

# Randomized trials and their observational emulations: a framework for benchmarking and joint analysis

Issa J. Dahabreh<sup>1-3</sup>, Jon A. Steingrimsón<sup>4</sup>, James M. Robins<sup>1-3</sup>, and Miguel A. Hernán<sup>1-3,5</sup>

<sup>1</sup>CAUSALab, Harvard T.H. Chan School of Public Health, Boston, MA USA

<sup>2</sup>Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, MA

<sup>3</sup>Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA

<sup>4</sup>Department of Biostatistics, School of Public Health, Brown University, Providence, RI

<sup>5</sup>Harvard-MIT Division of Health Sciences and Technology, Boston, MA

Tuesday 29<sup>th</sup> March, 2022

**Running head:** Framework for benchmarking and joint analysis

**Type of manuscript:** Original Research Article.

**Conflicts of interest:** None declared.

**Data and computing code availability:** Not applicable.

**Word count:** abstract = 248; main text  $\approx$  4000.

**Abbreviations that appear in the text:** No abbreviations.

## Abstract

A randomized trial and an analysis of observational data designed to emulate the trial sample observations separately, but have the same eligibility criteria, collect information on some shared baseline covariates, and compare the effects of the same treatments on the same outcomes. Treatment effect estimates from the trial and its emulation can be compared to benchmark observational analysis methods. In a simplified setting with complete adherence to the assigned treatment strategy and no loss-to-follow-up, we show that benchmarking relies on an exchangeability condition between the populations underlying the trial and its emulation, to account for differences in the distribution of covariates between them. When this exchangeability condition holds, and the usual conditions needed for the estimates from the trial and its emulation to have a causal interpretation also hold, we derive restrictions on the law of the observed data. When the data are compatible with the restrictions, joint analysis of the trial and its emulation is possible. When the data are incompatible with the restrictions, a discrepancy between (1) estimates based on extending inferences from the trial to the population underlying the emulation and (2) the emulation itself may reflect either inability to benchmark (e.g., due to selective participation into the trial) or a failure of the emulation (e.g., due to unmeasured confounding), but we cannot use the data to determine which is the case. Our analysis reveals how benchmarking attempts combine causal assumptions, data analysis methods, and substantive knowledge to examine the validity of observational analysis methods.

**Keywords:** randomized trials; observational analyses; target trial emulation; transportability; benchmarking.

# Introduction

Consider an index randomized trial and an analysis of observational data designed to emulate a target trial as similar as possible to the index trial [1]. The trial and the observational emulation sample observations separately, but have the same eligibility criteria, collect information on some shared baseline covariates, compare the same treatment recommendations (assigned randomly in the trial but not in the observational emulation), and examine the same outcomes [2,3]. Because of the commonalities of the two studies, it is natural to want to compare their results and jointly use their data to learn about the effects of the intervention.

Causal inference from observational data is typically considered more speculative than causal inference in trials due to the possibility of baseline confounding by unmeasured variables in the absence of randomization. Agreement between the results of a trial and its emulation might then be taken to indicate that the methods in the emulation have in some sense been “successful” in producing valid causal inferences [1,4]. We refer to the comparison of the results of the trial and the emulation as benchmarking the observational analyses [3]. By benchmarking in cases where data are conveniently available from pragmatic trials, we can develop confidence in the reliability of observational methods when used in settings where no trial data are available [1]. For example, if several benchmarking attempts in some clinical domain suggest that observational analyses successfully emulate the results of large well-run pragmatic trials, we might be willing to trust the results from similar observational analyses for treatments that have not been investigated in trials, or for follow-up times and patient subgroups for which the trial results are inadequate (e.g., due to short follow-up duration or limited sample size) [5].

Here, we examine a simplified setting with complete adherence to treatment and no loss to follow-up to show that benchmarking relies on an exchangeability assumption between the populations underlying the trial and its emulation to account for differences in the distribution of covariates between them and to allow transportability of inferences from the

trial to the population underlying the emulation. We show that, when this exchangeability condition holds, and the “usual” conditions needed for treatment effect estimates from the trial and its emulation to have a causal interpretation also hold [6], they imply restrictions for the observed data distribution, allowing investigators to examine whether the data are compatible with the assumptions. When the data are incompatible with these restrictions, a discrepancy between (1) estimates based on extending inferences from the trial to the population underlying the emulation and (2) the emulation itself may reflect either inability to benchmark (e.g., due to selective participation into the trial) or a failure of the emulation (e.g., due to unmeasured confounding), but we cannot use the data to decide which is the case. We also show that if the exchangeability conditions hold, joint analysis to estimate potential outcome means and treatment effects is feasible and can be more efficient than either relying on just the observational analysis or the transportability analysis from the trial to population represented by the emulation.

## Setup and causal estimands

Suppose that a trial and an analysis of observational data designed to emulate the trial measure some shared baseline (pre-treatment) covariates  $X$ , compare the same treatment strategies  $A$  (randomly assigned in the trial but not in the emulation), and examine the same outcome  $Y$ , measured at a single post-treatment time-point [2]. We use  $S$  as the binary indicator for participation status ( $S = 1$  for the trial and  $S = 0$  for the emulation).

