Biometrika (2025), **112**, 2, asae069 https://doi.org/10.1093/biomet/asae069
Advance Access publication 17 December 2024

Improving randomized controlled trial analysis via data-adaptive borrowing

By CHENYIN GAO, SHU YANG

Department of Statistics, North Carolina State University, 2311 Stinson Drive, Raleigh, North Carolina 27695, U.S.A.
cgao6@ncsu.edu syang24@ncsu.edu

MINGYANG SHAN, WENYU YE, ILYA LIPKOVICH and DOUGLAS FARIES

Eli Lilly & Company, Lilly Corporate Center, 893 Delaware Street, Indianapolis, Indiana 46285, U.S.A.
mingyang.shan@lilly.com ye_wendy_wenyu@lilly.com
ilya.lipkovich@lilly.com faries_douglas_e@lilly.com

Summary

In recent years, real-world external controls have grown in popularity as a tool to empower randomized placebo-controlled trials, particularly in rare diseases or cases where balanced randomization is unethical or impractical. However, as external controls are not always comparable to the trials, direct borrowing without scrutiny may heavily bias the treatment effect estimator. Our paper proposes a data-adaptive integrative framework capable of preventing unknown biases of the external controls. The adaptive nature is achieved by dynamically sorting out a comparable subset of external controls via bias penalization. Our proposed method can simultaneously achieve (a) the semiparametric efficiency bound when the external controls are comparable and (b) selective borrowing that mitigates the impact of the existence of incomparable external controls. Furthermore, we establish statistical guarantees, including consistency, asymptotic distribution and inference, providing Type-I error control and good power. Extensive simulations and two real-data applications show that the proposed method leads to improved performance over the trial-only estimator across various bias-generating scenarios.

Some key words: Adaptive lasso; Calibration weighting; Dynamic borrowing; Study heterogeneity.

1. Introduction

Randomized controlled trials have been considered the gold standard of clinical research to provide confirmatory evidence on the safety and efficacy of treatments. However, randomized placebo-controlled trials are expensive, require lengthy recruitment periods and may not always be ethical, feasible or practical in rare or life-threatening diseases. In response, quality patient-level real-world data from disease registries and electronic health records have become increasingly available and can generate fit-for-purpose real-world evidence to facilitate healthcare and regulatory decision-making (FDA, 2021). Studies using

real-world data may have advantages over randomized placebo-controlled trials, including longer observation windows, larger and more heterogeneous patient populations, and reduced burden on investigators and patients (Visvanathan et al., 2017; Colnet et al., 2020). There is interest in novel clinical trial designs that leverage external controls from real-world data to improve the efficiency of randomized placebo-controlled trials while maintaining robust evidence on the safety and efficacy of treatments (Silverman, 2018; FDA, 2019; Ghadessi et al., 2020). The focus of this paper is on hybrid control arm designs using real-world data, where the concurrent control arm is augmented with real-world external controls to form a hybrid comparator group.

The concept of hybrid controls dates back to Pocock (1976), who combined the trial data and historical controls by adjusting for data-source-level differences. Since then, numerous methods for using external controls have been developed. However, regulatory approvals of external control arm designs as confirmatory trials are rare and limited to ultra-rare diseases, pediatric trials or oncology trials (FDA, 2014, 2016; Odogwu et al., 2018). Concerns regarding the validity and comparability of the external controls have limited their use in a broader context. Guidance documents from regulatory agencies, including the recent FDA draft guidance (FDA, 2023), note several potential issues with the external controls, including selection bias, lack of concurrency, differences in the definitions of covariates, treatments or outcomes, and unmeasured confounding (FDA, 2001, 2019, 2023). Without proper scrutiny, each of these concerns may lead to biased treatment effect estimates and misleading conclusions.

Selection bias is a type of data heterogeneity often encountered in nonrandomized studies. In the context of external control augmentation, it arises when the real-world baseline subjects' characteristics differ from those in the trial data. Multiple methods are available to adjust for selection bias by balancing the baseline covariates' distributions across the different data sources. For example, matching and subclassification approaches select a subset of comparable external controls to construct the hybrid control arm (Stuart, 2010). Matching on the propensity score or the probability of trial inclusion can balance numerous baseline covariates simultaneously (Rosenbaum & Rubin, 1983). Weighting approaches that reweight external controls using the probability of trial inclusion or other balancing scores have also been proposed, e.g., empirical likelihood (Qin et al., 2015), entropy balancing (Lee et al., 2022b; Wu & Yang, 2022b; Chu et al., 2023), constrained maximum likelihood (Chatterjee et al., 2016; Zhang et al., 2020) and Bayesian power priors (Neuenschwander et al., 2010; van Rosmalen et al., 2018). Furthermore, matching or weighting can be combined with outcome modelling to enhance robustness against model misspecification in addressing selection bias of external controls (Li et al., 2023).

Differences in the outcomes may still exist between the concurrent controls and the external controls after matching or weighting due to differences in study settings, time frame, data quality or the definition of covariates or outcomes (Phelan et al., 2017). Methods were proposed to adaptively select the degree of borrowing or adjust the outcomes for external controls based on observed outcome differences with concurrent controls. Some researchers suggested first testing the heterogeneity in control outcomes before deciding whether to incorporate external subjects into the hybrid control arm (Viele et al., 2014; Li et al., 2023). More dynamic borrowing approaches were also proposed, including matching and bias adjustment (Stuart & Rubin, 2008), power priors (Ibrahim & Chen, 2000; Neuenschwander et al., 2009), Bayesian hierarchical models including meta-analytic predictive priors (Neuenschwander et al., 2010; Schoenfeld et al., 2019) and commensurate

priors (Hobbs et al., 2011). While these existing methods seem appealing, simulation studies could not identify a single approach that could perform well across all scenarios where hidden biases exist (Shan et al., 2022). The surveyed Bayesian methods often have inflated Type-I errors, while frequentist methods suffer lower power when hidden biases exist. Nearly all methods performed poorly in the presence of unmeasured confounding and could not simultaneously minimize bias and gain power. Furthermore, many existing methods rely on parametric assumptions that are sensitive to model misspecification and cannot capture complex relationships that are prevalent in practice.

In this paper, we propose an approach to achieve an efficient estimation of treatment effects that is robust to various potential discrepancies that may arise in the external controls. When handling the selection bias of external controls, our proposal is based on calibration weighting (Lee et al., 2022b) so that the covariate distribution of external controls matches with that of the trial subjects. Furthermore, leveraging semiparametric theory, we develop an integrative augmented calibration weighting estimator, motivated by the efficient influence function (Bickel et al., 1998; Tsiatis, 2006), which is semiparametrically efficient and doubly robust against model misspecification. Despite the potential to view the selection bias problem as a generalizability or transportability issue (Lee et al., 2022b), our framework fundamentally diverges from theirs as our context encompasses the outcomes from both the trial data and external controls, while Lee et al. (2022b) solely considered the trial outcomes.

To deal with potential outcome heterogeneity, we develop a selective borrowing framework to determine an optimal subset from the external controls for integration. Specifically, we introduce a bias parameter for each external subject entailing his or her comparability with the concurrent control. To prevent bias in the integrative estimator, the goal is to select the comparable external controls with zero bias and exclude any others with nonzero bias. Thus, this formulation recasts the selective borrowing strategy as a model selection problem, which can be solved by penalized estimation (e.g., the adaptive lasso penalty; Zou, 2006). Subsequent to the selection process, comparable external controls are utilized to construct the integrative estimator. Prior works such as those by Chen et al. (2021), Liu et al. (2021) and Zhai & Han (2022) although able to identify biases, exclude the entire external sample when confronted with incomparability. Moreover, compared to these existing selective borrowing approaches, our method leverages off-the-shelf machine learning models to achieve semiparametric efficiency and does not require stringent parametric assumptions on the distribution of outcomes.

