Journal of the Royal Statistical Society Series A: Statistics in Society, 2024, **00**, 1–28
https://doi.org/10.1093/jrsssa/qnae075

ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS
Journal of the Statistics Society
Series A
Statistics in Society

Original Article

Causal estimators for incorporating external controls in randomized trials with longitudinal outcomes

Xiner Zhou[1,2], Jiawen Zhu[2], Christiana Drake[1] and Herbert Pang[2,3]

[1]Department of Statistics, University of California - Davis, Davis, CA 95616, USA
[2]PD Data Sciences, Genentech, South San Francisco, CA 94080, USA
[3]Department of Biostatistics and Bioinformatics, Duke University School of Medicine, Durham, NC 27710, USA

Address for correspondence: Jiawen Zhu, PD Data Sciences, Genentech, South San Francisco, CA 94080, USA.
Email: jiawenxm@gmail.com

Abstract

Incorporating external data, such as external controls, holds the promise of improving the efficiency of traditional randomized controlled trials especially when treating rare diseases or diseases with unmet needs. To this end, we propose novel weighting estimators grounded in the causal inference framework. As an alternative framework, Bayesian methods are also discussed. From trial design perspective, operating characteristics including Type I error and power are particularly important and are assessed in our realistic simulation studies representing a variety of practical scenarios. Our proposed weighting estimators achieve significant power gain, while maintaining Type I error close to the nominal value of 0.05. An empirical application of the methods is demonstrated through a Phase III clinical trial in rare disease.

Keywords: causal inference, efficiency, external controls, propensity score weighting

1 Introduction

Randomized controlled trials (RCTs) are widely recognized as the gold standard for determining the effects of therapeutic products on specific outcomes. Yet, these trials are not without their limitations. Attaining adequate statistical power for hypothesis testing of the treatment effect at the primary endpoint may require sufficient number of patients, who are then assigned to either the treatment or control groups (Viele et al., 2014). Firstly, when the number of patients with any specific disease is small, recruiting enough participants can be impractical or even unfeasible. Secondly, the prospect of being assigned to the control group in high likelihood can be unappealing, or even ethically questionable, particularly in diseases with high unmet medical needs and no or limited effective treatments. These challenges often occur in the rare disease setting and result in smaller trial sizes, particularly affecting the size of the control arm, or in extreme cases, leading to the absence of a control arm in single-arm trials (Gross, 2021).

The concept of 'external controls' (Chen et al., 2021; Pocock, 1976; Yap et al., 2021), which involves using a comparison group of people external to the trial of interest who had not received the experimental treatment, is becoming more common (FDA, 2023) and industry (Burger et al., 2021). The use of external controls aims to harness the RCTs with above mentioned limitations by either fully or partially substituting the trial control arms. These designs are considered innovative (CID, 2023). Although various sources of data can serve as external controls, they are primarily

Received: June 1, 2023. **Revised:** May 26, 2024. **Accepted:** July 23, 2024

derived from patient-level data from other clinical trials or from real-world data (RWD) sources, like the example in the FDA CID program. The use of a control group from other trials is feasible if the eligible population and endpoints align closely with the current study.

This paper focuses on methodologies to analyse hybrid controlled trials, where the control group is composed of both trial and external controls. Incorporating external controls into RCTs presents uniform statistical challenges, regardless of their sources. Pooling external controls to estimate treatment effects on the trial population is subject to bias, as the randomized trial subjects and the nonrandomized external control subjects are not exchangeable.

Currently, there is a growing interest in addressing this problem through the lens of causal inference, as advocated by recent research [Ho et al., 2023]. Though much research has been done in combining experimental and observational data, it has primarily focussed on different goals: *generalizability* (Buchanan et al., 2018; Dahabreh et al., 2019), *representativeness* (Campbell, 1957), *external validity* (Stuart et al., 2018), *transportability* (Pearl & Bareinboim, 2011; Westreich et al., 2017), and *data fusion* (Bareinboim & Pearl, 2016). This line of research focuses on generalizing trial results to a target population for which the trial participants are not representative of. In contrast, this paper focuses on utilizing an external control population, to increase the statistical efficiency and power of hypothesis testing for the treatment effect in the trial population. For this goal, Bayesian dynamic borrowing methods (Fu et al., 2023) has been a popular framework. Bayesian methods provide a natural mechanism for information borrowing through the use of informative priors: the power prior (Ibrahim & Chen, 2000; Ibrahim et al., 2015), the commensurate prior and the commensurate power prior (Hobbs et al., 2011), and the robust metaanalytic-predictive prior (Schmidli et al., 2014).

The objective of this paper is to address the gap in current research on causal inference methods for hybrid controlled trials, aiming to enhance trial efficiency. The above mentioned Bayesian methods do not focus on using causal inference framework. By integrating trial data with external control data, the goal is to increase statistical power while maintaining the desired Type I error rate. When designing a future trial with external controls, fewer internal control subjects may be needed, which shortens the recruitment and possibly the duration the trial, and leads to more efficient medical product development and approval. In particular, we propose causal estimators based on weighting in two layers: the first layer tackles the distributional shift in confounders, while the second layer synthesizes evidence from both trial and external controls. Our focus is different from the literature in generalizability of trial results to another population. Instead, we focus on the reverse direction of using external data to enhance the efficiency of clinical trials. Our research work also establishes realistic data generating mechanisms likely encountered in actual clinical trials to support new study designs. Those simulations allow trial designers to evaluate the trial operating characteristics by assumption violation. Finally, all methodological and practical discussions are motivated and illustrated through a recent Phase III trial in a rare disease, illustrating the feasibility of the approach in practical applications.

The work is motivated by a recent study of the medicine risdiplam to treat spinal muscular atrophy (SMA). Spinal muscular atrophy is a rare neuromuscular disorder that results in the loss of motor neurones and progressive muscle wasting. The SUNFISH Trial (NCT02908685) is a Phase III, randomized, double-blind, placebo-controlled study of the efficacy and safety of the risdiplam treatment among patients aged 2–25 years with confirmed 5q autosomal recessive Type II or Type III SMA recruited from 42 hospitals in 14 countries across Europe, North America, South America, and Asia. Risdiplam is an oral small molecule that modifies pre-mRNA splicing of the SMN2 gene to increase production of functional SMN. Patients were stratified by age and randomly assigned (2:1) to receive either daily oral risdiplam or daily oral placebo. Patients were scheduled for follow-up visits roughly every 6 months to have their motor function measure (MFM) measured along with other clinically relevant indicators. The primary endpoint was the change in the MFM from baseline to the end of month 12. The results in the primary endpoint have shown a significant improvement in motor function compared with placebo. More information about the trial can be found in Mercuri et al. (2022).

The olesoxime trial (NCT01302600) (Berry et al., 2010) is a randomized, double-blind, placebo-controlled, Phase II study for the same disease population as SUNFISH. The olesoxime trial shares the same set of measurements and follow-up visits as the SUNFISH study, but with

a control arm that spans over 2 years. The control subjects from this study can serve as external controls to augment the SUNFISH study (McIver et al., 2023).

The remainder of this paper is organized as follows. In Section 2, we introduce the notations, estimand, and causal identifiability assumptions that explicit 'qualities' of the external controls. In Section 3, we propose an identifying functional and its doubly robust counterpart, with a weighting parameter that permits investigator control of the influence of the external controls. The two functionals are used to develop a suite of weighting estimators. Consistency and asymptotic normality are provided when parametric models are used to estimate nuisance parameters, given certain qualifying assumptions. In Section 4, we briefly discuss the power prior and commensurate prior methods. Simulation studies in Section 5 illustrate the finite-sample performance of the proposed methods as well as the Bayesian methods, with realistic simulations mimicking the SUNFISH trial and its external controls. In Section 6, application of the methods is demonstrated through the SUNFISH trial with external controls. Finally, Section 7 summarizes and discusses key takeaways.

2 Notations and assumptions for causal identifiability

2.1 Notations

The notations and assumptions used in this work are grounded in the potential outcome framework (Imbens & Rubin, 2015). We assume to have two data sets at hand:

- A RCT, denoted by $\mathcal{R}$ (for randomized), assessing the efficacy of a binary treatment A on longitudinal outcomes $\mathbf{Y} = (Y_1, \ldots, Y_T)$ repeatedly measured on T time points (if only a single primary endpoint is of interest, $T = 1$), conducted on n subjects. Each subject i, labelled from 1 to n, is sampled from a distribution $P_{\mathcal{R}}(\mathbf{X}, A, \mathbf{Y}^{(1)}, \mathbf{Y}^{(0)})$ the current RCT trial population, and also the target population of interest. For any subject i, $\mathbf{X}_i$ is a p-dimensional vector of measured baseline covariates, and potential confounders, accounting for individual characteristics. A_i denotes the binary treatment assignment (with $A_i = 1$ if treated and $A_i = 0$ if untreated), and $\mathbf{Y}^{(a)} = (Y_1^{(a)}, \ldots, Y_T^{(a)})$ are the potential outcomes had the subject i been given treatment a (for $a \in \{0, 1\}$). $\mathbf{Y}_i$ denotes the observed outcomes, by the Assumption 2, $\mathbf{Y}_i = A_i\mathbf{Y}_i^{(1)} + (1 - A_i)\mathbf{Y}_i^{(0)}$. Our interest is in those trials with small proportion of control subjects, hence the probability of treatment assignment $\pi_A = P_{\mathcal{R}}(A = 1)$, usually $>1/2$. Let n_1 be the number of treated subjects and n_0 be the number of control subjects.
- An external control sample, denoted by $\mathcal{E}$ (for external controls), containing m subjects sampled from a distribution $P_{\mathcal{E}}(\mathbf{X}, A = 0, \mathbf{Y}^{(0)})$. The available external control sample size m could potentially be large. We assume observe the same set of baseline covariates $\mathbf{X}_i$.

We assume that each trial subject $i \in \mathcal{R}$ is sampled from the current RCT trial population described by $P_{\mathcal{R}}(\mathbf{X}, A, \mathbf{Y}^{(1)}, \mathbf{Y}^{(0)})$ (also the target population of interest), while each external control subject $i \in \mathcal{E}$ is sampled from $P_{\mathcal{E}}(\mathbf{X}, A = 0, \mathbf{Y}^{(0)})$, labelled from $i = n + 1$ to $n + m$. $P_{\mathcal{E}}$ could represent the distribution of a larger disease population in the real world, or a population targeted by another trial. We use S_i to denote trial participation status, with $S_i = 1$ for $i \in \mathcal{R}$ and $S_i = 0$ for $i \in \mathcal{E}$.

Throughout the paper, we will use indices $\mathcal{R}$ and $\mathcal{E}$ to denote quantities (probability, expectation, variance, covariance) taken with respect to these populations, for example, $\mathbb{E}_{\mathcal{R}}()$ for an expectation over $P_{\mathcal{R}}$.

2.2 Causal estimand

We define the trial population (the target population) average treatment effect (ATE) as the causal estimand

$$\boldsymbol{\tau} \triangleq \mathbb{E}_{\mathcal{R}}\left[\mathbf{Y}^{(1)} - \mathbf{Y}^{(0)}\right] \tag{1}$$

where $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_T)$ is a vector of time-indexed ATE. In contrast, the estimand in the related generalizability literature would be the same expectation, except taken with respect to $P_{\mathcal{E}}$.

2.3 Identification assumptions

The identification assumptions needed are similar to those in the *generalizability* literature (Dahabreh et al., 2019; Stuart et al., 2018).

Assumption 1 (No direct effect of trial participation). The only way through which trial participation affects the outcome is through the treatment itself and there is no direct effect of trial participation on the outcome. It implies that there is no dependency of potential outcomes under no treatment on trial participation: $Y^{(s,a=0)} = Y^{(a=0)}$ for $s \in \{0, 1\}$.

Assumption 2 [Stable unit treatment value {SUTVA}]. If $i \in \mathcal{R}$, then $Y_i = Y_i^{(A_i)}$, and if $i \in \mathcal{E}$, then $Y_i = Y_i^{(0)}$. That is, the observed outcome for ith subject in the RCT equals to that individual's potential outcome under the treatment actually received, and the observed outcome for an external control subject equals to that individual's potential outcome under no treatment. Implicit in this notation is that there is a single version of 'no treatment' that is consistently defined across all subjects in the RCT and external controls.

Assumptions 1 and 2 assume that the outcomes for patients receiving no active treatment are assumed to be stable and not influenced by the specific conditions of the study they participate in. When external controls come from RWD, such a direct effect may arise if there are any substantial placebo effects. This assumption may be more reasonable when using external controls from a separate trial, as in our real data example. Then, it would be sufficient that the placebo effects are identical in the control groups for each trial.

Assumption 3 (Internal validity of the trial). (1) Treatment randomization holds for all subjects in the RCT: $(\mathbf{Y}_i^{(1)}, \mathbf{Y}_i^{(0)}) \perp A_i$ for $A_i \in \{0, 1\}, i \in \mathcal{R}$; and (2) Positivity of treatment assignment: $0 < \pi_A = P(A_i = 1 \mid S_i = 1) < 1$.

Assumption 4 (Conditional ignorability of trial participation). Given the measured covariates $\mathbf{X}$, the potential outcome under no treatment is independent of trial participation, i.e. trial participation is conditionally ignorable or exchangeable, $\mathbf{Y}_i^{(0)} \perp S_i \mid \mathbf{X}_i$ for $i \in \mathcal{R} \cup \mathcal{E}$

Assumption 4 is analogous to the no unmeasured confounding or ignorability assumption commonly used in causal inference literature, especially in the *generalizability* literature (Dahabreh et al., 2019; Stuart et al., 2018). In practical terms, this means both the trial and the external study must capture all risk factors of the outcomes that also influence study participation. This might include demographic, socioeconomic, and disease features.

Unlike the usual ignorability assumption based on counterfactual variables, Assumption 4 implies conditional independence among observed variables $\mathbf{Y}_i \perp S_i \mid \mathbf{X}_i, A_i = 0$, i.e. the trial participation is conditionally independent with outcomes given covariates among the trial controls and external controls. This assumption is falsifiable by testing for distributional differences in outcome given covariates between the trial controls and external controls. However, it is important to note that a violation of either Assumptions 1 or 4 (or both) could result in a discernible difference in outcomes that cannot be readily distinguished as being due to one specific assumption being violated.

Assumption 5 (Overlapping or positivity of trial participation). The support of the measured covariates in the RCT population is contained within that of the external control population: $\text{supp}(P_{\mathcal{R}}(\mathbf{X})) \subset \text{supp}(P_{\mathcal{E}}(\mathbf{X}))$, or equivalently, $Pr(S = 1 \mid \mathbf{X} = \mathbf{x}) < 1$ for all $\mathbf{x}$.