The trial’s eligibility criteria define the actual (finite) population of all trial-eligible individuals. We view the actual population as a simple random sample from a near-infinite underlying super-population. Typically, the actual population is not enumerated and thus the individuals who participate in the trial represent an unknown fraction of all trial-eligible individuals [2]. For simplicity, we assume that the observational emulation is conducted among a simple random sample of trial-eligible non-randomized individuals from the actual

population, with unknown sampling probability (this is the sampling scheme of non-nested trial designs [2, 7]). The simple random sampling assumption can be relaxed when at least partial information is available about the selection of non-randomized individuals [2, 7]; we do not consider such situations to maintain focus on concepts related to benchmarking.

The data from the trial are independent realizations of  $(X_i, S_i = 1, A_i, Y_i)$ ,  $i = 1, \dots, n_1$ , where  $n_1$  is the total number of trial participants; the data from the emulation are independent realizations of  $(X_i, S_i = 0, A_i, Y_i)$ ,  $i = 1, \dots, n_0$ , where  $n_0$  is the total number of participants in the emulation. The total sample size is  $n = n_0 + n_1$ . In non-nested trial designs sampling randomized and non-randomized individuals with different sampling fractions induces a biased sampling model [8, 9], in the sense that the relative sample size of the trial and the emulation,  $n_1/(n_0 + n_1)$ , does not necessarily reflect the population proportion of trial participation [2, 7]. Thus, the probability of trial participation  $\Pr[S = 1]$  is not identifiable in non-nested trial designs, even though the densities  $f(x|S = s)$  are identifiable for  $s = 0, 1$  [2].

Let  $Y^a$  denote the counterfactual (potential) outcome under intervention that sets treatment  $A$  to  $a$  [10, 11]. We are interested in the conditional potential outcome mean in the trial  $E[Y^a|X, S = 1]$  and the emulation  $E[Y^a|X, S = 0]$ , as well as the corresponding marginal (population-averaged) potential outcomes means  $E[Y^a|S = 1]$  and  $E[Y^a|S = 0]$ . Under separate sampling into the trial and the emulation, it is not possible to identify  $E[Y^a]$  without additional information on how individuals are sampled into the two studies; such information is usually unavailable except when the trial is nested in a broader cohort of trial-eligible individuals (see references [2, 12] for details).

## Identification in the trial and its emulation

**Identifiability conditions:** To focus on benchmarking concepts, we consider a trial and its emulation in the idealized case of complete adherence to assigned treatment strategies,

no loss-to-follow-up, and no measurement error (our approach can be generalized to address these complications).

We now list sufficient conditions under which the potential outcome means conditional on  $X$  under intervention that sets treatment  $A$  to  $a$  in the populations underlying the trial and its emulation are identifiable. Throughout this paper we take identifiability conditions that involve counterfactuals as primitive conditions (i.e., not derived) to focus on issues of design, analysis, and interpretation. We note, however, that the conditions can be derived from structural equation models for the data generating mechanism. In particular, we have recently discussed [3, 13] how the distributional exchangeability conditions invoked throughout the paper can be derived from non-parametric structural equation models with a finest fully randomized causally interpretable structured tree graph error structure, represented using causal directed acyclic graphs and single-world intervention graphs [14] (see also references [15, 16] for an alternative approach based on selection diagrams).

(i) *No data source effects and consistency of potential outcomes:* if  $A_i = a$ , then  $Y_i^a = Y_i$ , for every treatment strategy  $a$  and unit  $i$ , regardless of trial participation status.

(ii) *Conditional exchangeability over  $A$  in the trial:* for every treatment strategy  $a$ ,  $Y^a \perp\!\!\!\perp A | (X, S = 1)$ .

(iii) *Positivity of treatment in the trial:* for every  $x$  with  $f(x, S = 1) \neq 0$  and every treatment strategy  $a$ ,  $\Pr[A = a | X = x, S = 1] > 0$ .

(iv) *Conditional exchangeability over  $A$  in the emulation:* for every treatment strategy  $a$ ,  $Y^a \perp\!\!\!\perp A | (X, S = 0)$ .

(v) *Positivity of treatment in the emulation:* for every  $x$  with  $f(x, S = 0) \neq 0$  and every treatment strategy  $a$ ,  $\Pr[A = a | X = x, S = 0] > 0$ .

Condition (i) contains an exclusion restriction assumption that participation in the trial does not affect the outcome except through treatment assignment [13], and may be vi-

olated when trial participation has direct effects on the outcome (e.g., through ancillary non-protocol directed treatments or Hawthorne effects). Condition (i) also requires the interventions to assign treatment to be well-defined, and may be violated if there exist multiple outcome-relevant versions of treatment, especially if some versions are not available outside the experimental setting [17–19]. Conditions (ii) and (iii) are expected to hold because of randomization. Note that if the trial is marginally randomized, the exchangeability condition  $(Y^a, X) \perp\!\!\!\perp A|S = 1$  will hold; this condition implies but is not implied by condition (ii) and we use the weaker condition (ii) in the remainder of the paper. Condition (iv) is a strong untestable assumption about the emulation and needs to be evaluated in light of substantive background knowledge on a case-by-case basis; condition (v) is in principle testable, but empirical assessment can be challenging when  $X$  is high-dimensional [20].