2. Methodology

2.1. *Notation, assumptions and objectives*

Let $\mathcal{R}$ represent a randomized placebo-controlled trial and $\mathcal{E}$ represent an external control source, which contain $N_{\mathcal{R}}$ and $N_{\mathcal{E}}$ subjects, respectively. The total sample size is $N = N_{\mathcal{R}}+N_{\mathcal{E}}$. An extension to multiple external control groups is discussed in the Supplementary Material. A total of N_t and N_c subjects receive the active treatment and control treatment in $\mathcal{R}$, while we assume that all $N_{\mathcal{E}}$ subjects in $\mathcal{E}$ receive the control. Each observation $i \in \mathcal{R}$ comprises the outcomes Y_i, the treatment assignment A_i and a set of baseline covariates X_i. Similarly, each observation $i \in \mathcal{E}$ comprises Y_i, A_i and X_i. Let R_i represent a data source indicator, which is 1 for all subjects $i \in \mathcal{R}$ and 0 for all subjects $i \in \mathcal{E}$. To sum up, an independent and identically distributed sample $\{V_i : i \in \mathcal{R} \cup \mathcal{E}\}$ is observed, where

$V = (X, A, Y, R)$. Let $Y(a)$ denote the potential outcomes under treatment a (Rubin, 1974). The causal estimand of interest is defined as the average treatment effect among the trial population, $\tau = \mu_1 - \mu_0$, where $\mu_a = E\{Y(a) \mid R = 1\}$ for $a = 0, 1$. The clinical trials for treatment effect estimation satisfy the following assumption.

Assumption 1 (*Consistency, randomization and positivity*). Suppose that

(i) $Y = AY(1) + (1 - A)Y(0)$,
(ii) $Y(a) \perp\!\!\!\perp A \mid (X, R = 1)$ for $a = 0, 1$ and
(iii) the known treatment propensity score satisfies

$$1 > \pi_A(x) = \text{pr}(A = 1 \mid X = x,\ R = 1) > 0$$

for all x such that $\text{pr}(X = x, R = 1) > 0$.

Assumption 1 is standard in the causal inference literature (Rosenbaum & Rubin, 1983; Imbens, 2004) and holds for the well-controlled clinical trials guaranteed by the randomization mechanism. Under Assumption 1, τ is identifiable with the trial data.

Moreover, the external controls should ideally be comparable with the concurrent controls.

Assumption 2 (*External control compatibility*). Suppose that

(i) $E\{Y(0) \mid X = x,\ R = 0\} = E\{Y(0) \mid X = x,\ R = 1\}$ and
(ii) $\text{pr}(R = 1 \mid X = x) > 0$ for all x such that $\text{pr}(X = x,\ R = 0) > 0$.

Assumption 2 states that the conditional mean of $Y(0)$ is the same for the trial data and external controls. This assumption holds if X captures all the outcome predictors that are correlated with R. From the guidance in FDA (2023) for drug development in rare diseases, there are five main concerns regarding the use of external controls: (i) selection bias, (ii) unmeasured confounding, (iii) lack of concurrency, (iv) data quality and (v) outcome validity. Assumption 2 does not require the covariate distribution of external controls to be the same as that of the trial data, which is referred to as selection bias in the guidance. Under Assumption 2, borrowing external controls to improve treatment effect estimation is similar to a transportability or covariate shift problem. However, the presence of concerns (ii)–(v) can result in violation of Assumption 2. Our paper has two main objectives: (i) under Assumption 2, similarly to the work of Li et al. (2023), we develop a semiparametrically efficient and robust strategy to borrow external controls to improve estimation while correcting for selection bias (§ 2.2); (ii) considering that Assumption 2 can be potentially violated, we incorporate a selective borrowing procedure that will detect the biases and retain only a subset of comparable external controls for integration (§ 2.3).

2.2. *Semiparametric efficient estimation under the ideal situation*

From the semiparametric theory (Bickel et al., 1998), we derive efficient and robust estimators for τ under Assumptions 1 and 2. The derivation reaches the same estimator as Li et al. (2023), and will serve as the base for our selective borrowing strategy. The semiparametric model is attractive as it exploits the observed data without making assumptions about the nuisance parts of the data generation process that are not of substantive interest. We derive the efficient influence function of τ in Theorem 1 below, which shall serve as the foundational component of our proposed framework.

THEOREM 1. *Under Assumptions 1 and 2, the efficient influence function of* τ *is*

$$\psi_{\tau,\text{eff}}(V;\mu_1,\mu_0,q,r) = \frac{R}{\text{pr}(R=1)}\left[\{\mu_1(X)-\mu_0(X)-\tau\}+\frac{A\{Y-\mu_1(X)\}}{\pi_A(X)}\right] - \frac{R(1-A)+(1-R)r(X)}{\text{pr}(R=1)}\ \frac{q(X)\{Y-\mu_0(X)\}}{q(X)\{1-\pi_A(X)\}+r(X)},$$

where

$$\begin{aligned}\mu_1(X) &= E(Y \mid X, R=1, A=1),\\ \mu_0(X) &= E(Y \mid X, R=1, A=0) = E(Y \mid X, R=0),\\ r(X) &= \text{var}(Y \mid X, R=1, A=0)/\text{var}(Y \mid X, R=0),\\ q(X) &= \text{pr}(R=1 \mid X)/\text{pr}(R=0 \mid X).\end{aligned}$$

Based on Theorem 1, the semiparametric efficiency bound for τ is $\mathbb{V}_{\tau,\text{eff}} = E\{\psi^2_{\tau,\text{eff}}(V;\mu_1,\mu_0,q,r)\}$. Hence, a principled estimator can be motivated by solving the empirical analogue of $E\{\psi_{\tau,\text{eff}}(V;\mu_1,\mu_0,q,r)\}=0$ for τ.

Let the estimators of (μ_0,μ_1,q,r) be $(\hat\mu_0,\hat\mu_1,\hat q,\hat r)$, and define $\hat\epsilon_{a,i} = Y_i - \hat\mu_a(X_i)$ $(a = 0,1)$. Then, by solving the empirical version of the efficient influence function for τ, we have

$$\hat\tau = \frac{1}{N_{\mathcal{R}}}\sum_{i\in\mathcal{R}\cup\mathcal{E}} R_i\left\{\hat\mu_1(X_i)-\hat\mu_0(X_i)+\frac{A_i\hat\epsilon_{1,i}}{\pi_A(X_i)}\right\} - \frac{1}{N_{\mathcal{R}}}\sum_{i\in\mathcal{R}\cup\mathcal{E}}\frac{\{R_i(1-A_i)+(1-R_i)\hat r_i(X_i)\}\hat q(X_i)}{\hat q(X_i)\{1-\pi_A(X_i)\}+\hat r(X_i)}\hat\epsilon_{0,i}. \quad (1)$$

We now discuss the estimators for the nuisance functions (μ_0,μ_1,q,r). To estimate $\mu_0(X)$, $\mu_1(X)$ and $r(X)$, one can follow the standard approach by fitting parametric models based on the trial data.

For estimating weight $q(X)$, a direct approach is to predict $\text{pr}(R=0 \mid X)$, which however is unstable due to inverting probability estimates. To achieve stability of weighting, the key insight is based on the central role of $q(X)$ as balancing the covariate distribution between two groups: $E\{(1-R)q(X)g(X)\} = E\{Rg(X)\}$ for any $g(X) = \{g_1(X),\ldots,g_K(X)\}$, which is a K-dimensional function of X. Thus, we estimate $q(X)$ by calibrating the covariate balance between the trial data and external controls. In particular, we assign a weight q_i for each subject $i\in\mathcal{E}$, then solve the following optimization problem for $Q=\{q_i\colon i\in\mathcal{E}\}$:

$$\min_q L(Q) = \sum_{i\in\mathcal{E}} q_i\log q_i$$

subject to (i) $q_i \geqslant 0$, $i\in\mathcal{E}$, (ii) $\sum_{i\in\mathcal{E}} q_i g(X_i) = \sum_{i\in\mathcal{R}} g(X_i)$. First, $L(Q)$ is the entropy of the weights; thus, minimizing this criterion ensures that the calibration weights are not too far from uniform, so it minimizes the variability due to heterogeneous weights. Constraint (i) is a standard condition for the weights. Constraint (ii) forces the empirical moments of the covariates to be the same after calibration, leading to better-matched distributions of the trial data and external controls.

