ROYAL
STATISTICAL
SOCIETY
DATA | EVIDENCE | DECISIONS

Journal of the Statistics Society
Series B
Statistical Methodology

B

Original Article

Harmonized estimation of subgroup-specific treatment effects in randomized trials: the use of external control data

Daniel Schwartz[1,2] ⓘ, **Riddhiman Saha**[1], **Steffen Ventz**[3] ⓘ
and Lorenzo Trippa[1,2]

[1]Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA, USA
[2]Department of Data Science, Dana-Farber Cancer Institute, Boston, MA, USA
[3]Division of Biostatistics, School of Public Health, University of Minnesota, Minneapolis, MN, USA

Address for correspondence: Daniel Schwartz, Department of Biostatistics, Harvard T.H. Chan School of Public Health,
677 Huntington Ave, Boston, MA, USA. Email: daniels@ds.dfci.harvard.edu

Abstract

Subgroup analyses of randomized controlled trials (RCTs) constitute an important component of the drug development process in precision medicine. In particular, subgroup analyses of early-stage trials often influence the design and eligibility criteria of subsequent confirmatory trials and ultimately influence which subpopulations will receive the treatment after regulatory approval. However, subgroup analyses are often complicated by small sample sizes, which leads to substantial uncertainty about subgroup-specific treatment effects. We explore the use of external control (EC) data to augment RCT subgroup analyses. We define and discuss *harmonized estimators* of subpopulation-specific treatment effects that leverage EC data. Our approach can be used to modify any subgroup-specific treatment effect estimates that are obtained by combining RCT and EC data, such as linear regression. We alter these subgroup-specific estimates to make them coherent with a robust estimate of the average effect in the randomized population based only on RCT data. The weighted average of the resulting subgroup-specific harmonized estimates matches the RCT-only estimate of the overall effect in the randomized population. We discuss the proposed harmonized estimators through analytic results and simulations, and investigate standard performance metrics. The method is illustrated with a case study in oncology.

Keywords: historical control, model misspecification, real-world data, shrinkage estimation, subgroup analysis

1 Introduction

Evaluating the efficacy of experimental therapies in subpopulations defined by biomarkers that modulate the therapeutic mechanisms has become a major goal of precision medicine. However, the sample size of randomized controlled trials (RCTs) is typically chosen with the goal of estimating the therapy's average effect in the randomized population. Consequently, RCT sample sizes are often insufficient to accurately evaluate treatment effects in the subpopulations of interest. Yet subgroup analyses are often reported despite limited sample sizes. In early-phase trials these analyses are important for major decisions, such as determining the design and eligibility criteria for subsequent confirmatory trials (Pletcher & McCulloch, 2017; Sun et al., 2012). In late phase trials, if the therapy receives regulatory approval, subgroup analyses may influence how clinicians choose individualized treatment plans for their patients (Dijkman et al., 2009). Unfortunately, limited resources, time, and patient accrual rates often make it

infeasible to conduct large trials that can accurately estimate treatment effects within each subpopulation (Sully et al., 2013).

To address this limitation, we investigate potential improvements of RCT subgroup analyses by including external control (EC) data from patients that were treated with the RCT's control therapy. We consider EC data from previous trials or electronic health records. At completion of the RCT we augment the subgroup analyses with patient-level EC data, including relevant pretreatment clinical and demographic variables, and outcomes. Here, we introduce a method that balances the potential efficiency gains from the integration of external data and the risk of bias from unmeasured confounding or other distortion mechanisms.

Our work is motivated by two oncology trials in brain and breast cancer that differ substantially in terms of sample sizes and prevalence of the subgroups of interest. In both cases, subgroup analyses using only internal RCT data result in high uncertainty on the subgroup-specific treatment effects. This level of uncertainty affects several decisions, including the choice of designs and eligibility criteria (i.e. selected subgroups) of future confirmatory trials.

(a) *Temozolomide for glioblastoma (GBM).* A large RCT in GBM, a brain cancer with poor prognosis, showed that *temozolomide* used in combination with radiotherapy (TMZ + RT) is an effective first line treatment compared to radiotherapy alone (Stupp et al., 2005). As of today TMZ + RT remains the standard of care for GBM. After the RCT results were reported, the study investigators collected MGMT methylation data, an important biomarker in neuro-oncology. Subgroup analyses suggested that the treatment effect of TMZ + RT compared to RT alone may vary by methylation status, with methylation-positive patients benefiting from TMZ + RT and high uncertainty about whether or not methylation-negative patients benefit from the combination therapy (Stupp et al., 2009). The sample size of the biomarker negative subgroup was not sufficient to provide definitive evidence about treatment efficacy in this subpopulation.

(b) *Olapirib for metastatic breast cancer.* Olapirib is a promising drug approved to treat metastatic breast cancer. A recent RCT tested its efficacy in two populations defined by sets of somatic (population 1) and germline (population 2) mutations. We refer to Tung et al. (2020) for details. The trial was powered to test efficacy in these two populations. The analyses indicated that the benefit of olapirib might be limited to some of the patients in populations 1 and 2 with specific mutations. However, the study had limited power to test efficacy in narrow subgroups defined by the individual mutations without grouping them into the broad categories of somatic and germline alterations.

Our analyses (Section 4) are based on a recent data collection in GBM (Rahman et al., 2023). The aim is to illustrate the operating characteristics of subgroup-specific treatment effect estimators with comparisons tailored to a specific clinical setting.

Most methods for subgroup analyses use only internal data from the RCT. The simplest and perhaps most common approach in clinical research is to repeat the primary analysis, for instance a two-sample Z-test, in each of the subgroups of interest (Wang et al., 2007). This strategy is often underpowered for testing whether or not the experimental treatment is effective in the subgroups of interest (Pletcher & McCulloch, 2017). Regression methods that model the outcome conditional on biomarkers and other covariates have been proposed to reduce the uncertainty of subgroup-specific results (Morita & Müller, 2017; Sivaganesan et al., 2017).

The use of EC data for inference at completion of RCTs has been previously discussed (Corrigan-Curay et al., 2018; Di Maio et al., 2020). However, these approaches leverage RCT and EC data primarily to estimate average treatment effects in the *randomized* population, not in subpopulations. Bayesian methods in this area include meta-analytic models (Schmidli et al., 2014), power priors (Ibrahim et al., 2015), and commensurate priors (Hobbs et al., 2012). Other approaches include the use of propensity scores (Li et al., 2022; Lim et al., 2018) and adaptive procedures that compare the RCT control data with the EC data to decide whether to pool them in the analyses or use only the RCT data for inference (Viele et al., 2014). EC data have also been used for sample size re-estimation (Hobbs et al., 2013; Schmidli et al., 2014) and early stopping for futility (Ventz et al., 2022).

We focus on patient subgroups and potential treatment effects' variations across subgroups. We investigate if incorporating EC data can produce a substantial reduction of the uncertainty on the subgroup-specific treatment effects. At the same time, we account for the fact that most RCTs are required by their protocol to report an estimate $\hat{\theta}^{(r)}$ of the average treatment effect that is based only on trial data and *does not* incorporate EC data. This estimate and any subgroup-specific estimate $\hat{\theta}_{1:K}^{(r+e)}$ based on both RCT and EC data could be discordant, for example $\hat{\theta}^{(r)}$ might be negative and the subgroup-specific estimates could be all positive. This discordance can arise because $\hat{\theta}^{(r)}$ and $\hat{\theta}_{1:K}^{(r+e)}$ are based on different sets of data. We introduce harmonization to reduce or remove this misalignment.

Our method requires two inputs: (i) an estimate of the average treatment effect in the randomized population $\hat{\theta}^{(r)}$, which is based only on the RCT data, and (ii) an estimate of subgroup-specific effects $\hat{\theta}_{1:K}^{(r+e)}$ in K nonoverlapping subgroups, which is based on both RCT and EC data, with some risk of bias. We use the superscript $(r + e)$ for estimators that use both RCT and EC data, and (r) for estimators that use only RCT data. The method is based on a simple linear transformation (Equation (5)) to modify the subgroup-specific estimates $\hat{\theta}_{1:K}^{(r+e)}$ and harmonize them (i.e. make them coherent) with respect to the average treatment effect estimate $\hat{\theta}^{(r)}$. The term *harmonization* refers to an intuitive relation between the average treatment effect $E(Y \mid T = 1) - E(Y \mid T = 0)$ in the trial population, and the treatment effects $E(Y \mid T = 1, W = k) - E(Y \mid T = 0, W = k)$ in the subgroups $k = 1, \ldots, K$ that partition the randomized population:

$$E(Y \mid T = 1) - E(Y \mid T = 0) = \sum_{k=1}^{K} \pi_k [E(Y \mid T = 1, W = k) - E(Y \mid T = 0, W = k)].$$

Here, W indicates the subpopulation and $\pi_1, \ldots, \pi_K$ are the subpopulation prevalences in the trial population. In the RCT, patients are randomized and the variables T and W are independent. We say that a vector of subgroup-specific estimates $\hat{\theta}_{1:K}$ is fully harmonized to the primary analysis if $\sum_{k=1}^{K} \pi_k \hat{\theta}_k = \hat{\theta}^{(r)}$. When a subgroup analysis that leverages ECs is not harmonized and there is a marked discrepancy between $\sum_{k=1}^{K} \pi_k \hat{\theta}_k^{(r+e)}$ and $\hat{\theta}^{(r)}$ it can be challenging to interpret the trial results.