**Identification:** In the trial, using conditions (i), (ii), and (iii) [6], we obtain

$$E[Y^a|X, S = 1] = E[Y^a|X, S = 1, A = a] = E[Y|X, S = 1, A = a], \quad (1)$$

and, by the law of total expectation,  $E[Y^a|S = 1] = E[E[Y|X, S = 1, A = a]|S = 1]$ .

In the emulation, using conditions (i), (iv), and (v) [6], we obtain

$$E[Y^a|X, S = 0] = E[Y^a|X, S = 0, A = a] = E[Y|X, S = 0, A = a]. \quad (2)$$

By the law of total expectation, the above result implies that the potential outcome mean under treatment assignment  $a$  in the population underlying the emulation,  $E[Y^a|S = 0]$ , is identified by the observed data functional

$$\phi(a) \equiv E[E[Y|X, S = 0, A = a]|S = 0]. \quad (3)$$

Nothing in the above results suggests that the marginal or conditional potential outcome

means in the two studies are related in any way and, in general, we should expect  $E[Y^a|S = 1] \neq E[Y^a|S = 0]$  and, consequently,  $E[E[Y|X, S = 1, A = a]|S = 1] \neq E[E[Y|X, S = 0, A = a]|S = 0]$ , even if conditions (i) through (v) hold. Up to this point in our exposition, the results simply pertain to the different populations underlying the trial and the emulation.

## Benchmarking and joint analysis

In order to connect the causal quantities in the trial and its emulation we can use concepts from the emerging literature on extending (generalizing or transporting) causal inferences from trials to a new population [3, 21]. Suppose that we believe that selection into the trial is independent of potential outcomes, given baseline covariates, so that we can re-calibrate the trial results to the population underlying the emulation (similar ideas have appeared in references [22–24]). More formally, suppose that we are also willing to assume the following two conditions:

(vi) *Conditional exchangeability over S*: for every treatment strategy  $a$ ,  $Y^a \perp\!\!\!\perp S|X$ .

(vii) *Positivity of participation*: for every  $x$  with positive density  $f(x) \neq 0$  and every  $s$ ,  $\Pr[S = s|X = x] > 0$ .

Note that condition (vi) essentially requires the baseline covariates  $X$  to be adequate for addressing selective participation in the trial [25], in addition to being sufficient for addressing confounding in the emulation. For simplicity, here, we assume that the set of baseline covariates in conditions (vi) and (vii) is the same as in conditions (iv) and (v). In the Appendix, we consider the slightly more complicated case where a different set of covariates is needed for for each pair of conditions.

## Benchmarking

We will now argue that conditions  $(vi)$  and  $(vii)$ , when combined with conditions  $(i)$  through  $(v)$ , impose restrictions on the law of the observed data and allow the formal benchmarking of observational analysis methods against trials.

**Restrictions on the law of the observed data:** Under conditions  $(vi)$  and  $(vii)$  we have  $E[Y^a|X, S = 1] = E[Y^a|X, S = 0]$ , that is, the far left-hand-sides in equations (1) and (2) are equal. Combined with conditions  $(i)$  through  $(v)$ , this result implies that the far right-hand-sides of equations (1) and (2) are also equal,  $E[Y|X, S = 1, A = a] = E[Y|X, S = 0, A = a]$ . Informally, *if* the conditions needed for *both* the trial and emulation results to have a causal interpretation, *and if* results are transportable between the populations underlying the two studies, *then* the conditional (observed) outcome mean among randomized individuals assigned to treatment  $a$  in the trial has to equal the conditional outcome mean among individuals who chose to receive treatment  $a$  in the emulation.

Because the equality  $E[Y|X, S = 1, A = a] = E[Y|X, S = 0, A = a]$  only involves the observed variables, investigators can examine whether it is compatible with the data. Investigators can use various qualitative or quantitative approaches to examine whether this implication of the assumptions is compatible with the data [26–28]. A detailed discussion of such approaches is beyond the scope of this paper; however, we advise against over-reliance on formal statistical testing of the observed data implications of the identifiability conditions. We also note that under our identifiability conditions, when the outcome is not binary, it is possible to deduce restrictions on the law of the observed data that are stronger than the equality of the two conditional expectations. Specifically, in the Appendix, we show that exchangeability conditions  $(ii)$ ,  $(iv)$ , and  $(vi)$ , together with the consistency condition  $(i)$ , imply that  $Y \perp\!\!\!\perp S|(X, A = a)$  for every treatment  $a$ . Thus, the distribution of the observed outcome  $Y$  is independent of  $S$ , conditional on  $X$ , within each level of treatment  $A$ . In

other words, the implications of the conditions extend beyond mean independence, to independence of the entire distribution of the outcome from the indicator for trial participation, conditional on baseline covariates and within levels of treatment. Here, we only consider mean independence.