The optimization problem can be solved using constrained convex optimization. The estimated calibration weight is $\hat{q}_i = q(X_i; \hat{\eta}) = \exp\{\hat{\eta}^{\mathrm{T}} g(X_i)\}$, and $\hat{\eta}$ solves $U(\eta) = \sum_{i\in\mathcal{E}} \exp\{\eta^{\mathrm{T}} g(X_i)\} g(X_i) - \sum_{i\in\mathcal{R}} g(X_i) = 0$, which is the Lagrangian dual problem to the optimization problem. The dual problem also entails that the calibration weighting approach makes a log regression model for $q(X)$. We refer to $\hat{\tau}$ with calibration weights as the augmented calibration weighting estimator $\hat{\tau}_{\mathrm{acw}}$.

Remark 1. The variance ratio $r(X)$ quantifies the relative residual variability of $Y(0)$ given X between the trial data and external controls. In general, estimating the conditional variance ratio involves nonparametric regression, which can be challenging; see Shen et al. (2020) and the references therein. Fortunately, the consistency of $\hat{\tau}_{\mathrm{acw}}$ does not rely on the correct specification of $r(X)$. For example, if $\hat{r}(X)$ is set to be zero, $\hat{\tau}_{\mathrm{acw}}$ reduces to the trial-only estimator without borrowing any external information, which is always consistent. In order to leverage external information and estimate $r(X)$ practically, we can make a simplifying homoscedasticity assumption that the residual variances of $Y(0)$ after addressing X are constant over studies. In this case, $r(X)$ can be estimated by $\hat{r} = N_{\mathcal{E}} N_c^{-1} \sum_{i\in\mathcal{R}} (1 - A_i)\{Y_i - \hat{\mu}_0(X_i)\}^2 / \sum_{i\in\mathcal{E}} \{Y_i - \hat{\mu}_0(X_i)\}^2$.

We show that $\hat{\tau}_{\mathrm{acw}}$ has the following desirable properties. (i) Local efficiency: $\hat{\tau}_{\mathrm{acw}}$ achieves the semiparametric efficiency bound if the nuisance functions are correctly specified. (ii) Double robustness: $\hat{\tau}_{\mathrm{acw}}$ is consistent for τ if either the model for $\mu_a(X)$ or that for $q(X)$ is correct; see the proof in the Supplementary Material.

The doubly robust estimators were initially developed to gain robustness to parametric misspecification, but are now known to also be robust to approximation errors using machine learning methods (e.g., Chernozhukov et al., 2018). We investigate this new doubly robust feature for the proposed estimator $\hat{\tau}_{\mathrm{acw}}$, and use flexible semiparametric or nonparametric methods to estimate both $\mu_a(X)$ $(a = 0, 1)$, $r(X)$ and $q(X)$ in (1). First, we consider the method of sieves (Chen, 2007) for $q(X)$. In comparison with other nonparametric methods such as kernels, the method of sieves is particularly well suited for calibration weighting. We consider general sieve basis functions such as power series, Fourier series, splines, wavelets and artificial neural networks; see Chen (2007) for a comprehensive review. The number of bases can be selected by cross-validation. Second, we consider flexible outcome models, e.g., generalized additive models, kernel regression and the method of sieves for $\mu_a(X)$ $(a = 0, 1)$. Using flexible methods alleviates bias from the misspecification of parametric models. The following regularity conditions are required for the nuisance function estimators.

Assumption 3. For a function $f(X)$ with a generic random variable X, define its L_2 norm as $\|f(X)\| = \{\int f(x)^2 \,\mathrm{dpr}(x)\}^{1/2}$. Assume that

(i) $\|\hat{\mu}_a(X) - \mu_a(X)\| = o_p(1)$, $a = 0, 1$ and $\|\hat{q}(X) - q(X)\| = o_p(1)$,
(ii) $\|\hat{\mu}_0(X) - \mu_0(X)\| \times \|\hat{q}(X) - q(X)\| = o_p(N^{-1/2})$,
(iii) $\|\hat{r}(X) - r^*(X)\| = o_p(1)$ for some $r^*(X)$,
(iv) the additional regularity conditions Assumptions S1 and S2 in the Supplementary Material hold.

Assumption 3 is a set of typical regularity conditions for M-estimation to achieve rate double robustness (Van der Vaart, 2000). Under these regularity conditions, our proposed

framework can incorporate flexible methods for estimating the nuisance functions, while maintaining parametric rate consistency for $\hat{\tau}_{\text{acw}}$.

THEOREM 2. *Under Assumptions 1–3, we have* $N^{1/2}(\hat{\tau}_{\text{acw}} - \tau) \xrightarrow{\text{D}} N(0, \mathbb{V}_\tau)$, *where* $\mathbb{V}_\tau = E\{\psi^2_{\tau,\text{eff}}(V; \mu_1, \mu_0, q, r^*)\}$. *If* $r^*(X) = r(X)$, $\hat{\tau}_{\text{acw}}$ *achieves semiparametric efficiency.*

Theorem 2 motivates variance estimation by $\hat{\mathbb{V}}_r = N^{-1} \sum_{i \in \mathcal{R} \cup \mathcal{E}} \psi^2_{\tau,\text{eff}}(V_i; \hat{\mu}_1, \hat{\mu}_0, \hat{q}, \hat{\tau}_{\text{acw}})$, which is consistent for V_τ under Assumptions 1–3.

2.3. *Bias detection and selective borrowing*

In practical situations, Assumption 2 may not hold, and the augmentation in (1) can be biased. We develop a selective borrowing framework to select external subjects that are comparable with the concurrent controls for integration. To account for potential violations, we introduce a vector of bias parameters $b_0 = (b_{1,0}, \ldots, b_{N_\mathcal{E},0})$ for all $i \in \mathcal{E}$, where $b_{i,0} = b_0(X_i) = E(Y_i \mid X_i, A_i = 0, R_i = 0) - E(Y_i \mid X_i, A_i = 0, R_i = 1) = \mu_{0,\mathcal{E}}(X_i) - \mu_0(X_i)$. When Assumption 2 holds, we have $b_0 = 0$. Otherwise, there exists at least one $i \in \mathcal{E}$ such that $b_{i,0} \neq 0$. To prevent bias in $\hat{\tau}_{\text{acw}}$ from incomparable external controls, the goal is to select the comparable subset with $b_{i,0} = 0$ and exclude any others with $b_{i,0} \neq 0$.

Let $\hat{b}_i = \hat{\mu}_{0,\mathcal{E}}(X_i) - \hat{\mu}_0(X_i)$ be a consistent estimator for $b_{i,0}$, where $\hat{\mu}_{0,\mathcal{E}}(X_i)$ is a consistent estimator for $\mu_{0,\mathcal{E}}(X_i)$. Let $\hat{b} = (\hat{b}_1, \ldots, \hat{b}_{N_\mathcal{E}})$ be an initial estimator for b_0. We propose a refined estimator of b_0 by penalized estimation:

$$\tilde{b} = \arg\min_b \left\{ (\hat{b} - b)^{\text{T}} \hat{\Sigma}_b^{-1} (\hat{b} - b) + \lambda_N \sum_{i \in \mathcal{E}} p(|b_i|) \right\}. \tag{2}$$

Here $\hat{\Sigma}_b$ is the estimated variance of $\hat{b}$, $p(|b_i|) = |b_i|/|\hat{b}_i|^\nu$ is the adaptive lasso penalty term and (λ_N, ν) are two tuning parameters. Intuitively, if $\hat{b}_i$ is close to zero, the associated penalty will be large, which further shrinks estimate $\tilde{b}_i$ towards zero. According to Zou (2006), Huang et al. (2008) and Lin et al. (2009), the adaptive lasso penalty can lead to a desirable property under the following regularity conditions.