In contrast to positivity assumption in the generalizability literature (Buchanan et al., 2018; Dahabreh et al., 2019), which generally requires $Pr(S = 1 \mid \mathbf{X} = \mathbf{x}) > 0$ for all $\mathbf{x}$, Assumption 5 requires that, in all areas of the covariate distribution, it is possible to have external controls.

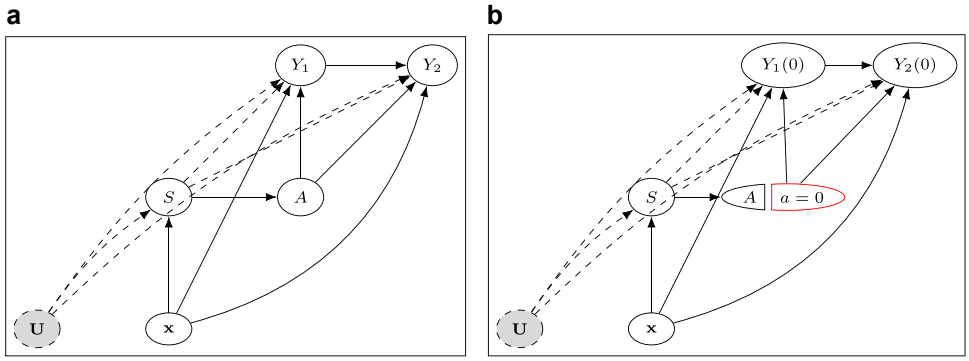

Figure 1. Graphical model representing the pooled dataset. Here, we illustrate two repeated outcomes, in general, there could be arbitrary number of repeated outcomes, where the dashed node and edges are assumed to be absent. a) Directed Acyclic Graph (DAG); b) Equivalent Single World Intervention Graph (SWIG).

Assumption 5 is empirically falsifiable by comparing the empirical distributions of the covariates in the trial sample and the external controls.

The Directed Acyclic Graph (DAG) in Figure 1a illustrates the possible mechanism of the setting, and the equivalent Single World Intervention Graph (SWIG) (Richardson & Robins, 2013) in Figure 1b shows the relationship between covariates, interventions and potential outcomes under no treatment, where the dashed node and edges are assumed to be absent. The existence of U would violate the Conditional Ignorability Assumption 4. Two mechanisms are assumed to be absent: (1) the arrow from trial participation S to outcome $\mathbf{Y}$, and (2) unmeasured confounding U. (1) assumes that there is no direct effect of trial participation, it frees the dependency of potential outcomes on trial participation, which we formalize in Assumption 1. (2) is analogous to the no unmeasured confounding or ignorability assumption common in causal inference literature and we formalize it in Assumption 4.

3 Causal inference methods

Without external controls, ATE can be estimated from the trial data, via (covariate-adjusted) difference in means, by internal validity.

A brute-force incorporation of the external controls, treating them equally as trial controls, could run the risk of introducing bias (see Section 3.2 for a discussion on bias), outweighing the benefit of gaining efficiency by accessing to additional data. Granting that Assumptions 1, 2, 3, 4, 5 hold, we weight the external controls so that the weighted distribution of the confounder $\mathbf{X}$ in $\mathcal{E}$ is equal to that of the target trial population $\mathcal{R}$. As a result, the weighted observed outcomes of the external controls can be thought of as a representative sample from the distribution of potential outcome under no treatment in the trial population, similar to the trial controls.

We are in possession of two representative samples to inform the potential outcome under no treatment in the trial population: (1) the trial control sample, and (2) the weighted external control sample, which is the augmented 'control' subjects to the trial without actually recruiting additional trial controls, with the goal to increase the precision and efficiency in the estimation of and inference about the ATE in the trial population. The trial controls, from source (1), could give an unbiased estimate of $\mathbb{E}_{\mathcal{R}}[\mathbf{Y}^{(0)}]$, (assuming the internal validity), with limited samples, on the contrary, and the external controls, from source (2), give another estimate of $\mathbb{E}_{\mathcal{R}}[\mathbf{Y}^{(0)}]$ that could suffer from bias, but requires fewer resources to obtain. The intuitive idea to combine multiple sources of information to estimate the same quantity is a convex combination, that is, to assign a weight w $(0 \leq w \leq 1)$ to the external controls, and synthesizes multiple estimates (here only two), by a convex combination as the hybrid control estimate of $\boldsymbol{\mu}_{\mathcal{R}}^{(0)}$.

3.1 Weighting estimators for external controls to enhance trial findings

In this work, we focus on the weighting strategy that weights samples in two layers, with the first layer of weights tackling the distribution shift between the external control population and the

trial population and can be thought of as a special case of the *balancing weights* in Li et al. (2018) and similar to the *inverse probability of sampling weighting* estimators (Cole & Stuart, 2010; Stuart et al., 2011), and with the second layer of weights synthesizing the external controls estimate with that of the trial controls estimate to form a single hybrid estimate. We name the novel estimators *External Controls Enhanced Inverse Probability Weighting (EC-IPW)*, and its doubly robust (DR) version *External Controls Enhanced Augmented Inverse Probability Weighting (EC-AIPW)*, derived from identification formulae that share a general structure

$$\tau = \mu_{11} - \underbrace{[(1-w)\mu_{10} + w\mu_{00}]}_{\text{synthesis of external control and trial control}},$$

for any $w \in [0, 1]$, where μ_{sa} denote expectation of appropriately weighted outcome of the subjects in the sample $s \in \{0, 1\}$ and received treatment $a \in \{0, 1\}$ to be specified in Theorems 1 and 2.

Theorem 1 (Identification via EC-IPW). The estimand in equation (1) can be identified using the observed data, combining trial and external controls, for any $w \in [0, 1]$,

$$\boldsymbol{\tau}^{EC-IPW} \triangleq \underbrace{\frac{\mathbb{E}_{\mathcal{R}}[A\mathbf{Y}W_{11}]}{\mathbb{E}_{\mathcal{R}}[AW_{11}]}}_{\mu_{11}} - \left[(1-w)\underbrace{\frac{\mathbb{E}_{\mathcal{R}}[(1-A)\mathbf{Y}W_{10}]}{\mathbb{E}_{\mathcal{R}}[(1-A)W_{10}]}}_{\mu_{10}} + w\underbrace{\frac{\mathbb{E}_{\mathcal{E}}[\mathbf{Y}W_{00}]}{\mathbb{E}_{\mathcal{E}}[W_{00}]}}_{\mu_{00}}\right] \tag{2}$$

where the trial treated, trial control, and external control subjects receive weights $W_{11} = \frac{1}{\pi_A}$, $W_{10} = \frac{1}{1-\pi_A}$, and $W_{00} = \frac{P_{\mathcal{R}}(\mathbf{X})}{P_{\mathcal{E}}(\mathbf{X})} = \frac{\pi_S(\mathbf{X})(1-\pi_S)}{(1-\pi_S(\mathbf{X}))\pi_S}$. $\pi_S(\mathbf{X}) = Pr(S = 1 \mid \mathbf{X})$ denote the conditional probability of trial participation given covariates, $\pi_S = Pr(S = 1)$ denote the marginal probability of trial participation. The weights associated with the trial subjects, W_{11} and W_{00}, based on covariates can increase the efficiency of the estimator (Tsiatis, 2006) and are optional in a completely randomized trial. W_{00} is the balancing weights as in Li et al. (2018) which weights the external control samples to represent the target trial population.

In practice, $\pi_S(\mathbf{X})$ is unknown, which suggests fitting a model for the probability of trial participation $\pi_S(\mathbf{X})$ based on a logistic regression model $logit\{\pi_S(\mathbf{X}; \alpha)\} = \boldsymbol{\alpha}'\mathbf{X}$ with an unknown p-dimensional parameter $\boldsymbol{\alpha}$.

If the trial participation model $\pi_S(\mathbf{X}; \alpha)$ for $\pi_S(\mathbf{X})$ is correct,

$$\boldsymbol{\tau}^{EC-IPW} = \tau, \tag{3}$$

see proof in Appendix A.

When $w = 0$, $\boldsymbol{\tau}^{EC-IPW}$ reduces to the *normalized inverse probability of treatment weighting (IPTW)* estimator using only the trial data, i.e. the *Hajek* estimator, that is considered more stable than the *unnormalized IPTW*, i.e. the *Horvitz–Thompson* estimator (Horvitz & Thompson, 1952).

Theorem 2 (Identification via EC-AIPW). To improve upon Theorem 1, τ^{EC-IPW} in equation (2) be robustified similarly as the *augmented IPW* for estimating ATE

using observational data, by replacing $\mathbf{Y}$ with $\widetilde{\mathbf{Y}}$, the residual after projection $\widetilde{\mathbf{Y}} = \mathbf{Y} - \mu(\mathbf{X})$, where $\mu(\mathbf{X}) = \mathbb{E}[\mathbf{Y} \mid \mathbf{X}, A = 0]$.

$$\boldsymbol{\tau}^{EC-AIPW} \triangleq \underbrace{\frac{\mathbb{E}_{\mathcal{R}}\left[A\widetilde{\mathbf{Y}}W_{11}\right]}{\mathbb{E}_{\mathcal{R}}[AW_{11}]}}_{\mu_{11}} - \left[(1-w)\underbrace{\frac{\mathbb{E}_{\mathcal{R}}\left[(1-A)\widetilde{\mathbf{Y}}W_{10}\right]}{\mathbb{E}_{\mathcal{R}}[(1-A)W_{10}]}}_{\mu_{10}} + w\underbrace{\frac{\mathbb{E}_{\mathcal{E}}\left[\widetilde{\mathbf{Y}}W_{00}\right]}{\mathbb{E}_{\mathcal{E}}[W_{00}]}}_{\mu_{00}}\right], \tag{4}$$

In practice, both $\pi_S(\mathbf{X})$ and $\mu(\mathbf{X})$ are unknown, which suggests (i) fitting a model for the probability of trial participation $Pr(S = 1 \mid \mathbf{X})$ based on a logistic regression model $logit\{\pi_S(\mathbf{X}; \alpha)\} = \boldsymbol{\alpha}'\mathbf{X}$ with an unknown parameter $\boldsymbol{\alpha}$, and (ii) fitting a parametric model $\mu(\mathbf{X}; \boldsymbol{\beta}) = g(\boldsymbol{\beta}'\mathbf{X})$ for $\mu(\mathbf{X})$ with g^{-1} a known link function and $\boldsymbol{\beta}$ is an unkown parameter. For continuous outcome, the canonical link function g^{-1} is identity so that $\mu(\mathbf{X}; \boldsymbol{\beta}) = \boldsymbol{\beta}'\mathbf{X}$.

If either the trial participation model $\pi_S(\mathbf{X}; \alpha)$ for $\pi_S(\mathbf{X})$ or the outcome model $\mu(\mathbf{X}; \boldsymbol{\beta})$ for $\mu(\mathbf{X})$ is correct,

$$\boldsymbol{\tau}^{EC-AIPW} = \tau, \tag{5}$$

see proof in Appendix A.

The identification formulae in Theorems 1 and 2, result in two novel estimators presented in Definitions 1 and 2 that share a common structure

$$\hat{\tau} = \widehat{\mu}_{11} - \underbrace{[(1-w)\widehat{\mu}_{10} + w\widehat{\mu}_{00}]}_{\text{synthesis of external control and trial control}},$$

for any $w \in [0, 1]$, where $\widehat{\mu}_{sa}$ denote appropriately weighted average outcome of the subjects in the sample $s \in \{0, 1\}$ and received treatment $a \in \{0, 1\}$ to be specified in Definitions 1 and 2.

Definition 1 (EC-IPW Estimator). The EC-IPW estimator, when nuisance parameters are unknown and estimated, is given by

$$\widehat{\boldsymbol{\tau}}^{EC-IPW}_{n,m} \triangleq \underbrace{\frac{\sum_{i\in\mathcal{R}} A_i\mathbf{Y}_i\widehat{W}_{11i}}{\sum_{i\in\mathcal{R}} A_i\widehat{W}_{11i}}}_{\widehat{\mu}_{11}} - \left[(1-w)\underbrace{\frac{\sum_{i\in\mathcal{R}}(1-A_i)\mathbf{Y}_i\widehat{W}_{10i}}{\sum_{i\in\mathcal{R}}(1-A_i)\widehat{W}_{10i}}}_{\widehat{\mu}_{10}} + w\underbrace{\frac{\sum_{i\in\mathcal{E}}\mathbf{Y}_i\widehat{W}_{00i}}{\sum_{i\in\mathcal{E}}\widehat{W}_{00i}}}_{\widehat{\mu}_{00}}\right] \tag{6}$$

where a hat indicates an estimated quantity.

Definition 2 (EC-AIPW Estimator). The EC-AIPW estimator, when nuisance parameters are unknown and estimated, is given by

$$\widehat{\tau}_{n,m}^{EC-AIPW} \triangleq \underbrace{\frac{\sum_{i\in\mathcal{R}} A_i \widehat{\widetilde{\mathbf{Y}}}_i \widehat{W}_{11i}}{\sum_{i\in\mathcal{R}} A_i \widehat{W}_{11i}}}_{\widehat{\mu}_{11}} - \left[(1-w) \underbrace{\frac{\sum_{i\in\mathcal{R}} (1-A_i) \widehat{\widetilde{\mathbf{Y}}}_i \widehat{W}_{10i}}{\sum_{i\in\mathcal{R}} (1-A_i) \widehat{W}_{10i}}}_{\widehat{\mu}_{10}} + w \underbrace{\frac{\sum_{i\in\mathcal{E}} \widehat{\widetilde{\mathbf{Y}}}_i \widehat{W}_{00i}}{\sum_{i\in\mathcal{E}} \widehat{W}_{00i}}}_{\widehat{\mu}_{00}} \right] \tag{7}$$

where a hat indicates an estimated quantity.

3.2 Bias-variance trade-off

The EC-IPW and EC-AIPW estimators in Definitions 1 and 2 indexed by a weight $w \in [0, 1]$ are motivated by the need to combine an unbiased but small trial data, with a potentially biased but large external data, with the goal to improve efficiency in estimating ATE in the RCT. The synthesizing weight gives rise to a class of EC-IPSW estimators (similarly EC-AIPW), with infinitely many choices of weights. The weight represents the degree of trust we put on the external controls, where $w < 1/2$ amounts to discounting the influence of external controls on the final estimate. It is similar to the power parameter in the Bayesian power prior approach (essentially uses weighted likelihood), which assign weights through a power parameter to discount the external controls. Ideally, the choice of the synthesize weight should strike a balance between bias and variance trade-off of the final estimator: larger weight for external controls could potentially introduce more bias through $\widehat{\mu}_{00}$ while reduce the variance of the hybrid estimate; smaller weight minimizes such bias but lost the gain of efficiency.