Our method modifies initial subgroup-specific estimates $\hat{\theta}_{1:K}^{(r+e)}$ and harmonizes them to $\hat{\theta}^{(r)}$; therefore, the subgroup analyses become coherent with a primary analysis that reports $\hat{\theta}^{(r)}$ for the randomized population without using EC data. We illustrate that the subgroup-specific harmonized estimates present an attractive balance between efficiency and bias in settings where the investigator expects identical or similar differences $\gamma_k \approx \gamma$ between the outcomes in the EC population and the outcomes in the RCT control arm across all subgroups $k = 1, \ldots, K$. Formal definitions of these discrepancies γ_k are provided in the next sections.

2 Harmonized estimates of subgroup-specific treatment effects

We consider a two-arm RCT. The analysis of the trial can leverage EC data of patients receiving the same therapy as in the RCT control group. Before treatment patients are categorized into non-overlapping biomarker subgroups $k = 1, \ldots, K$. When necessary we use the superscript (r) to indicate the RCT samples, (e) to denote the ECs, and $(r + e)$ for their union. For example the sample sizes are $n^{(r)}$, $n^{(e)}$ and $n^{(r+e)}$. Also, the subgroup- and treatment-specific sample sizes are $n_{k,t}^{(s)}$, where $t = 0$ for the control, $t = 1$ for the experimental treatment, and $s \in \{r, e, r + e\}$. Similarly, $n_{k,\cdot}^{(r)}$ is the total number of RCT patients in subgroup k and $n_{\cdot,0}^{(r+e)}$ is the number of control patients in the RCT and EC datasets. The notation is summarized in Table 1.

For the ith patient in the RCT, we have the vector $(T_i^{(r)}, W_i^{(r)}, Y_i^{(r)}, X_i^{(r)})$, where $T_i^{(r)} \in \{0, 1\}$ is the treatment indicator (0 for control, 1 for treatment), $W_i^{(r)} \in \{1, \ldots, K\}$ is the biomarker subgroup, $Y_i^{(r)}$ is the outcome, and $X_i^{(r)}$ is a vector of pretreatment covariates. Similarly, for the ith patient in the EC data, we have $(T_i^{(e)}, W_i^{(e)}, Y_i^{(e)}, X_i^{(e)})$, with $T_i^{(e)} = 0$ because all patients in the external

Table 1. Notation

Variable	Notation	Definition
Outcome	Y_i	
Treatment	T_i	0 for control arm, 1 for treatment arm
Study		(r) for RCT, (e) for EC
Subgroup	W_i	$W_i = k$ if patient i is in subgroup k
Covariates	X_i	Vector of pretreatment covariates
Datasets	$D^{(r)}, D^{(e)}$	(T_i, W_i, Y_i, X_i) for all RCT and EC patients, respectively
Group-specific sample sizes	$n^{(s)}_{k,t}$	In subgroup k, arm t, and study s
Subgroup prevalences in the RCT population	$\pi = (\pi_1, \ldots, \pi_K)$	$\pi_k = p(W_i^{(r)} = k)$
Expected control outcomes in the RCT subgroups	$\mu_{1:K} = (\mu_1, \ldots, \mu_K)$	$\mu_k = E(Y_i^{(r)} \mid W_i^{(r)} = k, T_i^{(r)} = 0)$
Expected control outcomes in the RCT population	μ	$E(Y_i^{(r)} \mid T_i^{(r)} = 0) = \sum_{k=1}^{K} \pi_k \mu_k$
Treatment effects in the RCT subpopulations	$\theta_{1:K} = (\theta_1, \ldots, \theta_K)$	$\theta_k = E(Y_i^{(r)} \mid W_i^{(r)} = k, T_i^{(r)} = 1) - \mu_k$
Treatment effect in the RCT population	θ	$E(Y_i^{(r)} \mid T_i^{(r)} = 1) - \mu = \sum_{k=1}^{K} \pi_k \theta_k$
Initial estimator of $\theta_{1:K}$	$\hat{\theta}_{1:K}^{(r+e)}$	A function of the RCT data $D^{(r)}$ and the EC data $D^{(e)}$
Estimator of θ	$\hat{\theta}^{(r)}$	A function of the RCT data $D^{(r)}$
Harmonized estimator of $\theta_{1:K}$	$\hat{\theta}_{1:K}^{b}$	Defined by Equation (1)

dataset received the control therapy. The RCT and EC datasets are indicated as $D^{(r)}$ and $D^{(e)}$, respectively.

In the RCT population, within subgroup k, the expected outcome under the control therapy is $\mu_k = E(Y_i^{(r)} \mid W_i^{(r)} = k, T_i^{(r)} = 0)$, and the treatment effect is

$$\theta_k = E\left(Y_i^{(r)} \mid W_i^{(r)} = k, T_i^{(r)} = 1\right) - E\left(Y_i^{(r)} \mid W_i^{(r)} = k, T_i^{(r)} = 0\right).$$

The expected outcome and treatment effect in the RCT population are $\mu = E(Y_i^{(r)} \mid T_i^{(r)} = 0)$ and

$$\theta = E\left(Y_i^{(r)} \mid T_i^{(r)} = 1\right) - E\left(Y_i^{(r)} \mid T_i^{(r)} = 0\right) = \sum_{k=1}^{K} \pi_k \theta_k,$$

where $\pi_k = p(W_i^{(r)} = k)$ is the prevalence of the subgroup k in the trial population.

2.1 Harmonized estimates

The subgroup-specific treatment effect estimates that we propose are harmonized to the primary analysis, which utilizes only RCT data. The definition of the harmonized estimator proceeds in two steps. In the first step we compute an initial estimate of the subgroup-specific treatment effects $\hat{\theta}_{1:K}^{(r+e)}$, using both the RCT data $D^{(r)}$ and the EC data $D^{(e)}$. The analyst chooses the estimator $\hat{\theta}_{1:K}^{(r+e)}$. For example, $\hat{\theta}_{1:K}^{(r+e)}$ can be the result of a regression analysis that merges $D^{(r)}$ and $D^{(e)}$ into a single dataset under the assumption that the RCT and EC populations are identical, or the primary findings of a Bayesian analysis that models the RCT and EC populations. Throughout the article, we

consider several definitions of $\hat{\theta}_{1:K}^{(r+e)}$. In the second step we modify $\hat{\theta}_{1:K}^{(r+e)}$ and harmonize it with respect to an overall treatment effect estimate $\hat{\theta}^{(r)}$ based only on $D^{(r)}$. Here $\hat{\theta}^{(r)}$ is an estimate associated with low risks of bias, confounding, and various distortion mechanisms, and is based only on RCT data. We say that a vector of estimates $\hat{\theta}_{1:K}$ is fully harmonized if $\sum_{k=1}^{K} \pi_k \hat{\theta}_k = \hat{\theta}^{(r)}$, i.e. if its weighted average equals the estimate of the average treatment effect. Recall that $\pi = (\pi_1, \ldots, \pi_K)$ are the subgroup prevalences in the RCT population. The harmonized estimator is defined as

$$\hat{\theta}_{1:K}^{b} = \arg\min_{v \in \mathbb{R}^K}\left[\left(v - \hat{\theta}_{1:K}^{(r+e)}\right)^\top \Sigma^{-1} \left(v - \hat{\theta}_{1:K}^{(r+e)}\right) + \lambda\left(\pi^\top v - \hat{\theta}^{(r)}\right)^2 \right], \tag{1}$$

where Σ is a positive-definite $K \times K$ matrix and λ is a nonnegative parameter that controls how closely the weighted average $\sum_{k=1}^{K} \pi_k \hat{\theta}_k^{b}$ matches the RCT estimate $\hat{\theta}^{(r)}$. When $\lambda = 0$ the harmonized estimator is equal to $\hat{\theta}_{1:K}^{(r+e)}$, while when $\lambda = \infty$ it becomes fully harmonized and $\sum_{k=1}^{K} \pi_k \hat{\theta}_k^{b} = \hat{\theta}^{(r)}$. This article examines the class of harmonized estimators defined by Equation (1), with particular attention to harmonized estimators with $\lambda = \infty$.