Assumption 4. Suppose that

(i) $a_N \max_i\{\hat{\mu}_0(X_i) - \mu_0(X_i)\} = O_p(1)$ and $a_N \max_i\{\hat{\mu}_{0,\mathcal{E}}(X_i) - \mu_{0,\mathcal{E}}(X_i)\} = O_p(1)$ for all $i \in \mathcal{E}$,
(ii) there exist constants τ_1 and τ_2 such that $0 < \tau_1 \leqslant \tau_{b,\min} \leqslant \tau_{b,\max} \leqslant \tau_2$, where $\tau_{b,\min}$ and $\tau_{b,\max}$ are the smallest and largest eigenvalues of $\hat{\Sigma}_b$,
(iii) $a_N b_{\min} \to \infty$, where $b_{\min} = \min\{b_{i,0}, i \notin \mathcal{A}\}$, and
(iv) $\lambda_N / b_{\min}^{\nu+1} \to 0$ and $\lambda_N a_N^\nu \to \infty$.

LEMMA 1. *Suppose that the assumptions in Theorem 2 and Assumption 4 hold except that Assumption 2 may be violated. We have* $\lim_{N\to\infty} \text{pr}(\tilde{\mathcal{A}} = \mathcal{A}) = 1$.

Lemma 1 shows that the adaptive lasso penalty has the ability to select zero-valued parameters consistently when using an a_N-consistent initial estimator $\hat{b}_i$ and proper choices of (λ_N, ν), provided that the minimum of the nonzero bias $b_{\min}$ does not diminish too fast and the initial estimator $\hat{b}_i$ is sufficiently good. In practice, the initial estimator $\hat{b}_i$ can be obtained by leveraging off-the-shelf machine learning models with a guaranteed

convergence rate, and (λ_N, ν) are selected by minimizing the mean square error using cross-validation. Given $\tilde{b}$, the selected set of comparable external controls is $\tilde{\mathcal{A}} = \{i\colon \tilde{b}_i = 0\}$. The modified integrative estimator is

$$\hat{\tau}^{\text{alasso}}_{\text{acw}} = \frac{1}{N_{\mathcal{R}}} \sum_{i \in \mathcal{R} \cup \mathcal{E}} R_i \left[\hat{\mu}_1(X_i) - \hat{\mu}_0(X_i) + \frac{A_i \hat{\epsilon}_{1,i}}{\pi_A(X_i)} \right] - \frac{1}{N_{\mathcal{R}}} \sum_{i \in \mathcal{R} \cup \mathcal{E}} \frac{\{R_i(1 - A_i) + (1 - R_i)\hat{r}_b(X_i)\mathbb{1}(\tilde{b}_i = 0)\}\hat{q}(X_i)}{\hat{q}(X_i)\{1 - \pi_A(X_i)\} + \hat{r}_b(X_i)\text{pr}(\tilde{b}_i = 0 \mid X, R = 0)} \hat{\epsilon}_{0,i}, \tag{3}$$

where $\hat{r}_b(X)$ is the estimated function of $r_b(X) = \text{var}(Y \mid X, R = 1, A = 0)/\text{var}(Y \mid X, R = 0, b_0 = 0)$, which is used to adjust for changes in the covariate distribution from all external controls in $\mathcal{E}$ to $\tilde{\mathcal{A}}$.

Following the suggestions of Ho et al. (2007) to improve the finite-sample performances, nearest-neighbour matching based on the estimated probability of trial inclusion $e(X) = \text{pr}(R = 1 \mid X)$ is performed after selecting the comparable subset $\tilde{\mathcal{A}}$, which ensures a more balanced allocation ratio between the treated group and the hybrid control arm; see Algorithm 1 below for an overview of our selective borrowing framework.

Algorithm 1. Proposed selective integrative estimator.

Input: a randomized controlled trial with size $N_{\mathcal{R}} = N_t + N_c$ and external controls.

Step 1. Fit the models for the outcome means $\mu_1, \mu_0, \mu_{0,\mathcal{E}}$ and weights q.
Step 2. Construct the initial estimator $\hat{b}$ for the bias parameter b_0.
Step 3. Select the comparable subset $\tilde{\mathcal{A}} = \{i\colon \tilde{b}_i = 0\}$ via the bias penalization (2).
Step 4. If $|\tilde{\mathcal{A}}| > N_t - N_c$ then perform the nearest-neighbour matching to select $N_t - N_c$ external controls as the final $\tilde{\mathcal{A}}$; otherwise, jump to step 5.
Step 5. Compute $\hat{\tau}^{\text{alasso}}_{\text{acw}}$ in (3) using the selected external controls in $\tilde{\mathcal{A}}$.

We show the efficiency gain of the proposed estimator compared to the trial-only estimator.

Theorem 3. *Suppose that the assumptions in Theorem 2 and Assumption 4 hold except that Assumption 2 may be violated. Let* $r^*_b(X) = r_b(X)$. *The reduction of the asymptotic variance of* $\hat{\tau}^{\text{alasso}}_{\text{acw}}$ *compared to the trial-only estimator is*

$$\frac{1}{\text{pr}^2(R = 1)} E\left[\frac{\text{pr}(R = 1 \mid X) r_b(X) \mathbb{1}(b_0 = 0) \text{var}(Y \mid X, R = 1, A = 0)}{[q(X)\{1 - \pi_A(X)\} + r_b(X)\text{pr}(b_0 = 0 \mid X, R = 0)]\{1 - \pi_A(X)\}} \right], \tag{4}$$

which is strictly positive unless $r_b(x) = 0$ *or* $b_0 \neq 0$ *or* $\text{var}(Y \mid X, R = 1, A = 0) = 0$ *for all* x *such that* $\text{pr}(X = x) > 0$.

We derive (4) using orthogonality of the efficient influence function of τ to the nuisance tangent space, and relegate the details to the should be highlighted. Theorem 3 showcases the advantage of including external controls in a data-adaptive manner, where the asymptotic variance of $\hat{\tau}^{\text{alasso}}_{\text{acw}}$ should be strictly smaller than the trial-only estimator unless the external controls all suffer exceeding noise, i.e., $r_b(X_i) = 0$, or the compatible subset $\mathcal{A}$ of the external

controls is an empty set, i.e., $b_0 \neq 0$, or the covariate X captures all the variability of $Y(0)$ in the trial data, i.e., $\text{var}(Y \mid X, R = 1, A = 0) = 0$. Below, we establish the asymptotic properties and provide a valid inferential framework for the proposed integrative estimator; more details are provided in the Supplementary Material.

THEOREM 4. *Suppose that the assumptions in Theorem 2 and Assumption 4 hold except that Assumption 2 may be violated. We have* $N^{1/2}(\hat{\tau}^{\text{alasso}}_{\text{acw}} - \tau) \to N(0, \mathbb{V}^{\text{alasso}}_{\tau})$. *Furthermore, the* $(1-\alpha) \times 100\%$ *confidence interval* $[L_\tau, U_\tau]$ *for* τ *can be constructed as*

$$[L_\tau, U_\tau] = [\hat{\tau}^{\text{alasso}}_{\text{acw}} - z_{\alpha/2}(\hat{\mathbb{V}}^{\text{alasso}}_{\tau}/N)^{1/2}, \hat{\tau}^{\text{alasso}}_{\text{acw}} + z_{\alpha/2}(\hat{\mathbb{V}}^{\text{alasso}}_{\tau}/N)^{1/2}],$$

where $\hat{\mathbb{V}}^{\text{alasso}}_{\tau}$ *is a variance estimator of* $\mathbb{V}^{\text{alasso}}_{\tau}$, $z_{\alpha/2}$ *is the* $1-\alpha/2$ *quantile for the standard normal distribution and* $[L_\tau, U_\tau]$ *satisfies* $\text{pr}(\tau \in [L_\tau, U_\tau]) \to 1-\alpha$ *as* $N \to \infty$.