Since $\mathbb{E}[\widehat{\boldsymbol{\mu}}_{10}] = \mathbb{E}_{\mathcal{R}}[\mathbf{Y}^{(0)}]$, it is unbiased, but $\widehat{\boldsymbol{\mu}}_{00}$ could be biased, denote $\mathbb{E}[\widehat{\boldsymbol{\mu}}_{00}] = \mathbb{E}_{\mathcal{R}}[\mathbf{Y}^{(0)}] + B$ where B characterizes the bias of estimating $\mathbb{E}_{\mathcal{R}}[\mathbf{Y}^{(0)}]$ using the external control. Then the mean square error (MSE) indexed by weight

$$\begin{aligned} \mathrm{MSE}(w) &\triangleq \mathbb{E}\left[\widehat{\tau}(w) - \tau \right)^2 \Big] \\ &= \mathrm{Var}[\widehat{\boldsymbol{\mu}}_{11}] + (1-w)^2 \mathrm{Var}[\widehat{\boldsymbol{\mu}}_{10}] \\ &\quad + w^2 \ \mathrm{Var}[\widehat{\boldsymbol{\mu}}_{00}] + B^2) \\ &\quad - 2(1-w)\mathrm{Cov}[\widehat{\boldsymbol{\mu}}_{11}, \widehat{\boldsymbol{\mu}}_{10}] - 2w\mathrm{Cov}[\widehat{\boldsymbol{\mu}}_{11}, \widehat{\boldsymbol{\mu}}_{00}] \\ &\quad + 2w(1-w)\mathrm{Cov}[\widehat{\boldsymbol{\mu}}_{10}, \widehat{\boldsymbol{\mu}}_{00}] \end{aligned} \tag{8}$$

where $\widehat{\tau}(w)$ is a generic representation of either $\widehat{\tau}_{n,m}^{EC-IPW}$ or $\widehat{\tau}_{n,m}^{EC-AIPW}$ in Definitions 1 and 2.

The MSE is quadratic in weight w, and the optimal weight minimizing equation (8) is

$$w^* = \frac{\begin{pmatrix} \mathrm{Var}[\widehat{\boldsymbol{\mu}}_{10}] + \mathrm{Cov}[\widehat{\boldsymbol{\mu}}_{11}, \widehat{\boldsymbol{\mu}}_{00}] \\ -\mathrm{Cov}[\widehat{\boldsymbol{\mu}}_{11}, \widehat{\boldsymbol{\mu}}_{10}] - \mathrm{Cov}[\widehat{\boldsymbol{\mu}}_{10}, \widehat{\boldsymbol{\mu}}_{00}] \end{pmatrix}}{\begin{pmatrix} \mathrm{Var}[\widehat{\boldsymbol{\mu}}_{10}] + \mathrm{Var}[\widehat{\boldsymbol{\mu}}_{00}] \\ +B^2 - 2\mathrm{Cov}[\widehat{\boldsymbol{\mu}}_{10}, \widehat{\boldsymbol{\mu}}_{00}] \end{pmatrix}} \tag{9}$$

The optimal weight in equation (9) can be simplified when the nuisance components are known,

$$w^* = \frac{\mathrm{Var}[\widehat{\boldsymbol{\mu}}_{10}]}{\mathrm{Var}[\widehat{\boldsymbol{\mu}}_{10}] + \mathrm{Var}[\widehat{\boldsymbol{\mu}}_{00}] + B^2} \tag{10}$$

Note that the optimal weight depends on the variances of trial controls' estimate $\mathrm{Var}[\widehat{\boldsymbol{\mu}}_{10}]$ and external controls' estimate $\mathrm{Var}[\widehat{\boldsymbol{\mu}}_{00}]$, as well as the bias term B: If the variance of the external control' estimate is large relative to the external control' estimate, or the bias is large, the optimal weight is close to 0, reflecting that the incorporation of external controls tends to inflate unnecessary bias that cannot be compensated by a decrease in variance; on the other hand, when the variance of external controls estimate is small relative to the external controls' estimate, and its bias is negligible, then the optimal weight is larger, reflecting that the incorporation of external controls does not incur extra bias and can increase precision.

Bias could arise through the incorporation of external controls for several reasons: (1) When there is a direct effect of trial participation such that Assumption 1 is not satisfied. For example, patients in the clinical trial might be monitored more closely, receive better care, or simply be measured differently. (2) Assumption 4 is violated by the existence of unmeasured confounding. For example, rare disease patients may differ in terms of access to high-quality care, financial resources, or general living conditions, that might make some patients less likely to participate in the RCT, and coincidentally, these same conditions could exacerbate the progression of the disease. Therefore, the population participating in the trial could be self-selected in a way that differs from the external control population in manners that investigators are unaware of. (3) The correct models to estimate the nuisance parameters $\pi_S(\mathbf{X})$ and $\mu(\mathbf{X})$ are never known and subject to misspecification. Though (1) and (2) are collectively falsifiable by testing for distributional differences in outcome between the trial controls and external controls given covariates, such tests are subject to limitation of Type I and Type II errors.

Additional simulations are presented in the Web-based online supplementary materials to demonstrate the role of the synthesizing weight w in the bias-variance trade-off, facing violations of the causal Assumptions 1, 4, and 5, as well as model misspecification for nuisance components.

3.3 The choice of weight

The weight w needs to be chosen at the design stage. Though desirable, the weight that minimizes MSE in equation (9) cannot be estimated without access to the outcome data. One option is to select the weight at the design stage based on subjective knowledge about the size and quality of the external controls: if the sample size of the external controls is large and it is likely that the external controls are similar to the trial population, then a larger weight can be assigned; a smaller weight can be assigned if the opposite is true. However, this choice is subjective and vulnerable to misconception that may end up with too much bias or no precision gain.

An alternative approach is to choose a weight that approximately minimizes the variability of the resulting estimator without access to the outcome data. This aligns with Rubin's principles in designing observational studies (Rubin, 2008), that the design phase (refers to employing propensity scores to help create distributional balance of covariates between the two treatment groups) should be done without access to any outcome data. With a few simplifying assumptions, equation (10) can be approximated by a 'variance ratio' (shown in Appendix B), which is estimable without any outcome data.

$$\widehat{w}^* = \frac{\dfrac{\sum_i S_i(1-A_i)\widehat{W}_{10i}^2}{\left(\sum_i S_i(1-A_i)\widehat{W}_{10i}\right)^2}}{\dfrac{\sum_i S_i(1-A_i)\widehat{W}_{10i}^2}{\left(\sum_i S_i(1-A_i)\widehat{W}_{10i}\right)^2} + \dfrac{\sum_i (1-S_i)\widehat{W}_{00i}^2}{\left(\sum_i (1-S_i)\widehat{W}_{00i}\right)^2}} \tag{11}$$

Equation (11) can be interpreted as the proportion of variance in covariate balancing weights associated with the trial controls relative to the total variance in these weights across both trial and external controls. Intuitively, a greater disparity in the covariate distribution $\mathbf{X}$ leads to increased variability in the covariate balancing weights for the external controls, consequently elevating the

variability of the hybrid estimator. A higher value of $\widehat{w}$ suggests a smaller weight assigned to the external controls. Conversely, if the external controls have a covariate distribution similar to that of the trial subjects, the variability in their covariate balancing weights decreases. In such cases, $\widehat{w}$ would indicate a larger weight assigned to the external controls, effectively treating them almost on par with the trial data in terms of influence and relevance.

We will refer to estimators with weight $\widehat{w}^*$ as the *Optimally weighted EC-IPW-OPT* and *Optimally weighted EC-AIPW-OPT*, respectively.

3.4 Large-sample properties of the weighting estimators

The results of the section provide the asymptotic distributions of the EC-IPSW and EC-AIPSW estimators, which is necessary for carrying out inference tasks. To study the (asymptotic) behaviour of the weighting estimators (EC-IPW and EC-AIPW) presented in Section 3.1, we express them as the solutions to corresponding estimating equations to establish asymptotic normality and to provide consistent sandwich estimators for the variances (Tsiatis, 2006).

We consider the practical case when the nuisance parameters $\pi_S(\mathbf{X})$ and $\mu(\mathbf{X})$ are estimated via parametric models that satisfy Assumption 6. These requirements are standard and satisfied when the outcome regression and propensity score models are estimated by maximum likelihood methods.

Assumption 6 $g(\mathbf{X}) = g(\mathbf{X}; \boldsymbol{\theta})$ is a parametric model, where $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^k$, Θ is compact; (ii) $g(\mathbf{X}; \boldsymbol{\theta})$ is a.s. continuous at each $\boldsymbol{\theta} \in \Theta$; (iii) there exists a unique pseudo-true parameter $\boldsymbol{\theta}^* \in \text{int}\,(\Theta)$; (v) the estimator $\widehat{\boldsymbol{\theta}}$ is consistent for $\boldsymbol{\theta}^*$.

Theorem 3 (Asymptotics of $\widehat{\tau}^{EC-IPW}_{n,m}$). Let $\boldsymbol{\theta}^* = (\boldsymbol{\mu}_{11}, \boldsymbol{\mu}_{10}, \boldsymbol{\mu}_{00}, \boldsymbol{\alpha}^*)$, $\widehat{\boldsymbol{\theta}} = (\widehat{\boldsymbol{\mu}}_{11}, \widehat{\boldsymbol{\mu}}_{10}, \widehat{\boldsymbol{\mu}}_{00}, \widehat{\boldsymbol{\alpha}})$, and note that $\widehat{\boldsymbol{\theta}}$ is the solution for $\boldsymbol{\theta}^*$ of the estimating equation

$$\sum_{i \in \mathcal{R} \cup \mathcal{E}} \boldsymbol{\Psi}(\mathcal{O}_i; \theta) = 0,$$

where $\mathcal{O}_i = (\mathbf{X}_i, S_i, A_i, \mathbf{Y}_i)$, with the influence function

$$\underbrace{\boldsymbol{\Psi}(\mathcal{O}_i; \theta)}_{(3T+p)\times 1} = \begin{pmatrix} \boldsymbol{\Psi}_1(\mathcal{O}_i; \theta) \\ \boldsymbol{\Psi}_2(\mathcal{O}_i; \theta) \\ \boldsymbol{\Psi}_3(\mathcal{O}_i; \theta) \\ \boldsymbol{\Psi}_4(\mathcal{O}_i; \theta) \end{pmatrix} = \begin{pmatrix} \dfrac{S_i A_i (\mathbf{Y}_i - \boldsymbol{\mu}_{11})}{\pi_S \pi_A} \\ \dfrac{S_i (1 - A_i)(\mathbf{Y}_i - \boldsymbol{\mu}_{10})}{\pi_S (1 - \pi_A)} \\ \dfrac{(1 - S_i) W_{00i} (\mathbf{Y}_i - \boldsymbol{\mu}_{00})}{1 - \pi_S} \\ \left(S_i - \dfrac{e^{\mathbf{X}_i^T \alpha}}{1 + e^{\mathbf{X}_i^T \alpha}} \right) \mathbf{X}_i \end{pmatrix}$$

that satisfies $\mathbb{E}[\boldsymbol{\Psi}(\mathcal{O}; \theta)] = 0$.

Then as $n, m \to \infty$, $\widehat{\boldsymbol{\theta}}$ converges in probability to $\boldsymbol{\theta}^*$, and $\sqrt{n+m}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)$ converges in distribution to $N(0, \Sigma^{EC-IPW}_{\boldsymbol{\theta}})$, where

$$\begin{aligned} \underbrace{\Sigma^1}_{(3T+p)\times(3T+p)} &= \mathbf{A}_1^{-1} \mathbf{B}_1 \mathbf{A}_1^{-T} \\ \underbrace{\mathbf{A}_1}_{(3T+p)\times(3T+p)} &= \mathbb{E}\left[\frac{\partial}{\partial \boldsymbol{\theta}} \boldsymbol{\Psi}\ \mathcal{O}; \theta^*) \right] \\ &= \begin{pmatrix} -\mathbb{I}_T & 0 & 0 & 0 \\ 0 & -\mathbb{I}_T & 0 & 0 \\ 0 & 0 & -\mathbb{E}_{\mathcal{E}}[W_{00}]\mathbb{I}_T & \mathbb{E}\left[\dfrac{(1 - S)\pi_S(\mathbf{X})}{(1 - \pi_S(\mathbf{X}))\pi_S} \mathbf{Y} - \boldsymbol{\mu}_{00}) \mathbf{X}^T \right] \\ 0 & 0 & 0 & \mathbb{E}[-\pi_S(\mathbf{X})(1 - \pi_S(\mathbf{X}))\mathbf{X}\mathbf{X}^T] \end{pmatrix} \\ \underbrace{\mathbf{B}_1}_{(3T+p)\times(3T+p)} &= \text{Var}[\boldsymbol{\Psi}\ \mathcal{O}; \theta^*)] = \mathbb{E}\left[\boldsymbol{\Psi}\ \mathcal{O}; \theta^*) \boldsymbol{\Psi}\ \mathcal{O}; \theta^*)^T \right] \end{aligned} \tag{12}$$

Assuming that the logistic regression model $\pi_S(\mathbf{X}; \boldsymbol{\alpha})$ for $\pi_S(\mathbf{X})$ is correct, then by equation (3), Slusky's theorem and the delta method, $\widehat{\boldsymbol{\tau}}^{EC-IPW}_{n,m}$ is a consistent estimator of $\boldsymbol{\tau}$, for any fixed $w \in [0, 1]$, meaning that

1. As $n, m \to \infty$, $\widehat{\boldsymbol{\tau}}^{EC-IPW}_{n,m}$ converges in probability to $\boldsymbol{\tau}$.
2. $\sqrt{n+m}(\widehat{\boldsymbol{\tau}}^{EC-IPW}_{n,m} - \boldsymbol{\tau})$ converges in distribution to $\mathcal{N}(0, \Sigma^{EC-IPW})$, where $\Sigma^{EC-IPW} = \Sigma^1_{11} + (1-w)^2\Sigma^1_{22} + w^2\Sigma^1_{33}$ and $\Sigma^1_{11}, \Sigma^1_{22}$, and Σ^1_{33} are block matrices in Σ^1 corresponding to the var-covariance of $\widehat{\boldsymbol{\mu}}_{11}, \widehat{\boldsymbol{\mu}}_{10}$, and $\widehat{\boldsymbol{\mu}}_{00}$, respectively. A consistent estimate of the variance can be obtained by plugging in sample analogue of population expectations into the expression for matrices $\mathbf{A}_1$ and $\mathbf{B}_1$:

$$\widehat{\mathrm{Var}}\left[\hat{\tau}^{EC-IPW}\right] = \frac{1}{n+m}\left(\widehat{\Sigma}^1_{11} + (1-w)^2\widehat{\Sigma}^1_{22} + w^2\widehat{\Sigma}^1_{33}\right) \tag{13}$$

We can use this result to build Gaussian confidence intervals:

$$\mathbb{P}\left[\tau \in \left\{\widehat{\boldsymbol{\tau}}^{EC-IPW}_{n,m} \pm z_{1-\alpha/2}\widehat{\mathrm{Var}}\left[\hat{\tau}^{EC-IPW}\right]^{1/2}\right\}\right] \to 1-\alpha \tag{14}$$

Next we show that $\widehat{\tau}^{EC-AIPW}_{n,m}$ is DR or doubly protected in that it remains consistent when either the trial participation model or the outcome model is correctly specified.