As a simple example of harmonized estimator, consider the following:

(Step 1) Use the pooled estimates

$$\hat{\theta}_k^{(r+e)} = \bar{Y}_{k,1}^{(r)} - \bar{Y}_{k,0}^{(r+e)}, \tag{2}$$

for each subgroup k. Here, $\bar{Y}_{k,1}^{(r)}$ is the mean outcome in subgroup k under the experimental treatment, and $\bar{Y}_{k,0}^{(r+e)}$ is the mean outcome under the control therapy, including both the EC data $D^{(r)}$ and the external data $D^{(e)}$.

(Step 2) The estimate of the treatment effect in the RCT population is

$$\hat{\theta}^{(r)} = \bar{Y}_{\cdot,1}^{(r)} - \bar{Y}_{\cdot,0}^{(r)}, \tag{3}$$

the difference between the mean outcomes in the experimental and control RCT arms, excluding the EC data. In this example, if Σ is the identity matrix then $\left(v - \hat{\theta}_{1:K}^{(r+e)}\right)^\top \Sigma^{-1} \left(v - \hat{\theta}_{1:K}^{(r+e)}\right)$ in Equation (1) is the squared Euclidean distance between $\hat{\theta}_{1:K}^{(r+e)}$ and v. Similarly, if Σ is an estimate of the covariance matrix of $\hat{\theta}_{1:K}^{(r+e)}$, then the first term in the right part of Equation (1) becomes the squared Mahalanobis distance.

We mention two practical aspects of the harmonized estimator. First, it ensures that the subgroup and primary analyses provide coherent inference about treatment effects. Indeed, we infer subgroup-specific treatment effects with a weighted average close to the point estimate $\hat{\theta}^{(r)}$ in the primary analysis. Other methods that leverage both $D^{(r)}$ and $D^{(e)}$ (e.g. expression (2)) may produce subgroup-specific results that are clearly incongruous with $\hat{\theta}^{(r)}$. Second, a broad range of estimators $\hat{\theta}_{1:K}^{(r+e)}$ can be used, for example:

(a) the simple difference between $\bar{Y}_{k,1}^{(r+e)}$ and $\bar{Y}_{k,0}^{(r+e)}$ as in Equation (2),
(b) the results of a Bayesian analysis based on a flexible regression model where the outcome distribution varies across biomarker subgroups, treatments, populations (RCT or EC), and patient-level pretreatment covariates (e.g. Sivaganesan et al., 2017), and
(c) or estimates based on a matching procedure (e.g. Lin et al., 2019) applied within each subgroup, to account for different distributions of potential confounders in the RCT and EC populations.

The solution of Equation (1) is

$$\hat{\theta}^b_{1:K} = \left(\Sigma^{-1} + \lambda\pi\pi^\top\right)^{-1}\left[\Sigma^{-1}\hat{\theta}^{(r+e)}_{1:K} + \lambda\pi\pi^\top\left(\hat{\theta}^{(r)}1_{K\times1}\right)\right], \tag{4}$$

where the column vector $1_{K\times1} = (1, \ldots, 1)$ has length K. This equation shows that the harmonized estimator is a *matrix weighted average* (see Chamberlain & Leamer, 1976) of $\hat{\theta}^{(r+e)}_{1:K}$ and $\hat{\theta}^{(r)}1_{K\times1}$. The solution can be derived based on the derivatives of the objective function in (1)

$$\Sigma^{-1}v - \Sigma^{-1}\hat{\theta}^{(r+e)}_{1:K} + \lambda\pi^\top v\pi - \lambda\hat{\theta}^{(r)}\pi,$$

which are equal to zero when

$$\left(\Sigma^{-1} + \lambda\pi\pi^\top\right)v = \Sigma^{-1}\hat{\theta}^{(r+e)}_{1:K} + \lambda\hat{\theta}^{(r)}\pi.$$

Moreover, the harmonized estimator can be rewritten as

$$\hat{\theta}^b_{1:K} = \hat{\theta}^{(r+e)}_{1:K} + c\frac{\lambda}{\lambda + c}\left(\hat{\theta}^{(r)} - \sum_{k=1}^{K}\pi_k\hat{\theta}^{(r+e)}_k\right)\Sigma\pi, \tag{5}$$

where $c = (\pi^\top\Sigma\pi)^{-1}$. This representation of $\hat{\theta}^b_{1:K}$ shows that the harmonized estimator is obtained by shifting $\hat{\theta}^{(r+e)}_{1:K}$ in a fixed direction, $\Sigma\pi$. The distance between $\hat{\theta}^b_{1:K}$ and $\hat{\theta}^{(r+e)}_{1:K}$ is proportional to the discrepancy between $\hat{\theta}^{(r)}$ and $\pi^\top\hat{\theta}^{(r+e)}_{1:K}$. For example, if $\lambda = \infty$ and $\Sigma\pi \propto 1_{K\times1}$, then in each subgroup $\hat{\theta}^b_k - \hat{\theta}^{(r+e)}_k = \hat{\theta}^{(r)} - \pi^\top\hat{\theta}^{(r+e)}_{1:K}$. We derived Equation (5) using the inversion formula $(\Sigma^{-1} + \lambda\pi\pi^\top)^{-1} = \Sigma - \lambda(1 + \lambda\pi^\top\Sigma\pi)^{-1}\Sigma\pi\pi^\top\Sigma$.

   The choice of Σ. In this article, we discuss two approaches for selecting Σ, which lead to what we call bias-directed (BD) harmonization and variance-directed (VD) harmonization.

   BD harmonization: In several cases, Σ can be chosen by leveraging an estimate of the bias of $\hat{\theta}^{(r+e)}_{1:K}$, which we denote as $\psi_{1:K} = E\left(\hat{\theta}^{(r+e)}_{1:K}\right) - \theta_{1:K}$. See for example expression (11) in the next subsection. Harmonization with $\lambda = \infty$ and a Σ matrix chosen to satisfy $\Sigma\pi \propto \psi_{1:K}$ provides unbiased estimates of $\theta_{1:K}$. We refer to $\hat{\theta}^b_{1:K}$ with this choice of Σ and λ as the *BD harmonized estimator*. Indeed, when $\lambda = \infty$ and $\Sigma\pi \propto \psi_{1:K}$, Equation (5) can be rewritten as

$$\hat{\theta}^b_{1:K} = \hat{\theta}^{(r+e)}_{1:K} - \left(\pi^\top\hat{\theta}^{(r+e)}_{1:K} - \hat{\theta}^{(r)}\right)\left(\pi^\top\Sigma\pi\right)^{-1}\Sigma\pi,$$

and the second term on the right-hand side is an unbiased estimate of $\psi_{1:K}$. We note that, if $E(\hat{\theta}^{(r)}) = \theta = \pi^\top\theta_{1:K}$ holds, then $E(\pi^\top\hat{\theta}^{(r+e)}_{1:K} - \hat{\theta}^{(r)}) = \pi^\top\psi_{1:K}$ and

$$E\left[\left(\pi^\top\hat{\theta}^{(r+e)}_{1:K} - \hat{\theta}^{(r)}\right)\left(\pi^\top\Sigma\pi\right)^{-1}\Sigma\pi\right] = \pi^\top\psi_{1:K}\left(\pi^\top\psi_{1:K}\right)^{-1}\psi_{1:K} = \psi_{1:K}.$$

We will use similar arguments in later sections to derive BD harmonized estimators when the input $\hat{\theta}^{(r+e)}_{1:K}$ is based on regression models. For example, in Section 3.2, we consider logistic regression.

   VD harmonization: Another interpretable choice is Σ proportional to an estimate of the covariance matrix of $\hat{\theta}^{(r+e)}_{1:K}$. When $\lambda = \infty$, we refer to $\hat{\theta}^b_{1:K}$ as the *VD harmonized estimator*. A Bayesian interpretation is discussed in Section 2.3.