3. SIMULATION

In this section, we evaluate the finite-sample performance of the proposed framework to estimate treatment effects under potential bias scenarios via plasmode simulations. First, a set of $d = 12$ baseline covariates $X \in \mathbb{R}^d$ is generated by mimicking the correlation structure and the moments (up to the sixth) of variables from an oncology randomized placebo-controlled trial (i.e., the trial data) and the Flatiron Health Spotlight Phase 2 cohort (© 2020 Flatiron Health, all rights reserved; external controls).

Next, we generate the data source indicator R_i as $R_i \mid X_i, U_i \sim \text{Ber}\{\pi_R(X_i, U_i)\}$ given the sample sizes $(N_{\mathcal{R}}, N_{\mathcal{E}})$, where U_i represents an unmeasured confounder. The treatment assignment for the trial data is completely at random (i.e., $\text{pr}(A_i = 1 \mid R_i = 1) = N_t/N_{\mathcal{R}}$), while all external subjects receive the control (i.e., $\text{pr}(A_i = 0 \mid R_i = 0) = 1$). The outcomes Y_i are generated as

$$Y_i \mid (X_i, A_i, U_i, R_i = 1) \sim N\{\mu_0(X_i, U_i, A_i), \sigma^2_Y\},$$
$$Y_i \mid (X_i, U_i, R_i = 0) \sim N\{\mu_{0,\mathcal{E}}(X_i, U_i), \sigma^2_Y\}.$$

We consider three data-generating scenarios in Table 1(a), where η_0 is chosen adaptively to ensure the desired sample sizes $(N_{\mathcal{R}}, N_{\mathcal{E}})$, and $(\eta, \beta, \tilde{\eta}, \tilde{\beta}, \sigma^2_Y)$ are chosen empirically based on the model fits using the observed oncology clinical trial data. In all the scenarios, we use the linear predictor of X to fit $(q, \mu_0, \mu_{0,\mathcal{E}})$, and thus the models are correctly specified under the model choices C, where the linear predictor of X governs the true data generation, but are misspecified under choices W, where the data generation depends on a new set of covariates $\tilde{X}$, which include the quadratic and cubic terms of the $(d-1)$th and dth covariates (i.e., $X^2_{d-1}, X^2_d, X^3_{d-1}, X^3_d$) addition to the baseline covariate X. Moreover, we utilize the cross-fitting procedure to select tuning parameters for the gradient boosting model.

The proposed framework is evaluated on imbalanced trial data, where $N_c = (20, 30, 40, 50, 75, 100)$ and $N_t = 200$ with an external control group of size $N_{\mathcal{E}} = 3000$. We investigate the performance of our proposed estimator under two levels of unmeasured confounding ($\omega = 0$ and 0.3) by comparing with other estimators in Table 1(b). The trial-only augmented inverse probability weighting estimator $\hat{\tau}_{\text{aipw}}$ (Cao et al., 2009) and the augmented calibration weighting estimator $\hat{\tau}_{\text{acw}}$ with full borrowing (Li et al., 2023) are used as benchmarks. Two data-adaptive integrative estimators, $\hat{\tau}^{\text{alasso}}_{\text{acw}}$ and $\hat{\tau}^{\text{alasso}}_{\text{acw,gbm}}$, are

Table 1. *Simulation settings: (a) model choices (C and W), where* $\tilde{X} = [X, X_{d-1}^2, X_d^2, X_{d-1}^3, X_d^3]$, *and (b) descriptions of the five estimators*

(a) Model choices

	$\text{logit}\{\pi_R(X,U)\}$	$\mu_0(X,U,A)$	$\mu_{0,\mathcal{E}}(X,U)$
C	$\eta_0 + \eta^{\mathrm{T}} X + \omega U$	$\beta^{\mathrm{T}} X + A\alpha^{\mathrm{T}}(1,X) + \omega U \sigma_Y$	$\beta^{\mathrm{T}} X + \omega U \sigma_Y + \omega \sigma_Y$
W	$\eta_0 + \tilde{\eta}^{\mathrm{T}} \tilde{X} + \omega U$	$\tilde{\beta}^{\mathrm{T}} \tilde{X} + A\alpha^{\mathrm{T}}(1,X) + \omega U \sigma_Y$	$\tilde{\beta}^{\mathrm{T}} \tilde{X} + \omega U \sigma_Y + \omega \sigma_Y$

(b) Estimators

$\hat{\tau}_{\text{aipw}}$	The augmented inverse probability weighting estimator without borrowing (Cao et al., 2009)
$\hat{\tau}_{\text{acw}}$	The integrative augmented calibration weighting estimator with full borrowing (Li et al., 2023)
$\hat{\tau}_{\text{acw}}^{\text{alasso}}$	The data-adaptive integrative estimator using the linear regressions for $(\mu_0, \mu_{0,\mathcal{E}})$
$\hat{\tau}_{\text{acw,gbm}}^{\text{alasso}}$	The data-adaptive integrative estimator using the tree-based gradient boosting for $(\mu_0, \mu_{0,\mathcal{E}})$
$\hat{\tau}_{\text{ppp}}$	The Bayesian predictive p-value power prior estimator (Kwiatkowski et al., 2023)

considered, where linear regressions and tree-based gradient boosting are used to estimate the nuisance models. Other machine learning algorithms that satisfy pointwise consistency, such as the generalized additive model, can also be utilized to select a comparable subset of external controls consistently. The Bayesian predictive p-value power prior estimator, $\hat{\tau}_{\text{ppp}}$, is an extension of the power prior, which discounts each external control according to its outcome compatibility using Box's p-value (Kwiatkowski et al., 2023).

Figure 1 displays the average bias, variance, mean squared error and Type-I error when $E\{\tau(X) \mid R = 1\} = 0$, and power for testing $\tau > 0$ when $E\{\tau(X) \mid R = 1\} = 0.3$ based on 1000 sets of data replications. Over the three model scenarios, the trial-only estimator $\hat{\tau}_{\text{aipw}}$ is always consistent, but lacks efficiency as it only utilizes the concurrent controls for estimation, especially when N_c is small. When the conditional mean exchangeability in Assumption 2 holds (i.e., $\omega = 0$), the full-borrowing estimator $\hat{\tau}_{\text{acw}}$ is most efficient, shown by its low mean squared error and high power for detecting a significant treatment effect. Our proposed selective integrative estimators, $\hat{\tau}_{\text{acw}}^{\text{alasso}}$ and $\hat{\tau}_{\text{acw,gbm}}^{\text{alasso}}$, may be less efficient than $\hat{\tau}_{\text{acw}}$ due to finite-sample selection error. However, they maintain smaller variance and improved power compared to $\hat{\tau}_{\text{aipw}}$, regardless of whether the nuisance models are misspecified. When Assumption 2 is violated (i.e., $\omega = 0.3$), $\hat{\tau}_{\text{acw}}$ becomes biased, leading to an inflated Type-I error and low power. The Bayesian estimator $\hat{\tau}_{\text{ppp}}$ requires correct parametric specification of the outcome model and performs poorly when the model omits a key confounder that is imbalanced between data sources. In our simulations, high weights were assigned to the external control subjects, which led to some bias in the treatment effect estimates when N_c was small. However, both $\hat{\tau}_{\text{acw}}^{\text{alasso}}$ and $\hat{\tau}_{\text{acw,gbm}}^{\text{alasso}}$ achieve smaller mean squared errors than the trial-only estimator by incorporating external control subjects. In cases where the outcome model is incorrectly specified and $\omega = 0.3$, the benefit of using machine learning methods becomes apparent. Specifically, the flexibility of the gradient boosting model ensures the convergence rate assumption for $\hat{b}_i$, i.e., $a_N(\hat{b}_i - b_{i,0}) = O_p(1)$ for a certain sequence a_N (Zhang & Yu, 2005). By incorporating compatible external controls more accurately, $\hat{\tau}_{\text{acw,gbm}}^{\text{alasso}}$ better controls bias and achieves comparable power levels to $\hat{\tau}_{\text{acw}}^{\text{alasso}}$. However, the adaptive lasso estimation based on the misspecified linear model lacks such properties and may not provide gains in power. One notable trade-off of our proposed estimators is the slight Type-I error inflation when N_c is small and Assumption 2 is violated, which can be attributed to finite-sample selection error and was also observed by Viele et al. (2014).