Theorem 4 (Asymptotics of $\widehat{\tau}^{EC-AIPW}_{n,m}$). Let $\boldsymbol{\theta}^* = (\boldsymbol{\mu}_{11}, \boldsymbol{\mu}_{10}, \boldsymbol{\mu}_{00}, \boldsymbol{\alpha}^*, \boldsymbol{\beta}^*)$, $\widehat{\boldsymbol{\theta}} = (\widehat{\boldsymbol{\mu}}_{11}, \widehat{\boldsymbol{\mu}}_{10}, \widehat{\boldsymbol{\mu}}_{00}, \widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\beta}})$, and note that $\widehat{\boldsymbol{\theta}}$ is the solution for $\boldsymbol{\theta}^*$ of the estimating equation

$$\sum_{i \in \mathcal{R} \cup \mathcal{E}} \boldsymbol{\Psi}(\mathcal{O}_i; \theta) = 0,$$

where $\mathcal{O}_i = (\mathbf{X}_i, S_i, A_i, \mathbf{Y}_i)$, with the influence function

$$\boldsymbol{\Psi}(\mathcal{O}_i; \theta) = \begin{pmatrix} \boldsymbol{\Psi}_1(\mathcal{O}_i; \boldsymbol{\theta}) \\ \boldsymbol{\Psi}_2(\mathcal{O}_i; \boldsymbol{\theta}) \\ \boldsymbol{\Psi}_3(\mathcal{O}_i; \boldsymbol{\theta}) \\ \boldsymbol{\Psi}_4(\mathcal{O}_i; \boldsymbol{\theta}) \\ \boldsymbol{\Psi}_5(\mathcal{O}_i; \boldsymbol{\theta}) \end{pmatrix} = \begin{pmatrix} \dfrac{S_i A_i(\widetilde{\mathbf{Y}}_i - \boldsymbol{\mu}_{11})}{\pi_S \pi_A} \\ \dfrac{S_i(1-A_i)(\widetilde{\mathbf{Y}}_i - \boldsymbol{\mu}_{10})}{\pi_S(1-\pi_A)} \\ \dfrac{(1-S_i)W_{00i}(\widetilde{\mathbf{Y}}_i - \boldsymbol{\mu}_{00})}{1-\pi_S} \\ \left(S_i - \dfrac{e^{\mathbf{X}^T\boldsymbol{\alpha}}}{1+e^{\mathbf{z}^T\boldsymbol{\alpha}}}\right)\mathbf{X} \\ \frac{1-A_i}{1-\pi_A}(\mathbf{Y} - \mathbf{X}^T\boldsymbol{\beta})\mathbf{X} \end{pmatrix}$$

that satisfies $E[\boldsymbol{\Psi}(\mathcal{O}; \theta)] = 0$

Then as $n, m \to \infty$, $\widehat{\boldsymbol{\theta}}$ converges in probability to $\boldsymbol{\theta}^*$, and $\sqrt{n+m}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)$ converges in distribution to $\mathcal{N}(0, \Sigma_{\boldsymbol{\theta}}^{EC-AIPW})$, where

$$\begin{aligned}
\underbrace{\Sigma^2}_{(3T+p+q)\times(3T+p+q)} &= \mathbf{A}_2^{-1}\mathbf{B}_2\mathbf{A}_2^{-T} \\
\underbrace{\mathbf{A}_2}_{(3T+p+q)\times(3T+p+q)} &= \mathbb{E}\left[\frac{\partial}{\partial\boldsymbol{\theta}}\boldsymbol{\Psi}(\mathcal{O};\boldsymbol{\theta})\right] \\
&= \left[\begin{array}{cccc|c} & & & & -\mathbb{E}\left[\frac{SA}{\pi_S\pi_A}\mathbf{X}^T\right] \\ & & \mathbf{A}_1(\boldsymbol{\theta}) & & -\mathbb{E}\left[\frac{S(1-A)}{\pi_S(1-\pi_A)}\mathbf{X}^T\right] \\ & & & & -\mathbb{E}\left[\frac{1-S}{1-\pi_S}W_{00}\mathbf{X}^T\right] \\ \hline 0 & 0 & 0 & 0 & -\mathbb{E}\left[\frac{1-A_i}{1-\pi_A}\mathbf{X}\mathbf{X}^T\right] \end{array}\right] \\
\underbrace{\mathbf{B}_2}_{(3T+p+q)\times(3T+p+q)} &= \mathrm{Var}[\boldsymbol{\Psi}\ \mathcal{O};\boldsymbol{\theta})] = \mathbb{E}\left[\boldsymbol{\Psi}\ \mathcal{O};\boldsymbol{\theta})\boldsymbol{\Psi}\ \mathcal{O};\boldsymbol{\theta})^T\right]
\end{aligned} \tag{15}$$

where $\mathbf{A}_1(\boldsymbol{\theta})$ is defined in Theorem 3

Assuming that either the logistic regression model $\pi_S(\mathbf{X};\boldsymbol{\alpha})$ for $\pi_S(\mathbf{X})$ or the outcome model $\mu(\mathbf{X};\boldsymbol{\beta})$ for $\mu(\mathbf{X})$ is correct, then by equation (5), Slusky's theorem and the delta method, $\widehat{\boldsymbol{\tau}}_{n,m}^{EC-AIPW}$ is a consistent estimator of $\boldsymbol{\tau}$, for any fixed $w \in [0, 1]$, meaning that

1. As $n, m \to \infty$, $\widehat{\boldsymbol{\tau}}_{n,m}^{EC-AIPW}$ converges in probability to $\boldsymbol{\tau}$.
2. $\sqrt{n+m}(\widehat{\boldsymbol{\tau}}_{n,m}^{EC-AIPW} - \boldsymbol{\tau})$ converges in distribution to $\mathcal{N}(0, \Sigma^{EC-AIPW})$, where $\Sigma^{EC-AIPW} = \Sigma_{11}^2 + (1-w)^2\Sigma_{22}^2 + w^2\Sigma_{33}^2$ and Σ_{11}^2, Σ_{22}^2, and Σ_{33}^2 are block matrices in Σ^2 corresponding to the var-covariance of $\widehat{\boldsymbol{\mu}}_{11}$, $\widehat{\boldsymbol{\mu}}_{10}$, and $\widehat{\boldsymbol{\mu}}_{00}$, respectively. A consistent estimate of the variance can be obtained by plugging in sample analogue of population expectations into the expression for matrices $\mathbf{A}_2$ and $\mathbf{B}_2$:

$$\begin{aligned}
\widehat{\mathrm{Var}}\left[\hat{\tau}^{EC-AIPW}\right] = \frac{1}{n+m}(&\widehat{\Sigma}_{11}^2 \\
&+ (1-w)^2\widehat{\Sigma}_{22}^2 + w^2\widehat{\Sigma}_{33}^2)
\end{aligned} \tag{16}$$

We can use this result to build Gaussian confidence intervals:

$$\mathbb{P}\left[\tau \in \left\{\widehat{\boldsymbol{\tau}}_{n,m}^{EC-AIPW} \pm z_{1-\alpha/2}\widehat{\mathrm{Var}}\left[\hat{\tau}^{EC-AIPW}\right]^{1/2}\right\}\right] \to 1-\alpha \tag{17}$$

4 Other methods: Bayesian dynamic borrowing

The goal of Bayesian methods, similar to our proposal, is to increase precision when external controls are 'compatible' and simultaneously control bias when not 'compatible'. In this work, we do not intend to do an exhaustive study of the Bayesian methods, instead, we compare the performance of two widely used approaches from this category with our proposed methods.

Let θ denote model parameters (regression parameters for the relation between covariates $\mathbf{X}$ and the outcomes) and $L(\theta \mid S)$ denote a general likelihood function associated with a given outcome model [such as linear and generalized linear model (GLM)], and a population ($S = 1$ for trial population, $S = 0$ for external control population).

Formulation of the power prior. Following (Ibrahim & Chen, 2000), the power prior is formulated as

$$\pi\ \theta \mid S=0, a_0) \propto L\ \theta \mid S=0)^{a_0} \pi_0(\theta_0) \tag{18}$$

where $0 \leq a_0 \leq 1$ is a discounting parameter for the external controls data likelihood, and $\pi_0(\theta)$ is the initial prior for θ. The parameter a_0 allows researchers to control the influence of the external controls: with $a_0 = 0$, external control information is discarded, and with $a_0 = 1$, the external controls contribute equally as the trial data to the likelihood. Priors can be specified for a_0, such as the beta distribution, and the choice is discussed in Ibrahim et al. (2015).

Formulation of the commensurate priors. Following (Hobbs et al., 2011), the informative prior is constructed using a hierarchical model that incorporates commensurate priors as the primary mechanism for weighting the influence of prior information relative to its consistency with the trial data. Specifically, the informative prior is constructed in a hierarchical model by specifying the prior for the trial parameters $\boldsymbol{\theta}$ to be 'centred' at the external control parameters $\boldsymbol{\theta}_0$ with precision $\tau > 0$, where τ is the commensurability parameter that quantifies the 'similarity' among the external controls and the trial controls in terms of model parameters $\boldsymbol{\theta}$ and $\boldsymbol{\theta}_0$,

$$\pi\ \theta \mid S=0, \theta_0, \tau) \propto L\ \theta_0 \mid S=0)\pi\ \theta \mid \theta_0, \tau)\pi_0(\theta) \tag{19}$$

As $\tau \to 0$, the external controls data is discarded. On the other hand, as $\tau \to \infty$, the external controls are treated equivalently as the trial data.

5 Simulations

In choosing a simulation to comparison the finite-sample performance of the methods discussed in Sections 3 and 4, we design our simulations to have a few properties. First, the data generating mechanism should reflect the particular problem under study. Second, an important property of the simulation is the realism of the relationship between the confounding variables and the outcome, as well as a realistic level of confounding. Third, as all methods require either an outcome model or propensity score model (or both), important aspect of these estimators to comparison is how they respond to possible model misspecifications.

The simulations are based on the following data generating processes that is consistent with the DAG in Figure 1:

$$\begin{cases} S = Bernoulli(\pi_S(\mathbf{W})) \\ A = \begin{cases} Bernoulli(\pi_A), & if\ S=1 \\ 0, & if\ S=0 \end{cases} \\ Y_t(a) = \mu(\mathbf{W}, t) + a\tau_t + \epsilon_t, a \in \{0,1\}, t \in [0, T] \end{cases} \tag{20}$$

where τ_t represents the true time dependent ATE.

Let $\mathbf{W}$ denote a vector of baseline covariates that includes both the measured ($\mathbf{X}$) and unmeasured ($\mathbf{U}$) confounders. We simulate $\mathbf{W}$ based on the empirical distribution of the measured baseline covariates in the combined SUNFISH and external controls data set. There are five baseline covariates that $\mathbf{X}$ and U emulate: scoliosis (binary, yes or no) as W_1 , SMN2 copy number (binary, 2, 3, or 4) as W_2, baseline MFM (continuous) as W_3, age at enrolment (continuous) as W_4, and SMA type (binary, Type II or III) as W_5. Different subsets of ($W_1, \ldots, W_5$) are designated as measured $\mathbf{X}$ and unmeasured $\mathbf{U}$ based on the needs of each simulation setting. We then sample the trial participation on the basis of a propensity score model learned from fitting a logistic regression model on the real data. The treatment assignment is based on the trial participation and 2:1 treatment-to-control ratio for trial participants. Lastly, the outcome is sampled based on a linear model learned from the real data. The sample sizes are set to be similar as the real data, with total $n+m$ to be 220, with $n_1 : n_0 : m$ roughly equals to 2 : 1 : 1 ratio, resulting in $n_1 \approx 110, n_0 \approx 55, m \approx 55$. This approach ensures that the joint distribution of ($\mathbf{W}$, S, A, $\mathbf{Y}(0)$, $\mathbf{Y}(1)$) are as realistic as possible, coming directly from real data.

To demonstrate the performance of proposed methods in different practical scenarios, we simulate five settings with various levels of selection bias due to unobserved U (ranging from the most ideal case where U does not exist, the more practical case where U exist) and various levels of difficulty in modelling the nuisance components: (1) No causal assumption violated, models for both nuisance parameters well-specified, (2) outcome model misspecified, (3) propensity score model misspecified, (4) both outcome model and propensity score model misspecified, and (5) No unmeasured confounding assumption violated.

To simulate these settings, the true models are

$$logit\{\pi_S(\mathbf{W})\} = \begin{cases} \boldsymbol{\alpha}'(W_1, W_2, W_3, W_4, W_5), & \text{for Settings 1, 2, 5} \\ \boldsymbol{\alpha}'(W_1, W_2, W_3, W_4, W_4^2, W_5), & \text{for Settings 3, 4} \end{cases}$$

$$\mu(\mathbf{W}, t) = \begin{cases} \boldsymbol{\beta}'(W_1, W_2, W_3, W_4, W_5), & \text{for Settings 1, 3, 5} \\ \boldsymbol{\beta}'(W_1, W_2, W_3, W_4, W_4^2, W_5), & \text{for Settings 2, 4} \end{cases}$$

The observed and omitted confounding variables are

$$\mathbf{X} = \begin{cases} (W_1, W_2, W_3, W_4, W_5), & \text{for Settings 1, 2, 3, 4} \\ (W_1, W_2, W_3, W_4), & \text{for Setting 5} \end{cases}$$

$$U = \begin{cases} \text{None}, & \text{for Settings 1, 2, 3, 4} \\ W_5, & \text{for Setting 5} \end{cases}$$

We then specify the propensity of trial participation model and the component of the outcome model related to measured confounders as

$$logit\{\pi_S(\mathbf{W})\} = \begin{cases} \boldsymbol{\alpha}'(W_1, W_2, W_3, W_4, W_5), & \text{for Settings 1, 2, 3, 4} \\ \boldsymbol{\alpha}'(W_1, W_2, W_3, W_4) & \text{for Setting 5} \end{cases}$$

$$\mu(\mathbf{W}, t) = \begin{cases} \boldsymbol{\beta}'(W_1, W_2, W_3, W_4, W_5), & \text{for Settings 1, 2, 3, 4} \\ \boldsymbol{\beta}'(W_1, W_2, W_3, W_4), & \text{for Setting 5} \end{cases}$$

For each design, we perform 3,000 Monte Carlo simulations.