2.2 Bias and variance of the harmonized estimator

When the joint distribution of $\hat{\theta}^{(r+e)}_{1:K}$ and $\hat{\theta}^{(r)}$ is known, we can derive the bias and variance of $\hat{\theta}^b_{1:K}$, for any choice of λ and Σ, because the harmonized estimator is a linear function of $\hat{\theta}^{(r+e)}_{1:K}$ and $\hat{\theta}^{(r)}$

(Equation (5)). In this section, we specify scenarios (i.e. data-generating models), consider simple estimators $\hat{\theta}_{1:K}^{(r+e)}$ and $\hat{\theta}^{(r)}$, and illustrate the bias, variance, and MSE of $\hat{\theta}_{1:K}^{b}$ based on analytic expressions. In Sections 3 and 4, we will discuss harmonization in other settings, with different scenarios and definitions of $\hat{\theta}_{1:K}^{(r+e)}$ and $\hat{\theta}^{(r)}$.

We consider data generated from the following model:

$$Y_i^{(r)} \mid W_i^{(r)}, T_i^{(r)} \stackrel{iid}{\sim} N\left(\mu_{W_i^{(r)}} + \theta_{W_i^{(r)}} T_i^{(r)}, \phi^2\right), \quad \text{and}$$
$$Y_i^{(e)} \mid W_i^{(e)} \stackrel{iid}{\sim} N\left(\mu_{W_i^{(e)}} + \gamma_{W_i^{(e)}}, \phi^2\right). \tag{6}$$

Here, μ_k is the expected outcome under the control therapy for subgroup k in the RCT population, θ_k is the subgroup-specific treatment effect, and γ_k is a bias term, the difference between the expected outcomes of the external and RCT controls in subgroup k. All patients in the external data receive the control therapy. Model (6) does not include covariates X.

As inputs for $\hat{\theta}_{1:K}^{b}$, we consider the estimators $\hat{\theta}_k^{(r+e)} = \bar{Y}_{k,1}^{(r)} - \bar{Y}_{k,0}^{(r+e)}$ and $\hat{\theta}^{(r)} = \bar{Y}_{\cdot,1}^{(r)} - \bar{Y}_{\cdot,0}^{(r)}$ as in Equations (2) and (3). For the study design, we assume that the vector π is known and that for each subgroup k

$$\frac{n_{k,0}^{(r)}}{n_{\cdot,0}^{(r)}} = \frac{n_{k,1}^{(r)}}{n_{\cdot,1}^{(r)}} = \pi_k.$$

This condition is satisfied if recruitment is stratified by subgroup and the trial adopts block randomization. For simplicity, we also assume the same proportions in the EC dataset, i.e. for each subgroup k

$$\frac{n_{k,\cdot}^{(e)}}{n^{(e)}} = \pi_k.$$

After Proposition 1, all other theoretical results in the article will not assume equal subgroup proportions in the RCT and ECs, see for example the last paragraphs of this subsection where we consider settings with $n_{k,0}^{(e)}/n^{(e)} \neq \pi_k$.

We derive the sampling distribution of the harmonized estimator in this setting. We first restrict attention to the case where randomization is balanced, i.e. $n_{\cdot,1}^{(r)} = n_{\cdot,0}^{(r)}$, and $\pi = (1/K, \ldots, 1/K)$. We will later relax the assumptions of this stylized example. Let $q = n_{\cdot,0}^{(e)}/(n_{\cdot,0}^{(r)} + n_{\cdot,0}^{(e)})$ be the proportion of controls who are in the external data.

Proposition 1   In the outlined setting, if $\Sigma^{-1} = I_K$, then $\hat{\theta}_{1:K}^{b} \sim N(\theta_{1:K} + B^b, V^b)$, where the bias vector B^b is

$$B^b = -q\left(I_K - \frac{1}{K}\frac{\lambda}{\lambda + K} 1_{K\times K}\right)\gamma_{1:K}, \tag{7}$$

and the variance–covariance matrix is

$$V^b = K\frac{1}{n_{\cdot,0}^{(r)}}(2 - q)\phi^2 I_K + \left(\frac{\lambda}{\lambda + K}\right)^2 q\frac{1}{n_{\cdot,0}^{(r)}}\phi^2 1_{K\times K}. \tag{8}$$

Here, $1_{K\times K}$ is a $K \times K$ matrix with all elements equal to 1. As expected the harmonized estimator is normally distributed; indeed, it is a linear transformation of the outcome data (see Equation (4)). The derivations of this proposition and other results are provided in the online supplementary material. If the outcome data in $D^{(r)}$ and $D^{(e)}$ are not normal but have means and variances identical to those in (6), then the bias and variance expressions in Proposition 1 still hold.

Systematic distortion mechanisms. In several contexts, it is implausible that the distortion mechanisms (e.g. measurement errors) cause differences between the conditional outcome distributions in the RCT control arm and the EC arm that vary across subgroups. Throughout this article, we assume that if there is a discrepancy between the conditional outcome distributions in the control arm of the trial and the EC group, then the distortion mechanism is systematic across subgroups. In the context of model (6), the systematic distortion mechanism (SDM) assumption is formalized by the equation

$$\gamma_k = \gamma, \tag{9}$$

for all subgroups $k = 1, \ldots, K$ and some $\gamma \in \mathbb{R}$. More generally, SDMs can be represented through statistical models in which the subgroup-specific conditional outcome distributions $p(Y_i^{(e)} \mid W_i^{(e)} = k, X_i^{(e)} = x)$ for the ECs are distorted versions of $p(Y_i^{(r)} \mid W_i^{(r)} = k, X_i^{(r)} = x, T_i^{(r)} = 0)$.

> **Definition** We have an SDM if there exists a *distortion* function, with domain and range equal to the space of probability distributions on the real line, such that
>
> $$p\left(Y_i^{(e)} \mid W_i^{(e)} = k, X_i^{(e)} = x\right) = \text{distortion}\left(p\left(Y_i^{(r)} \mid W_i^{(r)} = k, X_i^{(r)} = x\right)\right) \tag{10}$$
>
> for all subgroups $k = 1, \ldots, K$ and pretreatment profiles x.

Several examples of distortion functions will be discussed in the remainder of the article. The SDM assumption holds when the difference between the distributions of the outcomes in the EC group and the RCT control arm, conditional on pretreatment covariates, is identical across subgroups. These differences can be summarized by comparing conditional means (here and Section 3.1), odds ratios (Section 3.2), or other metrics. In previous work, we identified systematic distortion mechanisms in collections of datasets in neuro-oncology (Rahman et al., 2023; Ventz et al., 2022).

There are several potential causes of systematic distortions. For example, the measurement of the outcomes in the external group might include a systematic shift, in all K subgroups, compared to the measurements in the RCT. In particular, when we consider oncology RCTs, different procedures may be used to measure tumor sizes in the trial patients and the ECs. Similarly, the definition of the first day of follow-up, or the schedule of radiographic measurements can vary across studies and cause systematic discrepancies in the assessment of response to therapy and tumor progression (Dabush et al., 2021). We note that in these settings the experimental therapy might produce positive effects in one or a few of the K subpopulations. Systematic distortions might also arise with subtle differences between the administration of the control therapy in the RCT and the external group, such as different levels of adherence to the schedule of the chemoradiation plan.

When the SDM assumption is satisfied and the assumptions of Proposition 1 hold, if we set $\lambda = \infty$, then the harmonized estimator is *unbiased*. Indeed B^h becomes $-q(I_K - (1/K)1_{K \times K})\gamma 1_{K \times 1} = 0_{K \times 1}$. If the subgroups are unbalanced (i.e. $\pi_k \neq 1/K$ for some k), then we can choose Σ to produce an unbiased harmonized estimator. We provide details at the end of this subsection.
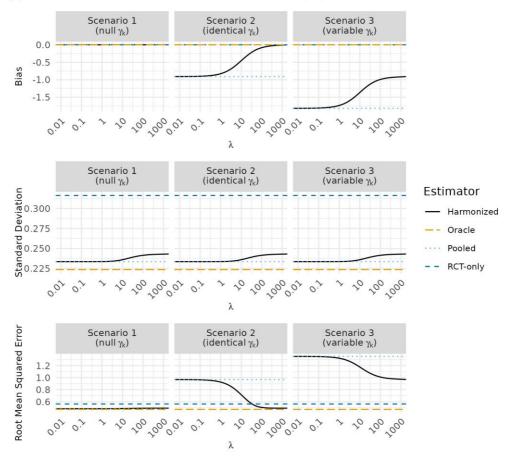
Figure 1 illustrates the bias, variance, and mean squared error of the harmonized estimator (see Equations (7) and (8) in Proposition 1). It also shows the operating characteristics of other estimators of $\theta_{1:K}$:

(a) the *RCT-only* estimator $\hat{\theta}_k^{(r)} = \bar{Y}_{k,1}^{(r)} - \bar{Y}_{k,0}^{(r)}$, for $k = 1, \ldots, K$, which is unbiased, and has variance equal to $K(1/n_{\cdot,0}^{(r)} + 1/n_{\cdot,1}^{(r)})\phi^2 I_K$,

(b) the *pooled estimator* in Equation (2), which has bias equal to $-q\gamma_{1:K}$ and variance equal to $K(1/n_{\cdot,0}^{(r)})(2-q)\phi^2 I_K$, and

(c) an *oracle estimator* $\hat{\theta}_{1:K}^{\text{ora}}$ (with known $\mu_{1:K}$ values),

$$\hat{\theta}_k^{\text{ora}} = \bar{Y}_{k,1}^{(r)} - \mu_k,$$

which is unbiased and has variance equal to $(K/n_{\cdot,0}^{(r)})\phi^2 I_K$.