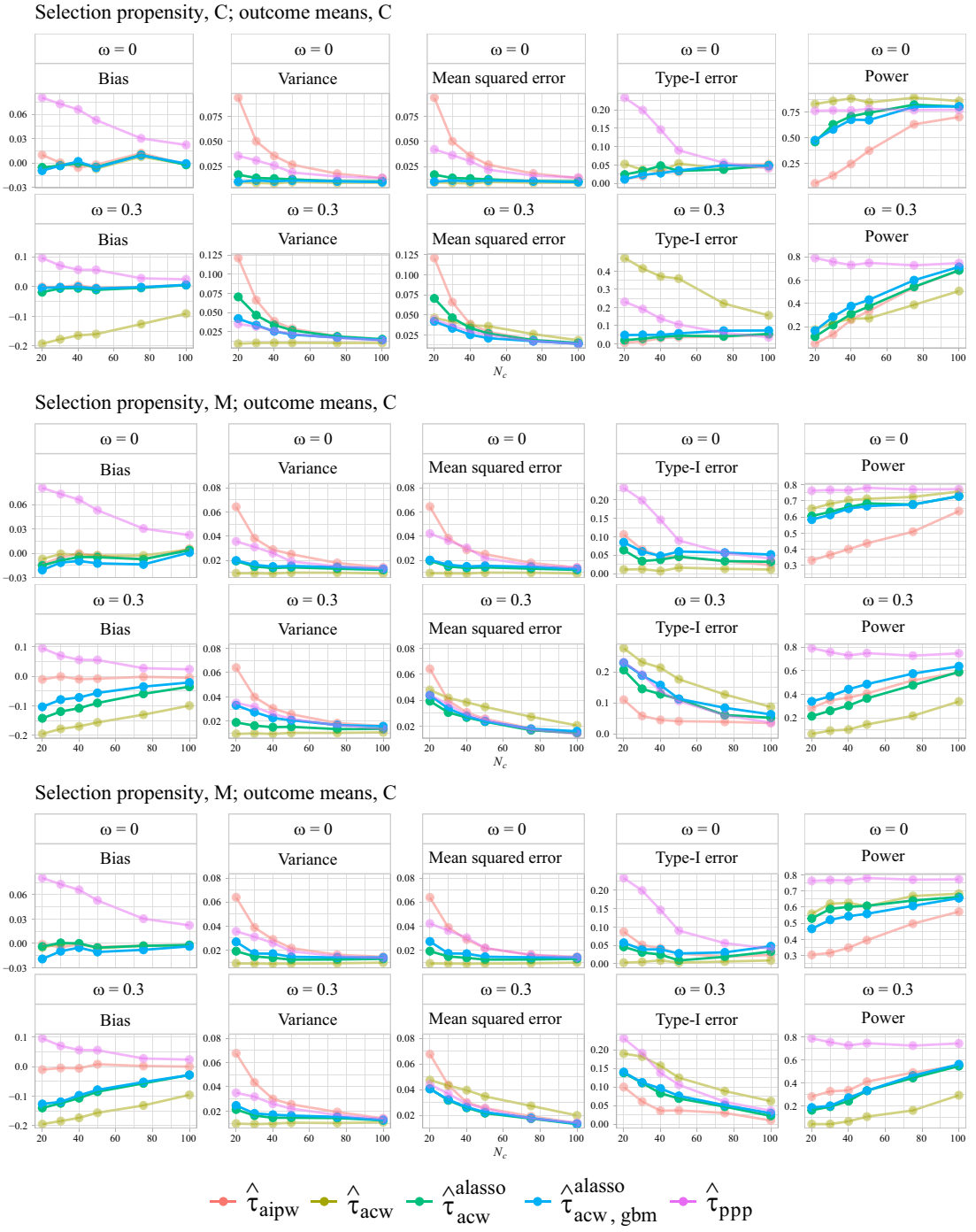

Fig. 1. Simulation results under various levels of ω, and different model choices of $q(X)$ and $\mu_0(X)$.

4. Real-data application

In this section, we present an application of the proposed methodology to investigate the effectiveness of basal insulin lispro against regular insulin glargine in patients with Type-I diabetes. When combined with preprandial insulin lispro, basal insulin lispro and insulin

Table 2. *Point estimates, standard errors and 95% confidence intervals of the treatment effect of BIL against regular GL based on the IMAGINE-1 and IMAGINE-3 studies*

	$\hat{\tau}_{\text{aipw}}$	$\hat{\tau}_{\text{acw}}$	$\hat{\tau}_{\text{acw}}^{\text{alasso}}$	$\hat{\tau}_{\text{acw,gbm}}^{\text{alasso}}$	$\hat{\tau}_{\text{ppp}}$
Est. (SE)	−0.25 (0.072)	−0.22 (0.057)	−0.24 (0.065)	−0.25 (0.070)	−0.27 (0.062)
CI	(−0.39, −0.11)	(−0.33, −0.11)	(−0.37, −0.08)	(−0.39, −0.12)	(−0.39, −0.15)

Est., estimate; SE, standard error; CI, confidence interval; BIL, basal insulin lispro; GL, regular insulin glargine.

glargine are two long-acting insulin formulations used for patients with Type-I diabetes mellitus. We analyse the IMAGINE-1 study, a randomized controlled trial where participants were unevenly assigned to either basal insulin lispro (treatment group) or insulin glargine (control group). Additionally, external control subjects from the IMAGINE-3 trial were used. In the Supplementary Material we also explore the effectiveness of solanezumab versus the placebo in slowing Alzheimer's disease progression using external observational data.

Our primary objective is to test the hypothesis of whether basal insulin lispro is superior to regular insulin glargine at glycemic control for patients with Type-I diabetes mellitus. This can be achieved by comparing the deviation of the hemoglobin A1c level from baseline after 52 weeks of treatment. Both studies contain a rich set of baseline covariates X, such as age, gender, baseline hemoglobin A1c (%), baseline fasting serum glucose (mmol/L), baseline triglycerides (mmol/L), baseline low-density lipoprotein cholesterol (mmol/L) and baseline alanine transaminase (U/L). The primary analysis population in IMAGINE-1 was the randomized patients who received at least one treatment dose. To mimic the full-analysis population from IMAGINE-1, external control subjects with missing baseline assessments are discarded from IMAGINE-3. The last observation carried forward is used to impute missing postbaseline outcomes. The IMAGINE-1 study consists of $N_{\mathcal{R}} = 439$ subjects with 286 in the treated group and 153 in the control group, while the IMAGINE-3 study includes $N_{\mathcal{E}} = 444$ patients in the control arm. In our statistical analysis, we first use the baseline covariates X to model the trial inclusion probability by calibration weighting under the entropy loss function. Next, we assume a linear heterogeneity treatment effect function for the outcomes with X as the treatment modifier, and compare the same set of estimators in the simulation study.

Table 2 reports the estimated results. The trial-only estimator $\hat{\tau}_{\text{aipw}}$ shows that basal insulin lispro has a significant treatment effect on reducing the glucose level solely based on the IMAGINE-1 study. Because of potential population bias, the naively integrative estimators $\hat{\tau}_{\text{acw}}$ and $\hat{\tau}_{\text{ppp}}$, albeit significant, are slightly different from $\hat{\tau}_{\text{aipw}}$, which may be subject to possible biases of the external controls. After filtering out the incompatible patients from the external controls by our adaptive lasso selection, the final integrative estimates $\hat{\tau}_{\text{acw}}^{\text{alasso}}$ and $\hat{\tau}_{\text{acw,gbm}}^{\text{alasso}}$ are closer to the benchmark, but have narrower confidence intervals. According to our adaptive analysis result, basal insulin lispro is significantly more effective than regular insulin glargine at glycemic control when used for patients with Type-I diabetes mellitus.

