BIOMETRIC PRACTICE

Biometrics WILEY

Improving efficiency of inference in clinical trials with external control data

Xinyu Li[1] | **Wang Miao**[1] | **Fang Lu**[2] | **Xiao-Hua Zhou**[3]

[1] School of Mathematical Sciences & Center for Statistical Science, Peking University, Beijing, China

[2] Xiyuan Hospital, China Academy of Chinese Medical Sciences, Beijing, China

[3] Department of Biostatistics & Beijing International Center for Mathematical Research, Peking University, Beijing, China

Correspondence
Wang Miao, School of Mathematical Sciences, and Center for Statistical Science, Peking University, Beijing, China.
Email: mwfy@pku.edu.cn

Abstract

Suppose we are interested in the effect of a treatment in a clinical trial. The efficiency of inference may be limited due to small sample size. However, external control data are often available from historical studies. Motivated by an application to Helicobacter pylori infection, we show how to borrow strength from such data to improve efficiency of inference in the clinical trial. Under an exchangeability assumption about the potential outcome mean, we show that the semiparametric efficiency bound for estimating the average treatment effect can be reduced by incorporating both the clinical trial data and external controls. We then derive a doubly robust and locally efficient estimator. The improvement in efficiency is prominent especially when the external control data set has a large sample size and small variability. Our method allows for a relaxed overlap assumption, and we illustrate with the case where the clinical trial only contains a treated group. We also develop doubly robust and locally efficient approaches that extrapolate the causal effect in the clinical trial to the external population and the overall population. Our results also offer a meaningful implication for trial design and data collection. We evaluate the finite-sample performance of the proposed estimators via simulation. In the Helicobacter pylori infection application, our approach shows that the combination treatment has potential efficacy advantages over the triple therapy.

KEYWORDS
data combination, double robustness, efficiency, external control

1 | INTRODUCTION

Although randomized clinical trials are the gold standard for evaluation of a treatment effect, the efficiency of inference may be limited due to small sample size. However, external data sets containing control arm data are increasingly available from historical clinical databases. As a motivating application, we consider a clinical study on *Helicobacter pylori* (*H. pylori*) infection, which is a leading world-wide infectious disease. A randomized clinical trial is conducted at the traditional Chinese medicine (TCM) hospital, with the aim of examining whether additional use of TCM can lead to better efficacy than only the triple therapy (clarithromycin, amoxicillin, and omeprazole), a standard of care for *H. pylori* infection. But the sample size of this clinical trial is small. In parallel, external control data are available from a single-arm study at the Western-style hospital where all patients receive the triple therapy; see Section 5 for more details about the application. These external control data have the potential to improve the power of statistical inference about the treatment effect.

Borrowing strength from external control data to improve inference has long been a demand in the design and analysis of clinical trials since Pocock (1976), termed as *external control*, *historical control*, or *historical borrowing* in the literature; it is also mentioned in clinical guidance documents (e.g., U.S. Food and Drug Administration, 2017). For recent practice of using external control data in clinical and pharmaceutical studies, see Viele *et al.* (2014), van Rosmalen *et al.* (2018), and Schmidli *et al.* (2019). As a straightforward approach, synthesizing clinical trial and external control data as if they were from the same population may lead to bias, due to heterogeneity across study populations. Several other methods, such as common meta-analytic, Bayesian and frequentist methods (e.g., Viele *et al.*, 2014; Weber *et al.*, 2018; Zhang *et al.*, 2019) are also shown by Li and Song (2020) to fail to adequately address population heterogeneity.

In recent years, there has been an increasing amount of literature on causal inference methods that integrate information from multiple data sources, particularly when both randomized trials and observational studies are available. Most of these studies focus on *extrapolating* the causal inference from the randomized trials, such as drawing causal inference over a target population (Stuart *et al.*, 2011; Hartman *et al.*, 2015; Zhang *et al.*, 2016; Rudolph and van der Laan, 2017; Buchanan *et al.*, 2018; Rosenman *et al.*, 2018; Chen *et al.*, 2020; Dahabreh *et al.*, 2020a, 2019, 2020b; Li *et al.*, 2021), or validating observational methods that may suffer from confounding bias (Kallus *et al.*, 2018; Lodi *et al.*, 2019; Athey *et al.*, 2020; Yang *et al.*, 2020). Much of the previous research in this field is developed in nested designs; for example, Dahabreh *et al.* (2019) investigate the case where trials are nested in cohort studies that collect baseline data from all eligible individuals, and Lu *et al.* (2019) consider a comprehensive cohort study; yet recently nonnested designs have received growing attention where observations in different data sets are sampled separately. Pearl and Bareinboim (2011) and Pearl *et al.* (2014) also consider the data integration using directed acyclic graphs. Different terms are often used in literature for similar issues, for example, *generalizability*, *external validity*, *transportability*, and *data fusion*, see Colnet *et al.* (2020) for a systematic review.

In this article, we show how to improve efficiency of inference in the clinical trial by *borrowing* strength from external control data. Analogous to Dahabreh *et al.* (2019, 2020b), we maintain a mean exchangeability assumption, which reveals that the mean efficacy of the placebo or a standard of care is the same across studies within the same baseline level. Under the mean exchangeability, we derive the semiparametric efficiency bound for estimating the average treatment effect in the clinical trial when external controls are available. It is lower than the bound obtained without external controls. For estimation, we develop a semiparametric estimator, which is doubly robust and locally efficient. The efficiency gain over the trial-based doubly robust estimator is significant especially when the external control data set has a large sample size and small noise and when there is a sufficient overlap between the two data sets regarding covariates. Our method allows for a relaxed overlap assumption, which is weaker than that required by the trial-based estimator. We also develop doubly robust estimators that generalize the causal inference from the trial population to the external population and to the overall population of the study. The two estimators are more efficient than the estimator proposed by Rudolph and van der Laan (2017), Dahabreh *et al.* (2020b), and the one proposed by Dahabreh *et al.* (2019), respectively, as we make full use of the outcome observations contained in the external data set. We illustrate with simulations and an application to the *H. pylori* study, which is of a nonnested design. In the application, our estimators result in smaller variances and show that the combination treatment has potential efficacy advantages over the triple therapy.