(a) Performance of four estimators in Scenarios 1, 2, and 3

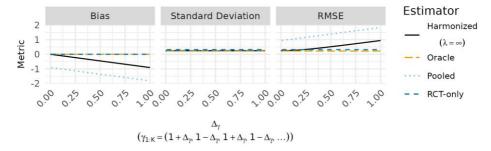**(b)** Comparison of the four estimators in Scenario 3

Figure 1. (a) Performance summaries of the harmonized estimator (see Proposition 1) and other estimators (analogous analytic results) for subgroup $k = 1$. The λ value used to define the harmonized estimator varies on the x-axis. We illustrate the bias, standard deviation, and the RMSE of the estimator as a function of λ. The scenarios have different $\gamma_{1:K}$ values: $0_{K \times 1}$ (scenario 1), constant across subgroups (scenario 2), or varying across subgroups, with $\gamma_{1:K} = (1 + \Delta_\gamma, 1 - \Delta_\gamma, 1 + \Delta_\gamma, 1 - \Delta_\gamma, \ldots)$ and $\Delta_\gamma = 1$ (scenario 3). (b) The performance of the harmonized estimator (with $\lambda = \infty$) for subgroup $k = 1$ degrades when the subgroup-specific differences $\gamma_{1:K}$ vary substantially across subgroups, i.e. as Δ_γ grows.

In Figure 1, we examine three scenarios with different values of the subgroup-specific differences $\gamma_{1:k}$ between the EC and RCT control means: $\gamma_k = 0$ for each $k = 1, \ldots, K$ (scenario 1, no distortions), $\gamma_k = 1$ for each $k = 1, \ldots, K$ (scenario 2, SDM assumption holds), and $\gamma_{1:K} = (1 + \Delta_\gamma, 1 - \Delta_\gamma,$

$1 + \Delta_\gamma, \dots$) (scenario 3, SDM assumption violated). For the sample sizes we use $n^{(r)} = 100$ RCT patients and $n^{(e)} = 500$ EC patients, we have $K = 10$ subgroups, and $\phi^2 = 1$.

Figure 1a shows how the bias of $\hat{\theta}^b_{1:K}$ converges to 0 as $\lambda \to \infty$ in scenario 2. The estimator $\hat{\theta}^b_{1:K}$ is suboptimal in settings where the bias terms $\gamma_{1:K}$ vary substantially across subgroups, such as scenario 3. Recall that we introduced the harmonized estimator for inference on the treatment effects in small patient subgroups in disease settings where heterogeneous bias mechanisms across subgroups (i.e. relevant variations of $\gamma_{1:k}$ across subgroups) appear implausible. We emphasize that in Figure 1a, scenario 3 involves a high level of variability of γ_k across subgroups to illustrate that when the SDM assumption is markedly violated harmonization produces biased estimates. Figure 1b shows that when the γ_k vary *moderately* across subgroups ($\Delta_\gamma < 0.25$), the estimator $\hat{\theta}^b_{1:K}$ has lower root mean squared error (RMSE) than $\hat{\theta}^{(r)}$, despite the violation of the SDM assumption. With $\Delta_\gamma < 0.25$ the violation of this assumption is less marked than in scenario 3, where the γ_k parameters are highly heterogeneous ($\Delta_\gamma = 1$ in Figure 1a). The estimate $\hat{\theta}^b_{1:K}$ has lower variance than $\hat{\theta}^{(r)}_{1:K}$. Also, the variance of $\hat{\theta}^b_{1:K}$ is similar to the variance of $\hat{\theta}^{(r+e)}_{1:K}$.

We can relax some of the assumptions in Proposition 1. First, the positive-definite matrix Σ used in (1) to define the estimator $\hat{\theta}^b_{1:K}$ can be arbitrary. Second, the randomization ratio and the subgroup proportions in the RCT can be unbalanced, i.e. we consider $n^{(r)}_{\cdot,1} \neq n^{(r)}_{\cdot,0}$ and $\pi \neq (1/K, \dots, 1/K)$. We still assume that $n^{(r)}_{k,1}/n^{(r)}_{k,0}$ is constant across subgroups and $n^{(r)}_{k,\cdot}/n^{(r)} = \pi_k$, yet we allow the subgroup proportions in the EC data to differ from those in the RCT; that is $n^{(e)}_{k,0}/n^{(e)}$ can be different from π_k. In this case, the harmonized estimator $\hat{\theta}^b_{1:K}$ remains normally distributed, and the bias and variance in expressions (7) and (8) become

$$
\begin{aligned}
B^b &= -\left(I_K - c\frac{\lambda}{\lambda+c}\Sigma\pi\pi^\top\right)Q\gamma_{1:K}, \quad \text{and} \\
V^b &= \phi^2\left[\frac{1}{n^{(r)}_{\cdot,1}}I_K - \frac{1}{n^{(r)}_{\cdot,0}}(I_K - Q)\right]\Pi^{-1} + \left(c\frac{\lambda}{\lambda+c}\right)^2\bar{q}\frac{1}{n^{(r)}_{\cdot,0}}\phi^2\Sigma\pi\pi^\top\Sigma
\end{aligned}
\tag{11}
$$

respectively. Here, $c = (\pi^\top\Sigma\pi)^{-1}$ as before, Q is a diagonal $K \times K$ matrix with $Q_{k,k} = n^{(e)}_{k,\cdot}/n^{(r+e)}_{k,0}$, $\bar{q} = \sum_{k=1}^K \pi_k Q_{k,k}$, and Π is a diagonal $K \times K$ matrix with $\Pi_{k,k} = \pi_k$. When $\lambda = \infty$ and the SDM assumption holds with $\gamma_{1:K} = \gamma 1_{K \times 1}$ for some $\gamma \in \mathbb{R}$, the bias is $B^b = 0_{K \times 1}$ if we choose $\Sigma = \mathrm{diag}(Q_{k,k}/\pi_k)$. This choice of λ and Σ gives a BD harmonized estimator. Indeed, $\Sigma\pi$ becomes equal to $Q1_{K \times 1}$, and B^b can be rewritten as

$$
\left(I_K - \left(\pi^\top\Sigma\pi\right)^{-1}\Sigma\pi\pi^\top\right)Q\gamma_{1:K} = \gamma\left(I_K - \left(\pi^\top\Sigma\pi\right)^{-1}\Sigma\pi\pi^\top\right)\Sigma\pi = 0_{K \times 1}.
$$

For the kth subgroup, the difference between the mean squared errors of the pooled estimator and the harmonized estimator is

$$
\mathrm{MSE}\left(\hat{\theta}^{(r+e)}_k\right) - \mathrm{MSE}\left(\hat{\theta}^b_k\right) = q_k^2\left[(\gamma_k - \bar{\gamma})^2 - \gamma_k^2\right] - \frac{q_k^2\phi^2}{n^{(r)}_{\cdot,0}\left(\sum_{k=1}^K \pi_k q_k\right)},
\tag{12}
$$

where $\bar{\gamma} = \sum_{k=1}^K \pi_k q_k \gamma_k / \sum_{k=1}^K \pi_k q_k$ is the weighted average of the subgroup-specific EC distortion parameters. The first term in the right hand of Equation (12) is the difference between the squared biases of $\hat{\theta}^{(r+e)}_{1:K}$ and $\hat{\theta}^b_{1:K}$, and the second term is the difference between their variances. Pooling has a slightly lower variance compared to harmonization. In the absence of distortions (i.e. scenario 1, with $\gamma_k = 0$ for all k) both pooling and harmonization are unbiased. In contrast, in the presence of systematic distortions across subgroups (scenario 2, $\gamma_k = \gamma \in \mathbb{R}$ for all k) pooling is biased while harmonization is unbiased.