Next, we compare the performances of $\hat{\tau}_{\text{aipw}}$ with our data-adaptive integrative estimates to highlight the advantages of our dynamic borrowing framework. To this end, we retain the size of the treatment group, but create 100 subsamples by randomly selecting N_c^s patients from its control group, where $N_c^s = 10, \ldots, 153$. Then, the patients treated with regular insulin glargine in the IMAGINE-3 study are augmented to each selected subsample and

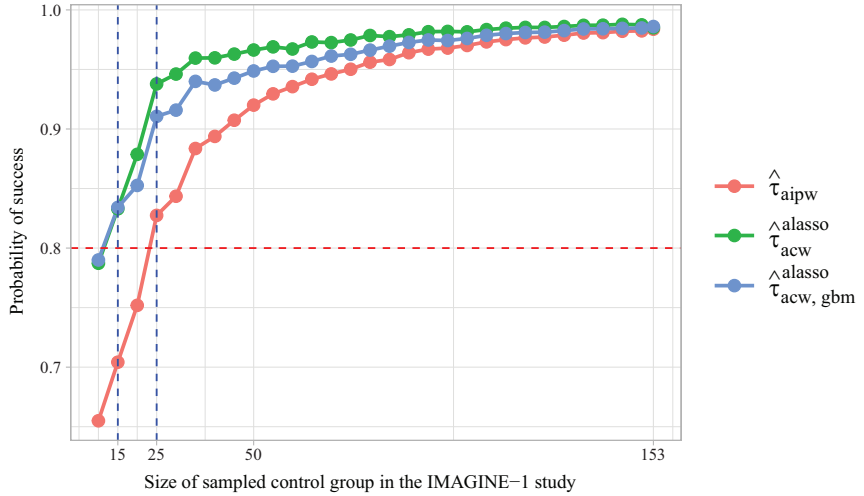

Fig. 2. Probability of success for detecting $\tau < -0.1$ by $\hat{\tau}_{\text{aipw}}$, $\hat{\tau}^{\text{alasso}}_{\text{acw}}$ and $\hat{\tau}^{\text{alasso}}_{\text{acw,gbm}}$ with varying control group sizes in the IMAGINE-1 study.

the treatment effect is evaluated upon the hybrid control arm design. Figure 2 presents the average probabilities of successfully detecting $\tau < -0.1$, the so-called probability of success, against the size of subsamples. When solely utilizing patients from the IMAGINE-1 study, $\hat{\tau}_{\text{aipw}}$ produces a probability of success larger than 0.8 only if the size of the control group is larger than 25. Combined with the IMAGINE-3 study, $\hat{\tau}^{\text{alasso}}_{\text{acw}}$ and $\hat{\tau}^{\text{alasso}}_{\text{acw,gbm}}$ refine the treatment effect estimation and only 15 patients are needed in the concurrent control group to attain a probability of success higher than 0.8. Therefore, by properly leveraging the external controls, we may accelerate drug development by decreasing the number of patients on the concurrent control, thereby reducing the duration and cost of the clinical trial.

5. Discussion

Interest in the use of external control arms for drug development is becoming more common. However, concerns regarding their quality and validity have limited their use for healthcare decision-making thus far, necessitating careful and appropriate assessment. To adjust for potential selection bias, our proposed method calibrates the covariate moments across two data sources, ensuring that the covariate distributions in both sources match each other. Alternative predictive model-based strategies are applicable when only a subset of covariates is shared (Stuart et al., 2011; Tipton, 2014). To address differences in outcomes, we select comparable external subsets based on the adaptive lasso penalty. Alternative penalties can be considered if the selection consistency property is attained, such as the smoothly clipped absolute deviation penalty (Fan & Li, 2001). Moreover, our framework can be easily extended to augment observational studies with external data, which may require additional modelling and assumptions to achieve double robustness. Slight Type-I error inflation is observed in our simulations when the concurrent control group is small, attributed to selection error in finite samples. One future direction will be to rigorously construct a data-adaptive confidence interval to account for finite-sample selection uncertainty without being overly conservative (Lee et al., 2016; Tibshirani et al., 2016). Other future directions include extending the proposed integrated inferential framework to survival outcomes (Lee et al., 2022a), estimating heterogeneous treatment effects (Wu & Yang, 2022a;

Yang et al., 2022) and combining probability and nonprobability samples (Yang et al., 2020; Gao & Yang, 2023).

Acknowledgement

This project was supported by the Food and Drug Administration (FDA) of the U.S. Department of Health and Human Services (HHS) (U01FD007934) and the National Institute on Aging of the National Institutes of Health (R01AG06688). The views and opinions expressed herein are those of the authors and do not necessarily represent those of, nor endorsement by, FDA/HHS, the National Institutes of Health or the U.S. Government.

Supplementary material

The Supplementary Material includes all technical proofs, additional simulation results and other real-data applications. An open-source software `R` package (R Development Core Team, 2025) is available for implementing our proposed methodology at `https://github.com/IntegrativeStats/SelectiveIntegrative`.

References

Bickel, P. J., Klaassen, C., Ritov, Y. & Wellner, J. (1998). *Efficient and Adaptive Inference in Semiparametric Models*, vol. 50. Baltimore, MD: Johns Hopkins University Press.

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**, 723–34.

Chatterjee, N., Chen, Y.-H., Maas, P. & Carroll, R. J. (2016). Constrained maximum likelihood estimation for model calibration using summary-level information from external big data sources. *J. Am. Statist. Assoc.* **111**, 107–17.

Chen, X. (2007). Large sample sieve estimation of semi-nonparametric models. In *Handbook of Econometrics*, vol. 6, Ed. J.J. Heckman and E. E. Leamer, pp. 5549–5632. Amsterdam: Elsevier.

Chen, Z., Ning, J., Shen, Y. & Qin, J. (2021). Combining primary cohort data with external aggregate information without assuming comparability. *Biometrics* **77**, 1024–36.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W. & Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. *Econom. J.* **21**, 1–68.

Chu, J., Lu, W. & Yang, S. (2023). Targeted optimal treatment regime learning using summary statistics. *Biometrika* **110**, 913–31.

Colnet, B., Mayer, I., Chen, G., Dieng, A., Li, R., Varoquaux, G., Vert, J.-P., Josse, J. & Yang, S. (2020). Causal inference methods for combining randomized trials and observational studies: a review. *Statist. Sci.* **39**, 165–91.

Fan, J. & Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Am. Statist. Assoc.* **96**, 1348–60.

FDA. (2001). *E10 Choice of Control Group and Related Issues in Clinical Trials*. `https://www.fda.gov/regulatory-information/search-fda-guidance-documents/e10-choice-control-group-and-related-issues-clinical-trials`

FDA. (2014). *Blinatumomab Drug Approval Package*. `https://www.accessdata.fda.gov/drugsatfda_docs/nda/2014/125557Orig1s000TOC.cfm`

FDA. (2016). *Avelumab Drug Approval Package*. `https://www.fda.gov/drugs/resources-information-approved-drugs/avelumab-bavencio`

FDA. (2019). *Rare Diseases: Natural History Studies for Drug Development*. `https://www.fda.gov/regulatory-information/search-fda-guidance-documents/rare-diseases-natural-history-studies-drug-development`

FDA. (2021). *Real-World Data: Assessing Registries to Support Regulatory Decision-Making for Drug and Biological Products Guidance for Industry*. `https://www.fda.gov/regulatory-information/search-fda-guidance-documents/real-world-data-assessing-registries-support-regulatory-decision-making-drug-and-biological-products`

FDA. (2023). *Considerations for the Design and Conduct of Externally Controlled Trials for Drug and Biological Products Guidance for Industry*. `https://www.fda.gov/regulatory-information/search-`

fda-guidance-documents/considerations-design-and-conduct-externally-controlled-trials-drug-and-biological-products

Gao, C. & Yang, S. (2023). Pretest estimation in combining probability and non-probability samples. *Electron. J. Statist.* **17**, 1492–546.