**The formal logic of benchmarking:** Under conditions (i), (ii), (iii), (vi), and (vii) the potential outcome mean in the population underlying the emulation,  $E[Y^a|S = 0]$ , can be identified by the observed data functional

$$\chi(a) \equiv E[E[Y|X, S = 1, A = a]|S = 0]. \quad (4)$$

Note that the right-hand-sides of displays (3) and (4) are different and the equality in each of these displays depends on different sets of conditions – only condition (i) is needed for both results.

These identification results provide the formal logic of benchmarking: under conditions (i), (iv), and (v),  $\phi(a)$  can be given a causal interpretation as  $E[Y^a|S = 0]$ ; under conditions (i), (ii), (iii), (vi), and (vii), a different quantity,  $\chi(a)$ , has the same interpretation. Thus, to the extent that the data can be used to estimate both  $\phi(a)$  and  $\chi(a)$ , when the estimates are substantially different (i.e., beyond what would be expected by sampling variability) we might reasonably conclude that at least one of the conditions (i) through (vii) are violated.

Conversely, if estimates of  $\chi(a)$  are approximately equal to those of  $\phi(a)$  it is reasonable to think that the conditions are not grossly violated. This is a sensible inference, though not logically necessary: It is possible to have  $\phi(a) \approx \chi(a)$  when  $E[Y^a|X, S = 0, A = a] = E[Y^a|X, S = 0]$  and  $E[Y^a|X, S = 1] = E[Y^a|X, S = 0]$  hold, even if the distributional exchangeability conditions are violated. Furthermore, even if  $E[Y^a|X, S = 0, A = a] \neq E[Y^a|X, S = 0]$  or  $E[Y^a|X, S = 1] \neq E[Y^a|X, S = 0]$ , it is still possible to have  $\chi(a) \approx \phi(a)$  because these conditional expectations are averaged over the covariate distribution of the

population underlying the emulation (so differences can “cancel-out”). Furthermore, it is possible, though unlikely, that multiple conditions fail in a way that still leads to  $\chi(a) \approx \phi(a)$ .

In the section of this paper on estimation, we show how data from a trial and its emulation can be used to efficiently and robustly estimate  $\phi(a)$  and  $\chi(a)$  and how to use the estimates for benchmarking purposes.

**The heuristic logic of benchmarking, revisited:** We now revisit the logic of benchmarking, sketched in the Introduction, in view of our formal results. To the extent that condition *(vi)* is deemed plausible, the leading (but, as discussed in the preceding section, not the only possible) explanation for  $\chi(a)$  to be different from  $\phi(a)$  is violation of condition *(iv)*. If estimates of these two quantities are similar we might believe that in some sense the methods in the observational emulation were successful. Such informal benchmarking of observational analyses is methodologically interesting [4], as evidenced by numerous meta-epidemiological studies comparing independently conducted trial and observational analyses (e.g., [29–33]) and systematic attempts to emulate trials using observational data [34]. It is also substantively important because it allows investigators to develop trust in observational analyses in order to use them in contexts where trial data are not available. We note, however, that if condition *(vi)* is violated, as would be the case, for instance, if the populations underlying the trial and its emulation differ with respect to an unmeasured important outcome predictor, estimates of  $\chi(a)$  and  $\phi(a)$  may disagree, even in the absence of unmeasured confounding in the emulation.

Because of the multitude of assumptions involved in benchmarking, disagreement between a trial and its emulation will not necessarily mean that the emulation produced estimates that do not have a causal interpretation. To see this suppose that conditions *(i)* through *(iii)*, *(v)*, and *(vii)* all hold. In that situation, the only assumptions in question pertain to the critical exchangeability conditions *(iv)*  $Y^a \perp\!\!\!\perp A|(X, S = 0)$  (i.e., no confounding in the population underlying the emulation) and *(vii)*  $Y^a \perp\!\!\!\perp S|X$  (i.e., transportability among the

two populations). In Table 1 we summarize how the possible truth values of these conditions relate to the observed data independence condition  $Y \perp\!\!\!\perp S|(X, A = a)$ , and for interpreting the results of the emulation and benchmarking.

Table 1: Truth values of conditions  $(iv)$   $Y^a \perp\!\!\!\perp A|(X, S = 0)$  and  $(vi)$   $Y^a \perp\!\!\!\perp S|X$  and their implications for the observed data independence  $Y \perp\!\!\!\perp S|(X, A = a)$  and for interpreting benchmarking results, provided conditions  $(i)$  through  $(iii)$ ,  $(v)$ , and  $(vii)$  hold.

$(iv)$	$(vi)$	$Y \perp\!\!\!\perp S (X, A = a)$	Interpretation		
			Do emulation estimates have a causal interpretation?	Is benchmarking possible?	Will benchmarking show agreement between the trial and its emulation?
T	T	has to hold	yes	yes	yes
F	T	does not have to hold	no	yes	no
T	F	does not have to hold	yes	no	NA
F	F	does not have to hold	no	no	NA

T = true; F = false; NA = not applicable.

