection (extensions to cases where only a subset of treatments are evaluated in each trial are straightforward but require more cumbersome notation). From each trial $s \in \mathscr{S}$, the data are independent and identically distributed random tuples $(X_i, S_i = s, A_i, Y_i)$, $i = 1, \ldots, n_s$, where n_s is the total number of randomized individuals in trial s.

We also obtain a simple random sample from the target population (of individuals not participating in the trials), in which treatment assignment may not be under the control of investigators (e.g., individuals may self-select into treatment). We use the convention that $S = 0$ for individuals from the target population. The data from the sample of the target population consists of independent and identically distributed tuples $(X_i, S_i = 0, A_i, Y_i)$, $i = 1, \ldots, n_0$, where n_0 is the total number of individuals sampled from the target population. The total number of observations from the trials and the sample of the target population is $n = \sum_{s=0}^{m} n_s$. We define a new random variable R, such that $R = 1$, if $S \in \mathscr{S}$; and $R = 0$, if $S = 0$.

Informally, we assume that the observations from the trials in the collection $\mathscr{S}$ and from the target population $S = 0$ are Bernoulli-type (independent random) samples (Breslow and Wellner, 2007; Saegusa and Wellner, 2013) from (near-infinite) underlying populations (Robins, 1988). Specifically, we assume that the sample from the target population is representative of a well-specified population of substantive interest. In contrast, we do not require that the trial samples are obtained through a formal sampling process; rather, that the investigators are willing to model the data from each trial *as if* sampled from a hypothetical underlying population. This modeling choice is consistent with the fairly standard super-population approach to the analysis of individual randomized trials (Robins, 1988). Though alternative frameworks (e.g., randomization-based inference) can be appealing in some cases, we find the super-population approach attractive when the goal is to extend inferences from a collection of trials to a new target population (Dahabreh and Hernán, 2019). Investigators conducting meta-analyses usually do not have control over the selection of participants of each trial or the relative sample size of different trials. Thus, we view the observations available for analysis as obtained by sampling the corresponding underlying populations with sampling fractions that are not under the control of the investigators and are unknown to them. In Section 4 we further formalize this sampling scheme; for now it suffices to say that all expectations and probabilities below are defined under the sampling scheme. Throughout, we use $f(\,\cdot\,)$ to generically denote densities. In Section 4 we discuss the implications of the sampling scheme for identifiability of these densities.

To define causal estimands, let Y^a denote the potential (counterfactual) outcome under intervention to set treatment to $a \in \mathscr{A}$ (Rubin, 1974; Robins and Greenland, 2000). We are interested in the potential outcome mean in the target population, $\mathrm{E}[Y^a \mid R = 0]$ for each $a \in \mathscr{A}$, as well as the average causal effect, $\mathrm{E}[Y^a - Y^{a'} \mid R = 0]$ for each pair of treatments $a \in \mathscr{A}$ and $a' \in \mathscr{A}$. The treatments used in the target population need not be the same as the treatments used in the trials (e.g., some treatments may not be available outside experimental

settings). Furthermore, as we show below, data on treatments and outcomes from the target population are not necessary for identification and estimation of the potential outcome means and average treatment effects of interest.

3.   Identification

3.1   Identifiability conditions

The following are sufficient conditions for identifying the potential outcome mean in the target population $E[Y^a \mid R = 0]$.

A1.   *Consistency of potential outcomes*: if $A_i = a$, then $Y_i^a = Y_i$, for every individual i and every treatment $a \in \mathscr{A}$.—Implicit in condition *A1* are assumptions that (i) there is no direct effect of participation in any trial ($R = 1$) or participation in some specific trial ($S = s$) on the outcome (Dahabreh et al., 2019; Dahabreh and Hernán, 2019); (ii) there are no multiple versions of treatment (or that treatment variation is irrelevant with respect to the outcome (VanderWeele, 2009)); and (iii) there is no treatment-outcome interference (Rubin, 1986, 2010). These assumptions may be most plausible in the case of large pragmatic trials where fidelity to the assigned treatment can be high across settings. Note that consistency of potential outcomes is distinct from the notion of consistency used in so-called network meta-analyses; consistency of potential outcomes refers to the relationship between observed (factual) and potential (counterfactual) outcomes under each treatment, not the relationship between contrasts ("direct" and "indirect", in meta-analytic parlance) of different treatments.

A2.   *Exchangeability over treatment A in each trial*: for each trial $s \in \mathscr{S}$ and each $a \in \mathscr{A}, Y^a \perp\!\!\!\perp A \mid (X, S = s)$.—Condition *A2* is typically plausible because of randomization (marginal or conditional on X) in each of the trials . The condition is not equivalent to the condition $Y^a \perp\!\!\!\perp A \mid (X, R = 1)$ because the trials may have different randomization ratios and trial participation may have direct effects on the outcome or share unmeasured common causes with the outcome (the last two possibilities, however, are precluded by conditions *A1* and *A4*, respectively). It is also worth noting that condition *A2* would also hold if $\mathscr{S}$ was a collection of observational studies in which the covariates X were sufficient to adjust for baseline confounding. Thus, our results can also apply to pooled analyses of observational studies, provided that background knowledge suggests "adjustment" for X is sufficient to control confounding.

A3.   *Positivity of treatment assignment in each trial*: for each treatment $a \in \mathscr{A}$ and for each trial $s \in \mathscr{S}$, if $f(x, S = s) \neq 0$, then $\Pr[A = a \mid X = x, S = s] > 0$.—Condition *A3* is also plausible in marginally and conditionally randomized trials.

A4.   *Exchangeability over S*: for each $a \in \mathscr{A}, Y^a \perp\!\!\!\perp S \mid X$.—In applied work, condition *A4* will typically be a critical assumption connecting the randomized trials in the

collection $\mathcal{S}$, between them and with the target population ($S = 0$). On its own, condition *A4* is not testable; we will show, however, that it has testable implications when combined with conditions *A1* and *A2*). In practice, condition A4 will need to be examined in light of substantive knowledge and examining the impact of violations of the condition may require undertaking sensitivity analyses (Robins et al., 2000).

A5. *Positivity of trial participation*: **for each** $s \in \mathcal{S}$, **if** $f(x, R = 0) \neq 0$, **then** $\mathbf{Pr}[S = s \mid X = x] > 0$.—Informally, condition *A5* states that covariate patterns in the target population can also occur in each trial. This condition only involves the observable data and thus is in principle testable. Formal examination of the condition, however, is challenging when X is high dimensional (Petersen et al., 2012).

3.2 Implications of the identifiability conditions

We now explore some implications of the identifiability conditions. By Lemma 4.2 of Dawid (1979), condition *A4* implies exchangeability over S among the collection of trials, that is,

$$Y^a \per\!\!\!\perp S \mid X \Longrightarrow Y^a \per\!\!\!\perp S \mid (X, R = 1), \text{ for every } a \in \mathcal{A},$$

(1)

and also implies exchangeability of participants in the collection of trials and the target population,

$$Y^a \per\!\!\!\perp S \mid X \Longrightarrow Y^a \per\!\!\!\perp I(S = 0) \mid X \Longleftrightarrow Y^a \per\!\!\!\perp R \mid X.$$

(2)

The first of these results will be useful to derive restrictions on the law of the observed data imposed by the identifiability conditions; the second result will be useful in obtaining identification results for causal quantities in the target population, using information from the collection of trials.

By Lemma 4.3 of Dawid (1979), the result in (1) and condition *A2* imply conditional exchangeability of potential outcomes over (S, A) in the collection of trials, that is,

$$\left\{ \begin{array}{l} Y^a \perp\!\!\!\perp A \mid (X, R = 1, S) \\ Y^a \perp\!\!\!\perp S \mid (X, R = 1) \end{array} \right\} \Longrightarrow Y^a \perp\!\!\!\perp \left(S, A \right) \middle| (X, R = 1).$$

(3)

Noting that $Y^a \perp\!\!\!\perp (S, A) \mid (X, R = 1)$ implies $Y^a \perp\!\!\!\perp S \mid (X, R = 1, A = a)$, and using condition *A1*, we obtain

$$Y \perp\!\!\!\perp S \mid (X, R = 1, A = a).$$

(4)

Thus, among individuals participating in any trial ($R = 1$), the observed outcome Y is independent of the trial random variable S, within treatment groups ($A = a$) and conditional on covariates X. Because this condition does not involve potential outcomes, it is testable using the observed data (e.g., using methods for comparing conditional densities). Furthermore, because $\{R = 0\} \iff \{R = 0, S = 0\}$, we also obtain that $Y \perp\!\!\!\perp S \mid (X, R, A = a)$. The independence condition in (4) implies that for every $a \in \mathcal{A}$ and every x such that $f(x, S = 0) \neq 0$,

$$\mathrm{E}[Y \mid X = x, S = 1, A = a] = \dots = \mathrm{E}[Y \mid X = x, S = m, A = a].$$

(5)

This restriction is testable using widely available parametric or non-parametric approaches for modeling the conditional expectations (e.g., (Racine et al., 2006; Luedtke et al., 2019)). Such statistical testing may be a useful adjunct to assessments of the causal assumptions based on substantive knowledge.