Ghadessi, M., Tang, R., Zhou, J., Liu, R., Wang, C., Toyoizumi, K., Mei, C., Zhang, L., Deng, C. & Beckman, R. A. (2020). A roadmap to using historical controls in clinical trials – by Drug Information Association Adaptive Design Scientific Working Group (DIA-ADSWG). *Orphanet J. Rare Dis.* **15**, 1–19.

Ho, D. E., Imai, K., King, G. & Stuart, E. A. (2007). Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Polit. Anal.* **15**, 199–236.

Hobbs, B. P., Carlin, B. P., Mandrekar, S. J. & Sargent, D. J. (2011). Hierarchical commensurate and power prior models for adaptive incorporation of historical information in clinical trials. *Biometrics* **67**, 1047–56.

Huang, J., Ma, S. & Zhang, C.-H. (2008). Adaptive lasso for sparse high-dimensional regression models. *Statist. Sinica* **18**, 1603–18.

Ibrahim, J. G. & Chen, M.-H. (2000). Power prior distributions for regression models. *Statist. Sci.* **15**, 46–60.

Imbens, G. W. (2004). Nonparametric estimation of average treatment effects under exogeneity: a review. *Rev. Econ. Statist.* **86**, 4–29.

Kwiatkowski, E., Zhu, J., Li, X., Pang, H., Lieberman, G. & Psioda, M. A. (2023). Case weighted adaptive power priors for hybrid control analyses with time-to-event data. *arXiv:* 2305.05913v1.

Lee, D., Yang, S., Dong, L., Wang, X., Zeng, D. & Cai, J. (2022b). Improving trial generalizability using observational studies. *Biometrics* **79**, 1213–25.

Lee, D., Yang, S. & Wang, X. (2022a). Doubly robust estimators for generalizing treatment effects on survival outcomes from randomized controlled trials to a target population. *J. Causal Infer.* **10**, 415–40.

Lee, J. D., Sun, D. L., Sun, Y. & Taylor, J. E. (2016). Exact post-selection inference, with application to the lasso. *Ann. Statist.* **44**, 907–27.

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.

Lin, Z., Xiang, Y. & Zhang, C. (2009). Adaptive lasso in high-dimensional settings. *J. Nonparam. Statist.* **21**, 683–96.

Liu, M., Bunn, V., Hupf, B., Lin, J. & Lin, J. (2021). Propensity-score-based meta-analytic predictive prior for incorporating real-world and historical data. *Statist. Med.* **40**, 4794–808.

Neuenschwander, B., Branson, M. & Spiegelhalter, D. J. (2009). A note on the power prior. *Statist. Med.* **28**, 3562–6.

Neuenschwander, B., Capkun-Niggli, G., Branson, M. & Spiegelhalter, D. J. (2010). Summarizing historical information on controls in clinical trials. *Clin. Trials* **7**, 5–18.

Odogwu, L., Mathieu, L., Blumenthal, G., Larkins, E., Goldberg, K. B., Griffin, N., Bijwaard, K., Lee, E. Y., Philip, R., Jiang, X. et al. (2018). FDA approval summary: dabrafenib and trametinib for the treatment of metastatic non-small cell lung cancers harboring BRAF V600E mutations. *The Oncologist* **23**, 740–5.

Phelan, M., Bhavsar, N. A. & Goldstein, B. A. (2017). Illustrating informed presence bias in electronic health records data: how patient interactions with a health system can impact inference. *J Electron. Health Data Meth.* **5**, 22–36.

Pocock, S. J. (1976). The combination of randomized and historical controls in clinical trials. *J. Chronic Dis.* **29**, 175–88.

Qin, J., Zhang, H., Li, P., Albanes, D. & Yu, K. (2015). Using covariate-specific disease prevalence information to increase the power of case-control studies. *Biometrika* **102**, 169–80.

R Development Core Team (2025). *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0. http://www.R-project.org

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

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

Schoenfeld, D. A., Finkelstein, D. M., Macklin, E., Zach, N., Ennist, D. L., Taylor, A. A., Atassi, N. & Pooled Resource Open-Access ALS Clinical Trials Consortium. (2019). Design and analysis of a clinical trial using previous trials as historical control. *Clin. Trials* **16**, 531–8.

Shan, M., Faries, D., Dang, A., Zhang, X., Cui, Z. & Sheffield, K. M. (2022). A simulation-based evaluation of statistical methods for hybrid real-world control arms in clinical trials. *Statist. Biosci.* **14**, 259–84.

Shen, Y., Gao, C., Witten, D. & Han, F. (2020). Optimal estimation of variance in nonparametric regression with random design. *Ann. Statist.* **48**, 3589–618.

Silverman, B. (2018). A baker's dozen of US FDA efficacy approvals using real world evidence. *Pharma Intelligence Pink Sheet*, 7 August.

Stuart, E. A. (2010). Matching methods for causal inference: a review and a look forward. *Statist. Sci.* **25**, 1–21.

Stuart, E. A. & Rubin, D. B. (2008). Matching with multiple control groups with adjustment for group differences. *J. Educ. Behav. Statist.* **33**, 279–306.

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. *J. R. Statist. Soc.* A **174**, 369–86.
Tibshirani, R. J., Taylor, J., Lockhart, R. & Tibshirani, R. (2016). Exact post-selection inference for sequential regression procedures. *J. Am. Statist. Assoc.* **111**, 600–20.
Tipton, E. (2014). How generalizable is your experiment? An index for comparing experimental samples and populations. *J. Educ. Behav. Statist.* **39**, 478–501.
Tsiatis, A. (2006). *Semiparametric Theory and Missing Data*. New York: Springer.
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? *Statist. Meth. Med. Res.* **27**, 3167–82.
Viele, K., Berry, S., Neuenschwander, B., Amzal, B., Chen, F., Enas, N., Hobbs, B., Ibrahim, J. G., Kinnersley, N., Lindborg, S. et al. (2014). Use of historical control data for assessing treatment effects in clinical trials. *Pharm. Statist.* **13**, 41–54.
Visvanathan, K., Levit, L. A., Raghavan, D., Hudis, C. A., Wong, S., Dueck, A. & Lyman, G. H. (2017). Untapped potential of observational research to inform clinical decision making: American Society of Clinical Oncology research statement. *J. Clin. Oncol.* **35**, 1845–54.
Wu, L. & Yang, S. (2022a). Integrative *R*-learner of heterogeneous treatment effects combining experimental and observational studies. In *Proc. 1st Conf. Causal Learn. Reason.*, pp. 904–26. PMLR.
Wu, L. & Yang, S. (2022b). Transfer learning of individualized treatment rules from experimental to real-world data. *J. Comp. Graph. Statist.* **32**, 1036–45.
Yang, S., Kim, J. K. & Song, R. (2020). Doubly robust inference when combining probability and non-probability samples with high dimensional data. *J. R. Statist. Soc.* B **82**, 445–65.
Yang, S., Zeng, D. & Wang, X. (2022). Elastic integrative analysis of randomized trial and real-world data for treatment heterogeneity estimation. *arXiv:* 2005.10579v3.
Zhai, Y. & Han, P. (2022). Data integration with oracle use of external information from heterogeneous populations. *J. Comp. Graph. Statist.* **31**, 1001–12.
Zhang, H., Deng, L., Schiffman, M., Qin, J. & Yu, K. (2020). Generalized integration model for improved statistical inference by leveraging external summary data. *Biometrika* **107**, 689–703.
Zhang, T. & Yu, B. (2005). Boosting with early stopping: convergence and consistency. *Ann. Statist.* **33**, 1538–79.
Zou, H. (2006). The adaptive LASSO and its oracle properties. *J. Am. Statist. Assoc.* **101**, 1418–29.

[*Received on* 29 *December* 2023. *Editorial decision on* 21 *October* 2024]