2 | INFERENCE ABOUT THE TREATMENT EFFECT IN THE CLINICAL TRIAL

2.1 | Study design and data structure

In general, the study designs for multiple observational data sets fall into two categories: (i) nested trial designs, (ii) nonnested trial designs, see Dahabreh *et al.* (2021, 2020b) and Colnet *et al.* (2020). Our motivating *H. pylori* application concerns a nonnested composite data set design. To understand the sampling model in the composite data set design, we consider the following setup. Suppose that there exists an underlying population, that is, the study-eligible patient population of *H. pylori* infection to whom research studies would be applicable. The patient preference defines two different sub-populations, namely the study-eligible patient population at the TCM hospital and that at the Western-style hospital. For simplicity, we refer to them as the trial population and external population, respectively. Then the clinical trial observations and external data observations can be viewed as simple random samples from the corresponding sub-populations with unknown sampling probabilities. Therefore, each sub-population is identified by its sampled observations, but the underlying patient population is not due to the unknown sampling probabilities. The overall population corresponding to all observed samples is a mixture of the two sub-populations as we discuss later, yet is different from the underlying population in general. See the

TABLE 1 Observed data structure. "✓" and "?" indicate observed and unobserved, respectively

		Data set D	Treatment T	Covariates X	Potential Outcome		Observed Outcome Y
					Y_1	Y_0	
	1	1	1	✓	✓	?	✓
$\mathcal{O}_1$	⋮	⋮	⋮	⋮	⋮	⋮	⋮
	n_1	1	0	✓	?	✓	✓
	$n_1 + 1$	0	0	✓	?	✓	✓
$\mathcal{O}_2$	⋮	⋮	⋮	⋮	⋮	⋮	⋮
	n	0	0	✓	?	✓	✓

Supporting Information for detailed descriptions of the design and discussions about the relationship to other designs. Note that our proposed methods also apply to other designs as long as the data structure and assumptions described below are satisfied; we will return to this issue in the discussion.

Now we formally describe the observed data structure in our setting. We let Y denote the outcome of interest, T the indicator for treatment assignment with $T = 1$ for the treated group and $T = 0$ the control group, D the indicator for data source with $D = 1$ for the clinical trial and $D = 0$ for external control, and X a set of pre-treatment covariates. By the design, the observed data set $\mathcal{O}$ consists of n independent and identically distributed samples of (Y, X, T, D) from the overall population; it is divided into two separate subsets, the clinical trial data set $\mathcal{O}_1$ containing n_1 observations of $(Y, X, T, D = 1)$, and the external control data set $\mathcal{O}_2$ containing $n_2 = n - n_1$ observations of $(Y, X, T = 0, D = 0)$. Table 1 illustrates the observed data structure. The limitation of n_1/n as $n \to \infty$ approaches a positive constant $q = \mathrm{pr}(D = 1)$. We maintain the stable unit treatment value assumption that no interference between units and no hidden variations of treatments occur. We let Y_t denote the potential outcome that would be observed had T been set to t for $t \in \{0, 1\}$. We make the consistency assumption that the observed outcome is a realization of the potential outcome under the exposure actually received.

Assumption 1.

(i) $Y = TY_1 + (1 - T)Y_0$ for observations in $\mathcal{O}_1$;
(ii) $D = 0 \implies T = 0$ and thus $Y = Y_0$ for observations in $\mathcal{O}_2$.

In this section, the causal parameter of interest is $\tau = E(Y_1 - Y_0 \mid D = 1)$, the average treatment effect in the trial population.

2.2 | Inference solely based on the clinical trial data

Identification of the treatment effect τ is guaranteed based on the trial data under Assumption 1 and the following strong ignorability assumption (Rosenbaum and Rubin, 1983), which is commonly required in observational studies and can be met in randomized trials.

Assumption 2 (Strong ignorability).

(i) Ignorability: $(Y_1, Y_0) \perp\!\!\!\perp T \mid X, D = 1$;
(ii) Overlap: $0 < \mathrm{pr}(T = 1 \mid X = x, D = 1) < 1$ for all x such that $\mathrm{pr}(X = x \mid D = 1) > 0$.

We let $m_t(X) = E(Y \mid X, D = 1, T = t)$ for $t \in \{0, 1\}$ and $p(X) = \mathrm{pr}(T = 1 \mid X, D = 1)$ denote the outcome model and the treatment propensity score model, respectively. The residual of Y by subtracting $m_t(X)$ is $R_t = Y - m_t(X)$, and the difference between $m_t(X)$ is $\Delta(X) = m_1(X) - m_0(X)$. Under Assumptions 1 and 2, the asymptotic variance of any regular and asymptotic linear estimator of τ based solely on the trial data can be no smaller than the efficiency bound $\tilde{B}_\tau = E(\tilde{\mathrm{IF}}_\tau^2)$ (see Robins et al., 1994; Hahn, 1998; van der Laan and Robins, 2003), where

$$\tilde{\mathrm{IF}}_\tau = \frac{D}{q}\left\{\Delta(X) - \tau + \frac{T}{p(X)}R_1 - \frac{1-T}{1-p(X)}R_0\right\},$$

is the efficient influence function for τ when only the trial data are used for estimation.

We let $m_t(X; \beta_t)$ and $p(X; \phi)$ denote parametric working models for $m_t(X)$ and $p(X)$, respectively. A doubly robust estimator of τ can be obtained by first fitting $m_t(X; \tilde{\beta}_t)$ and $p(X; \tilde{\phi})$ with the clinical trial data and then using $\tilde{\mathrm{IF}}_\tau = 0$ as an estimating equation

$$\tilde{\tau}_{\mathrm{dr}} = \hat{E}\left[\left\{\tilde{\Delta}(X) + \frac{T}{p(X; \tilde{\phi})}\tilde{R}_1 - \frac{1-T}{1-p(X; \tilde{\phi})}\tilde{R}_0\right\} \mid D = 1\right],$$

where $\hat{E}$ is the empirical mean operator, $\tilde{\Delta}(X) = m_1(X; \tilde{\beta}_1) - m_0(X; \tilde{\beta}_0)$, and $\tilde{R}_t = Y - m_t(X; \tilde{\beta}_t)$ for $t \in \{0, 1\}$. This estimator only rests on the observed data from the clinical trial, and thus we refer to $\tilde{\tau}_{dr}$ as the trial-based doubly robust estimator.