In practical terms, violations of the restrictions in displays (4) or (5) can occur, for example, when there are unmeasured common causes of trial participation S and the outcome Y (failure of condition *A4*), when trial participation directly effects the outcomes (failure of condition *A1*), or when there exists outcome-relevant variation in the treatments evaluated across trials (again, failure of condition *A1*). Though formal statistical assessments of the restrictions are possible, they will often be challenging when X is high-dimensional. Furthermore, formal assessments of the restrictions in (4) and (5) will in general not be able to pinpoint the specific identifiability condition whose failure explains the violation of the observed data restrictions. Thus, considerable background knowledge and scientific judgment will be needed to decide on the appropriateness of combining information across trials to learn about the target population.

3.3 Identification of potential outcome means using the collection of trials

When conditions *A1* through *A5* hold, the potential outcome mean in the target population can be identified using covariate, treatment, and outcome data from the collection of trials, and baseline covariate data from the target population.

Theorem 1 (Identification of potential outcome means): *Under conditions A1 through A5, the potential outcome mean in the target population under treatment $a \in \mathcal{A}$, $\mathrm{E}[Y^a \mid R = 0]$, is identifiable by the observed data functional*

$$\psi(a) \equiv \mathrm{E}[\mathrm{E}[Y \mid X, R = 1, A = a] \mid R = 0],$$

(6)

which can be equivalently expressed as

$$\psi(a) = \frac{1}{\Pr[R = 0]} \mathrm{E}\left[\frac{I(R = 1, A = a) Y \Pr[R = 0 \mid X]}{\Pr[R = 1 \mid X] \Pr[A = a \mid X, R = 1]} \right].$$

(7)

The proof is given in Web Appendix A.

It also follows that, under the conditions of Theorem 1, the average treatment effect in the target population comparing treatments a and a' in $\mathscr{A}$ is also identifiable:

$$\mathrm{E}\left[Y^a - Y^{a'} \mid R = 0\right] = \mathrm{E}\left[Y^a \mid R = 0\right] - \mathrm{E}\left[Y^{a'} \mid R = 0\right] = \psi(a) - \psi(a').$$

3.4   Identification under weaker conditions

Practitioners of meta-analysis believe that by combining evidence from multiple randomized trials it should be possible to draw inferences about target populations that are – in some vague sense – broader than the population underlying each randomized trial. To give a concrete example: suppose that we have data from two trials of the same medications, one recruiting individuals with mild disease and the other recruiting individuals with severe disease. Suppose also that our target population includes some individuals with mild and some with severe disease and that effectiveness varies by disease severity. Intuition suggests that if inferences are "transportable" from each of the two trials to each subset of the target population defined by disease severity, then it should be possible to draw some conclusions about treatment effectiveness in the target population. Yet, in this setting, condition *A5* is grossly violated (e.g., individuals with mild disease have zero probability of participating in one of the two trials). In the following two sub-sections, we give identification results that rely on weaker identifiability conditions and capture the intuition that when combining evidence from multiple trials we can draw inferences about target populations that are broader than the population underlying each randomized trial.

3.4.1   Weakening the positivity conditions of Theorem 1.—Suppose that the independence conditions $Y^a \per\!\!\!\perp R \mid X$ and $Y^a \per\!\!\!\perp A \mid (X, R = 1)$ hold, taken either as primitive conditions or as implications of identifiability conditions *A2* and *A4* (as was done for Theorem 1).

Furthermore, consider the following positivity conditions:

A3*. Positivity of the probability of treatment in the collection of trials: for each treatment $a \in \mathscr{A}$, if $f(x, R = 1) \neq 0$, then $\Pr[A = a \mid X = x, R = 1] > 0$.

A5*. Positivity of the probability of participation in the collection of trials: if $f(x, R = 0) \neq 0$, then $\Pr[R = 1 \mid X = x] > 0$.

It is worth noting that these positivity conditions are weaker than the corresponding conditions *A3* and *A5*, in the sense that *A3** is implied by, but does not imply, *A3*; and *A5** is implied by, but does not imply, *A5*. The difference between the two sets of assumptions is practically important: informally, condition *A3* requires every treatment to be available in every trial and for every covariate pattern that can occur in that trial, and condition *A5* requires every covariate pattern that can occur in the target population to be represented in every trial. In contrast, condition *A3** requires every treatment to be available in the aggregated collection of trials and for any covariate pattern that can occur in the collection

(but not necessarily in every trial); condition *A5** requires every covariate pattern that can occur in the target population to occur in the aggregated collection of trials (but not necessarily in every trial).

Using these modified conditions, we obtain the following result (we give the formal arguments in Web Appendix A and Web Appendix B):

Theorem 2 (Identification of potential outcome means under weaker positivity conditions): *If $Y^a \perp\!\!\!\perp R \mid X$ and $Y^a \perp\!\!\!\perp A \mid (X, R = 1)$ and conditions A1, A3*, and A5* hold, then the potential outcome mean in the target population under treatment $a \in \mathscr{A}, \mathrm{E}[Y^a \mid R = 0]$, is identifiable by $\psi(a)$.*

3.4.2 Identification under weaker overlap and exchangeability conditions.—
We will now show that the exchangeability conditions for potential outcome means can also be relaxed, alongside the positivity conditions; to do so, we need to introduce some additional notation in order to keep track of the subsets of the collection of trials where different covariate patterns can occur.

Let $\mathscr{X}_j$ for $j \in \{0, 1, \dots, m\}$ denote the support of the random vector X in the subset of the population with $S = j$. That is, $\mathscr{X}_j \equiv \{x : f(x \mid S = j) > 0\}$. For each covariate pattern $X = x$ that can occur in the collection of trials, that is, for each $x \in \bigcup_{s \in \mathscr{S}} \mathscr{X}_s$, define $\mathscr{S}_x$ as the subset of trials in $\mathscr{S}$ such that x belongs in their support. That is, $\mathscr{S}_x \equiv \{s : s \in \mathscr{S}, x \in \mathscr{X}_s\}$. Intuitively, $\mathscr{S}_x$ denotes the subset of trials in the collection $\mathscr{S}$ where the covariate pattern $X = x$ can occur. Using this additional notation, consider the following identifiability conditions:

A4†. Exchangeability in mean over: For every x such that $f(x, S = 0) \neq 0$ and every $s \in \mathscr{S}_x, \mathrm{E}[Y^a \mid X = x, S = 0] = \mathrm{E}[Y^a \mid X = x, S = s]$.

A5†. Overlap of the collection $\mathscr{S}$ with the target population: $\bigcup_{s \in \mathscr{S}} (\mathscr{X}_s \cap \mathscr{X}_0) = \mathscr{X}_0$. Using the above two conditions in the place of conditions *A4* and *A5* still allows for identification of the potential outcome means, using the following result.

Theorem 3 (Identification of potential outcome means under weaker conditions): *Under identifiability conditions A1 through A3, A4†, and A5†, for every $a \in \mathscr{A}, \mathrm{E}[Y^a \mid R = 0]$ is identifiable by*

$$\phi(a) \equiv \int \mathrm{E}[Y \mid X = x, I(S \in \mathscr{S}_x) = 1, A = a] f(x \mid S = 0) dx.$$

The proof is given in Web Appendix B.

In simple cases, as in our hypothetical example of effect modification by disease severity where the positivity violation was due to a single binary covariate, the above result simply suggests that the conditional mean of the outcome Y can be modeled separately in each group of trials defined by that binary covariate. For more complicated violations of positivity

condition *A5*, especially those involving multiple continuous covariates, the above result is mostly useful, not for suggesting a particular modeling strategy, but for showing that some degree of interpolation when modeling the conditional expectation may be relatively benign, when there is adequate overlap between the collection of trials and the target population, in the sense of condition $A5^{\dagger}$.

3.5 Identification under exchangeability in effect measure

Up to this point, we have examined conditions that are sufficient for identifying both potential outcome means and average treatment effects. If interest is restricted to average treatment effects, identification is possible under a weaker condition of exchangeability in effect measure over S given baseline covariates, essentially, a requirement that conditional average treatment effects, but not necessarily potential outcome means, can be transported from each trial in the collection S to the target population; see Dahabreh et al. (2018, 2020) for a similar argument in the context of transporting inferences from a single trial. Under this weaker condition, however, the individual potential outcome means are not identifiable. Specifically, consider the following condition:

$A4^{\ddagger}$. Exchangeability in effect measure over S: for every pair of treatments a and a', with $a \in \mathcal{A}$ and $a' \in \mathcal{A}$, for every $s \in \mathcal{S}$, and for every x such that $f(x, S = s) \neq 0$, we have $\mathrm{E}\Big[Y^a - Y^{a'} \mid X = x, S = s\Big] = \mathrm{E}\Big[Y^a - Y^{a'} \mid X = x, R = 0\Big]$.