2.3 A quasi-Bayesian interpretation of the harmonized estimator

We provide an interpretation of the harmonized estimator that builds on the modular Bayesian framework in Jacob et al. (2017) and Bayarri et al. (2009). In this section, we focus on a simple model to exemplify the relationship between harmonization and Bayesian analyses. We consider two analysts: **Analyst 1** conducts a primary analysis and estimates θ, the overall effect, using only the RCT data $D^{(r)}$ (without considering subgroups). The analyst uses the following Bayesian model:

$$
\begin{aligned}
\text{Likelihood:} \quad & Y_i^{(r)} \mid T_i^{(r)}, \mu, \theta \overset{\text{ind.}}{\sim} N\big(\mu + \theta T_i^{(r)}, \sigma^2\big), \\
\text{Prior:} \quad & (\mu, \theta) \sim N_2(m^{(A1)}, \tau^{(A1)}).
\end{aligned}
\tag{13}
$$

To simplify the presentation the analyst chooses some fixed value for σ^2. The result of the analysis is the posterior $p^{(A1)}(\theta \mid D^{(r)}) = p^{(A1)}(\theta \mid Y_{1:n^{(r)}}^{(r)}, T_{1:n^{(r)}}^{(r)})$, conditional on the outcomes $Y_{1:n^{(r)}}^{(r)}$ and treatment assignments $T_{1:n^{(r)}}^{(r)}$ in the RCT. We use the superscript $(A1)$ to identify the posterior inference of Analyst 1. Note that $p^{(A1)}(\theta \mid D^{(r)})$ does not involve the EC data.

Analyst 2 conducts a subgroup analysis and estimates $\theta_{1:K}$ using the RCT and EC data $D^{(r+e)}$, thus augmenting the subgroup-specific sample sizes of the RCT. The analyst uses the following Bayesian model:

$$
\begin{aligned}
\text{Likelihood:} \quad & Y_i^{(r)} \mid T_i^{(r)}, W_i^{(r)}, \mu_{1:K}, \theta_{1:K} \overset{\text{ind.}}{\sim} N\big(\mu_{W_i^{(r)}} + \theta_{W_i^{(r)}} T_i^{(r)}, \phi^2\big), \\
& Y_i^{(e)} \mid W_i^{(e)}, \mu_{1:K}, \theta_{1:K} \overset{\text{ind.}}{\sim} N\big(\mu_{W_i^{(e)}}, \phi^2\big), \\
\text{Prior:} \quad & (\mu_{1:K}, \theta_{1:K}) \sim N_{2K}\big(m^{(A2)}, \tau^{(A2)}\big).
\end{aligned}
\tag{14}
$$

This model assumes identical outcome distributions for the RCT controls and the EC population. Also, in this case, the analyst chooses some fixed value for ϕ^2. The analyst assumes that $(n_{1,\cdot}^{(r)}/n^{(r)}, \ldots, n_{K,\cdot}^{(r)}/n^{(r)}) = \pi$. The posterior distribution is $p^{(A2)}(\theta_{1:K} \mid D^{(r)}, D^{(e)})$. If the RCT controls and ECs have different outcome distributions then the model is misspecified.

We can combine Analyst 1's inference on the overall effect θ and Analyst 2's inference on the subgroup-specific effects $\theta_{1:K}$ using the modular approach discussed in Jacob et al. (2017). Analyst 2 has a joint posterior on $\theta_{1:K}$ and $\theta = \sum_{k=1}^{K} \pi_k \theta_k$, that can be factorized into the product of (i) the posterior of θ and (ii) the posterior of $\theta_{1:K}$ given θ:

$$
p^{(A2)}\big(\theta, \theta_{1:K} \mid D^{(r)}, D^{(e)}\big) = p^{(A2)}\big(\theta \mid D^{(r)}, D^{(e)}\big) \times p^{(A2)}\big(\theta_{1:K} \mid \theta, D^{(r)}, D^{(e)}\big).
\tag{15}
$$

Based on the modular framework we obtain the following *cut* distribution (we use the term *cut* as in Jacob et al., 2017):

$$
p^{\text{cut}}\big(\theta_{1:K}, \theta \mid D^{(r)}, D^{(e)}\big) = p^{(A1)}\big(\theta \mid D^{(r)}\big) \times p^{(A2)}\big(\theta_{1:K} \mid \theta, D^{(r)}, D^{(e)}\big).
\tag{16}
$$

This joint distribution is obtained by replacing $p^{(A2)}(\theta \mid D^{(r)}, D^{(e)})$ in (15) with $p^{(A1)}(\theta \mid D^{(r)})$, which only uses the RCT data without subgroup information. The cut distribution of $\theta_{1:K}$ is based on both $D^{(r)}$ and $D^{(e)}$, and is coherent with the inference of the first analyst on θ, which is based only on $D^{(r)}$.

A simpler plug-in distribution can be obtained by ignoring Analyst 1's uncertainty about θ, using $\theta^{(A1)} = E^{(A1)}(\theta \mid Y_{1:n^{(r)}}^{(r)}, T_{1:n^{(r)}}^{(r)})$:

$$
p^{\text{plug-in}}\big(\theta_{1:K} \mid D^{(r)}, D^{(e)}\big) = p^{(A2)}\big(\theta_{1:K} \mid \theta = \theta^{(A1)}, D^{(r)}, D^{(e)}\big).
\tag{17}
$$

Both the cut distribution in (16) and the plug-in distributions in (17) produce point estimates of $\theta_{1:k}$ that coincide with the harmonized estimator. That is,
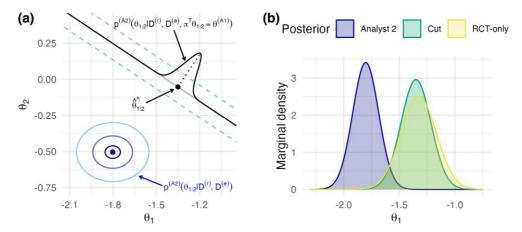
Figure 2. A quasi-Bayesian interpretation of the harmonized estimator with $K = 2$ subgroups. (a) Analyst 2's posterior $p^{(A2)}(\theta_1, \theta_2 \mid D^{(r)}, D^{(e)})$ in the bottom-left in blue (mean and isolines). The plug-in posterior $p^{(A2)}(\theta_1, \theta_2 \mid D^{(r)}, D^{(e)}, \pi^\top \theta_{1:2} = \theta^{(A1)})$ is the density in the top right black, which has support equal to the linear subspace $\pi^\top \theta_{1:2} = \theta^{(A1)}$. The mean of the plug-in posterior (black point) is equal to the harmonized estimator. The dashed lines show the linear subspaces $\pi^\top \theta_{1:2} = \theta$ with θ equal to the 25% and 75% quantiles of Analyst 1's posterior $p^{(A1)}(\theta \mid D^{(r)})$. (b) The posterior density of θ_1 under Analyst 2's joint model for RCT and EC data (left-most curve, blue), the cut distribution (Equation (16), middle curve, green), and a Bayesian subgroup analysis using only RCT data (right-most curve, yellow).

$$E^{\text{cut}}(\theta_{1:K} \mid D^{(r)}, D^{(e)}) = E^{\text{plug-in}}(\theta_{1:K} \mid D^{(r)}, D^{(e)}) = \hat{\theta}^b_{1:K}, \tag{18}$$

where $\hat{\theta}^b_{1:K}$ is obtained with inputs $\lambda = \infty$, $\hat{\theta}^{(r)} = E^{(A1)}(\theta \mid D^{(r)})$, $\hat{\theta}^{(r+e)}_{1:K} = E^{(A2)}(\theta_{1:K} \mid D^{(r)}, D^{(e)})$, and $\Sigma = Var^{(A2)}(\theta_{1:K} \mid D^{(r)}, D^{(e)})$ in the right-hand side of Equation (1). Thus, the Bayesian view of our work suggests VD harmonization, where $\lambda = \infty$ and Σ measures the variability associated with $\hat{\theta}^{(r+e)}_{1:K}$.

The described relationship between the harmonized estimator and Bayesian analyses is illustrated in Figure 2a. In Figure 2b, we show the posterior density of θ_1 under Analyst 2's joint model for the RCT and EC data (blue), the cut distribution (Equation (16), green), and a Bayesian subgroup analysis using only RCT data (i.e. Analyst 2's inference if $D^{(e)}$ is not included, yellow). We refer to the online supplementary material for the details on the prior models and the simulation illustrated in this figure.