Let us now discuss the practical implications of this table for benchmarking efforts. Suppose the data are consistent with the independence condition  $Y \perp\!\!\!\perp S|(X, A = a)$  (first row of the table). Then our (subjective) belief in the following three hypotheses would be strengthened: (1) conditions  $(iv)$  and  $(vi)$  both hold; (2) the emulation estimates have a causal interpretation; (3) the treatment effect is the same in the emulation as in transportability analyses from the trial to the population underlying the emulation. We are not aware of any systematic empirical assessments of independence conditions such as  $Y \perp\!\!\!\perp S|(X, A = a)$  in ongoing efforts to emulate trials using observational analyses; such assessments may be an interesting extension of ongoing research.

Next, suppose that the data are inconsistent with the independence condition  $Y \perp\!\!\!\perp S|(X, A = a)$ . The leading explanation for this finding would be a violation of condition  $(iv)$ ,  $(vi)$ , or both (bottom three rows of the table). Unfortunately, the data cannot distinguish between these three possibilities. In other words, we cannot tell whether the emulation

estimates do not have a causal interpretation and this manifests as a benchmarking discrepancy (second row: confounding in the emulation; exchangeable populations underlying the trial and the emulation); the emulation estimates have a causal interpretation but formal benchmarking is not possible (third row: no confounding in the emulation, non-exchangeable populations underlying the trial and the emulation); or the emulation estimates do not have a causal interpretation and formal benchmarking is impossible (fourth row: confounding in the emulation, non-exchangeable populations underlying the trial and the emulation). This is a somewhat sobering result for the interpretation of benchmarking attempts that uncover a discrepancy between (1) estimates based on extending inferences from the trial to the population underlying the emulation and (2) the emulation itself: the discrepancy may reflect a failure of the emulation (e.g., due to unmeasured confounding) or inability to benchmark (e.g., due to selective participation into the trial), but we cannot use the data to decide which is the case.

## Joint analysis

If conditions *(i)* through *(vii)* all hold, then it is clear that we can identify  $E[Y^a|S = 0]$  using either  $\phi(a)$  or  $\chi(a)$ . We now argue that a third identification result is also available. As we saw above, when conditions *(i)* through *(vii)* hold,

$$E[Y|X, S = 1, A = a] = E[Y|X, S = 0, A = a] = E[Y|X, A = a].$$

Consequently, if conditions *(i)* through *(vii)* hold, we can identify  $E[Y^a|S = 0]$  by the observed data functional

$$\psi(a) \equiv E [ E[Y|X, A = a] | S = 0 ].$$

This result suggests that we can completely pool the data from the trial and its emulation

when modeling the conditional outcome mean (because  $E[Y|X, A = a]$  in the formula does not condition on trial participation status  $S$ ) but still obtain inferences about the population underlying the emulation, which may be more representative of routine clinical practice (by standardizing the conditional outcome mean over the distribution of covariates in  $S = 0$ ). Thus,  $\psi(a)$  provides a way to combine data from the trial and its emulation in a joint analysis that may have a more natural causal interpretation than standard meta-analyses combining estimates from trial and observational analyses [35].

## Estimation and inference

In this Section we informally describe some results about estimation and inference for benchmarking and joint analyses using a trial and its emulation. The focus here is on the intuition behind our results; additional details are provided in the Appendix.

### Benchmarking

As noted, benchmarking involves a comparison of  $\phi(a)$  versus  $\chi(a)$ . Such a comparison should ideally use efficient and robust methods to statistically compare these quantities using the data.

Because  $\phi(a)$  is a version of the well-known g-formula identification result for observational analyses, we can use a well-known doubly robust estimator to estimate it [36]; we give a formula for this estimator in the Appendix and denote it as  $\hat{\phi}(a)$ . The estimator depends on estimating the probability of treatment in the emulation,  $\Pr[A = a|X, S = 0]$ , and the expectation of the outcome in each treatment group in the emulation,  $E[Y|X, S = 0, A = a]$ . The estimator  $\hat{\phi}(a)$  is doubly robust [37] in the sense that it remains consistent when either the estimator for  $\Pr[A = a|X, S = 0]$  or the estimator for  $E[Y|X, S = 0, A = a]$  is consistent (but not necessarily both).

We propose to estimate  $\chi(a)$  using an estimator for transporting inferences from the trial

to a target population [12]; we also give a formula for this estimator in the Appendix and denote it as  $\widehat{\chi}(a)$ . This estimator depends on estimating the probability of participation in the trial,  $\Pr[S = 1|X]$ ; the expectation of the outcome in each treatment group in the trial,  $E[Y|X, S = 1, A = a]$ ; and the probability of treatment in the trial,  $\Pr[A = a|X, S = 1]$ . This estimator is doubly robust in the sense that it remains consistent when either the estimator for  $\Pr[S = 1|X]$  or the estimator for  $E[Y|X, S = 1, A = a]$  is consistent (but not necessarily both); the probability of treatment among trial participants is known and thus  $\Pr[A = a|X, S = 1]$  can always be consistently estimated.