It is easy to see that this condition, combined with conditions *A1* through *A3*, and *A5*, implies that the observed data trial-specific conditional mean difference comparing treatments $a \in \mathcal{A}$ and $a' \in \mathcal{A}, \mathrm{E}[Y \mid X, S = s, A = a] - \mathrm{E}[Y \mid X, S = s, A = a']$ does not vary over $s \in \mathcal{S}$ for all X values with positive density in the target population. Using this implication, we now give an identification result for the average treatment effect under the weaker condition of exchangeability in effect measure.

Theorem 4 (Identification under exchangeability in effect measure): *Under conditions A1 through A3, $A4^{\ddagger}$, and A5, the average treatment effect in the target population comparing treatments $a \in \mathcal{A}$ and $a' \in \mathcal{A}, \mathrm{E}\Big[Y^a - Y^{a'} \mid R = 0\Big]$, is identifiable by the observed data functional $\rho(a, a') \equiv \mathrm{E}[\tau(a, a'; X) \mid R = 0]$, where $\tau(a, a'; X) \equiv \mathrm{E}[Y \mid X, S = s, A = a] - \mathrm{E}[Y \mid X, S = s, A = a']$ does not vary over $s \in \mathcal{S}$. Furthermore, $\rho(a, a')$ can be re-expressed as $\rho(a, a') = \frac{1}{\Pr[R = 0]}\mathrm{E}[w(a, a'; R, X, S, A)Y]$, where*

$$w(a, a'; R, X, S, A) \equiv \left(\frac{I(R = 1, A = a)}{\Pr[A = a \mid X, S, R = 1]} - \frac{I(R = 1, A = a')}{\Pr[A = a' \mid X, S, R = 1]}\right)\frac{\Pr[R = 0 \mid X]}{\Pr[R = 1 \mid X]}.$$

The proof is given in Web Appenidx C. Note in passing that positivity condition *A5* in Theorem 4 can be relaxed in a way analogous to what was done in Theorem 3.

Because the potential outcome means $\mathrm{E}\Big[Y^a \mid R = 0\Big]$ in the target population often are of inherent interest, in the rest of this paper we focus on estimating the identifying functionals $\psi(a)$, for $a \in \mathcal{A}$, following Theorems 1 and 2.

4. Remarks on the sampling model

We now revisit the assumed sampling scheme that we described in Section 2 and introduce some notation to help us distinguish between the *population sampling model*, where data are obtained by simple random sampling from a common super-population, and a *biased sampling model*, where data are obtained by stratified random sampling from the target population and the populations underlying the trials, with unknown and possibly variable sampling fractions.

Under a nonparametric model for the observed data, the density of the law of the observable data $O = (X, S, R, A, Y)$, using p to generically denote densities, can be written as $p(r, x, s, a, y) = p(r)p(x \mid r)p(s \mid x, r)p(a \mid r, x, s)p(y \mid r, x, s, a)$. By definition, S contains all the information contained in R, which implies that $p(a \mid r, s, x) = p(a \mid s, x)$. Furthermore, by (4), $p(y \mid r = 1, x, a) = p(y \mid r = 1, s, x, a)$; and the equivalence $\{R = 0, S = 0\} \iff \{S = 0\}$ gives $p(y \mid r = 0, x, a) = p(y \mid r = 0, s, x, a)$; thus, $p(y \mid r, x, a) = p(y \mid r, s, x, a)$. We conclude that, under simple random sampling, the density of the observable data would be

$$p(r, x, s, a, y) = p(r)p(x \mid r)p(s \mid x, r)p(a \mid x, s)p(y \mid r, x, a).$$

This density can be viewed as reflecting a *population sampling model*.

The data collection approach described in Section 2, however, induces a *biased sampling model* (Bickel et al., 1993). What we mean here is that the sampling fraction from the sub-population underlying each randomized trial ($S = 1, \ldots, m$) and the target population sample ($R = S = 0$) is not under the control of the investigators (and, typically, unknown to them) and reflects the particular circumstances (e.g., recruiting practices) of how sampling was conducted (e.g., convenience sampling is typical in randomized trials). Thus, in the data, the ratios $\frac{n_j}{n}$, for $j \in \{0, 1, \ldots, m\}$, do not reflect the population probabilities of belonging to the subset of the population with $S = j$, because the sampling fraction from each subset is unknown (and possibly variable between subsets). As a technical condition, we require that as $n \longrightarrow \infty$, $\frac{n_j}{n} \longrightarrow \pi_j > 0$, for $j \in \{0, 1, \ldots, m\}$. Nevertheless, under the biased sampling model, the limiting values, π_j, are not necessarily equal to the super-population probabilities under the population sampling model.

Under this more plausible sampling model, the density of the observable data can be written as

$$
\begin{aligned}
q(r, x, s, a, y) &= q(r)q(x \mid r)q(s \mid x, r)p(a \mid x, s)p(y \mid r, x, a) \\
&= q(r)\{q(x \mid r = 1)\}^r \{q(x \mid r = 0)\}^{1-r} q(s \mid x, r)p(a \mid x, s)p(y \mid r, x, a) \\
&= q(r)\{q(x \mid r = 1)\}^r \{p(x \mid r = 0)\}^{1-r} q(s \mid x, r)p(a \mid x, s)p(y \mid r, x, a).
\end{aligned}
$$

Here, $q(r) \neq p(r)$ and $q(s \mid x, r) \neq p(s \mid x, r)$ to reflect the fact that the sampling fractions for participants in the trials and members of the target population are not equal to the super-population probabilities under the population sampling model, but instead depend on the complex processes of randomized trial design and conduct (i.e., the biased sampling

model). For the same reason, $q(x \mid r) \neq p(x \mid r)$, because in general we do not expect $q(x \mid r = 1) = p(x \mid r = 1)$; the distribution of X in the collection of trials depends on the sampling from the different population underlying each trial. Nevertheless, we expect $p(x \mid r = 0) = q(x \mid r = 0)$ because we take a simple random sample from the target population under the biased sampling model. In contrast, the terms $p(a \mid x, s)$ and $p(y \mid a, x, r)$ are the same in both densities, reflecting the property of stratified (by S) random sampling of the population underlying each trial or the target population, and the independence condition in (4).

The functional $\psi(a)$ in Theorems 1 and 2 only depends on components of $p(r, x, s, a, y)$ that are also identifiable under the distribution $q(r, s, x, a, y)$ induced by the biased sampling. Specifically, using the representation in (6), $\psi(a)$ depends on $p(y \mid r = 1, x, a)$, and $p(x \mid r = 0) = p(x \mid s = 0)$, both of which are identifiable under the biased sampling model.

5. Estimation and inference

We now turn our attention to the estimation of the functional $\psi(a)$, for $a \in \mathscr{A}$. The results presented in this section apply to data obtained under the population sampling model $p(r, x, s, a, y)$ and the biased sampling model $q(r, x, s, a, y)$, because $\psi(a)$ is identifiable under both. In fact, the results in Breslow et al. (2000) imply that influence functions for $\psi(a)$ under sampling from $q(r, x, s, a, y)$ are equivalent to those under sampling from $p(r, x, s, a, y)$, but with densities from $q(r, x, s, a, y)$ replacing those under $p(r, x, s, a, y)$; see Kennedy et al. (2015) for a similar argument in the context of matched cohort studies and Dahabreh et al. (2020) in the context of transporting inferences from a single trial. In the remainder of the paper, we assume that analysts will be working under $q(r, x, s, a, y)$, because the biased sampling model is more realistic for applied meta-analyses.