Under Assumptions 1 and 2 and regularity conditions (Newey and McFadden, 1994, theorems 2.6 and 3.4), $\tilde{\tau}_{dr}$ is doubly robust in the sense that it is consistent if either of the working models is correctly specified. In recent years, such estimators have grown in popularity for inference about missing data and treatment effects (Scharfstein *et al.*, 1999). Moreover, if both working models are correct, the asymptotic variance of the estimator $\tilde{\tau}_{dr}$ attains the efficiency bound $\tilde{B}_\tau$ (Bang and Robins, 2005; Cao *et al.*, 2009). However, in the presence of external control data, the efficiency bound for estimating τ can be lower than $\tilde{B}_\tau$, and thus $\tilde{\tau}_{dr}$ is no longer efficient because it does not take external controls into account. We aim to characterize the semiparametric efficiency bound of τ when external control data are available.

2.3 | Improving efficiency with external control data

In order to incorporate external control data, we maintain the following assumption.

Assumption 3 (Mean exchangeability for Y_0). $E(Y_0 \mid X, D = 0) = E(Y_0 \mid X, D = 1)$.

Assumption 3 reveals that given pretreatment covariates, the mean efficacy of the placebo or a standard of care remains the same in the clinical trial and external control studies. It implicitly excludes average engagement effects caused by participation in a particular study rather than by treatment, for example, the so-called Hawthorne effects. Such an assumption is adopted in many practical applications (e.g., Rudolph and van der Laan, 2017; Dahabreh *et al.*, 2019, 2020b). In our *H. pylori* application, it is plausible because the control treatment is a standard of care in clinical settings regardless of the type of hospital and participation of the study. Relating this to the transport formula developed by Pearl *et al.* (2014), this assumption means that the selection node (D) does not point into the outcome node (Y). When Assumptions 1 and 2 hold, this assumption implies $E(Y \mid X, D = 0, T = 0) = E(Y \mid X, D = 1, T = 0)$, which is a testable implication of the mean exchangeability. To assess Assumption 3, we can fit a parametric outcome model for $E(Y \mid X, D, T = 0)$ to test interactive effects between D and X, or see Luedtke *et al.* (2019) for a nonparametric omnibus test. The mean

exchangeability is weaker than the distribution exchangeability $Y_0 \perp\!\!\!\perp D \mid X$, considered by Stuart *et al.* (2011), Hartman *et al.* (2015), Buchanan *et al.* (2018), Breskin *et al.* (2019), Lu *et al.* (2019), Li and Song (2020), and Yang *et al.* (2020).

We aim to calculate the semiparametric efficiency bound for estimating τ in the nonparametric model, where the observed data distribution is only restricted by Assumptions 1–3. We let $\pi(X) = \text{pr}(D = 1 \mid X)$ denote the selection propensity score, which quantifies the difference between participants in the trial and external data, and let $r(X) = \text{var}(Y_0 \mid X, D = 1)/\text{var}(Y_0 \mid X, D = 0)$, which measures the relative variability of Y_0 between the trial and external data. Under Assumptions 1 and 2, the variance ratio $r(X)$ equals $\text{var}(Y \mid X, D = 1, T = 0)/\text{var}(Y \mid X, D = 0)$. In certain cases, $r(X)$ can be known; for a binary outcome, $r(X) = 1$ under Assumption 3. For the ease of notation, we let

$$W(X, D, T) = \frac{D(1 - T)\pi(X) + (1 - D)\pi(X)r(X)}{\pi(X)\{1 - p(X)\} + \{1 - \pi(X)\}r(X)}.$$

Proposition 1. *Under Assumptions 1–3, the efficient influence function for τ is*

$$\text{IF}_\tau = \frac{1}{q}\left[D\{\Delta(X) - \tau\} + \frac{DT}{p(X)}R_1 - W(X, D, T)R_0\right],$$

and the semiparametric efficiency bound is $B_\tau = E(\text{IF}_\tau^2)$.

We compare B_τ to $\tilde{B}_\tau$ to show how external control data can improve efficiency of inference about the treatment effect in the trial population. Let $V_d(X) = \text{var}(Y_0 \mid X, D = d)$ and in the Supporting Information we show that

$$
\tilde{B}_\tau - B_\tau
$$

$$
= E\left[\left\{\frac{1}{1 - p(X)} - \frac{1}{1 - p(X) + \frac{\text{pr}(X|D=0)}{\text{pr}(X|D=1)}\frac{1-q}{q}r(X)}\right\} \times \frac{V_1(X)}{q} \,\Big|\, D = 1\right], \tag{1}
$$

which is positive as long as there exist a nonzero measure set of overlapped X for which $\text{pr}(X \mid D = 1)\text{pr}(X \mid D = 0) > 0$.

From Equation (1), given a trial data set drawn from a fixed trial population, the efficiency gain brought by the external data set is determined by three key factors: the degree of overlap between covariate distributions in the trial and external populations, the proportion of external samples in the overall data set, and the variance

ratio, captured by $\text{pr}(X \mid D = 0)/\text{pr}(X \mid D = 1)$, q and $r(X)$, respectively. For an ideal case where $p(X)$ and conditional variances $V_d(X)$ are constants, in the Supporting Information we show that the optimal $\text{pr}(X \mid D = 0)$ that maximizes the efficiency gain is equal to $\text{pr}(X \mid D = 1)$. Besides, a smaller q and a larger $r(X)$ lead to a bigger gap between the two efficiency bounds. Therefore, it is advantageous to collect an external data set with a large sample size and small noise from an external population that has a sufficient overlap regarding covariates with the trial population. Moreover, the efficiency improvement also relates to the treatment assignment mechanism captured by $p(X)$; if the trial data set contains very few control units, the efficiency gain due to external control data is prominent. These results offer a meaningful implication for trial design and data collection. In the Supporting Information, we illustrate this with an ideal example in which $B_\tau/\tilde{B}_\tau$ is derived and more simulations.