To verify Equation (18), it is sufficient to note that

$$\int E^{(A2)}(\theta_{1:K} \mid \pi^\top \theta_{1:K} = \theta, D^{(r)}, D^{(e)}) \cdot p^{(A1)}(\theta \mid D^{(r)}) \, d\theta$$
$$= E^{(A2)}(\theta_{1:K} \mid \pi^\top \theta_{1:K} = \theta^{(A1)}, D^{(r)}, D^{(e)}) \quad \text{and,}$$
$$E^{(A2)}(\theta_{1:K} \mid \pi^\top \theta_{1:K} = \hat{\theta}^{(A1)}, D^{(r)}, D^{(e)}) = \underset{\pi^\top \theta_{1:K} = \hat{\theta}^{(A1)}}{\arg\max} \, p^{(A2)}(\theta_{1:K} \mid D^{(r)}, D^{(e)})$$
$$= \underset{\pi^\top \theta_{1:K} = \hat{\theta}^{(A1)}}{\arg\min} \left(\theta_{1:K} - \theta^{(A2)}_{1:K}\right)^\top \Sigma^{-1} \left(\theta_{1:K} - \theta^{(A2)}_{1:K}\right),$$

where $\theta^{(A2)}_{1:K} = E^{(A2)}(\theta_{1:K} \mid D^{(r)}, D^{(e)})$ and $\Sigma = Var^{(A2)}(\theta_{1:K} \mid D^{(r)}, D^{(e)})$. The first equality holds because $E^{(A2)}(\theta_{1:K} \mid \pi^\top \theta_{1:K} = \theta, D^{(r)}, D^{(e)})$ is linear in θ. The final two equalities hold because Analysts 1 and 2 use conjugate normal models. We note that the estimator in expression (18) is unbiased when $\gamma_1 = \gamma_2 \ldots = \gamma_K$ (SDM assumption) if the ratios $n^{(e)}_{k,0}/n^{(r)}_{k,0}$ and $n^{(r)}_{k,1}/n^{(r)}_{k,0}$ remain the same across subgroups $k = 1, \ldots, K$ (see online supplementary material Section S2.4.5).

Up to this point, we have shown how the cut distribution and the harmonized estimator are connected when Analysts 1 and 2 have normal posterior distributions on θ and $\theta_{1:K}$. This discussion can be extended beyond the conjugate model that we considered. In particular, Analyst 2 may use a

Bayesian joint regression model for $D^{(r)}$ and $D^{(e)}$, for example a generalized linear model (Hobbs et al., 2012), in which $\theta_{1:K}$ are subgroup-specific treatment effects (i.e. the differences between the expected outcomes under experimental and control therapy across the RCT subpopulations). Analyst 1 may use another Bayesian model (with or without biomarker subgroups and covariate information) for $D^{(r)}$ alone, and θ is the average treatment effect in the overall population. If Bernstein–von Mises conditions (Ghosal & van der Vaart, 2017) are satisfied—as in most regression models—then with sufficiently large $n^{(r)}$ and $n^{(e)}$ the relationships between posterior distributions and the harmonized estimators that we discussed (Figure 2a and Equation (18)) are preserved.

2.4 Interval estimation

We discuss $(1 - \alpha)$-level intervals obtained through harmonization. We consider $\hat{\theta}_k^{(r+e)} = \bar{Y}_{k,1}^{(r)} - \bar{Y}_{k,0}^{(r+e)}$ and $\hat{\theta}^{(r)} = \bar{Y}_{\cdot,1}^{(r)} - \bar{Y}_{\cdot,0}^{(r)}$ (Equations (2) and (3)). The confidence intervals of θ_k are computed assuming that $D^{(r)}$ and $D^{(e)}$ are distributed according to model (6). We describe three distinct approaches.

(a) We can use the analytic expression of the variance $V_{k,k}^b$ (Equation (11)) and compute

$$\left(\hat{\theta}_k^b - z_{1-\alpha/2} \sqrt{V_{k,k}^b}, \quad \hat{\theta}_k^b + z_{1-\alpha/2} \sqrt{V_{k,k}^b} \right), \tag{19}$$

where $z_{1-\alpha/2}$ is the $(1 - \alpha/2) \times 100\%$ quantile of the standard normal distribution. Here, we are assuming implicitly that $\hat{\theta}_k^b$ is an unbiased estimate (see Section 2.2). Standard central limit arguments imply that the confidence interval can still be considered in settings where the assumption of normally distributed outcomes in model (6) is violated.

(b) Based on the connection between the harmonized estimator and Bayesian analyses illustrated in Section 2.3, we can use the cut distribution in (16) to compute the interval

$$\left(\hat{\theta}_k^b - z_{1-\alpha/2} \sqrt{V_{k,k}^{(\text{cut})}}, \quad \hat{\theta}_k^b + z_{1-\alpha/2} \sqrt{V_{k,k}^{(\text{cut})}} \right), \tag{20}$$

where $V_{k,k}^{(\text{cut})}$ is the variance of θ_k under the cut distribution. If Analysts 1 and 2 use the conjugate models described in Section 2.3, then the covariance matrix is

$$V^{(\text{cut})} = \Sigma_{1:K}^{(A2)} + \left(\frac{1}{\Sigma_\theta^{(A2)}} \right)^2 \left(\Sigma_\theta^{(A1)} - \Sigma_\theta^{(A2)} \right) \Sigma_{1:K}^{(A2)} \pi \pi^\top \Sigma_{1:K}^{(A2)},$$

where $\Sigma_{1:K}^{(A2)}$ is the posterior covariance matrix of $\theta_{1:K}$ under Analyst 2's model, $\Sigma_\theta^{(A2)}$ is the posterior variance of θ under Analyst 2's model, and $\Sigma_\theta^{(A1)}$ is the posterior variance of θ under Analyst 1's model. Moreover, $V^{(\text{cut})} = V^b$ under the following conditions: (1) Analysts 1 and 2 use flat priors ($\tau^{(A1)}$ and $\tau^{(A2)}$ are diagonal with diverging diagonal entries), (2) the variance parameters σ^2 and ϕ^2 in models (13), (14), and (6) (underlying V^b) are identical, and (3) $\lambda = \infty$ and Σ is diagonal with entries proportional to $(1/n_{k,1}^{(r)}) + (1/n_{k,0}^{(r+e)})$. Thus, the intervals in expressions (20) and (19) become identical.

(c) The bootstrap method (Efron & Tibshirani, 1986) can also be used to construct confidence intervals. We can generate $a = 1, \ldots, R$ replicates $(D_a^{(r)}, D_a^{(e)})$ through a parametric bootstrap (using model 6). For each replicate, the harmonized estimator $\hat{\theta}_{1:K,a}^b$ is computed. This allows us to estimate the variability of $\hat{\theta}_k^b$ through the empirical distribution $\hat{G}_k$ of the $\hat{\theta}_{k,a}^b$ replicates. We can report a confidence interval for θ_k that is centred at $\hat{\theta}_k^b$ and has width that matches the distance between the quantiles $\hat{G}_k^{-1}(\alpha/2)$ and $\hat{G}_k^{-1}(1 - \alpha/2)$.
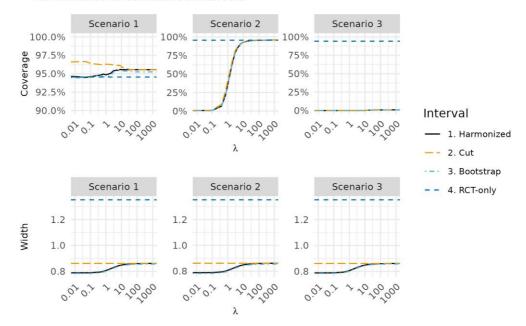
Performance of confidence intervals

Figure 3. Coverage and width of 95% confidence intervals. We use the procedures described in Section 2.4 to compute the intervals. We consider the same scenarios as in Figure 1. Coverage and average width are computed using 2,000 Monte Carlo replicates for each scenario. On the x-axis, we vary the λ value that defines the harmonized estimator.