5.1 Proposed estimator

In Web Appendix D we show that the first-order influence function (Bickel et al., 1993) of $\psi(a)$, for $a \in \mathscr{A}$, under the nonparametric model for the observable data, is

$$\Psi_{q_0}^1\!\left(a\right) = \frac{1}{\pi_{q_0}}\left\{\frac{I(R = 1, A = a)\mathrm{Pr}_{q_0}[R = 0 \mid X]}{\mathrm{Pr}_{q_0}[R = 1 \mid X]\mathrm{Pr}_{q_0}[A = a \mid X, R = 1]}\{Y - \mathrm{E}_{q_0}[Y \mid X, R = 1, A = a]\right.$$
$$\left.\} + I(R = 0)\{\mathrm{E}_{q_0}[Y \mid X, R = 1, A = a] - \psi_{q_0}(a)\}\right\},$$

where $\pi_{q_0} = \mathrm{Pr}_{q_0}[R = 0]$ and the subscript q_0 denotes that all quantities are evaluated at the "true" data law. Specifically, we show that $\Psi_{q_0}^1(a)$ satisfies $\left.\dfrac{\partial \psi_{q_t}(a)}{\partial t}\right|_{t = 0} = \mathrm{E}\big[\Psi_{q_0}^1(a)u(O)\big]$, where $u(O)$ denotes the score of the observable data $O = (R, X, S, A, Y)$ and the left hand side of the above equation is the pathwise derivative of the target parameter $\psi(a)$. Theorem 4.4 in (Tsiatis, 2007) shows that $\Psi_{q_0}^1(a)$ lies in the tangent set; it follows (see, e.g., (Van der Vaart, 2000), page 363) that $\Psi_{q_0}^1(a)$ is the efficient influence function under the nonparametric model (it is in fact the unique influence function under that model). For a more thorough discussion of semiparametric efficiency theory and precise definitions of

pathwise derivatives and the tangent space we refer to Chapter 25 in (Van der Vaart, 2000). Furthermore, in Web Appendix D we show that $\Psi_{q_0}^1(a)$ is the efficient influence function under semiparametric models incorporating the restriction $Y \perp\!\!\!\perp S \mid (X, A = a, R)$ or where the probability of treatment conditional on covariates and trial participation status is known (e.g., if all trials have the same randomization probabilities).

The influence function above suggests the estimator

$$
\widehat{\psi}_{\mathrm{aug}}(a) = \frac{1}{n\widehat{\pi}} \sum_{i=1}^{n} \left\{ I(R_i = 1, A_i = a)\frac{1 - \widehat{p}(X_i)}{\widehat{p}(X_i)\widehat{e}_a(X_i)}\{Y_i - \widehat{g}_a(X_i)\} + I(R_i = 0)\widehat{g}_a(X_i) \right\},
$$

(8)

where $\widehat{\pi} = n^{-1}\sum_{i=1}^{n} I(R_i = 0), \widehat{g}_a(X)$ is an estimator for $\mathrm{E}[Y \mid X, R = 1, A = a], \widehat{e}_a(X)$ is an estimator for $\Pr[A = a \mid X, R = 1]$, and $\widehat{p}(X)$ is an estimator for $\Pr[R = 1 \mid X]$. Note that $\widehat{\psi}_{\mathrm{aug}}(a)$ involves data on $(R = 1, X, A, Y)$ from trial participants and data on $(R = 0, X)$ from the sample of the target population; thus, treatment and outcome data from the target population are not necessary. Furthermore, it is natural to estimate the average treatment effect in the target population for comparing treatments a and a' in $\mathscr{A}$, that is, $\mathrm{E}\left[Y^a - Y^{a'} \mid R = 0\right]$, using the contrast estimator $\widehat{\delta}(a, a') = \widehat{\psi}_{\mathrm{aug}}(a) - \widehat{\psi}_{\mathrm{aug}}(a')$.

In practical applications, analysts may be able to improve the performance of the estimator by normalizing the weights so that their sum is equal to the total number of observations in the target population sample (Dahabreh et al., 2018, 2019, 2020). This normalization, which was originally proposed for survey analyses (Hájek, 1971), may be particularly useful when $\frac{1 - \widehat{p}(X)}{\widehat{p}(X)\widehat{e}_a(X)}$ is highly variable over the observations in the trials (Robins et al., 2007).

Asymptotic properties of estimators for potential outcome means: Let $g_a^*(X), e_a^*(X)$, and $p^*(X)$ denote the asymptotic limits (assumed to exist) of $\widehat{g}_a(X), \widehat{e}_a(X)$, and $\widehat{p}(X)$, respectively. Finally, define $\widehat{\gamma} = \left\{\frac{1}{n}\sum_{i=1}^{n} I(R_i = 0)\right\}^{-1} = \widehat{\pi}^{-1}$ as an estimator for $\gamma^* = \Pr[R = 0]^{-1}$.

For general functions $\gamma', g_a'(X), e_a'(X)$, and $p'(X)$ define

$H(\gamma', g_a'(X), e_a'(X), p'(X)) = \gamma'\left\{I\left(R = 0\right)g_a'\left(X\right) + I\left(R = 1, A = a\right)\frac{1 - p'(X)}{p'(X)e_a'(X)}\{Y - g_a'(X)\}\right\}$. Using notation from van der Vaart and Wellner (1996), define $\mathbb{P}_n\left(v\left(W\right)\right) = n^{-1}\sum_{i=1}^{n} v(W_i)$ and $\mathbb{G}_n\left(v\left(W\right)\right) = \sqrt{n}(\mathbb{P}_n(v(W)) - \mathrm{E}[v(W)])$, for some function v and a random variable W. Using this notation, $\widehat{\psi}_{\mathrm{aug}}(a) = \mathbb{P}_n(H(\widehat{\gamma}, \widehat{g}_a(X), \widehat{e}_a(X), \widehat{p}(X)))$.

To establish asymptotic properties of $\widehat{\psi}(a)$, we make the following assumptions:

i. (i) The sequence $H(\widehat{\gamma}, \widehat{g}_a(X), \widehat{e}_a(X), \widehat{p}(X))$ and its limit $H(\gamma^*, g_a^*(X), e_a^*(X), p^*(X))$ fall in a Donsker class (van der Vaart and Wellner, 1996).

ii. (ii) $\| H(\widehat{\gamma}, \widehat{g}_a(X), \widehat{e}_a(X), \widehat{p}(X)) - H(\gamma^*, g_a^*(X), e_a^*(X), p^*(X)) \|_2 \xrightarrow{\text{a.s.}} 0$.

iii. (iii) $E\left[H\left(\gamma^*, g_a^*(X), e_a^*(X), p^*(X)\right)^2\right] < \infty$.

iv. (iv) At least one of the following two assumptions holds: (a)
$\widehat{p}(X) \xrightarrow{\text{a.s.}} p^*(X) = \Pr\left[R = 1 \mid X\right]$ and $\widehat{e}_a(X) \xrightarrow{\text{a.s.}} e_a^*(X) = \Pr\left[A = a \mid X, R = 1\right]$; or
(b)$\widehat{g}_a(X) \xrightarrow{\text{a.s.}} g_a^*(X) = E\left[Y \mid X, R = 1, A = a\right]$, $p^*(X) \geqslant \varepsilon$, and $e_a^*(X) \geqslant \varepsilon$ a.s. for some
$\varepsilon > 0$.

The following Theorem gives the asymptotic properties of $\widehat{\psi}_{\text{aug}}(a)$; a detailed proof is given in Web Appendix E.

Theorem 5: *If assumptions* (i) *through* (iv) *hold, then*

(1) $\widehat{\psi}_{\text{aug}}(a) \xrightarrow{\text{a.s.}} \psi(a)$; *and*

(2) $\widehat{\psi}_{\text{aug}}(a)$ *has the asymptotic representation*

$$\sqrt{n}(\widehat{\psi}_{\text{aug}}(a) - \psi(a)) = \mathbb{G}_n\left(H\left(\gamma^*, g_a^*(X), e_a^*(X), p^*(X)\right)\right) + Rem + o_P(1),$$

(9)

where $\mathbb{G}_n\left(H\left(\gamma^*, g_a^*(X), e_a^*(X), p^*(X)\right)\right)$ *is asymptotically normal and*

$$Rem \leqslant \sqrt{n} O_P\left(\left(\| \Pr[R = 1 \mid X] - \widehat{p}(X) \|_2 + \| \Pr[A = a \mid X, R = 1] - \widehat{e}_a(X) \|_2\right) \times \| \widehat{g}_a(X) - E[Y \mid X, R = 1, A = a] \|_2\right).$$

(10)

If $\widehat{g}_a(X), \widehat{p}(X), \widehat{e}_a(X)$ and $g_a^*(X), p^*(X), e_a^*(X)$ are Donsker; $p^*(X)$ and $e_a^*(X)$ are uniformly bounded away from zero – assumption (iv); and $g_a^*(X)$ and Y are uniformly bounded, then $H(\widehat{\gamma}, \widehat{g}_a(X), \widehat{e}_a(X), \widehat{p}(X))$ and its limit are Donsker (Kosorok, 2008).

Assumptions (i), (ii), and (iii) are standard assumptions used to show asymptotic normality of M-estimators (Van der Vaart, 2000). Assumption (i) restricts the flexibility of the models used to estimate the nuisance parameters (and their corresponding limits). But, it still covers a wide range of commonly used estimators such as parametric Lipschitz classes and VC classes (van der Vaart and Wellner, 1996). For data-adaptive estimators, Donsker assumptions can be relaxed using sample splitting (Robins et al., 2008).