Bayesian specifications:. The power prior and commensurate prior for generalized linear models accommodating our simulation setup is presented in Appendix C.

5.1 Results

At the trial design stage, operating characteristics including Type I error rate, power, and coverage are important. Theoretically, if two estimators are consistent and their variances are correctly specified, then at least for large samples, both should have well-controlled Type I error, the one with smaller standard error (SE) should have higher power. It has been shown in Theorem 3 that, when the probability of trial participation model is correctly specified, EC-IPW is consistent, and when either the probability of trial participation or the outcome model is correctly specified, EC-AIPW is consistent (doubly robust). Generally, DR estimators also enjoy the benefit of variance reduction and therefore efficiency gain. We expect to see more variance reduction with EC-AIPW as a DR estimator. This supports the validity of our proposed estimators towards the goal of trial design with correct Type I error control and power boost, especially the EC-AIPW estimator.

We compare methods in terms of bias, SE, MSE, 95% coverage probability, power (at a specific nonnull treatment effect similar to the observed effect in SUNFISH) and Type I error (at null treatment effect), estimated empirically across 3,000 simulations. We use the mixed models for

repeated measures on trial data (within trial MMRM) as the reference to compare methods in Sections 3 and 4, as it is the standard method of reporting trial results with only RCT data.

The degree of borrowing from external controls is indicated by the 'weight' column in Table 1; however, the values are not directly comparable across methods. In the power prior method, the weight represents the posterior mean of the discounting parameter in equation (19). For the commensurate prior method, it corresponds to the posterior mean of the commensurability parameter in equation (20). In the case of EC-IPW-OPT and EC-AIPW-OPT, the weight w is defined in Definitions 1 and 2, respectively. Across all settings, the power prior method exhibits minimal borrowing from external controls, as its power parameters have posterior distributions centred around a value close to zero. The commensurate prior method displays varying degrees of borrowing, with more significant borrowing in Settings 1–4 and less borrowing in Setting 5. A larger commensurability parameter implies greater agreement between trial controls and external controls in terms of parameter values, resulting in increased borrowing. For EC-IPW-OPT and EC-AIPW-OPT, the weight $\hat{w}^*$ is determined in a data-driven manner without using outcome data. Greater weight is given to external controls when they offer more precise estimates, which occurs when a large number of similar external controls are available in terms of observed confounders. In Settings 1 and 2, larger weights result in minimal bias and significant efficiency gains from incorporating external controls. In Settings 3 and 4, smaller weights lead to minor bias increases and still notable power gains. In Setting 5, where an unmeasured confounder is present, larger weights are assigned because external controls are similar to trial subjects in terms of measured attributes. This still leads to efficiency gains that outweigh the bias introduced by the unmeasured confounder.

The bias column in Table 1 displays the estimated bias across 3,000 simulations. Generally, both the weighting estimators and the Bayesian methods effectively control the additional bias introduced by utilizing external controls under various mechanisms.

Figure 2 and the SE column in Table 1 show the potential benefit of incorporating external controls using the proposed estimators as well as the Bayesian methods, which is to increase the precision or reduce the variability. For EC-IPW-OPT, and EC-AIPW-OPT, SEs were estimated using the asymptotic sandwich variance estimators given in Section 3.4, and for the power prior and commensurate prior methods, although no frequentist SE is defined, we use the standard deviation of the posterior sampling to make the comparison. The EC-IPW-OPT and EC-AIPW-OPT estimators show greater reduction across all settings, with EC-AIPW-OPT having the greatest reduction.

The last column in Table 1 shows the empirical Type I error rates, for testing the null hypothesis $H_0 : \tau = 0$ versus alternative $H_0 : \tau \neq 0$. For EC-IPW-OPT, and EC-AIPW-OPT, 95% Wald confidence intervals were constructed based on the estimated SEs and the asymptotic normality results established in Section 3.4. For the power prior and commensurate prior methods, the posterior 95% posterior credible intervals were constructed. The confidence or credible intervals were used to perform inference. We notice that both the Bayesian methods and our proposed weighting estimators have Type I error close to the nominal value of 0.05.

The power column in Table 1 shows the estimated power at one specific alternative that is most relevant to the SUNFISH trial $\tau = 1.5$, for testing the null hypothesis $H_0 : \tau = 0$ versus alternative $H_0 : \tau \neq 0$. Again, the confidence or credible intervals were then used to perform inference. The within trial MMRM are powered slightly over 70% depending on settings. The 'power gain' column shows the power increase using the *within trial MMRM* as the reference. The four methods can be ranked by their power gain as follows: the EC-AIPW-OPT has the highest power gain, followed by the EC-IPW-OPT as the second-highest, then the commensurate prior method as the third-highest power gain, and finally, the power prior method shows little to no power gain. In addition, the estimated power curves over a range of nonnull treatment effects are shown in Figure 3 to augment Table 1. The pattern of relative power performance remains consistent across a variety of nonnull treatment effects, encompassing the specific alternative that holds the most relevance for the SUNFISH trial. In summary, the proposed causal weighting estimators, particularly EC-AIPW-OPT, exhibit a more substantial and consistent power gain compared to other methods.

The coverage column in Table 1 shows the empirical 95% CI coverage probabilities. Due to our use of 3,000 simulations, empirical coverage rates between 0.943 and 0.957, can be considered close to 0.95. EC-IPW-OPT and EC-AIPW-OPT have slight under coverage in Settings 3 and 4, while the Bayesian posterior credible intervals achieve the frequentist nominal coverage rate.

Table 1. Estimated bias, standard error (SE), root mean square error (RMSE), 95% coverage probability, power (at nonnull treatment effect 1.5), power gain using the 'within trial MMRM' as reference, and Type I error (at null treatment effect), across 3,000 simulations[c]

Method	Weight[a]	Bias	SE	RMSE	Coverage (%)	Power (%)	Power gain[b] (%)	Type I error (%)
Setting 1								
Within trial MMRM	–	−0.007	0.591	0.591	94.3	72.6	0.0	4.7
Power prior	0.001	−0.010	0.578	0.579	**95.3**	72.4	−0.1	**4.7**
Commensurate prior	2.370	0.018	0.547	0.547	95.1	78.9	6.3	**4.7**
EC-IPW-OPT	0.402	**0.002**	0.536	0.536	95.0	81.0	8.4	**4.7**
EC-AIPW-OPT	0.404	−0.003	**0.493**	**0.493**	94.7	**85.9**	**13.4**	5.1
Setting 2								
Within trial MMRM	–	0.002	0.586	0.586	94.8	73.5	0.0	5.6
Power prior	0.001	0.006	0.573	0.573	**95.5**	73.4	−0.1	6.9
Commensurate prior	2.363	0.003	0.544	0.544	95.4	78.0	4.5	5.5
EC-IPW-OPT	0.404	0.007	0.548	0.549	93.7	79.8	6.3	4.5
EC-AIPW-OPT	0.403	**0.002**	**0.497**	**0.497**	94.7	**85.2**	**11.7**	**4.0**
Setting 3								
Within trial MMRM	–	−0.008	0.578	0.578	94.7	72.0	0.0	4.9
Power prior	0.001	−0.004	0.577	0.577	94.4	71.6	−0.4	5.0
Commensurate prior	2.349	−0.015	0.534	0.535	**95.6**	76.6	4.6	**4.3**
EC-IPW-OPT	0.404	0.006	0.530	0.530	94.9	81.2	9.2	5.0
EC-AIPW-OPT	0.403	**−0.002**	**0.505**	**0.505**	94.0	**84.9**	**12.9**	5.1
Setting 4								
Within trial MMRM	–	0.004	0.597	0.597	94.8	72.7	0.0	5.8
Power prior	0.001	**0.004**	0.577	0.577	94.5	73.6	0.8	**3.6**
Commensurate prior	2.350	−0.006	0.548	0.548	94.8	77.8	5.0	4.2
EC-IPW-OPT	0.402	**0.004**	0.529	0.529	**94.9**	80.4	7.7	4.4
EC-AIPW-OPT	0.403	−0.011	**0.506**	**0.507**	94.4	**84.3**	**11.6**	5.1
Setting 5								
Within trial MMRM	–	0.004	0.565	0.565	95.7	73.8	0.0	5.6
Power prior	0.001	−0.026	0.574	0.575	94.9	71.9	−1.9	5.8
Commensurate prior	2.245	0.014	0.536	0.537	95.1	78.6	4.8	**4.6**
EC-IPW-OPT	0.415	**0.006**	0.527	0.527	**95.3**	81.4	7.6	5.2
EC-AIPW-OPT	0.413	**0.006**	**0.506**	**0.506**	93.6	**85.9**	**12.1**	5.1

[a] The weight column represents different parameters for the four methods: for the *power prior*, it is posterior mean of discounting parameter in equation (19); for the *commensurate prior*, it is posterior mean of commensurability parameter in equation (20); and for *EC-IPW-OPT* and *EC-AIPW-OPT*, it is the weight w in Definitions 1 and 2, respectively.
[b] Power gain is calculated as the difference of power between the corresponding method and the *within trial MMRM*.
[c] The best number among the Bayesian methods and our proposed methods in each column (except for weight) is bolded.

Taking together, from the trial design perspective, the proposed causal weighting estimators have the advantage of significant power boost that are consistent across a range of favourable or unfavourable scenario, while maintaining the nominal Type I error.

6 Application: SUNFISH trial

In this section, the methods described in Sections 3 and 4 are applied to reanalyse the SUNFISH trial with the incorporation of an appropriately chosen external control group introduced in Section 1.

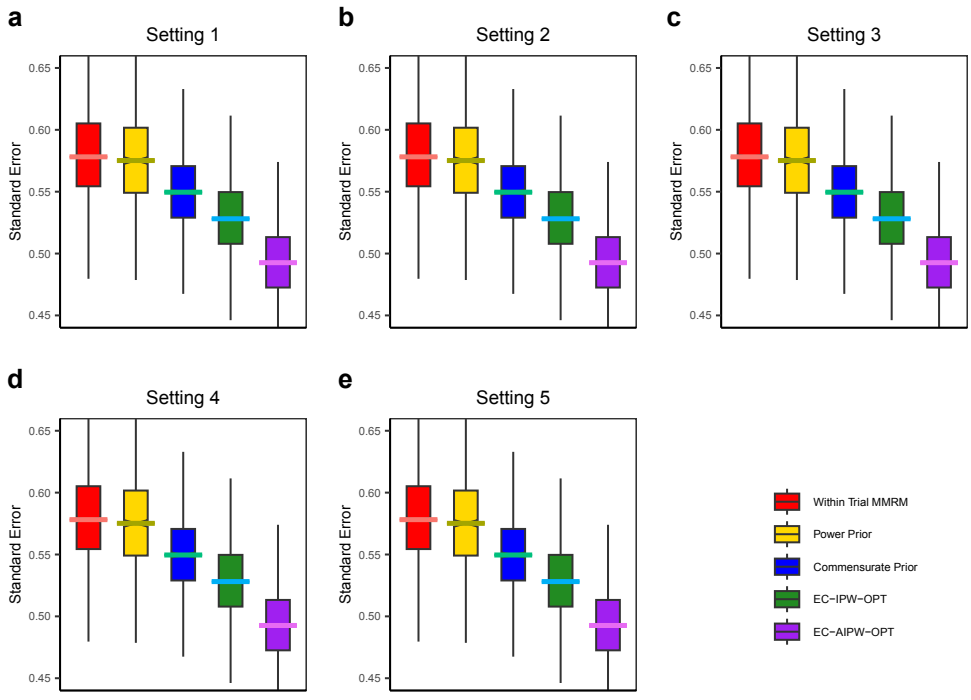

Figure 2. Boxplots of standard errors (SEs). For EC-IPW-OPT, and EC-AIPW-OPT, SEs were estimated using the asymptotic sandwich variance estimators given in Section 3.4, and for the power prior and commensurate prior methods, although no frequentist SE is defined, we use the standard deviation of the posterior sampling to make the comparison. a) No causal assumption violated, models for both nuisance parameters well-specified; b) Outcome model mis-specified; c) Propensity score model mis-specified; d) Both outcome model and propensity score model mis-specified; e) No unmeasured confounding assumption violated.

The EC-IPW-OPT and EC-IPW-OPT, along with the power prior and commensurate prior methods are employed to assess the ATE of risdiplam on MFM change from baseline, at months 6 and 12 (primary endpoint) for the SUNFISH trial population, using the olesoxime trial control subjects as the external controls. Whenever the propensity score model is needed (for EC-IPW-OPT and EC-AIPW-OPT), a logistic regression model was used, the outcome was trial participation and the possible covariates including Age at enrolment, SMA Type (II or III), SMN2 copy number (2,3,4), Scoliosis (Yes or No), and baseline MFM. Whenever the outcome model is needed (for power prior, commensurate prior, and EC-AIPW-OPT), a linear model was used, the outcome was MFM change from baseline and the possible covariates including time (categorical), Age at enrolment, SMA Type (II or III), SMN2 copy number (2,3,4), Scoliosis (Yes or No), and baseline MFM.

The distributions of the propensity of trial participation for the SUNFISH and external control subjects comparisons are shown in Figure 4 (a: Unweighted probability of trial participation). The slight imbalance in the two ends of the distributions is an indication of some covariate imbalance. The weighted comparison (c: Weighted probability of trial participation) in Figure 4 illustrates the improved balance by weighting the external controls using the EC-IPW and EC-AIPW weighting scheme. Figure 4b and d provides a closer look at one of the covariates, age at enrolment showing the unweighted and weighted age distributions under the EC-IPW and EC-AIPW weighting scheme. There is a group of relatively older patients in the external controls in the unweighted sample, but improved balance is achieved by weighting down this group.