2.4 | A novel doubly robust estimator incorporating external control data

The efficient influence function IF_τ described in Proposition 1 motivates an estimator that achieves the semiparametric efficiency bound by using IF_τ as an estimating equation. In addition to the models $m_t(X; \beta_t)$ and $p(X; \phi)$ used in the trial-based doubly robust estimator, we also specify parametric working models $\pi(X; \alpha)$ for the selection propensity score $\pi(X)$ and $r(X; \eta)$ for the variance ratio $r(X)$. Under Assumptions 1–3, $m_0(X) = E(Y \mid X, T = 0)$, and thus we fit the outcome model $m_0(X; \beta_0)$ with all control units available from both the trial and external data. We let $(\hat{\beta}_t, \hat{\phi}, \hat{\alpha}, \hat{\eta})$ denote a $n^{1/2}$-consistent estimator of the nuisance parameters. The efficient influence function motivates the following estimator of τ,

$$\hat{\tau}_{\text{dr}} = \frac{1}{\hat{q}} \hat{E} \left[D\widehat{\Delta}(X) + \frac{DT}{p(X; \hat{\phi})}\widehat{R}_1 - \widehat{W}(X, D, T)\widehat{R}_0 \right],$$

where $\hat{q} = n_1/n$, $\widehat{\Delta}(X) = m_1(X; \hat{\beta}_1) - m_0(X; \hat{\beta}_0)$ and $\widehat{R}_t = Y - m_t(X; \hat{\beta}_t)$ for $t \in \{0, 1\}$ and

$$\widehat{W}(X, D, T)$$
$$= \frac{D(1 - T) + (1 - D)r(X; \hat{\eta})}{\pi(X; \hat{\alpha})\{1 - p(X; \hat{\phi})\} + \{1 - \pi(X; \hat{\alpha})\}r(X; \hat{\eta})}\pi(X; \hat{\alpha}).$$

Under standard regularity conditions (van der Vaart, 2000, theorem 25.54), the proposed estimator $\hat{\tau}_{\text{dr}}$ is locally efficient, in the sense that its asymptotic variance attains the semiparametric efficiency bound B_τ when all working models are correctly specified. Moreover, $\hat{\tau}_{\text{dr}}$ is also

doubly robust, just as many locally efficient estimators are naturally robust for estimating parameters that arise in causal inference (see, e.g., Benkeser et al., 2017).

Proposition 2. *Under Assumptions 1–3 and regularity conditions described in theorems 2.6 and 3.4 of Newey and McFadden (1994), the estimator $\hat{\tau}_{\text{dr}}$ is consistent and asymptotically normal if either (i) the outcome models $m_t(X; \beta_t)$ for $t \in \{0, 1\}$ are correct, or (ii) the propensity score models $\pi(X; \alpha)$ and $p(X; \phi)$ are correct.*

The proposed estimator $\hat{\tau}_{\text{dr}}$ incorporates all data available from both the trial data set and the external controls, and thus we call it the full-data doubly robust estimator to distinguish from the trial-based one $\tilde{\tau}_{\text{dr}}$. Although we focus on parametric working models, the full-data doubly robust estimator can accommodate flexible working models and leads to a valid $n^{1/2}$ inference as long as estimators of nuisance parameters have a convergence rate of at least $n^{1/4}$ (Newey, 1990; Robins et al., 2017). Such a rate is achievable by a bunch of machine learning methods, such as random forests (Wager and Athey, 2018), neural networks (Chen and White, 1999), and the highly adaptive lasso (Benkeser and van der Laan, 2016).

Compared to the trial-based estimator $\tilde{\tau}_{\text{dr}}$, the full-data estimator $\hat{\tau}_{\text{dr}}$ entails two extra working models $\pi(X; \alpha)$ and $r(X; \eta)$. Consistency of $\hat{\tau}_{\text{dr}}$ rests on the selection propensity score model $\pi(X; \alpha)$, which is essential for incorporating the external control data. However, consistency of $\hat{\tau}_{\text{dr}}$ does not rest on the variance ratio model $r(X; \eta)$, although the efficiency does. The variance ratio $r(X)$ strikes a balance between the control units in the trial and external data by assigning more weights to those with smaller noise. A larger $r(X)$ attaches more importance to the external data. If $r(X)$ is replaced by zero and β_0 is estimated with only the trial data, then external control data are ignored and $\hat{\tau}_{\text{dr}}$ reduces to the trial-based estimator $\tilde{\tau}_{\text{dr}}$. In certain cases, the variance ratio is known and $\hat{\tau}_{\text{dr}}$ reduces to a simplified form. For instance, when the outcome is binary or $Y_0 \perp\!\!\!\perp D \mid X$ holds, we have $r(X) = 1$.

Although we require that $p(X) < 1$ in Assumption 2, it can be relaxed to $p(X)\pi(X) < 1$, which is sufficient for the denominator $\pi(X)\{1 - p(X)\} + \{1 - \pi(X)\}r(X)$ in IF_τ to be positive. It indicates that for the treated units, similar control units only need to exist in either the clinical trial or the external data. In the Supporting Information, we discuss this issue in detail and consider an extreme case where the trial data set only contains a treated group. In this case, we let $p(X) = 1$, then the full-data estimator $\hat{\tau}_{\text{dr}}$ reduces to an estimator that is doubly robust against misspecification of either $m_0(X; \beta_0)$ or $\pi(X; \alpha)$.

In addition to point estimates, standard errors and confidence intervals can be obtained based on normal

approximations under certain regularity conditions, which follows from the general theory for estimation equations (van der Vaart, 2000). Alternatively, bootstrap methods could also be implemented to obtain the variance estimate.

3 | EXTRAPOLATING INFERENCE TO OTHER POPULATIONS OF INTEREST

In Section 2, we have focused on the treatment effect in a clinical trial. However, the treatment effects in the external and the overall population may differ and thus extrapolation of inference is of interest. By the study design, the overall population can be considered as a mixture of the trial and external populations, with the mixing probability q. This mixed population characterizes patients in both types of hospitals and, although generally not equivalent to the underlying population, is relatively more representative than a single sub-population. In this section, we aim to develop locally efficient and doubly robust estimators of $\xi = E(Y_1 - Y_0 \mid D = 0)$ and $\psi = E(Y_1 - Y_0)$.

However, ξ and ψ are not identifiable without further assumptions, because the potential outcome Y_1 is not observed in the external control data. To aid in identification, we assume an additional mean exchangeability for Y_1 and an additional overlap condition.