Assumption (iv) indicates that the estimator $\widehat{\psi}_{\text{aug}}(a)$ is doubly robust, in the sense that it converges almost surely to $\psi(a)$ when either (1) the model for the conditional outcome mean $E[Y \mid X, R = 1, A = a]$ is correctly specified, so that $\widehat{g}_a(X)$ converges to the true conditional expectations almost surely; or (2) the model for the probability of participation in any trial $\Pr[R = 1 \mid X]$ and the model for the probability of treatment assignment $\Pr[A = a \mid X, R = 1]$ are correctly specified, so that $\widehat{p}(X)$ and $\widehat{e}_a(X)$ converge to the true conditional probabilities almost surely. When the treatment assignment is the same for all the trials $\Pr[A = a \mid X, R = 1]$ is known and when the treatment assignment only depends on

a few categorical covariates $\Pr[A = a \mid X, R = 1]$ is easy to estimate consistently. But, when the treatment assignment mechanism depends on continuous covariates, as would be the case in pooled analyses of observational studies, estimating $\Pr[A = a \mid X, R = 1]$ may be more challenging.

The asymptotic representation given by (9) gives several useful insights into how estimation of the nuisance parameters affect the asymptotic distribution of $\widehat{\psi}_{\text{aug}}(a)$. By the central limit theorem, the first term on the right hand side of (9) is asymptotically normal. Hence, the asymptotic distribution of the estimator relies on the behavior of the term given in equation (10). If both the pair of nuisance parameters $\widehat{p}(X)$ and $\widehat{g}_a(X)$ and the pair of nuisance parameters $\widehat{e}_a(X)$ and $\widehat{g}_a(X)$ converge combined at a rate fast enough such that term (10) is $o_P(1)$, then the estimator is $\sqrt{n}$-consistent, asymptotically normal, and has asymptotic variance equal to the variance of the efficient influence function.

The asymptotic representation gives the rate of convergence result

$$\| \widehat{\psi}_{\text{aug}}(a) - \psi(a) \| = O_P\bigg(\frac{1}{\sqrt{n}} + (\| \Pr[R = 1 \mid X] - \widehat{p}(X) \|_2 + \| \Pr[A = a \mid X, R = 1] - \widehat{e}_a(X) \|_2$$
$$) \times \| \widehat{g}_a(X) - \mathrm{E}[Y \mid X, R = 1, A = a] \|_2)$$

Thus, if both the combined rate of convergence of $\widehat{p}(X)$ and $\widehat{g}_a(X)$ and the combined rate of convergence of $\widehat{e}_a(X)$ and $\widehat{g}_a(X)$ are at least $\sqrt{n}$, then $\psi_{\text{aug}}(a)$ is $\sqrt{n}$ convergent.

5.2 Inference

To construct Wald-style confidence intervals for $\psi(a)$, when using parametric models, we can easily obtain the sandwich estimator (Stefanski and Boos, 2002) of the sampling variance for $\widehat{\psi}(a)$. Alternatively, we can use the non-parametric bootstrap. If the remainder term in equation (9) of Theorem 5 is $o_P(1)$, then the estimator $\widehat{\psi}_{\text{aug}}(a)$ is asymptotically normally distributed. The asymptotic normality combined with Assumption (iii) ensure that the confidence intervals calculated using the non-parametric bootstrap asymptotically have the correct coverage rate.

6. Simulation study

We performed simulation studies to evaluate the finite sample performance of the augmented estimators described in the previous section and compare them with alternative approaches.

6.1 Data generation

The data generation process involved six steps: generation of covariates, selection for trial participation, sampling of individuals from the target population, allocation of trial participants to specific trials, random treatment assignment, and potential/observed outcomes. Simulations contained either $n = 10{,}000$ or $n = 100{,}000$ individuals, including both trial participants and the sample of the target population.

1. *Covariates*: Three covariates X_1, X_2, X_3 for each individual in the target population were drawn from a mean-zero multivariate normal distribution with all marginal variances equal to 1 and all pairwise correlations equal to 0.5.

2. *Selection for trial participation*: We considered three trials, $\mathcal{S} = \{1, 2, 3\}$, a reasonable number given the requirement that all trials have examined the same treatments. We examined scenarios where the total number of trial participants, $\sum_{s=1}^{3} n_s$, was 1,000, 2,000, or 5,000. We "selected" observations for participation in any trial using a logistic-linear model, $R \sim \text{Bernoulli}(\Pr[R = 1 \mid X])$ with

    $$\Pr\left[R = 1 \mid X\right] = \frac{\exp\left(\beta X^T\right)}{1 + \exp\left(\beta X^T\right)}, X = (1, X_1, \ldots, X_3), \beta = (\beta_0, \ln(2), \ln(2), \ln(2)), \text{ where we}$$

    solved for β_0 to result (on average) in the desired total number of trial participants given the total cohort sample size (exact numerical values for β_0 are available for all scenarios in the code to reproduce the simulations in Web Appendix H).

3. *Sampling of individuals from the target population*: We used baseline covariate data from all remaining non-randomized individuals in the sample, $n_0 = n - \sum_{s=1}^{3} n_s$; this corresponds to taking a census of the non-randomized individuals in the simulated cohort and treating them as the sample from the target population.

4. *Allocation of trial participants to specific trials*: We allocated trial participants ($R = 1$) to one of the three randomized trials in $\mathcal{S}$ using a multinomial logistic model, $S \mid (X, R = 1) \sim \text{Multinomial}\left(p_1, p_2, p_3; \sum_{s=1}^{3} n_s\right)$, with

    $$p_1 = \Pr\left[S = 1 \mid X, R = 1\right] = 1 - p_2 - p_3, p_2 = \Pr\left[S = 2 \mid X, R = 1\right] = \frac{e^{\xi X^T}}{1 + e^{\xi X^T} + e^{\zeta X^T}},$$

    and $p_3 = \Pr\left[S = 3 \mid X, R = 1\right] = \dfrac{e^{\zeta X^T}}{1 + e^{\xi X^T} + e^{\zeta X^T}}$, where

    $\xi = (\xi_0, \ln(1.5), \ln(1.5), \ln(1.5))$ and $\zeta = (\zeta_0, \ln(0.75), \ln(0.75), \ln(0.75))$. We evaluated a scenario in which the trials had the same sample size and one where the sample size varied across trials. We used Monte Carlo methods to obtain intercepts ξ_0 and ζ_0 that resulted in approximately equal-sized trials or in unequal-sized trials with a 4:2:1 ratio of samples sizes (exact numerical values for ξ_0 and ζ_0 are available for all scenarios in the code to reproduce the simulations) (Robertson et al., 2021).

5. *Random treatment assignment*: We generated an indicator of unconditionally randomized treatment assignment, A, among randomized individuals. In one scenario the treatment assignment mechanism was marginally randomized and constant across trials, $A \sim \text{Bernoulli}(\Pr[A = 1 \mid S = s])$ with $\Pr[A = 1 \mid S = s] = 1/2$. In the second scenario, the treatment assignment mechanism varied across trials and was marginally randomized, with probabilities $\Pr[A = 1 \mid S = 1] = 1/2, \Pr[A = 1 \mid S = 2] = 1/3$, and $\Pr[A = 1 \mid S = 3] = 2/3$.

6. *Outcomes*: We generated potential outcomes as $Y^a = \theta^a X^T + \epsilon^a$, for $a \in \{0,1\}, \theta^0 = (1.5, \ 1, \ 1, \ 1)$, and $\theta^1 = (0.5, \ -1, \ -1, \ -1)$. In all simulations, ϵ^a had an independent standard normal distribution for $a = 0,1$. We generated observed outcomes under consistency, such that $Y = AY^1 + (1 - A)Y^0$.

6.2 Methods implemented and comparisons

Estimators: In each simulated dataset, we applied the estimators from Section 5. Heuristically, the identification results in Theorems 1 and 2 suggest two non-augmented (and non-doubly robust) estimators: the g-formula estimator,

$$\widehat{\psi}_{\mathrm{g}}(a) = \left\{ \sum_{i=1}^{n} I(R_i = 0) \right\}^{-1} \sum_{i=1}^{n} I(R_i = 0)\widehat{g}_a(X_i),$$

and the weighting estimator

$$\widehat{\psi}_{\mathrm{w}}(a) = \left\{ \sum_{i=1}^{n} I(R_i = 0) \right\}^{-1} \sum_{i=1}^{n} I(R_i = 1, A_i = a)\frac{1 - \widehat{p}(X_i)}{\widehat{p}(X_i)\widehat{e}_a(X_i)}Y_i,$$

where $\widehat{g}_a(X), \widehat{p}(X)$, and $\widehat{e}_a(X)$ were previously defined. The estimators $\widehat{\psi}_{\mathrm{g}}(a)$ and $\widehat{\psi}_{\mathrm{w}}(a)$ can be viewed as special cases of the augmented estimator $\widehat{\psi}_{\mathrm{aug}}(a)$: we obtain $\widehat{\psi}_{\mathrm{g}}(a)$ by identically setting $\widehat{p}(X)$ to one, and we obtain $\widehat{\psi}_{\mathrm{w}}(a)$ by identically setting $\widehat{g}_a(X)$ to zero, in equation (8). It follows that $\widehat{\psi}_{\mathrm{g}}(a)$ and $\widehat{\psi}_{\mathrm{w}}(a)$ are not robust to misspecification of the models used to estimate $\widehat{g}_a(X)$, or $\widehat{p}(X) \times \widehat{e}_a(X)$, respectively (Dahabreh et al., 2020).