The benchmarking contrast  $\phi(a) - \chi(a)$  can be estimated by taking the difference  $\widehat{\phi}(a) - \widehat{\chi}(a)$  as the basis for formal (statistical) comparisons between the trial and its emulation. Confidence intervals for the component quantities or for the difference contrast can be obtained using the usual sandwich methods [38]; the variance of the influence curves (based on the empirical analogs of the influence functions given in the Appendix; or the bootstrap [39]).

## Joint analysis

As noted above, the estimators for benchmarking are based on previously obtained results for observational analyses with no unmeasured confounding or for transportability analyses from a trial to a new target population. In the Appendix, we propose a novel estimator for  $\psi(a)$ , which we denote as  $\widehat{\psi}(a)$ . This estimator depends on models for the probability of trial participation,  $\Pr[S = 1|X]$ ; the probability of treatment in the pooled data,  $\Pr[A = a|X]$ ; and the expectation of the outcome in each treatment group,  $E[Y|X, A = a]$  in the pooled data. This estimator is also doubly robust, in the sense that it is consistent when either the pair of estimators for  $\Pr[S = 1|X]$  and  $\Pr[A = a|X]$  are consistent, or when the estimator for  $E[Y|X, A = a]$  is consistent.

When conditions (i) through (vii) hold, all three estimators –  $\widehat{\phi}(a), \widehat{\chi}(a), \widehat{\psi}(a)$  – are potentially useful for estimating potential outcome means in the target population. If all

models are correctly specified (and converge to the “true” model at a fast enough rate), then all estimators are consistent and asymptotically normal; if models are misspecified all three estimators enjoy certain robustness properties (see Appendix). A reasonable way to choose among estimators, then, is to use the one with the lowest large-sample variance. In the Appendix, we give expressions for the large-sample variance bounds of regular estimators for each of  $\phi(a)$ ,  $\chi(a)$ , and  $\psi(a)$ , and argue that the corresponding estimators attain them when models are correctly specified. We also show that the asymptotic variance bound for  $\psi(a)$  is smaller than or equal to the asymptotic variance bounds of  $\phi(a)$  and  $\chi(a)$ . This should make intuitive sense:  $\psi(a)$ , by taking advantage of the complete set of conditions (i) through (vii), allows analysts to use the covariate, treatment, and outcome data from all observations to estimate the necessary working models; when these models are correctly specified,  $\hat{\psi}(a)$  makes the most efficient use of the data. In the Appendix, we discuss some additional technical issues that arise when estimating the various models required for  $\hat{\phi}(a)$ ,  $\hat{\chi}(a)$ ,  $\hat{\psi}(a)$ .

## Discussion

We have provided a formal description of analyses that synthesize randomized and observational data in order to benchmark methods for analyzing the observational data and, when appropriate, perform joint analyses. Our description combines ideas from the large literature on conducting observational analyses that emulate target trials [1] with ideas from the emerging literature on analyses extending inferences from a trial to a target population [21], and provides a basic conceptual framework for thinking about analyses that combine aspects of both (e.g., [23, 24]). Our framework has some parallels with an independently proposed framework for replication studies and within-study comparisons [40, 41], though the study designs and methods we describe are different.

Using trial and observational data to benchmark methods has a long history, including seminal works in econometrics [42, 43]. Using data from a trial nested in a cohort of trial

eligible individuals, including those who were not randomized (i.e., a nested trial design [2]) in a joint analysis also has a long history (e.g., see [44] and the more modern treatment, similar to our approach, in [45]). Our contribution here is the consideration of these issues when using a modern causal and statistical approach in the much more common case where the trial and its emulation are conducted independently (i.e., non-nested trial design [2]). The non-nested design is more common for benchmarking because, unlike the nested design, it does not require prospective nesting of the trial in a well-defined cohort or the ability to retrospectively link data from trial participants to data sampled from a population that also includes non-randomized individuals.

To simplify exposition, we worked in a simple setting with complete adherence to treatment and outcomes measured at the end of follow-up for all individuals. This setting is adequate for illustrating the concepts of benchmarking and emulation; results from the simple setting can be extended to more realistic settings with non-adherence, longitudinal or failure time outcomes, and loss-to-followup [13], as well as to the consideration of multiple studies jointly [35]. Thus, analogous identification results should be fairly easy to obtain in more realistic settings. Nevertheless, we expect that estimation using real data in applied analyses will prove more challenging and that practical experience with such analyses will take time to accumulate. We hope that the results herein provide a reasonable conceptual basis for developing a broader framework and using the methods in practice.

Our results suggest that, even in a simplified setting, benchmarking combines causal assumptions, data analysis methods, and substantive knowledge to build qualitative arguments about the validity of observational analysis methods. Different investigators will find such arguments convincing to varying degrees. Some, including ourselves, will be convinced when (1) the observational analysis is explicitly designed to emulate a target trial with a protocol sufficiently similar to that of the index trial; (2) the emulation is based on sufficiently rich observational data; (3) both the index trial and the emulation have the same causal estimand

and are analyzed using comparable (and appropriate) methods; and (4) the results of the observational emulation and the index trial are similar. Naturally, in each clinical domain, different investigators will require different kinds of evidence (e.g., in different populations, or conducted concurrently with trials) and different amounts of evidence (e.g., more emulations, by independent teams, using different data sources) to increase trust in observational analyses.