Assumption 4 (Mean exchangeability for Y_1). $E(Y_1 \mid D = 0, X) = E(Y_1 \mid D = 1, X)$.

Assumption 5 (Population support overlap). $0 < \pi(X) < 1$.

Under Assumptions 1–5, the treatment effect in the external population ξ is identified by

$$\xi = E\{E(Y \mid X, D = 1, T = 1) - E(Y \mid X,$$
$$D = 1, T = 0) \mid D = 0\},$$

and the overall treatment effect ψ is identified by $\psi = \tau \cdot q + \xi \cdot (1 - q)$. For identification of ψ and ξ, it would be sufficient to assume the mean exchangeability for $Y_1 - Y_0$, under which the efficient estimators of ψ and ξ are the one proposed by Dahabreh *et al.* (2019) and the one by Rudolph and van der Laan (2017), and Dahabreh *et al.* (2020b), respectively. However, as we show later they cannot make use of the outcome information of external samples and thereby do not admit efficiency improvement with full external data. Under the nonparametric model where the observed data distribution is restricted only by Assumptions 1–5, we derive the efficient influence functions for ξ and ψ, respectively.

Proposition 3. *Under Assumptions 1–5, the efficient influence functions for ψ and ξ are*

$$\text{IF}_\psi = \Delta(X) - \psi + \frac{1}{\pi(X)}\left\{ \frac{DT}{p(X)}R_1 - W(X, D, T)R_0 \right\},$$

$$\text{IF}_\xi = \frac{1}{1-q}\left[(1-D)\{\Delta(X) - \xi\} \right.$$
$$\left. + \frac{1 - \pi(X)}{\pi(X)}\left\{ \frac{DT}{p(X)}R_1 - W(X, D, T)R_0 \right\} \right],$$

respectively, and the semiparametric efficiency bounds for estimating ψ and ξ are $B_\psi = E(\text{IF}_\psi^2)$ and $B_\xi = E(\text{IF}_\xi^2)$, respectively.

The efficient influence functions IF_ψ and IF_ξ motivate the following estimators,

$$\hat{\psi}_{\text{dr}} = \hat{E}\left\{ \hat{\Delta}(X) + \frac{DT}{\pi(X; \hat{\alpha})p(X; \hat{\phi})}\hat{R}_1 - \frac{\widehat{W}(X, D, T)}{\pi(X; \hat{\alpha})}\hat{R}_0 \right\},$$

$$\hat{\xi}_{\text{dr}} = \frac{1}{1-\hat{q}}\hat{E}\left[(1-D)\hat{\Delta}(X) + \frac{1 - \pi(X; \hat{\alpha})}{\pi(X; \hat{\alpha})} \right.$$
$$\left. \times \left\{ \frac{DT}{p(X; \hat{\phi})}\hat{R}_1 - \widehat{W}(X, D, T)\hat{R}_0 \right\} \right].$$

Semiparametric estimators $\hat{\psi}_{\text{dr}}$ and $\hat{\xi}_{\text{dr}}$ are both locally efficient, in the sense that the corresponding asymptotic variance attains the efficiency bound when all working models are correctly specified. In addition, they also enjoy the doubly robustness property.

Proposition 4. *Under Assumptions 1–5 and regularity conditions described in theorems 2.6 and 3.4 of Newey and McFadden (1994), the estimators $\hat{\psi}_{\text{dr}}$ and $\hat{\xi}_{\text{dr}}$ are consistent and asymptotically normal if either (i) the outcome models $m_t(X; \beta_t)$ for $t \in \{0, 1\}$ are correct, or (ii) the propensity score models $\pi(X; \alpha)$ and $p(X; \phi)$ are correct.*

Analogous to the estimator $\hat{\tau}_{\text{dr}}$, consistency of $\hat{\psi}_{\text{dr}}$ or $\hat{\xi}_{\text{dr}}$ does not depend on correct specification of $r(X)$. If $r(X)$ is replaced by zero and β_0 is estimated with only the trial data, $\hat{\psi}_{\text{dr}}$ and $\hat{\xi}_{\text{dr}}$ reduce to the estimator proposed by Dahabreh *et al.* (2019) and the one by Rudolph and van der Laan (2017), and Dahabreh *et al.* (2020b), hereafter referred to as $\tilde{\psi}_{\text{dr}}$ and $\tilde{\xi}_{\text{dr}}$, respectively. However, $\tilde{\psi}_{\text{dr}}$ and $\tilde{\xi}_{\text{dr}}$ only utilize the baseline covariates available in the external data, whereas our estimators can incorporate the information of both the outcome and covariates in the external data. Therefore, $\hat{\psi}_{\text{dr}}$ and $\hat{\xi}_{\text{dr}}$ are more efficient. When working models are correctly specified, the

asymptotic variance of $\hat{\psi}_{dr}$ is smaller than that of $\tilde{\psi}_{dr}$ by

$$
E\left[\left\{\frac{1}{\pi(X)\{1-p(X)\}} - \frac{1}{\pi(X)\{1-p(X)\}+\{1-\pi(X)\}r(X)}\right\} V_1(X)\right],
$$

and asymptotic variance of $\hat{\xi}_{dr}$ is smaller than that of $\tilde{\xi}_{dr}$ by

$$
E\left[\left\{\frac{\{1-\pi(X)\}^2}{\pi(X)\{1-p(X)\}} - \frac{\{1-\pi(X)\}^2}{\pi(X)\{1-p(X)\}+\{1-\pi(X)\}r(X)}\right\} \frac{V_1(X)}{(1-q)^2}\right].
$$

4 | SIMULATION STUDY

We evaluate the performance of the proposed methods via simulations. We consider four data-generating scenarios with a continuous response: (i) both the set of outcome models and the set of propensity score models are correct, (ii) only the set of outcome models are correct, (iii) only the set of propensity score models are correct, and (iv) both are incorrect. In our settings, $pr(D = 1)$ and $pr(T = 1 \mid D = 1)$ are around 50%. Details of the data-generating mechanisms and the specification of working models are described in the Supporting Information.

We use the following methods to estimate the average treatment effect τ: the full-data doubly robust estimator $\hat{\tau}_{dr}$ with a correct variance ratio, and the trial-based doubly robust estimator $\tilde{\tau}_{dr}$. Besides, to investigate the impact of misspecification of the variance ratio, we also report a full-data doubly robust estimator with a constant variance ratio.