Model specification: All working models included main effects for the observable covariates $X_j, j = 1,2,3$. Models for the probability of participation in the trial and models for the probability of treatment included main effects for all covariates. We used logistic regression to model the probability of participating in any trial ($R = 1$ vs 0). In our simulation scenarios, when the probability of treatment (the randomization ratio) varies across trials, a logistic regression model of treatment with main effects of the covariates (and no adjustment for trial S) is not correctly specified. To avoid misspecification of the model for the probability of treatment, we used the fact that $\Pr[A = a \mid X, R = 1] = \sum_{s=1}^{m} \Pr[A = a \mid X, S = s]\Pr[S = s \mid X, R = 1]$, and modeled the probability of treatment in each trial $\Pr[A = a \mid X, S = s]$ using logistic regression (separately in each trial) and the probability of participation in a particular trial among the collection of trials $\Pr[S = s \mid X, R = 1]$ using a multinomial logistic regression model (in the collection of trials). We refer to this approach as "averaging the trial-specific treatment probabilities." For comparison, in the Appendix, we report results using a (misspecified) logistic regression model with main effects only. In our simulation setup, when the randomization ratio varies across trials, use of this model is expected to have some bias when using $\widehat{\psi}_{\mathrm{w}}(a)$, should not affect $\widehat{\psi}_{\mathrm{g}}(a)$ at all, and should only affect the precision, but not induce bias for $\widehat{\psi}_{\mathrm{aug}}(a)$,

when the outcome model is correctly specified. Outcome models were fit separately in each treatment group using linear regression to allow effect modification by all covariates.

Comparisons: We compared the bias and variance of estimators over 10,000 runs for each scenario and each treatment.

6.3 Simulation results

Tables 1 and 2 present selected results from our simulation study for scenarios with $n = 10,000$ and total trial sample size, $\sum_{s=1}^{3} n_s$, of 2000 individuals, when averaging the trial-specific treatment probabilities; complete simulation results are presented in Web Appendix F, including results for modeling the probability of treatment using logistic regression across all trials. Bias estimates for $\widehat{\psi}_{\text{aug}}(a)$ and $\widehat{\psi}_{\text{g}}(a)$ were near-zero for all estimators, regardless of how the probability of treatment was modeled; $\widehat{\psi}_{\text{w}}(a)$ also had near-zero bias except when the randomization ratio varied across trials and the probability of treatment was modeled using logistic regression fit across all trials (see Appendix Table S3). The low bias of the g-formula estimator $\widehat{\psi}_{\text{g}}(a)$ and the weighting estimator $\widehat{\psi}_{\text{w}}(a)$ (when the model for the probability of treatment was correctly specified) indirectly verify the double robustness property of the augmented estimator $\widehat{\psi}_{\text{aug}}(a)$; this is also supported by the near-zero bias of $\widehat{\psi}_{\text{aug}}(a)$ even when the randomization ratio varied across trials and the model for the probability of treatment was misspecified (Appendix Table S3). Across all scenarios, the sampling variance of the augmented estimator $\widehat{\psi}_{\text{aug}}(a)$ was larger than that of the outcome model-based estimator $\widehat{\psi}_{\text{g}}(a)$, but substantially smaller than that of the weighting estimator $\widehat{\psi}_{\text{w}}(a)$.

7. Application of the methods to the HALT-C trial

7.1 Using the HALT-C trial to emulate meta-analyses

The HALT-C trial data: The HALT-C trial (Di Bisceglie et al., 2008) enrolled 1050 patients with chronic hepatitis C and advanced fibrosis who had not responded to previous therapy and randomized them to treatment with peginterferon alfa-2a ($a = 1$) versus no treatment ($a = 0$). Patients were enrolled at 10 research centers and followed up every 3 months after randomization. Here, we used the secondary outcome of platelet count at 9 months of follow-up as the outcome of interest; we report all outcome measurements as platelets$\times 10^3$/ml. We used the following baseline covariates: baseline platelet count, age, sex, previous use of pegylated interferon, race, white blood cell count, history of injected recreational drugs, ever received a transfusion, body mass index, creatine levels, smoking status, previous use of combination therapy (interferon and ribavirin), diabetes, serum ferritin, hemoglobin, aspartate aminotransferase levels, ultrasound evidence of splenomegaly, and ever drinking alcohol. For simplicity, we restricted our analyses to patients with complete baseline covariate and outcome data ($n = 948$).

Using the HALT-C trial to emulate a meta-analysis and evaluate the proposed methods: We treated observations from the largest center with complete data in the HALT-C trial, with sample size $n_0 = 199$, as a sample from the target population, $S = R = 0$. We then

used the data from the remaining 9 centers as a collection of "trials," $\mathcal{S} = \{1, \ldots, 9\}$, with a total sample size of $\sum_{j=1}^{9} n_j = 749$. Appendix Table S5 summarizes covariate information, stratified by R; Appendix Tables S6 and S7 summarize covariate information stratified by S. Because HALT-C used 1:1 randomized allocation, all trials had the same treatment assignment mechanism. Our goal was to transport causal inferences from the nine trials to the population represented by the target center. We were able to use the randomization in $R = 0$ to empirically benchmark the methods we propose using as a reference standard the unadjusted estimate of the average treatment effect in the target population. In view of how we configured the data for this illustrative analysis, our results should not be clinically interpreted.

Methods implemented and comparisons: We implemented the augmented estimator $\widehat{\psi}_{\mathrm{aug}}(a)$, along with the g-formula estimator $\widehat{\psi}_{\mathrm{g}}(a)$ and the weighting estimator $\widehat{\psi}_{\mathrm{w}}(a)$, defined in Section 6.2. These estimators use covariate, treatment, and outcome data from the collection of trials $\mathcal{S}$ but only covariate data from the sample of the target population $R = 0$. In these analyses we specified linear regression models for the conditional expectation of the outcome in each treatment group, and logistic regression models for the probability of trial participation (in any trial) and the probability of treatment among randomized individuals. All models included main effects for the baseline covariates listed above. For comparison, we also used an unadjusted linear regression analysis to estimate the potential outcome means and average treatment effects in $R = 0$ only; this analysis is justified by randomization and uses treatment and outcome data from the target population.

To provide an assessment of the observed data independence condition in equation (5), we used analysis of covariance (ANCOVA) to compare a linear regression model that included the main effects of baseline covariates, treatment, and all possible product terms between baseline covariates and treatment, against a linear regression model that additionally included the main effect of the trial indicators and all product terms between baseline covariates, treatment, and the trial indicators.

7.2 Results in the emulated meta-analysis

Table 3 summarizes results from the emulation of meta-analyses using HALT-C trial data. All transportability estimators, which only use baseline covariate data from $R = 0$ and covariate, treatment, and outcome data from $R = 1$, produced estimates that were very similar to the benchmark estimates from the unadjusted estimator that uses treatment and outcome data from the target center $R = 0$. The three transportability estimators produced similar point estimates, which suggests that the transported inferences are not driven by model specification choices. The three transportability estimators also produced estimates similar to those of the benchmark unadjusted estimator in $R = 0$, which suggests (but does not prove) that the identifiability conditions needed for the different transportability estimators hold, at least approximately, and that model specification is approximately correct. The ANCOVA p-value of 0.425 did not indicate gross violations of the observed data implication in display (5).

8. Discussion

We have described methods to transport causal inferences from a collection of randomized trials to a target population in which baseline covariate data are available but no experimentation has been conducted. Our results provide a solution to one of the leading problems in the field of "evidence synthesis": how to use a body of evidence consisting of multiple randomized trials to draw causal inferences about the effects of the interventions investigated in the trials for a new target population of substantive interest.

Traditional approaches to evidence synthesis, including meta-analysis, focus on modeling aspects of the distribution of study effects or producing statistically optimal summaries of available data (Higgins et al., 2009; Rice et al., 2018). In general, such analyses do not produce causally interpretable estimates of the effect of well-defined interventions on any target population because the contributions of individual studies to the meta-analysis summary are weighted by the precision of the study-specific estimates, without considering the relevance of the studies to any target population or the strong assumptions needed to combine information across studies to estimate treatment effects.