The estimated ATE of risdiplam MFM change from baseline and corresponding 95% CIs, from different methods are given in Table 2 and visualized in Figure 5, along with the within trial MMRM as the reference which can be viewed as unbiased estimate of the true treatment effect. All four estimators show similar ATE as the within trial estimate at both time points. In addition, the EC-IPW-OPT and EC-AIPW-OPT estimators have noticeably narrower confidence intervals,

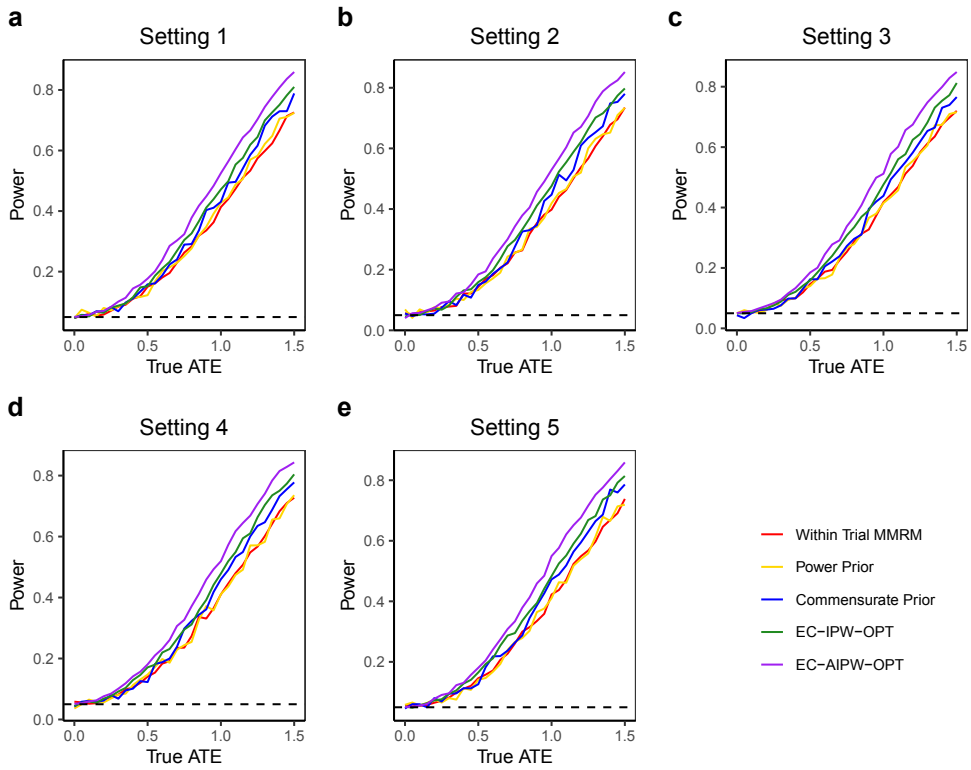

Figure 3. Estimated power at a range of treatment effects. a) No causal assumption violated, models for both nuisance parameters well-specified; b) Outcome model mis-specified; c) Propensity score model mis-specified; d) Both outcome model and propensity score model mis-specified; e) No unmeasured confounding assumption violated.

compared with within trial MMRM as well as the Bayesian methods. This shows that using data outside of the trial with effective adjustment methods have the potential to improve the efficiency of medical product approval by reducing the required trial subjects and shorten the length of the study.

7 Discussion

The primary aim of this work is to suggest methods based on causal inference for augmenting trial data with external control data to increase statistical power while preserving the desired Type I error rate. We have proposed two estimators, EC-IPW-OPT and its DR version EC-AIPW-OPT, that weight the trial and external controls combined data in two layers: the first layer of weights assigned to the external controls tackling the distribution shift between external control population and the trial population and can be thought of as a special case of the balancing weights in Li et al. (2018), and with second layer as a weighted combination of the external control evidence with the trial evidence. Through simulations representing a variety of practical scenarios, we have shown that the proposed weighting estimators achieve significant power gain, while maintaining Type I error close to the nominal value of 0.05, when certain assumptions are met. In particular, EC-AIPW-OPT has 'doubly robust' property about statistical model misspecifications similarly as other double robust estimators, and shows the most efficiency gain. In addition, the proposed weighting estimators are computationally less intensive compared to Bayesian methods, as the former have closed-form analytic expressions for both the point estimates and their confidence intervals, while the latter require intensive posterior sampling.

Whether the efficiency of the randomized trials can be improved by incorporating external controls hinges upon the 'compatibility' of external controls and the specification of statistical models for nuisance components in the proposed estimators. 'Compatibility' here refers to: (1) no direct

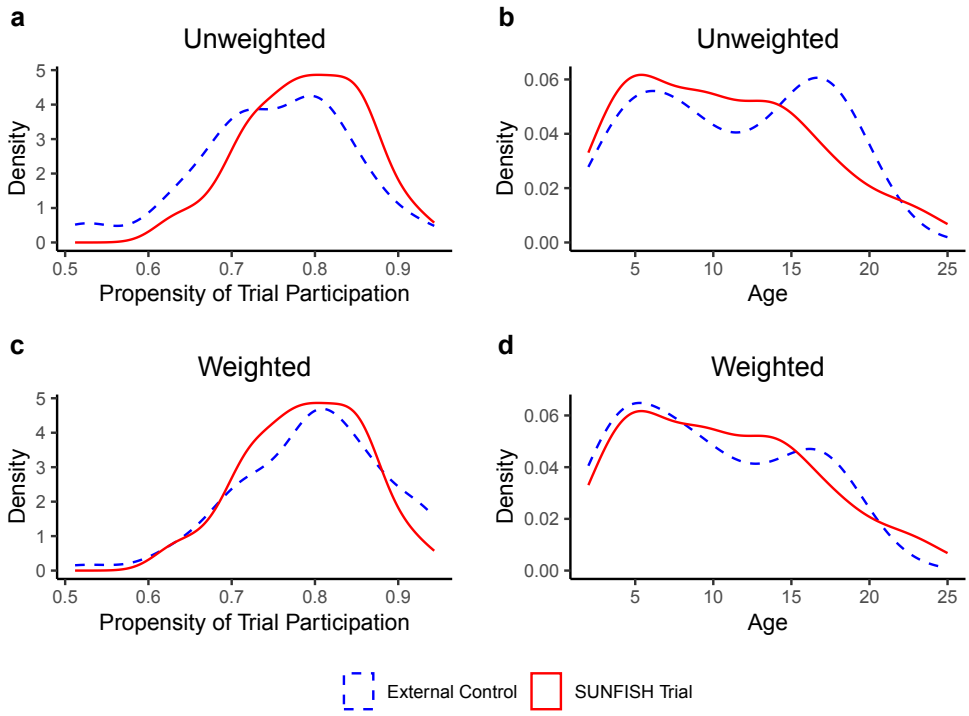

Figure 4. Comparisons of the distributions of the estimated probability of trial participation and age between SUNFISH and external control subjects. a) Unweighted distributions of the estimated probabilities of trial participation, c) weighted distributions of the estimated probabilities of trial participation, b) unweighted distributions of age, and d) weighted distributions of age.

effect of trial participation Assumption 1: a patient's potential outcome under no treatment would be the same, regardless of whether this patient is in the trial control group or the external control group (2) conditional ignorability Assumption 4: the outcome of trial controls and the external controls are exchangeable if they share the same characteristics, and (3) overlapping Assumption 5: for each trial subject, there is positive probability to have some external controls sharing the same characteristics. Therefore, the selection of suitable external controls is important when considering our approach. One may consider the Pocock criteria (Pocock, 1976) which has been used to evaluate the comparability between external controls and current trials. Food and Drug Administration's guideline on 'Choice of control group and related issues in clinical trials' (FDA, 2021) and 'Considerations for the Design and Conduct of Externally Controlled Trials for Drug and Biological Products Guidance for Industry' (FDA, 2023) provide discussion on the choice and data quality of external controls. When using external controls from a separate trial, as in our real data example, it might be reasonable to accept the assumptions if the placebo effects are similar or have low impact in both control groups and there is no unobserved selection bias between the two trials.

The practical implications of conditional ignorability Assumption 4 are essential during the design phase of externally controlled trials. The incorporation of external controls face significant concerns regarding the potential for differences in patient attributes that could impact outcomes between the external control group and the trial treated arm. Examples of such baseline attributes include demographic and related factors like age, sex, race, socioeconomic status, and geographic region. Other attributes that may differ but are often more difficult to address encompass disease features, such as severity, duration, specific signs and symptoms, comorbidities, and previous and ongoing treatments. These confounding factors should be consistently measured and captured in both sources of data. This is the conditional ignorability Assumption 4, upon which the validity of our proposed weighting estimators can be established.

In practice, it is important to note that the recorded data may not always be comprehensive enough. For instance, it might be the case that SMA patients differ in terms of economic conditions and access

Table 2. The estimated average treatment effect of risdiplam in motor function measure change from baseline and corresponding 95% CIs

Method	$\widehat{\tau}$ at primary endpoint	95% CI	CI width
Within trial MMRM	1.672	(0.293, 3.051)	1.379
Power prior	1.650	(0.271, 3.034)	1.381
Commensurate prior	1.527	(0.198, 2.820)	1.311
EC-IPW-OPT	1.712	(0.529, 2.895)	1.183
EC-AIPW-OPT	1.522	(0.343, 2.702)	1.179

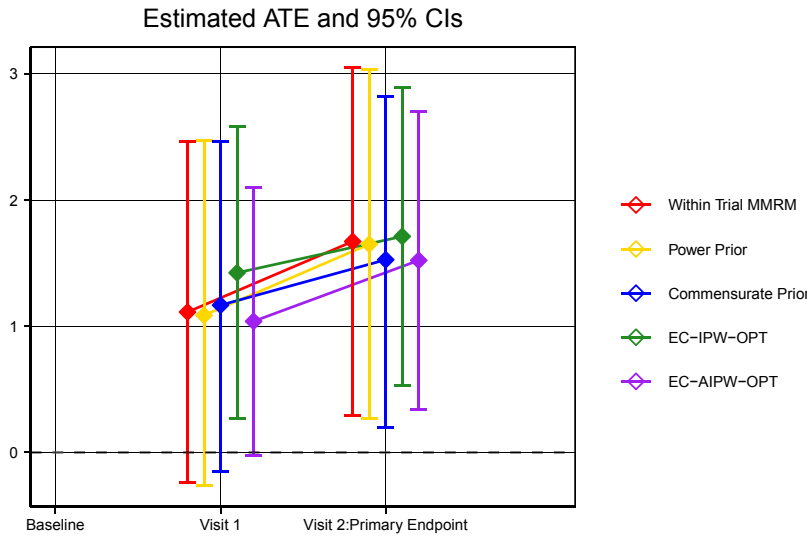

Figure 5. Estimated average treatment effect of risdiplam in motor function measure change from baseline and corresponding 95% CIs.

to quality care, which might prevent disadvantaged patients entering the SUNFISH trial, and those conditions tend to worsen the disease progression. That is, the trial population may be systematically different with the external controls in ways that are never known to investigators. Depending on the severity of unmeasured confounding, the ability to boost statistical power while maintain desirable Type I error might be limited. In our simulation Setting 5, we showed a case of weak unmeasured confounding, where 'weak' is suggested by the association in the real data. In future research, we could consider analytic methods that can accommodate unmeasured confounding bias.

The connection with platform trials is worth mentioning. Platform trials involve examining multiple experimental treatments that can join and leave the trial at different times, sharing a common control group. This sharing and borrowing of a control group to enhance statistical efficiency is also seen in hybrid controlled trials that use external controls, which is the focus of our work. This issue of borrowing from nonconcurrent controls is also relevant when using external controls (e.g. historical controls), which can introduce 'calendar time bias' as discussed in (Burger et al., 2021). If present, this time bias would violate the Conditional Ignorability Assumption 4.

A limitation of the proposed methods is their reliance on the validity of working models for nuisance parameters. Here, we assume those are parametric, which is a practical starting point, as parametric models are commonly used in clinical trials and practice, particularly when dealing with small sample sizes and low-dimensional covariates. On the other side, as shown in scenario E of the web-based supporting material, model misspecification can introduce bias when incorporating external controls, with the severity depending on the discrepancy between the working and true models. An alternative, more flexible approach might involve employing nonparametric methods for estimating nuisance parameters in DR estimators, such as EC-AIPW, as used in

this work. The *double/debiased machine learning* (Chernozhukov et al., 2018) and the *targeted maximum likelihood estimation* (Van Der Laan & Rubin, 2006) are two general approaches in this framework. These can be avenues for future research, particularly in terms of large-sample asymptotic results, and importantly, assessing how well theoretical results translate into practice, especially in small sample regime, within the context of hybrid controlled trials.

In this work, our focus is not on the Bayesian approach. The Bayesian dynamic borrowing methods (noncausal) integrate external controls via the use of informative priors. This is the major distinction with our proposed weighting-type estimators grounded in (frequentist) causal potential outcome framework. Nevertheless, it is worth mentioning the growing body of literature that explores the incorporation of propensity scores into Bayesian dynamic borrowing methods. The propensity score serves as a means to select external controls that are similar to trial subjects during the study's design phase. This can be achieved through various strategies, including propensity score stratification, regression adjustment for propensity scores, or matching based on propensity scores (Fu et al., 2023; Lewis et al., 2019; Li & Yue, 2023). This synthesis of Bayesian and propensity score methodologies offers a promising avenue. Future research can consider exploring Bayesian causal inference framework that fully takes into account the design and analysis stages.

The example of the SUNFISH trial incorporating external controls from the olesoxime trial is representative of the rare disease setting, where a randomized control group is less desirable or feasible due to limited alternative treatments and/or scarcity of patients. Spinal muscular atrophy is a rare neuromuscular disorder that results in the loss of motor neurones and progressive muscle wasting. A total of 180 patients with SMA enrolled in SUNFISH between 2017 and 2018 across 14 countries (Mercuri et al., 2022), which puts a limitation on the statistical power of the trial results. These participants were randomly assigned to receive either risdiplam or placebo in a 2:1 ratio to increase the likelihood of receiving risdiplam and to encourage the enrolment of patients who have limited alternative treatment options. Such design considerations are common in rare disease trials (Gross, 2021). The simulation studies, motivated by the real data example presented in Section 6, demonstrate the statistical efficiency gain of incorporating external controls using the proposed weighting estimators. Our example showcases that with appropriately chosen external controls and statistical methods, the challenges associated with relying on a randomized control arm can be alleviated. This is important given that the number of FDA drug approvals that used external control data (Jahanshahi et al., 2021) and rare disease trials (Gross, 2021) is increasing.