We simulate 1000 replicates under 1000 sample size for each scenario and summarize the results with bias boxplots in Figure 1. For estimation of τ, the proposed full-data estimators have smaller variability, which indicates that leveraging information from the external data improves the efficiency. As expected, all three estimators have very small bias if either the set of outcome models or the set of propensity score models are correct. But when both sets of working models are incorrect, these estimators may be biased; in this case, the proposed full-data estimators have smaller bias than the trial-based estimator in our simulations, although, this may not hold in general. Moreover, although the variance ratio is not a constant in the data-generating process, the performance of the full-data doubly robust estimator is not compromised by using a constant variance ratio. Table 2 shows coverage probabilities of the 95% confidence interval based on normal

TABLE 2 Coverage probability of 95% confidence interval

| Scenario	$\tilde{\tau}_{dr}$	$\hat{\tau}_{dr}$	$\hat{\tau}_{dr}^c$
(i)	0.955	0.954	0.959
(ii)	0.946	0.955	0.956
(iii)	0.960	0.956	0.952
(iv)	0.096	0.031	0.046

Note: $\hat{\tau}_{dr}^c$ is obtained by using a constant variance ratio in the full-data doubly robust estimation of τ.

approximation. The coverage probability approximates the nominal level of 0.95 provided that either set of the working models are correct. We also provide simulation results for estimation of ψ and ξ in the Supporting Information, showing that the full-data estimators $\hat{\psi}_{dr}$ and $\hat{\xi}_{dr}$ have very small bias even if one of the working models is incorrect, and have smaller variability than $\tilde{\psi}_{dr}$ and $\tilde{\xi}_{dr}$, respectively.

In the Supporting Information, we conduct additional numerical simulations particularly under the scenarios where the mean exchangeability is violated. We fit a parametric outcome model for $E(Y \mid D, X, T = 0)$ and then test interactive effects between D and X to assess the mean exchangeability. The simulation results show that in certain cases where the engagement effects are weak, the full-data estimator $\hat{\tau}_{dr}$ may still have a smaller MSE than the trial-based estimator $\tilde{\tau}_{dr}$. In the presence of strong engagement effects, the bias is no longer negligible, however, can be detected via the proposed test of the mean exchangeability.

5 | AN _H. PYLORI_ INFECTION EXAMPLE

We apply our methods to the _H. pylori_ infection data set for illustration. The trial data set is obtained from a two-arm clinical trial conducted at the TCM hospital; it consists of 362 observations, of which 180 patients are assigned to the triple therapy, and the rest are assigned to a combination treatment including both the triple therapy and TCM. The external data set is obtained from a single-arm clinical trial conducted at the Western-style hospital; it contains 110 observations, of which all patients are assigned to the triple therapy. To ensure the uniformity of treatment, for at least 2 weeks prior to treatment, patients in both trials are instructed not to take any other medications that may interact with a study drug or affect the clinical result, such as PPI, H2RA, and bismuth. The treatment process lasted for 4 weeks, and the outcome of interest is the binary disease status after the treatment detected by the C-14 urea breath test (UBT). The baseline covariates are the same in both external and trial data, including age, gender, height, BMI, work type, education level, marriage
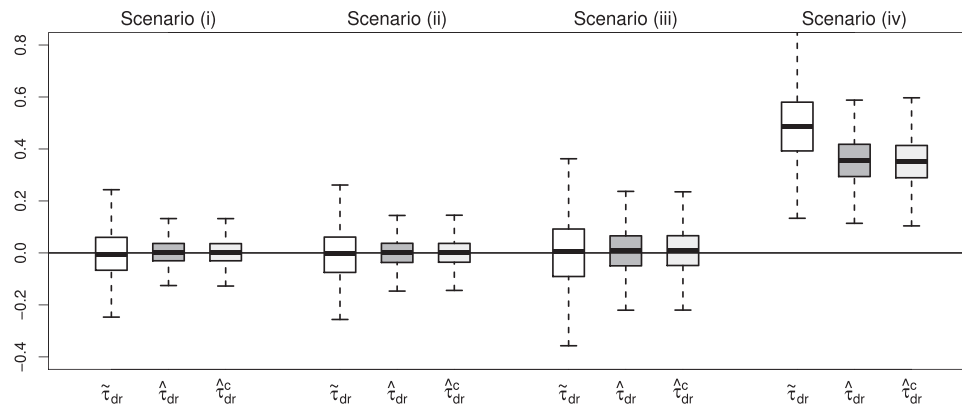
FIGURE 1 Bias boxplots for estimators of the treatment effect τ in the clinical trial. $\hat{\tau}_{dr}^{c}$ is obtained by using using a constant variance ratio in the full-data doubly robust estimation of τ

TABLE 3 Estimates of the causal effects in *H. pylori* application

	point est. $\times 10^2$	variance $\times 10^4$	*p*-value		point est. $\times 10^2$	variance $\times 10^4$	*p*-value
$\hat{\tau}_{dr}$	5.82	16.10	0.073	$\tilde{\tau}_{dr}$	5.43	19.55	0.110
$\hat{\psi}_{dr}$	6.72	14.98	0.041	$\tilde{\psi}_{dr}$	6.55	19.56	0.069
$\hat{\xi}_{dr}$	9.67	21.50	0.019	$\tilde{\xi}_{dr}$	10.24	38.05	0.048

status, and information on patients' symptoms; we provide descriptive statistics and more descriptions in the Supporting Information.

We are interested in whether the additional Chinese medicine would improve the cure rate in the clinical trial, particularly, testing the null hypothesis $H_0 : \tau \leq 0$ against $H_1 : \tau > 0$. All patients assigned to the triple therapy receive identical treatment and dose, and we assume the mean exchangeability for Y_0. We estimate τ with the full-data estimator $\hat{\tau}_{dr}$ and the trial-based estimator $\tilde{\tau}_{dr}$. The conditional outcome means and propensity scores are fitted with logistic models. We assess the mean exchangeability by testing whether there exist interactive effects between D and X on Y with the logistic outcome model built on control samples. The *p*-value of the test is 0.441, and hence under the significance level of 0.05, we cannot reject the assumption 3. Because the outcome is binary, the variance ratio $r(X)$ is set to one in our analysis. We also estimate the overall treatment effect ψ with $\hat{\psi}_{dr}$ and $\tilde{\psi}_{dr}$, and the treatment effect of the external data ξ with $\hat{\xi}_{dr}$ and $\tilde{\xi}_{dr}$. Table 3 presents the data analysis results, including point estimates, variance estimates, and *p*-values.