In contrast, our approach explicitly targets a well-defined population chosen on scientific or policy grounds, which may be different from the populations underlying the trials. In our experience, policy-makers who use evidence syntheses to inform their decisions are not interested in the populations represented by the completed trials and almost always have a different target population in mind. Thus, the statistical methods we propose can be viewed as a form of causally interpretable meta-analysis, consistent with the conceptual framework outlined in (Sobel et al., 2017). An important aspect of our approach is the specification of the target population and the use of sample data from that population to estimate treatment effects. In practical applications, obtaining data from a target population can be challenging (Barker et al., 2021), but we expect that challenge will become easier to address with increasing availability of administrative and registry data from populations of clinical or policy relevance. Another important aspect of our approach is the explicit statement of the assumptions needed for the causal interpretation of estimates produced by the estimators we propose. Though the assumptions are fairly strong even for the simple case we study here (point treatments with complete adherence and outcomes assessed at the end of the study), our approach naturally connect with the broad literature on causal inference and thus extensions to address time varying treatments, incomplete adherence, and longitudinal or failure-time outcomes are possible. We believe that making assumptions explicit allows reasoning about them and can be the basis of future work on sensitivity analyses. Future work may also address differential covariate measurement error in the randomized trials and the sample of the target population, systematically missing data (e.g., when some covariates are not collected in some trials), and variable selection methods (e.g., methods incorporating the independence in display (4)).

Our methods can are a generalization of methods for extending inferences from a single trial to a target population (e.g., Cole and Stuart (2010); Dahabreh et al. (2018, 2020); Rudolph and van der Laan (2017)). They also relate to the growing literature on matching-adjusted indirect comparison (e.g., Signorovitch et al. (2010, 2012); Cheng et al. (2020); Jackson et

al. (2021)). This literature has focused on the common situation where data on the target population are only available in summary form (e.g., estimates of the first moments of the covariates in X), typically does not use explicit causal (counterfactual) notation (with some recent exceptions, e.g., Cheng et al. (2020)), relies heavily on parametric assumptions (or other investigator-imposed constraints), and only applies to a single source trial (not multiple trials in meta-analysis) (Jackson et al., 2021). Our identification results suggest obvious extensions of matching-adjusted indirect comparison methods to the multiple trial case, provided investigators are willing to make the necessary causal assumptions. Conversely, methods from the literature on matching-adjusted indirect comparison and related matching methods (Jackson et al., 2021) can be used to relax the requirement of individual participant data from the target population inherent in our approach (at the cost of reliance on additional parametric assumptions or reduced efficiency). It may also be interesting to combine our approach with recent extensions of meta-regression methods to incorporate matching-adjusted indirect comparisons (Phillippo et al., 2020).

We caution that our methods will be most useful when the available trials are quite similar (e.g., in terms of treatments, outcomes, and follow-up protocols), as may be the case in prospective research programs investigating a particular intervention or close replication efforts. It should come as no surprise that a strong degree of similarity is needed in order for a "pooled" analysis to produce estimates with a clear causal interpretation with respect to a particular target population; after all, our approach focuses on differences in participant characteristics across the populations underlying the trials and the target population, not variations in treatment. Extensions of our approach to address variations in the randomized treatments, as well as non-randomized co-interventions, would be valuable. We suggest, however, that summarizing bodies of evidence much larger and diverse than a few closely related studies, particularly when individual participant data are only available from a minority of sources, is best undertaken with a descriptive or predictive attitude, focusing on modeling the response surface of study effects (Rubin, 1992), rather than on obtaining a single causally interpretable estimate. That said, we sincerely hope that future work will find ways to improve on this rather unambitious view.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

References

Barker DH, Dahabreh IJ, Steingrimsson JA, Houck C, Donenberg G, DiClemente R, and Brown LK (2021). Causally interpretable meta-analysis: Application in adolescent hiv prevention. Prevention Science pages 1–12.

Bickel PJ, Klaassen CA, Wellner JA, and Ritov Y (1993). Efficient and adaptive estimation for semiparametric models. Johns Hopkins University Press Baltimore.

Breslow NE, Robins JM, Wellner JA, et al. (2000). On the semi-parametric efficiency of logistic regression under case-control sampling. Bernoulli 6, 447–455.

Breslow NE and Wellner JA (2007). Weighted likelihood for semiparametric models and two-phase stratified samples, with application to Cox regression. Scandinavian Journal of Statistics 34, 86–102.

Cheng D, Ayyagari R, and Signorovitch J (2020). The statistical performance of matching-adjusted indirect comparisons: Estimating treatment effects with aggregate external control data. The Annals of Applied Statistics 14, 1806–1833.

Cole SR and Stuart EA (2010). Generalizing evidence from randomized clinical trials to target populations: the ACTG 320 trial. American Journal of Epidemiology 172, 107–115. [PubMed: 20547574]

Cooper H, Hedges LV, and Valentine JC (2009). The handbook of research synthesis and meta-analysis. Russell Sage Foundation.

Dahabreh I, Petito L, Robertson S, Hernán M, and Steingrimsson J (2020). Toward causally interpretable meta-analysis: Transporting inferences from multiple randomized trials to a new target population. Epidemiology (Cambridge, Mass.).

Dahabreh IJ and Hernán MA (2019). Extending inferences from a randomized trial to a target population. European Journal of Epidemiology 34, 719–722. [PubMed: 31218483]

Dahabreh IJ, Robertson SE, and Hernán MA (2019). On the relation between g-formula and inverse probability weighting estimators for generalizing trial results. Epidemiology (Cambridge, Mass.) 30, 807–812. [PubMed: 31517670]

Dahabreh IJ, Robertson SE, Steingrimsson JA, Stuart EA, and Hernán MA (2020). Extending inferences from a randomized trial to a new target population. Statistics in Medicine 39, 1999–2014. [PubMed: 32253789]

Dahabreh IJ, Robertson SE, Tchetgen Tchetgen EJ, Stuart EA, and Hernán MA (2018). Generalizing causal inferences from individuals in randomized trials to all trial-eligible individuals. Biometrics 75, 685–694.

Dahabreh IJ, Robins JM, Haneuse SJ-P, and Hernán MA (2019). Generalizing causal inferences from randomized trials: counterfactual and graphical identification. arXiv preprint arXiv:1906.10792.

Dawid AP (1979). Conditional independence in statistical theory. Journal of the Royal Statistical Society: Series B (Methodological) 41, 1–15.

Di Bisceglie AM, Shiffman ML, Everson GT, Lindsay KL, Everhart JE, Wright EC, Lee WM, Lok AS, Bonkovsky HL, Morgan TR, et al. (2008). Prolonged therapy of advanced chronic hepatitis C with low-dose peginterferon. New England Journal of Medicine 359, 2429–2441. [PubMed: 19052125]

Hájek J (1971). Comment on "An essay on the logical foundations of survey sampling by D. Basu". In Godambe VP and Sprott DA, editors, Foundations of statistical inference, page 236. Holt, Rinehart, and Winston, New York City, NY.

Higgins JP, Thompson SG, and Spiegelhalter DJ (2009). A re-evaluation of random-effects meta-analysis. Journal of the Royal Statistical Society: Series A (Statistics in Society) 172, 137–159. [PubMed: 19381330]

Jackson D, Rhodes K, and Ouwens M (2021). Alternative weighting schemes when performing matching-adjusted indirect comparisons. Research Synthesis Methods 12, 333–346. [PubMed: 33131206]

Kennedy EH, Sjölander A, and Small D (2015). Semiparametric causal inference in matched cohort studies. Biometrika 102, 739–746.

Kosorok MR (2008). Introduction to empirical processes and semiparametric inference. Springer.

Luedtke A, Carone M, and van der Laan MJ (2019). An omnibus non-parametric test of equality in distribution for unknown functions. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 81, 75–99. [PubMed: 31024219]

Petersen ML, Porter KE, Gruber S, Wang Y, and van der Laan MJ (2012). Diagnosing and responding to violations in the positivity assumption. Statistical Methods in Medical Research 21, 31–54. [PubMed: 21030422]

Phillippo DM, Dias S, Ades A, Belger M, Brnabic A, Schacht A, Saure D, Kadziola Z, and Welton NJ (2020). Multilevel network meta-regression for population-adjusted treatment comparisons. Journal of the Royal Statistical Society: Series A (Statistics in Society) 183, 1189–1210. [PubMed: 32684669]

Racine JS, Hart J, and Li Q (2006). Testing the significance of categorical predictor variables in nonparametric regression models. Econometric Reviews 25, 523–544.

Rice K, Higgins J, and Lumley T (2018). A re-evaluation of fixed effect (s) meta-analysis. Journal of the Royal Statistical Society: Series A (Statistics in Society) 181, 205–227.

Robertson SE, Steingrimsson JA, and Dahabreh IJ (2021). Using numerical methods to design simulations: revisiting the balancing intercept. American Journal of Epidemiology.