In order to generalize the proposed methods to accommodate broader studies using external controls, several directions can be undertaken. First, the weighting estimators offer a level of versatility that allows for potential extensions to accommodate other types of endpoints, such as time-to-event data. Second, sensitivity analysis could be developed to quantify how the estimates from the proposed methods vary as a function of the magnitude of unknown placebo effects (violating Assumption 1) and selection bias (violating Assumption 4). Third, extension of the proposed methods for examining heterogeneity treatment effect (HTE) is possible and would be useful especially for trials underpowered for HTE (Yang et al., 2023). Lastly, the weighting estimators can be adapted to include multiple external control samples originating from different sources. These potential extensions present promising avenues for future research in a broader range of contexts.

Acknowledgments

We would like to express our gratitude to Winnie Yeung and Tammy Mclver who provided valuable insights and expertize that greatly assisted the research.

Conflicts of interest: H.P. and J.Z. are employees of Genentech. They own Roche stocks. This work is methodological and not product focussed. H.P. and J.Z. are partially supported by the HHS U01 FDA grant for this work.

Funding

This project is partially supported by the Food and Drug Administration (FDA) of the US Department of Health and Human Services (HHS) as part of a research project cooperative agreement (U01 FD007206, H.P. and J.Z.). The contents are those of the author(s) and do not necessarily represent the official views of, nor an endorsement by, FDA/HHS or the US Government.

Data availability

The application example is based on data from the SUNFISH study. At the time of publication, the data for this paper are not yet publicly available because of privacy/ethical restrictions. Analysis and simulation codes are available by request.

Supplementary material

Supplementary material is available online at *Journal of the Royal Statistical Society: Series A*.

Author contributions

X.Z., J.Z., and H.P. designed the study, X.Z. conducted the simulations and analyses, J.Z., and H.P. provided supervision, All authors interpreted the results, X.Z., J.Z., and H.P. drafted and revised the manuscript, C.D. critically reviewed the manuscript. All authors approved of the final manuscript.

Appendix A

Proof of Theorem 1. (a) The ATE for the trial population with respect to base measure μ is defined as

$$\begin{aligned}\tau &= \int \mathbb{E}\left[Y^{(1)} - Y^{(0)} \mid X = x\right] p_{\mathcal{R}}(x)\mu(\mathrm{d}x) \\ &= \underbrace{\int \mathbb{E}\left[Y^{(1)} \mid X = x\right] p_{\mathcal{R}}(x)\mu(\mathrm{d}x)}_{I} \\ &\quad - \underbrace{\int \mathbb{E}\left[Y^{(0)} \mid X = x\right] p_{\mathcal{R}}(x)\mu(\mathrm{d}x)}_{II}\end{aligned}$$

Using the randomized trial data, both I and II can be easily identified through the (normalized) inverse propensity of treatment weighting:

$$I = \frac{\mathbb{E}_{\mathcal{R}}\left[\frac{YA}{\pi_A}\right]}{\mathbb{E}_{\mathcal{R}}\left[\frac{A}{\pi_A}\right]}$$

$$II = \frac{\mathbb{E}_{\mathcal{R}}\left[\frac{Y(1-A)}{1-\pi_A}\right]}{\mathbb{E}_{\mathcal{R}}\left[\frac{(1-A)}{1-\pi_A}\right]}$$

In addition, the external controls also provide another identification of II, provided that Assumptions 2, 3, 4, and 5 hold.

$$
\begin{aligned}
II &= \int \mathbb{E}\big[Y^{(0)} \mid X = x\big] p_{\mathcal{R}}(x)\mu(\mathrm{d}x) \\
&= \frac{\int \mathbb{E}\big[Y^{(0)} \mid X = x\big] \dfrac{p_{\mathcal{R}}(x)}{p_{\mathcal{E}}(x)} p_{\mathcal{E}}(x)\mu(\mathrm{d}x)}{\int \dfrac{p_{\mathcal{R}}(x)}{p_{\mathcal{E}}(x)} p_{\mathcal{E}}(x)\mu(\mathrm{d}x)} \\
&= \frac{\mathbb{E}_{\mathcal{E}}\big[Y^{(0)} \mid X = x\big] \dfrac{p_{\mathcal{R}}(x)}{p_{\mathcal{E}}(x)} p_{\mathcal{E}}(x)}{\mathbb{E}_{\mathcal{E}}\left[\dfrac{p_{\mathcal{R}}(x)}{p_{\mathcal{E}}(x)} p_{\mathcal{E}}(x)\right]} \\
&= \frac{\mathbb{E}_{\mathcal{E}}\big[Y^{(0)} W_{00}\big]}{\mathbb{E}_{\mathcal{E}}[W_{00}]} \\
&= \frac{\mathbb{E}_{\mathcal{E}}[Y W_{00}]}{\mathbb{E}_{\mathcal{E}}[W_{00}]}, \text{ by SUTVA}
\end{aligned}
$$

Therefore, $\tau^{EC-IPW} = \tau$. If the trial participation model $\pi_S(\mathbf{X}; \boldsymbol{\alpha})$ for $\pi_S(\mathbf{X})$ is correct, then $\tau^{EC-IPW} = \tau^{EC-IPW} = \tau$.

(b) If the trial participation model $\pi_S(\mathbf{X}; \boldsymbol{\alpha})$ for $\pi_S(\mathbf{X})$ is correct, then

$$
\begin{aligned}
\widetilde{\tau}^{EC-AIPW} &= \tau^{EC-IPW} \\
&\quad - \left[\frac{\mathbb{E}_{\mathcal{R}}\left[\dfrac{\mu(X;\beta)A}{\pi_A}\right]}{\mathbb{E}_{\mathcal{R}}\left[\dfrac{A}{\pi_A}\right]} - (1-w) \frac{\mathbb{E}_{\mathcal{R}}\left[\dfrac{\mu(X;\beta)(1-A)}{1-\pi_A}\right]}{\mathbb{E}_{\mathcal{R}}\left[\dfrac{(1-A)}{1-\pi_A}\right]} \right. \\
&\quad \left. - w \frac{\mathbb{E}_{\mathcal{E}}\big[\mu(X;\beta) W_{00}\big]}{\mathbb{E}_{\mathcal{E}}[W_{00}]} \right] \\
&= \tau^{EC-IPW} \\
&\quad - \underbrace{E_{\mathcal{E}}\big[\mu(X;\beta) - (1-w)\mu(X;\beta) + w\mu(X;\beta)\big]}_{=0} \\
&= \tau
\end{aligned}
$$

If the outcome model is $\mu(\mathbf{X};\boldsymbol{\beta})$ for $\mu(\mathbf{X})$ is correct,

$$
\begin{aligned}
\widetilde{\tau}^{EC-AIPW} &= \frac{\mathbb{E}_{\mathcal{R}}\left[\frac{\tilde{Y}A}{\pi_A}\right]}{\mathbb{E}_{\mathcal{R}}\left[\frac{A}{\pi_A}\right]} \\
&\quad - (1-w)\frac{\mathbb{E}_{\mathcal{R}}\left[\frac{\tilde{Y}(1-A)}{1-\pi_A}\right]}{\mathbb{E}_{\mathcal{R}}\left[\frac{(1-A)}{1-\pi_A}\right]} - w\frac{\mathbb{E}_{\mathcal{E}}\left[\tilde{Y}\frac{\pi_S(\mathbf{X};\alpha)(1-\pi_S)}{(1-\pi_S(\mathbf{X};\alpha)\pi_S}\right]}{\mathbb{E}_{\mathcal{E}}\left[\frac{\pi_S(\mathbf{X};\alpha)(1-\pi_S)}{(1-\pi_S(\mathbf{X};\alpha)\pi_S}\right]} \\
&= \frac{\mathbb{E}_{\mathcal{R}}\left[E[Y^{(1)} - \mu(X) \mid X=x]\right]}{\mathbb{E}_{\mathcal{R}}\left[\frac{A}{\pi_A}\right]} \\
&\quad - (1-w)\frac{\mathbb{E}_{\mathcal{R}}\left[\underbrace{\mathbb{E}[Y^{(0)} - \mu(X) \mid X=x]}_{=0}\right]}{\mathbb{E}_{\mathcal{R}}\left[\frac{(1-A)}{1-\pi_A}\right]} \\
&\quad - w\frac{\mathbb{E}_{\mathcal{E}}\left[\underbrace{\mathbb{E}[Y^{(0)} - \mu(X) \mid X=x]}_{=0}\frac{\pi_S(\mathbf{X};\alpha)(1-\pi_S)}{(1-\pi_S(\mathbf{X};\alpha)\pi_S}\right]}{\mathbb{E}_{\mathcal{E}}\left[\frac{\pi_S(\mathbf{X};\alpha)(1-\pi_S)}{(1-\pi_S(\mathbf{X};\alpha)\pi_S}\right]} \\
&= \frac{\mathbb{E}_{\mathcal{R}}\left[\mathbb{E}[Y^{(1)} - Y^{(0)} \mid X=x]\right]}{\mathbb{E}_{\mathcal{R}}\left[\frac{A}{\pi_A}\right]} \\
&= \mathbb{E}_{\mathcal{R}}\left[Y^{(1)} - Y^{(0)}\right], \text{ by Assumption 3} \\
&= \tau
\end{aligned}
$$

□

Appendix B

A simplifying approach is to consider baseline covariates, trial participation, treatment assignment, and propensity score as fixed at the design stage. Then

$$
\begin{aligned}
\mathrm{Var}\left[\widehat{\boldsymbol{\mu}}_{10}\right] &= \mathrm{Var}\left[\widehat{\boldsymbol{\mu}}_{10}|\mathbf{X}, \mathbf{S}, \mathbf{A}\right] \\
&= \frac{\sum_i S_i(1-A_i)\widehat{W}_{10i}^2}{\left(\sum_i S_i(1-A_i)\widehat{W}_{10i}\right)^2} \underbrace{\mathrm{Var}(Y|\mathbf{X}, S=1, A=0)}_{\text{Residual variance of trial control outcomes}} \\
\mathrm{Var}\left[\widehat{\boldsymbol{\mu}}_{00}\right] &= \mathrm{Var}\left[\widehat{\boldsymbol{\mu}}_{00}|\mathbf{X}, \mathbf{S}, \mathbf{A}\right] \\
&= \frac{\sum_i (1-S_i)\widehat{W}_{00i}^2}{\left(\sum_i (1-S_i)\widehat{W}_{00i}\right)^2} \underbrace{\mathrm{Var}(Y|\mathbf{X}, S=0)}_{\text{Residual variance of external control outcomes}}
\end{aligned}
$$

Since there is no straightforward way to estimate the bias term without accessing outcome data, if we are willing to accept that no bias arises due to the violation of causal assumptions outlined in

Section 2.3 and the validity of nuisance parameter models, i.e. assuming that the bias is negligible, we can then set $B \approx 0$. Additionally, if we assume that the residual variances of the trial control outcome and the external control outcome are the same, these two unknown terms can be cancelled out. Consequently, equation (10) can be approximated by the 'variance ratio' in equation (11) which is estimable without any outcome data.

Appendix C

We used the *GLM* for longitudinal data as the working model $Y_i = X_i\beta + Z_i\tau + \epsilon_i$ for subjects $i \in \mathcal{R}$, where the within-subject correlation over repeated measures is accounted by the variance–covariance matrix of ϵ_is by assuming $\epsilon_i \sim MVN(0, \Sigma)$. We assume unstructured covariance matrix Σ that allows the correlation to be different for each pair of time points. Here, $X = (1, \text{SMN2 Copy Number}, \text{SMA Type}, \text{Scoliosis}, \text{Time})'$, $Z = (A, \text{Time} \times A)$ where both treatment indicator A and visit indicator Time are binary variables (since we only consider $T = 2$). The coefficients $\tau = (\tau_1, \tau_2)$ represent the treatment effects for postintervention outcomes. Similarly, the external controls subjects $i \in \mathcal{E}$ follow the same outcome model $Y_i = X_i\beta + \epsilon_{0,i}$ with $\epsilon_i \sim MVN(0, \Sigma_{\mathcal{E}})$.

C.1 Power prior specification for simulation

We assume prior:

$$\begin{aligned}
\pi(\beta \mid S = 0, a_0) &\propto L(\beta \mid S = 0)^{a_0}\pi_0(\beta)\pi_0(a_0)\\
\pi(\Sigma) &\sim \text{Inverse-Wishart}\ (\nu, \mathbb{I}^{-1})\\
\pi(\Sigma_{\mathcal{E}}) &\sim \text{Inverse-Wishart}\ (\nu_{\mathcal{E}}, \mathbb{I}^{-1})\\
\pi_0(\tau) &\sim Gamma(1, 1)
\end{aligned}$$

where $L(\beta \mid S = 0) = \prod_{i\in\mathcal{E}} (2\pi)^{-1/2}|\Sigma_{\mathcal{E}}|^{-1/2}e^{-\frac{1}{2}(y_i - \mathbf{x}_i\beta)^T\Sigma_{\mathcal{E}}^{-1}(y_{0i} - \mathbf{x}_{0i}\beta)}$; $\nu = \nu_0 = p + 2$ and $\widehat{\Sigma}$ and $\widehat{\Sigma}_{\mathcal{E}}$ as identity matrix to represent weak prior belief.