For estimation of the treatment effect in the trial population τ, the point estimates $\hat{\tau}_{dr}$ and $\tilde{\tau}_{dr}$ are close; however, the variance of the full-data doubly roust estimator ($\hat{\tau}_{dr}$) is only 82% of that of the trial-based estimator ($\tilde{\tau}_{dr}$). As a result, $\hat{\tau}_{dr}$ leads to a smaller *p*-value (0.073) than $\tilde{\tau}_{dr}$ (*p*-value: 0.110). For estimation of the treatment effects ψ,

the proposed estimator $\hat{\psi}_{dr}$ has a smaller variance and also a smaller *p*-value (0.041) than $\tilde{\psi}_{dr}$ (*p*-value: 0.069). The claim also applies to the estimation of ξ. Therefore, under the significance level of 0.1, we reject the null hypothesis based on $\hat{\tau}_{dr}$, $\hat{\xi}_{dr}$, and $\hat{\psi}_{dr}$, and conclude that the combination treatment may have a better efficacy than the standard triple therapy in the trial population, the external population, and the overall population.

To assess the robustness of the findings, we incorporate only the top half of the external data into the analysis. The resulting estimator $\hat{\tau}_{dr}$ is 0.052 with variance 17.89×10^{-4}; as expected, the variance is smaller than that of the trial-based estimator (19.55×10^{-4}), but larger than that of the full-data estimator (16.10×10^{-4}).

6 | DISCUSSION

Our estimators are not restricted to the study design of the *H. pylori* application. The methods remain valid as long as the study design fulfills the requirement of the data structure and assumptions. However, under different study designs, the trial and external populations may differ, and so may the treatment effects. Therefore, the researchers need to first clarify the population of interest on which the inference is drawn. In the Supporting Information, we discuss the practical implications of the effects of interest when the trial and external data sets are collected under different study designs.

The use of external data may yield bias if the mean exchangeability is violated when there exist engagement effects. We denote the selection bias by

$$E(Y_0 \mid X, D = 1) - E(Y_0 \mid X, D = 0) = b(X),$$

which encodes the strength of engagement effects within each level of X. Assuming sufficiently flexible working models are employed such that no approximation error is introduced by model misspecification, we can show that the asymptotic bias of $\hat{\tau}_{\mathrm{dr}}$ is

$$\Lambda = E\left[\frac{\pi(X)}{q} \cdot \frac{\{1 - \pi(X)\}r(X)}{\pi(X)\{1 - p(X)\} + \{1 - \pi(X)\}r(X)} \cdot b(X) \right].$$

Suppose $b(X)$ is bounded with $|b(X)| \leq B$, then we have $|\Lambda| \leqslant B$, which states that the asymptotic bias of $\hat{\tau}_{\mathrm{dr}}$ does not exceed that the largest difference between the conditional means of two data sets. As a result, weak engagement effects would not negate our inference, although large ones can; see the Supporting Information for more discussions on this issue.

DATA AVAILABILITY STATEMENT
The data that support the findings in this paper are available from the corresponding author on reasonable request.

ORCID
Xinyu Li https://orcid.org/0000-0002-6429-4816
Xiao-Hua Zhou https://orcid.org/0000-0001-7935-1222

REFERENCES
Athey, S., Chetty, R. & Imbens, G. (2020) Combining experimental and observational data to estimate treatment effects on long term outcomes. [Preprint] Available from: http://arxiv.org/abs/2006.09676.

Bang, H.& Robins, J.M. (2005) Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61, 962–973.

Benkeser, D., Carone, M., van der Laan, M.J.& Gilbert, P.B. (2017) Doubly robust nonparametric inference on the average treatment effect. *Biometrika*, 104, 863–880.

Benkeser, D. & van der Laan, M. (2016) The highly adaptive Lasso estimator. *Proceedings of the International Conference on Data Science and Advanced Analytics*, 2016, 689–696.

Breskin, A., Westreich, D., Cole, S.R.& Edwards, J.K. (2019) Using bounds to compare the strength of exchangeability assumptions for internal and external validity. *American Journal of Epidemiology*, 188, 1355–1360.

Buchanan, A.L., Hudgens, M.G., Cole, S.R., Mollan, K.R., Sax, P.E., Daar, E.S. et al. (2018) Generalizing evidence from randomized trials using inverse probability of sampling weights. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 181, 1193.

Cao, W., Tsiatis, A.A.& Davidian, M. (2009) Improving efficiency and robustness of the doubly robust estimator for a population mean with incomplete data. *Biometrika*, 96, 723–734.

Chen, Y., Li, P.& Wu, C. (2020) Doubly robust inference with non-probability survey samples. *Journal of the American Statistical Association*, 115, 2011–2021.

Chen, X.& White, H. (1999) Improved rates and asymptotic normality for nonparametric neural network estimators. *IEEE Transactions on Information Theory*, 45, 682–691.

Colnet, B., Mayer, I., Chen, G., Dieng, A., Li, R., Varoquaux, G. et al. (2020) Causal inference methods for combining randomized trials and observational studies: a review. [Preprint] http://arxiv.org/abs/2011.08047.

Dahabreh, I.J., Haneuse, S.J.P., Robins, J.M., Robertson, S.E., Buchanan, A.L., Stuart, E.A. et al. (2021) Study designs for extending causal inferences from a randomized trial to a target population. *American Journal of Epidemiology*, 190, 1632–1642.

Dahabreh, I.J., Petito, L.C., Robertson, S.E., Hernán, M.A. & Steingrimsson, J.A. (2020a) Toward causally interpretable meta-analysis: transporting inferences from multiple randomized trials to a new target population. *Epidemiology*, 31, 334–344.

Dahabreh, I.J., Robertson, S.E., Steingrimsson, J.A., Stuart, E.A. & Hernán, M.A. (2020b) Extending inferences from a randomized trial to a new target population. *Statistics in Medicine*, 39, 1999–2014.