## References

- [1] Miguel A Hernán and James M Robins. Using Big Data to emulate a target trial when a randomized trial is not available. *American Journal of Epidemiology*, 183(8):758–764, 2016.
- [2] Issa J Dahabreh, Sebastien J-PA Haneuse, James M Robins, Sarah E Robertson, Ashley L Buchanan, Elisabeth A Stuart, and Miguel A Hernán. Study designs for extending causal inferences from a randomized trial to a target population. *American Journal of Epidemiology*, 2020 (in press).
- [3] Issa J Dahabreh, James M Robins, and Miguel A Hernán. Benchmarking observational methods by comparing randomized trials and their emulations. *Epidemiology*, 31(5):614–619, 2020.
- [4] Issa J Dahabreh and David M Kent. Can the learning health care system be educated with observational data? *JAMA*, 312(2):129–130, 2014.
- [5] Shaun P Forbes and Issa J Dahabreh. Benchmarking observational analyses against randomized trials: a review of studies assessing propensity score methods. *Journal of General Internal Medicine*, pages 1–9, 2020.
- [6] Miguel A Hernán and James M Robins. *Causal Inference: What If*. Chapman & Hall/CRC, Boca Raton, FL, 2020.
- [7] Issa J Dahabreh, Miguel A Hernán, Sarah E Robertson, Ashley Buchanan, and Jon A Steingrimsson. Generalizing trial findings in nested trial designs with sub-sampling of non-randomized individuals. *arXiv preprint arXiv:1902.06080*, 2019.
- [8] Peter J Bickel, Chris AJ Klaassen, Jon A Wellner, and Ya’acov Ritov. *Efficient and adaptive estimation for semiparametric models*. Johns Hopkins University Press Baltimore, 1993.

- [9] Norman E Breslow, James M Robins, Jon A Wellner, et al. On the semi-parametric efficiency of logistic regression under case-control sampling. *Bernoulli*, 6(3):447–455, 2000.
- [10] Donald B Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688, 1974.
- [11] James M Robins and Sander Greenland. Causal inference without counterfactuals: comment. *Journal of the American Statistical Association*, 95(450):431–435, 2000.
- [12] Issa J Dahabreh, Sarah E Robertson, Jon A Steingrímsson, Elizabeth A Stuart, and Miguel A Hernán. Extending inferences from a randomized trial to a new target population. *Statistics in Medicine*, 39(14):1999–2014, 2020.
- [13] Issa J Dahabreh, James M Robins, Sebastien J-PA Haneuse, and Miguel A Hernán. Generalizing causal inferences from randomized trials: counterfactual and graphical identification. *arXiv preprint arXiv:1906.10792*, 2019.
- [14] Thomas S Richardson and James M Robins. Single world intervention graphs (SWIGs): A unification of the counterfactual and graphical approaches to causality. Technical Report 128, Center for Statistics and the Social Sciences, University of Washington, 2013.
- [15] Judea Pearl and Elias Bareinboim. External validity: from do-calculus to transportability across populations. *Statistical Science*, 29(4):579–595, 2014.
- [16] Elias Bareinboim and Judea Pearl. Causal inference and the data-fusion problem. *Proceedings of the National Academy of Sciences*, 113(27):7345–7352, 2016.
- [17] Tyler J VanderWeele. Concerning the consistency assumption in causal inference. *Epidemiology*, 20(6):880–883, 2009.

- [18] Miguel A Hernán and Tyler J VanderWeele. Compound treatments and transportability of causal inference. *Epidemiology (Cambridge, Mass.)*, 22(3):368, 2011.
- [19] Tyler J VanderWeele and Miguel A Hernán. Causal inference under multiple versions of treatment. *Journal of Causal Inference*, 1(1):1–20, 2013.
- [20] Maya L Petersen, Kristin E Porter, Susan Gruber, Yue Wang, and Mark J van der Laan. Diagnosing and responding to violations in the positivity assumption. *Statistical Methods in Medical Research*, 21(1):31–54, 2012.
- [21] Issa J Dahabreh and Miguel A Hernán. Extending inferences from a randomized trial to a target population. *European Journal of Epidemiology*, 34(8):719–722, 2019.
- [22] Erin Hartman, Richard Grieve, Roland Ramsahai, and Jasjeet S Sekhon. From SATE to PATT: combining experimental with observational studies to estimate population treatment effects. *Journal of the Royal Statistical Society Series A (Statistics in Society)*, 10:1111, 2013.
- [23] Sara Lodi, Andrew Phillips, Jens Lundgren, Roger Logan, Shweta Sharma, Stephen R Cole, Abdel Babiker, Matthew Law, Haitao Chu, Dana Byrne, et al. Effect estimates in randomized trials and observational studies: comparing apples with apples. *American Journal of Epidemiology*, 188(8):1569–1577, 2019.
- [24] Michael Webster-Clark, Jennifer L. Lund, Til Stürmer, Charles Poole, Ross J. Simpson, and Jessie K. Edwards. Reweighting oranges to apples: Transported re-ly trial vs non-experimental effect estimates of anticoagulation in atrial fibrillation. *Epidemiology*, 31(5):605–613, 2020.
- [25] Philippe Gabriel Steg, José López-Sendón, Esteban Lopez de Sa, Shaun G Goodman, Joel M Gore, Frederick A Anderson, Dominique Himbert, Jeanna Allegrone, and Frans