The data for RCT subjects $i \in \mathcal{R}$ is distributed as

$$y_i \mid \beta, \tau, \Sigma_{\mathcal{R}} \sim \mathcal{N}(\mathbf{x}_i\beta + Z_i\tau, \Sigma_{\mathcal{R}})$$

The conditional posteriors can be derived and sampled due to conjugate prior specification

$$\begin{aligned}
&\beta \mid \beta_{trt}, \Sigma, \Sigma_{\mathcal{E}} \sim \mathcal{N}(A^{-1}b, A^{-1})\\
&A = \left(\sum_{i\in\mathcal{R}} \mathbf{x}_i\Sigma^{-1}\mathbf{x}_i^T + a_0\sum_{i\in\mathcal{E}} \mathbf{x}_i\Sigma_{\mathcal{E}}^{-1}\mathbf{x}_i^T\right)\\
&b = \left(\sum_{i\in\mathcal{R}} \mathbf{x}_i\Sigma^{-1}(y_i - S_i\beta_{trt}) + a_0\sum_{i\in\mathcal{E}} \mathbf{x}_i\Sigma_{\mathcal{E}}^{-1}y_i\right)\\
&\beta_{trt} \mid \beta, \Sigma, \Sigma_{\mathcal{E}} \sim \mathcal{N}\left(\left(\sum_{i\in\mathcal{R}} \mathbf{x}_i^T\Sigma^{-1}\mathbf{x}_i\right)^{-1}\left(\sum_{i\in\mathcal{E}} \mathbf{x}_i^T\Sigma^{-1}(y_i - \mathbf{x}_i\beta)\right),\right.\\
&\quad\left.\left(\sum_{i\in\mathcal{R}} \mathbf{x}_i^T\Sigma^{-1}\mathbf{x}_i\right)^{-1}\right)\\
&\Sigma \mid \beta, \beta_{trt}, \Sigma_{\mathcal{E}} \sim \text{Inverse-Wishart}\ (\nu + n, (\widehat{\Sigma} + \mathbf{S_1})^{-1})\\
&\mathbf{S_1} = \sum_{i\in\mathcal{E}} (y_i - \mathbf{x}_i\beta - Z_i\beta_{trt})(y_i - \mathbf{x}_i\beta - S_i\beta_{trt})^T\\
&\Sigma_{\mathcal{E}} \mid \beta, \beta_{trt}, \Sigma \sim \text{Inverse-Wishart}\ (\nu_0 + m, (\widehat{\Sigma} + \mathbf{S_0})^{-1})\\
&\mathbf{S_0} = \sum_{i\in\mathcal{E}} (y_i - \mathbf{x}_i\beta)(y_i - \mathbf{x}_i\beta)^T
\end{aligned}$$

C.2 Commensurate prior specification for simulation

Here, we adopt the approach proposed in Hobbs et al. (2012). We assume priors that incorporates likelihood for β_0 from external controls, a prior for β that centred around β_0 with variance equals to the inverse of commensurability parameter, the higher the commensurability, more borrowing from external controls.

$$\pi(\beta \mid S=0, a_0) \propto L(\beta \mid S=0) \times N\left(\beta \mid \beta_0, \frac{1}{\tau}\right) \times \pi_0(\tau)$$
$$\pi(\Sigma) \sim \text{Inverse-Wishart } (v, \widehat{\Sigma}^{-1})$$
$$\pi(\Sigma_{\mathcal{E}}) \sim \text{Inverse-Wishart } (v_{\mathcal{E}}, \widehat{\Sigma}_{\mathcal{E}}^{-1})$$

where prior for commensurability is assumed $\pi_0(\tau) \sim Gamma(1, 1)$.

The data for RCT subjects $i \in \mathcal{R}$

$$y_i \mid \beta, \Sigma \sim MVN(X_i\beta + A_i\beta_{trt}, \Sigma)$$

The conditional posteriors can be derived and sampled due to conjugate prior specification

$$\beta \mid \beta_{trt}, \beta_0, \Sigma, \Sigma_{\mathcal{E}} \sim \mathcal{N}(\mathbf{A}^{-1}b, \mathbf{A}^{-1})$$
$$\mathbf{A} = \left(\sum_{i\in\mathcal{R}} \mathbf{x}_i\Sigma^{-1}\mathbf{x}_i^T + diag(\tau)^{-1}\right)$$
$$b = \left(\sum_{i\in\mathcal{R}} \mathbf{x}_i\Sigma^{-1}(y_i - S_i\beta_{trt}) + diag(\tau)^{-1}\beta_0\right)$$
$$\beta_0 \mid \beta, \beta_{trt}, \Sigma, \Sigma_{\mathcal{E}} \sim \mathcal{N}(\mathbf{A}_0^{-1}b_0, \mathbf{A}_0^{-1})$$
$$\mathbf{A}_0 = \left(\sum_{i\in\mathcal{E}} \mathbf{x}_i\Sigma^{-1}\mathbf{x}_i^T + diag(\tau)^{-1}\right)$$
$$b_0 = \left(\sum_{i\in\mathcal{E}} \mathbf{x}_i\Sigma^{-1}(y_i - S_i\beta_{trt}) + diag(\tau)^{-1}\beta\right)$$
$$\tau \sim Gamma\left(1 + \frac{1}{2}, 1 + \frac{1}{2}\|\beta - \beta_0\|^2\right)$$
$$\beta_{trt} \mid \beta, \Sigma, \Sigma_{\mathcal{E}} \sim \mathcal{N}\left(\left(\sum_{i\in\mathcal{R}} \mathbf{x}_i^T\Sigma^{-1}\mathbf{x}_i\right)^{-1}\left(\sum_{i\in\mathcal{E}} \mathbf{x}_i^T\Sigma^{-1}(y_i - \mathbf{x}_i\beta)\right),\right.$$
$$\left.\left(\sum_{i\in\mathcal{R}} \mathbf{x}_i^T\Sigma^{-1}\mathbf{x}_i\right)^{-1}\right)$$
$$\Sigma \mid \beta, \beta_{trt}, \Sigma_{\mathcal{E}} \sim \text{Inverse-Wishart } (v + n, (\widehat{\Sigma} + \mathbf{S_1})^{-1})$$
$$\mathbf{S_1} = \sum_{i\in\mathcal{E}} (y_i - \mathbf{x}_i\beta - S_i\beta_{trt})(y_i - \mathbf{x}_i\beta - S_i\beta_{trt})^T$$
$$\Sigma_{\mathcal{E}} \mid \beta, \beta_{trt}, \Sigma \sim \text{Inverse-Wishart } (v_0 + m, (\widehat{\Sigma} + \mathbf{S_0})^{-1})$$
$$\mathbf{S_0} = \sum_{i\in\mathcal{E}} (y_i - \mathbf{x}_i\beta)(y_i - \mathbf{x}_i\beta)^T$$

References

Bareinboim E., & Pearl J. (2016). Causal inference and the data-fusion problem. *Proceedings of the National Academy of Sciences of the United States of America*, *113*(27), 7345–7352. https://doi.org/10.1073/pnas.1510507113

Berry S. M., Carlin B. P., Lee J. J., & Muller P. (2010). *Bayesian adaptive methods for clinical trials*. CRC Press.

Buchanan A. L., Hudgens M. G., Cole S. R., Mollan K. R., Sax P. E., Daar E. S., Adimora A. A., Eron J. J., & Mugavero M. J. (2018). Generalizing evidence from randomized trials using inverse probability of sampling weights. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *181*(4), 1193–1209. https://doi.org/10.1111/rssa.12357

Burger H. U., Gerlinger C., Harbron C., Koch A., Posch M., Rochon J., & Schiel A. (2021). The use of external controls: To what extent can it currently be recommended? *Pharmaceutical Statistics*, *20*(6), 1002–1016. https://doi.org/10.1002/pst.v20.6

Campbell D. T. (1957). Factors relevant to the validity of experiments in social settings. *Psychological Bulletin*, *54*(4), 297–312. https://doi.org/10.1037/h0040950

Chen J., Ho M., Lee K., Song Y., Fang Y., Goldstein B. A., He W., Irony T., Jiang Q., & van der Laan M. (2021). The current landscape in biostatistics of real-world data and evidence: Clinical study design and analysis. *Statistics in Biopharmaceutical Research*, *15*(1), 29–42. https://doi.org/10.1080/19466315.2021.1883474

Chernozhukov V., Chetverikov D., Demirer M., Duflo E., Hansen C., Newey W., & Robins J. (2018). Double/debiased machine learning for treatment and structural parameters. *Econometrics Journal*, *21*, C1–C68. https://doi.org/10.1111/ectj.12097

CID (2023). *Complex innovative trial design meeting program*. https://www.fda.gov/drugs/development-resources/complex-innovative-trial-design-meeting-program. Date accessed November 30, 2023.

Cole S. R., & Stuart E. A. (2010). Generalizing evidence from randomized clinical trials to target populations: The ACTG 320 trial. *American Journal of Epidemiology*, *172*(1), 107–115. https://doi.org/10.1093/aje/kwq084

Dahabreh I. J., Robertson S. E., Tchetgen E. J., Stuart E. A., & Hernán M. A. (2019). Generalizing causal inferences from individuals in randomized trials to all trial-eligible individuals. *Biometrics*, *75*(2), 685–694. https://doi.org/10.1111/biom.13009

FDA (2021). *E 10 choice of control group and related issues in clinical trials*. https://www.fda.gov/media/71349/download. Date accessed May 23, 2024.

FDA (2023). *Considerations for the design and conduct of externally controlled trials for drug and biological products guidance for industry*. https://www.fda.gov/media/164960/download. Date accessed November 30, 2023.

Fu C., Pang H., Zhou S., & Zhu J. (2023). Covariate handling approaches in combination with dynamic borrowing for hybrid control studies. *Pharmaceutical Statistics*, *22*(4), 619–632. https://doi.org/10.1002/pst.v22.4

Gross A. M. (2021). Using real world data to support regulatory approval of drugs in rare diseases: A review of opportunities, limitations & a case example. *Current Problems in Cancer*, *45*(4), 100769. https://doi.org/10.1016/j.currproblcancer.2021.100769

Ho M., van der Laan M., Lee H., Chen J., Lee K., Fang Y., He W., Irony T., Jiang Q., & Lin X. (2023). The current landscape in biostatistics of real-world data and evidence: causal inference frameworks for study design and analysis. *Statistics in Biopharmaceutical Research*, *15*(1), 43–56.

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*(3), 1047–1056. https://doi.org/10.1111/biom.2011.67.issue-3

Hobbs B. P., Sargent D. J., & Carlin B. P. (2012). Commensurate priors for incorporating historical information in clinical trials using general and generalized linear models. *Bayesian Analysis*, *7*(3), 639–674. https://doi.org/10.1214/12-BA722

Horvitz D. G., & Thompson D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, *47*(260), 663–685. https://doi.org/10.1080/01621459.1952.10483446

Ibrahim J. G., & Chen M. -H. (2000). Power prior distributions for regression models. *Statistical Science*, *15*(1), 46–60. https://doi.org/10.1214/ss/1009212673

Ibrahim J. G., Chen M. -H., Gwon Y., & Chen F. (2015). The power prior: Theory and applications. *Statistics in Medicine*, *34*(28), 3724–3749. https://doi.org/10.1002/sim.v34.28

Imbens G. W., & Rubin D. B. (2015). *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press.

Jahanshahi M., Gregg K., Davis G., Ndu A., Miller V., Vockley J., Ollivier C., Franolic T., & Sakai S. (2021). The use of external controls in FDA regulatory decision making. *Therapeutic Innovation & Regulatory Science*, *55*(5), 1019–1035. https://doi.org/10.1007/s43441-021-00302-y

Lewis C. J., Sarkar S., Zhu J., & Carlin B. P. (2019). Borrowing from historical control data in cancer drug development: A cautionary tale and practical guidelines. *Statistics in Biopharmaceutical Research*, *11*(1), 67–78. https://doi.org/10.1080/19466315.2018.1497533

Li F., Morgan K. L., & Zaslavsky A. M. (2018). Balancing covariates via propensity score weighting. *Journal of the American Statistical Association*, *113*(521), 390–400. https://doi.org/10.1080/01621459.2016.1260466

Li H., & Yue L. Q. (2023). Propensity score-based methods for causal inference and external data leveraging in regulatory settings: From basic ideas to implementation. *Pharmaceutical Statistics*, *22*(4), 721–738. https://doi.org/10.1002/pst.v22.4

McIver T., El-Khairi M., Yeung W. Y., & Pang H. (2023). The use of real-world data to support the assessment of the benefit and risk of a medicine to treat spinal muscular atrophy. In W. He, Y. Fang, H. Wang (Eds.), *Real-world evidence in medical product development* (pp. 287–411). Springer.

Mercuri E., Deconinck N., Mazzone E. S., Nascimento A., Oskoui M., Saito K., Vuillerot C., Baranello G., Boespflug-Tanguy O., Goemans N., Kirschner J., Kostera-Pruszczyk A., Servais L., Gerber M., Gorni K., Khwaja O., Kletzl H., Scalco R. S., Staunton H., …SUNFISH Study Group (2022). Safety and efficacy of once-daily risdiplam in type 2 and non-ambulant type 3 spinal muscular atrophy (SUNFISH part 2): A phase 3, double-blind, randomised, placebo-controlled trial. *The Lancet Neurology*, *21*(1), 42–52. https://doi.org/10.1016/S1474-4422(21)00367-7

Pearl J., & Bareinboim E. (2011). Transportability of causal and statistical relations: A formal approach. *Proceedings of the AAAI Conference on Artificial Intelligence*, *25*(1), 247–254. https://doi.org/10.1609/aaai.v25i1.7861

Pocock S. J. (1976). The combination of randomized and historical controls in clinical trials. *Journal of Chronic Diseases*, *29*(3), 175–188. https://doi.org/10.1016/0021-9681(76)90044-8

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

Rubin D. B. (2008). For objective causal inference, design trumps analysis. *Annals of Applied Statistics*, *2*(3), 808–840. https://doi.org/10.1214/08-AOAS187

Schmidli H., Gsteiger S., Roychoudhury S., O'Hagan A., Spiegelhalter D., & Neuenschwander B. (2014). Robust meta-analytic-predictive priors in clinical trials with historical control information. *Biometrics*, *70*(4), 1023–1032. https://doi.org/10.1111/biom.v70.4

Stuart E. A., Ackerman B., & Westreich D. (2018). Generalizability of randomized trial results to target populations: Design and analysis possibilities. *Research on Social Work Practice*, *28*(5), 532–537. https://doi.org/10.1177/1049731517720730

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*(2), 369–386. https://doi.org/10.1111/j.1467-985X.2010.00673.x

Tsiatis A. A. (2006). *Semiparametric theory and missing data*. Springer.

Van Der Laan M. J., & Rubin D. (2006). Targeted maximum likelihood learning. *The International Journal of Biostatistics*, *2*(1), Article 11. https://doi.org/10.2202/1557-4679.1043

Viele K., Berry S., Neuenschwander B., Amzal B., Chen F., Enas N., Hobbs B., Ibrahim J. G., Kinnersley N., Lindborg S., Micallef S., Roychoudhury S., & Thompson L. (2014). Use of historical control data for assessing treatment effects in clinical trials. *Pharmaceutical Statistics*, *13*(1), 41–54. https://doi.org/10.1002/pst.v13.1

Westreich D., Edwards J. K., Lesko C. R., Stuart E., & Cole S. R. (2017). Transportability of trial results using inverse odds of sampling weights. *American Journal of Epidemiology*, *186*(8), 1010–1014. https://doi.org/10.1093/aje/kwx164

Yang S., Gao C., Zeng D., & Wang X. (2023). Elastic integrative analysis of randomised trial and real-world data for treatment heterogeneity estimation. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *85*(3), 575–596. https://doi.org/10.1093/jrsssb/qkad017

Yap T. A., Jacobs I., Andre E. B., Lee L. J., Beaupre D., & Azoulay L. (2021). Application of real-world data to external control groups in oncology clinical trial drug development. *Frontiers in Oncology*, *11*, 695936. https://doi.org/10.3389/fonc.2021.695936