Dahabreh, I.J., Robertson, S.E., Tchetgen Tchetgen, E., Stuart, E.A.& Hernán, M.A. (2019) Generalizing causal inferences from individuals in randomized trials to all trial-eligible individuals. *Biometrics*, 75, 685–694.

Hahn, J. (1998) On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica*, 66, 315–331.

Hartman, E., Grieve, R., Ramsahai, R.& Sekhon, J.S. (2015) From sample average treatment effect to population average treatment effect on the treated: combining experimental with observational studies to estimate population treatment effects. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 178, 757–778.

Kallus, N., Puli, A.M. & Shalit, U. (2018) Removing hidden confounding by experimental grounding. *Advances in Neural Information Processing Systems*, 31, 10888–10897.

Li, F., Buchanan, A.L. & Cole, S.R. (2021) Generalizing trial evidence to target populations in non-nested designs: applications to aids clinical trials. [Preprint] http://arxiv.org/abs/2103.04907.

Li, X.H.& Song, Y. (2020) Target population statistical inference with data integration across multiple sources–an approach to mitigate information shortage in rare disease clinical trials. *Statistics in Biopharmaceutical Research*, 12, 322–333.

Lodi, S., Phillips, A., Lundgren, J., Logan, R., Sharma, S., Cole, S.R. et al. (2019) Effect estimates in randomized trials and

observational studies: comparing apples with apples. *American Journal of Epidemiology*, 188, 1569–1577.

Lu, Y., Scharfstein, D.O., Brooks, M.M., Quach, K. & Kennedy, E.H. (2019) Causal inference for comprehensive cohort studies. [Preprint] http://arxiv.org/abs/1910.03531.

Luedtke, A., Carone, M.& van der Laan, M.J. (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.

Newey, W.K. (1990) Semiparametric efficiency bounds. *Journal of Applied Econometrics*, 5, 99–135.

Newey, K. & McFadden, D. (1994) Large sample estimation and hypothesis testing. In: Engle, R. & McFadden, D. (Eds.) *Handbook of econometrics*, vol. 4. New York, NY: Elsevier Science, pp. 2112–2245.

Pearl, J. & Bareinboim, E. (2011) Transportability of causal and statistical relations: a formal approach. *Proceedings of the AAAI Conference on Artificial Intelligence*, 25, 247–254.

Pearl, J.& Bareinboim, E. (2014) External validity: from do-calculus to transportability across populations. *Statistical Science*, 29, 579–595.

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

Robins, J., Li, L., Mukherjee, R., Tchetgen Tchetgen, E.& van der Vaart, A. (2017) Higher order estimating equations for high-dimensional models. *Annals of Statistics*, 45, 1951–1987.

Robins, J.M., Rotnitzky, A.& Zhao, L.P. (1994) Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89, 846–866.

Rosenbaum, P.R.& Rubin, D.B. (1983) The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70, 41–55.

Rosenman, E., Owen, A.B., Baiocchi, M. & Banack, H. (2018) Propensity score methods for merging observational and experimental datasets. [Preprint] http://arxiv.org/abs/1804.07863.

Rudolph, K.E.& van der Laan, M.J. (2017) Robust estimation of encouragement design intervention effects transported across sites. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79, 1509.

Scharfstein, D.O., Rotnitzky, A.& Robins, J.M. (1999) Adjusting for nonignorable drop-out using semiparametric nonresponse models. *Journal of the American Statistical Association*, 94, 1096–1120.

Schmidli, H., Häring, D.A., Thomas, M., Cassidy, A., Weber, S.& Bretz, F. (2019) Beyond randomized clinical trials: use of external controls. *Clinical Pharmacology & Therapeutics*, 107, 806–816.

Stuart, E.A., Cole, S.R., Bradshaw, C.P.& Leaf, P.J. (2011) The use of propensity scores to assess the generalizability of results from randomized trials. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 174, 369–386.

U.S. Food and Drug Administration (2017) *Antibacterial therapies for patients with an unmet medical need for the treatment of serious bacterial diseases*. Rockville, MD: U.S. Department of Health and Human Services, Food and Drug Administration, Center for Drug Evaluation and Research.

van der Laan, M.J. & Robins, J.M. (2003) *Unified methods for censored longitudinal data and causality*. New York, NY: Springer Science & Business Media.

van der Vaart, A.W. (2000) *Asymptotic statistics*, vol. 3. Cambridge: Cambridge University Press.

van Rosmalen, J., Dejardin, D., van Norden, Y., Löwenberg, B. & Lesaffre, E. (2018) Including historical data in the analysis of clinical trials: is it worth the effort? *Statistical Methods in Medical Research*, 27, 3167–3182.

Viele, K., Berry, S., Neuenschwander, B., Amzal, B., Chen, F., Enas, N., et al. (2014) Use of historical control data for assessing treatment effects in clinical trials. *Pharmaceutical Statistics*, 13, 41–54.

Wager, S.& Athey, S. (2018) Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113, 1228–1242.

Weber, K., Hemmings, R. & Koch, A. (2018) How to use prior knowledge and still give new data a chance? *Pharmaceutical Statistics*, 17, 329–341.

Yang, S., Kim, J.K.& Song, R. (2020) Doubly robust inference when combining probability and non-probability samples with high dimensional data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82, 445–465.

Zhang, J., Ko, C.W., Nie, L., Chen, Y.& Tiwari, R. (2019) Bayesian hierarchical methods for meta-analysis combining randomized-controlled and single-arm studies. *Statistical Methods in Medical Research*, 28, 1293–1310.

Zhang, Z., Nie, L., Soon, G.& Hu, Z. (2016) New methods for treatment effect calibration, with applications to non-inferiority trials. *Biometrics*, 72, 20–29.

SUPPORTING INFORMATION

Web Appendices, Tables, and Figures referenced in Sections 2–6 are available with this paper at the Biometrics website on Wiley Online Library. R codes to replicate the simulation results are also provided.

How to cite this article: Li, X., Miao, W., Lu, F., & Zhou, X.-H. (2023) Improving efficiency of inference in clinical trials with external control data. *Biometrics*, 79, 394–403. https://doi.org/10.1111/biom.13583