Robins J, Li L, Tchetgen Tchetgen E, and van der Vaart A (2008). Higher order influence functions and minimax estimation of nonlinear functionals. In Probability and statistics: essays in honor of David A. Freedman, pages 335–421. Institute of Mathematical Statistics.

Robins JM (1988). Confidence intervals for causal parameters. Statistics in Medicine 7, 773–785. [PubMed: 3406605]

Robins JM and Greenland S (2000). Causal inference without counterfactuals: comment. Journal of the American Statistical Association 95, 431–435.

Robins JM, Rotnitzky A, and Scharfstein DO (2000). Sensitivity analysis for selection bias and unmeasured confounding in missing data and causal inference models. In Statistical models in epidemiology, the environment, and clinical trials, pages 1–94. Springer.

Robins JM, Sued M, Lei-Gomez Q, and Rotnitzky A (2007). Comment: Performance of double-robust estimators when "inverse probability" weights are highly variable. Statistical Science 22, 544–559.

Rubin DB (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. Journal of Educational Psychology 66, 688–701.

Rubin DB (1986). Statistics and causal inference: Comment: Which ifs have causal answers. Journal of the American Statistical Association 81, 961–962.

Rubin DB (1992). Meta-analysis: literature synthesis or effect-size surface estimation? Journal of Educational Statistics 17, 363–374.

Rubin DB (2010). Reflections stimulated by the comments of Shadish (2010) and West and Thoemmes (2010). Psychological Methods 15, 38–46. [PubMed: 20230101]

Rudolph KE and van der Laan MJ (2017). Robust estimation of encouragement design intervention effects transported across sites. Journal of the Royal Statistical Society. Series B (Statistical Methodology) 79, 1509–1525. [PubMed: 29375249]

Saegusa T and Wellner JA (2013). Weighted likelihood estimation under two-phase sampling. Annals of Statistics 41, 269–295. [PubMed: 24563559]

Signorovitch JE, Sikirica V, Erder MH, Xie J, Lu M, Hodgkins PS, Betts KA, and Wu EQ (2012). Matching-adjusted indirect comparisons: a new tool for timely comparative effectiveness research. Value in Health 15, 940–947. [PubMed: 22999145]

Signorovitch JE, Wu EQ, Andrew PY, Gerrits CM, Kantor E, Bao Y, Gupta SR, and Mulani PM (2010). Comparative effectiveness without head-to-head trials. Pharmacoeconomics 28, 935–945. [PubMed: 20831302]

Sobel M, Madigan D, and Wang W (2017). Causal inference for meta-analysis and multi-level data structures, with application to randomized studies of Vioxx. Psychometrika 82, 459–474. [PubMed: 27388291]

Stefanski LA and Boos DD (2002). The calculus of m-estimation. The American Statistician 56, 29–38.

Tsiatis A (2007). Semiparametric theory and missing data. Springer Science & Business Media.

Van der Vaart AW (2000). Asymptotic statistics, volume 3. Cambridge University Press.

van der Vaart AW and Wellner JA (1996). Weak Convergence and Empirical Processes. Springer.

VanderWeele TJ (2009). Concerning the consistency assumption in causal inference. Epidemiology 20, 880–883. [PubMed: 19829187]

Vo T-T, Porcher R, Chaimani A, and Vansteelandt S (2019). A novel approach for identifying and addressing case-mix heterogeneity in individual participant data meta-analysis. Research Synthesis Methods 10, 582–596. [PubMed: 31682071]

Vo T-T, Porcher R, and Vansteelandt S (2021). Assessing the impact of case-mix heterogeneity in individual participant data meta-analysis: Novel use of I^2 statistic and prediction interval. Research Methods in Medicine & Health Sciences 2, 12–30.

Westreich D, Edwards JK, Lesko CR, Stuart E, and Cole SR (2017). Transportability of trial results using inverse odds of sampling weights. American Journal of Epidemiology 186, 1010–1014. [PubMed: 28535275]

Table 1:

Bias estimates based on 10,000 simulation runs; selected sample size scenarios; the probability of treatment in the collection of trials was estimated by averaging the trial-specific treatment probabilities.

a	n	$\sum_{s=1}^{3} n_s$	Balanced	TxAM varies	$\widehat{\psi}_{\text{aug}}(a)$	$\widehat{\psi}_{\text{g}}(a)$	$\widehat{\psi}_{\text{w}}(a)$
1	10000	2000	Yes	No	0.0022	0.0003	−0.0011
1	10000	2000	Yes	Yes	0.0004	−0.0003	−0.0026
1	10000	2000	No	No	−0.0006	−0.0001	−0.0082
1	10000	2000	No	Yes	0.0011	0.0002	−0.0067
1	100000	2000	Yes	No	−0.0003	0.0002	−0.0104
1	100000	2000	Yes	Yes	−0.0002	0.0008	0.0102
1	100000	2000	No	No	−0.0012	−0.0002	−0.0074
1	100000	2000	No	Yes	0.0009	0.0013	0.0039
0	10000	2000	Yes	No	0.0001	0.0008	0.0083
0	10000	2000	Yes	Yes	0.0006	0.0004	0.0060
0	10000	2000	No	No	0.0010	−0.0003	0.0002
0	10000	2000	No	Yes	0.0002	0.0005	0.0106
0	100000	2000	Yes	No	0.0013	0.0005	0.0016
0	100000	2000	Yes	Yes	−0.0009	−0.0000	0.0034
0	100000	2000	No	No	−0.0009	−0.0003	0.0064
0	100000	2000	No	Yes	0.0012	0.0006	0.0109

In the column titled *Balanced*, *Yes* denotes scenarios in which the trials had on average equal sample sizes; *No* denotes scenarios with unequal trial sample sizes. In the column titled *TxAM varies*, *Yes* denotes scenarios in which the treatment assignment mechanism varied across trials; *No* denotes scenarios in which the mechanism did not vary.

Table 2:

Variance estimates based on 10,000 simulation runs; the probability of treatment in the collection of trials was estimated by averaging the trial-specific treatment probabilities.

a	n	$\sum_{s=1}^{3} n_s$	Balanced	TxAM varies	$\widehat{\psi}_{\text{aug}}(a)$	$\widehat{\psi}_{\text{g}}(a)$	$\widehat{\psi}_{\text{w}}(a)$
1	10000	2000	Yes	No	0.0105	0.0038	0.2893
1	10000	2000	Yes	Yes	0.0084	0.0035	0.1945
1	10000	2000	No	No	0.0101	0.0038	0.2167
1	10000	2000	No	Yes	0.0088	0.0038	0.1925
1	100000	2000	Yes	No	0.0144	0.0039	0.3316
1	100000	2000	Yes	Yes	0.0120	0.0035	0.3610
1	100000	2000	No	No	0.0143	0.0039	0.4028
1	100000	2000	No	Yes	0.0128	0.0037	0.2925
0	10000	2000	Yes	No	0.0103	0.0039	0.1165
0	10000	2000	Yes	Yes	0.0131	0.0042	0.1509
0	10000	2000	No	No	0.0101	0.0039	0.1136
0	10000	2000	No	Yes	0.0118	0.0040	0.1551
0	100000	2000	Yes	No	0.0151	0.0039	0.1873
0	100000	2000	Yes	Yes	0.0198	0.0043	0.2699
0	100000	2000	No	No	0.0146	0.0038	0.1663
0	100000	2000	No	Yes	0.0218	0.0040	0.3243

In the column titled *Balanced, Yes* denotes scenarios in which the trials had on average equal sample sizes; *No* denotes scenarios with unequal trial sample sizes. In the column titled *TxAM varies, Yes* denotes scenarios in which the treatment assignment mechanism varied across trials; *No* denotes scenarios in which the mechanism did not vary.

Table 3:

Results from analyses using the HALT-C trial data.

Estimator	$a = 1$	$a = 0$	Mean difference
$R = 0$ (benchmark)	121.6 (111.0, 133.3)	164.4 (151.2, 178.1)	−42.8 (−60.2, −25.7)
$\widehat{\psi}_{\mathrm{aug}}(a)$	124.5 (116.3, 133.5)	167.1 (157.3, 177.7)	−42.6 (−51.2, −34.4)
$\widehat{\psi}_{\mathrm{g}}(a)$	124.6 (116.6, 133.0)	167.1 (157.3, 177.3)	−42.5 (−50.5, −34.9)
$\widehat{\psi}_{\mathrm{w}}(a)$	123.2 (113.6, 136.0)	165.9 (152.6, 182.9)	−42.7 (−60.1, −26.2)

$R = 0$ denotes analyses using data only from the sample of the target population. Numbers in parentheses are 95% quantile-based confidence intervals from 10,000 bootstrap samples; 95% normal confidence intervals from the sandwich estimator were similar.