- Van de Werf. External validity of clinical trials in acute myocardial infarction. *Archives of Internal Medicine*, 167(1):68–73, 2007.
- [26] Miguel A Delgado. Testing the equality of nonparametric regression curves. *Statistics & Probability Letters*, 17(3):199–204, 1993.
- [27] Natalie Neumeyer, Holger Dette, et al. Nonparametric comparison of regression curves: an empirical process approach. *The Annals of Statistics*, 31(3):880–920, 2003.
- [28] Jeffery S Racine, Jeffrey Hart, and Qi Li. Testing the significance of categorical predictor variables in nonparametric regression models. *Econometric Reviews*, 25(4):523–544, 2006.
- [29] John Concato, Nirav Shah, and Ralph I Horwitz. Randomized, controlled trials, observational studies, and the hierarchy of research designs. *New England Journal of Medicine*, 342(25):1887–1892, 2000.
- [30] Kjell Benson and Arthur J Hartz. A comparison of observational studies and randomized, controlled trials. *New England Journal of Medicine*, 342(25):1878–1886, 2000.
- [31] Issa J Dahabreh, Radley C Sheldrick, Jessica K Paulus, Mei Chung, Vasileia Varvarigou, Haseeb Jafri, Jeremy A Rassen, Thomas A Trikalinos, and Georgios D Kitsios. Do observational studies using propensity score methods agree with randomized trials? A systematic comparison of studies on acute coronary syndromes. *European Heart Journal*, 33(15):1893–1901, 2012.
- [32] Georgios D Kitsios, Issa J Dahabreh, Sean Callahan, Jessica K Paulus, Anthony C Campagna, and James M Dargin. Can we trust observational studies using propensity scores in the critical care literature? a systematic comparison with randomized clinical trials. *Critical Care Medicine*, 43(9):1870–1879, 2015.

- [33] Guillaume Lonjon, Isabelle Boutron, Ludovic Trinquart, Nizar Ahmad, Florence Aim, Rémy Nizard, and Philippe Ravaud. Comparison of treatment effect estimates from prospective nonrandomized studies with propensity score analysis and randomized controlled trials of surgical procedures. *Annals of Surgery*, 259(1):18–25, 2014.
- [34] Jessica M Franklin, Elisabetta Patorno, Rishi J Desai, Robert J Glynn, David Martin, Kenneth Quinto, Ajinkya Pawar, Lily G Bessette, Hemin Lee, Elizabeth M Garry, et al. Emulating randomized clinical trials with nonrandomized real-world evidence studies: first results from the rct duplicate initiative. *Circulation*, 143(10):1002–1013, 2021.
- [35] Issa J Dahabreh, Lucia C Petito, Sarah E Robertson, Miguel A Hernán, and Jon A Steingrimsson. Toward causally interpretable meta-analysis: Transporting inferences from multiple randomized trials to a new target population. *Epidemiology*, 31(3):334–344, 2020.
- [36] Anastasios Tsiatis. *Semiparametric theory and missing data*. Springer Science & Business Media, 2007.
- [37] Heejung Bang and James M Robins. Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61(4):962–973, 2005.
- [38] Dennis D Boos and Leonard A Stefanski. *Essential statistical inference: theory and methods*, volume 120. Springer Science & Business Media, 2013.
- [39] Bradley Efron and Robert J Tibshirani. *An introduction to the bootstrap*, volume 57 of *Monographs on Statistics and Applied Probability*. Chapman & Hall/CRC, 1994.
- [40] Vivian C Wong and Peter M Steiner. Replication designs for causal inference. Technical report, EdPolicyWorks Working Paper Series, 2018.
- [41] Vivian C Wong and Peter M Steiner. Designs of empirical evaluations of nonexperimental methods in field settings. *Evaluation Review*, 42(2):176–213, 2018.

- [42] Robert J LaLonde. Evaluating the econometric evaluations of training programs with experimental data. *The American Economic Review*, pages 604–620, 1986.
- [43] Thomas Fraker and Rebecca Maynard. The adequacy of comparison group designs for evaluations of employment-related programs. *Journal of Human Resources*, pages 194–227, 1987.
- [44] Manfred Olschewski, Martin Schumacher, and Kathryn B Davis. Analysis of randomized and nonrandomized patients in clinical trials using the comprehensive cohort follow-up study design. *Controlled Clinical Trials*, 13(3):226–239, 1992.
- [45] Yi Lu, Daniel O Scharfstein, Maria M Brooks, Kevin Quach, and Edward H Kennedy. Causal inference for comprehensive cohort studies. *arXiv preprint arXiv:1910.03531*, 2019.