In Figure 3, we compare the coverage and average length of these intervals through simulations. We also consider the usual confidence interval of θ_k based only on RCT data:

$$\left(\hat{\theta}_k^{(r)} - z_{1-\alpha/2} \times \sqrt{\mathrm{Var}\left(\hat{\theta}_k^{(r)}\right)}, \quad \hat{\theta}_k^{(r)} + z_{1-\alpha/2} \times \sqrt{\mathrm{Var}\left(\hat{\theta}_k^{(r)}\right)} \right),$$

where $\hat{\theta}_k^{(r)} = \bar{Y}_{k,1}^{(r)} - \bar{Y}_{k,0}^{(r)}$. We include this interval for comparison. We simulate data using the same scenarios used for Figure 1. For the harmonized estimator, we use $\Sigma = \mathrm{diag}(Q_{k,k}/\pi_k)$, and $(\hat{\theta}^{(r)}, \hat{\theta}_{1:K}^{(r+e)})$ as defined by expressions (3) and (2). To compute the cut distribution we set $\tau^{(A1)} = \mathrm{diag}(10^4)$ and $\tau^{(A2)} = \mathrm{diag}(10^4)$. For each scenario, we simulated 2,000 trials. We illustrate results for subgroup $k = 1$. Figure 3 shows that in scenarios 1 and 2, in which $\hat{\theta}_1^b$ is unbiased for $\lambda = \infty$, the intervals have coverage nearly identical to the nominal $1 - \alpha$ level (when $\lambda > 10$). In contrast in scenario 3, where the SDM assumption is markedly violated and $\hat{\theta}_1^b$ is biased, as expected the coverage is far below the nominal value (close to 0%). The harmonized and cut intervals centred on $\hat{\theta}_1^b$ are substantially shorter than the intervals that we compute using only RCT data. The poor coverage in scenario 3 emphasizes that harmonization is not sufficient to remove the bias of $\hat{\theta}_{1:K}^{(r+e)}$ when the SDM assumption (i.e. $\gamma_1 = \ldots = \gamma_K$, Equation (9)) is violated. Importantly, data analyses can detect violations of the SDM assumption (see for example online supplementary material Section S3.8). Figure 3 confirms that the analyst should consider harmonization if (i) high-quality EC data (see Corrigan-Curay et al., 2018 for explicit criteria) are available, (ii) the violation of the SDM assumption based on context-specific considerations is implausible, and (iii) the datasets do not provide evidence of a violation of the SDM assumption.

When an analytic expression of the interval, based on exact or asymptotic arguments (e.g. Equation (19)), is available, the computing time is lower than bootstrapping. However, if the distribution of $\hat{\theta}_{1:K}^b$ is not available, then the bootstrap interval is an easy-to-interpret and practical

alternative. In Figure 3, we observe negligible differences between the approaches that we considered to provide confidence intervals.

3 Harmonized estimators and regression models

In this section, we discuss harmonized estimators that allow the investigator to account for pretreatment covariates X. The distributions of relevant prognostic variables, such as patient age, may be different in the RCT and EC populations. Recall that the datasets are $D^{(r)} = (Y_i^{(r)}, T_i^{(r)}, W_i^{(r)}, X_i^{(r)})_{i \leq n^{(r)}}$ for the RCT and $D^{(e)} = (Y_i^{(e)}, T_i^{(e)}, W_i^{(e)}, X_i^{(e)})_{i \leq n^{(e)}}$ for the EC. Here, X includes d pretreatment covariates, such as demographics. Also, the harmonized estimator $\hat{\theta}_{1:K}^{b}$ is a transformation of (i) $\hat{\theta}_{1:K}^{(r+e)}$, an estimator of $\theta_{1:K}$ (the subgroup-specific treatment effects) based on $D^{(r)}$ and $D^{(e)}$, and (ii) $\hat{\theta}^{(r)}$, an estimator of $\theta = \pi^{\top} \theta_{1:K}$ (the average treatment effect in the randomized population) based only on $D^{(r)}$. The examples in this section include linear regression, logistic regression, and propensity score weighted logistic regression. In each case we show how to obtain a BD harmonized estimator that is approximately unbiased when the SDM assumption holds. These results are unified in one proposition (see online supplementary material Proposition S2) on weighted generalized linear models (GLMs) in Section S3.6 of the online supplementary material. The analytic expressions of Σ for BD harmonization in this section are special cases of Propositions S3 and S5 in the online supplementary material.

3.1 Linear model

We first discuss the harmonized estimator in a simple setting, where the data are analysed through linear models. In particular, we consider $\hat{\theta}_{1:K}^{b}$ defined by Equation (1) with the following inputs.

(a) The estimator $\hat{\theta}^{(r)}$ is an unbiased estimator, like the simple difference of means in Equation (3). In the remainder of this section, we consider $\hat{\theta}^{(r)}$ based on ordinary least squares (OLS), assuming that

$$E\left(Y_i^{(r)} \mid T_i^{(r)}, X_i^{(r)}\right) = \mu + \theta T_i^{(r)} + \beta^{\top} X_i^{(r)}. \tag{21}$$

Sufficient conditions for $\hat{\theta}^{(r)}$ to be an unbiased estimator of $E(Y_i^{(r)} \mid T_i^{(r)} = 1) - E(Y_i^{(r)} \mid T_i^{(r)} = 0)$ are provided in the online supplementary material.

(b) The estimator $\hat{\theta}_{1:K}^{(r+e)}$ is computed with OLS, assuming

$$\begin{aligned}
E\left(Y_i^{(r)} \mid W_i^{(r)}, T_i^{(r)}, X_i^{(r)}\right) &= \mu_{W_i^{(r)}} + \theta_{W_i^{(r)}} T_i^{(r)} + \beta^{\top} X_i^{(r)}, \\
E\left(Y_i^{(e)} \mid W_i^{(e)}, X_i^{(e)}\right) &= \mu_{W_i^{(e)}} + \beta^{\top} X_i^{(e)}, \quad \text{and} \\
Var\left(Y_i^{(r)} \mid W_i^{(r)}, T_i^{(r)}, X_i^{(r)}\right) &= Var\left(Y_i^{(e)} \mid W_i^{(e)}, X_i^{(e)}\right) = \phi^2.
\end{aligned} \tag{22}$$

In this model, conditional on the covariates X, the outcomes in the external and RCT control groups have identical mean and variance.

We evaluate the resulting harmonized estimator when the assumptions of model (22) are violated. In particular, we consider a *true* data-generating model with conditional moments

$$\begin{aligned}
E\left(Y_i^{(r)} \mid W_i^{(r)}, T_i^{(r)}, X_i^{(r)}\right) &= \mu_{W_i^{(r)}} + \theta_{W_i^{(r)}} T_i^{(r)} + \beta^{\top} X_i^{(r)}, \\
E\left(Y_i^{(e)} \mid W_i^{(e)}, X_i^{(e)}\right) &= \mu_{W_i^{(e)}} + \gamma_{W_i^{(e)}} + \beta^{\top} X_i^{(e)}, \quad \text{and} \\
Var\left(Y_i^{(r)} \mid W_i^{(r)}, T_i^{(r)}, X_i^{(r)}\right) &= Var\left(Y_i^{(e)} \mid W_i^{(e)}, X_i^{(e)}\right) = \phi^2.
\end{aligned} \tag{23}$$

This conditional distribution differs from model (22) and the parameters $\gamma_{1:K}$ allow the external and RCT controls to have different conditional means. Here, the SDM assumption is formalized by the following equality

$$\gamma_k = E\left(Y_i^{(e)} \mid W_i^{(e)} = k, X_i^{(e)} = x\right) - E\left(Y_i^{(r)} \mid W_i^{(r)} = k, X_i^{(r)} = x, T_i^{(r)} = 0\right) = \gamma, \qquad (24)$$

for all subgroups $k = 1, \ldots, K$ and some $\gamma \in \mathbb{R}$.

We evaluate the harmonized estimators conditioning on individual covariates X and treatment assignment. It can be shown under minimal design assumptions that $\hat{\theta}^{(r)}$ is an unbiased estimator of $\theta = \pi^\top \theta_{1:K}$ and that

$$\begin{aligned} Bias\left(\hat{\theta}_{1:K}^{(r+e)}, \theta_{1:K}\right) &= B\gamma_{1:K}, \quad \text{with} \\ B &= \begin{bmatrix} 0_{K\times K}, & I_K, & 0_{K\times d} \end{bmatrix} \left(M_1^\top M_1\right)^{-1} M_1^\top M_2, \end{aligned} \qquad (25)$$

where $[0_{K\times K}, I_K, 0_{K\times d}]$ is a block matrix of dimension $K \times (2K + d)$, I_K is the $K \times K$ identity matrix, and (M_1, M_2) form design matrices. In particular M_1 is the $n^{(r+e)} \times (2K + d)$ design matrix for model (22); it contains columns corresponding to $\mu_{1:K}$, $\theta_{1:K}$, and β. The full design matrix for model (23) is $[M_1, M_2]$; M_2 contains the columns corresponding to $\gamma_{1:K}$ and has dimension $n^{(r+e)} \times K$. Moreover,

$$\begin{aligned} \text{Var}\left(\hat{\theta}_{1:K}^{(r+e)}\right) &= \phi^2 \begin{bmatrix} 0_{K\times K}, & I_K, & 0_{K\times d} \end{bmatrix} \left(M_1^\top M_1\right)^{-1} \begin{bmatrix} 0_{K\times K}, & I_K, & 0_{K\times d} \end{bmatrix}^\top, \\ \text{Var}\left(\hat{\theta}^{(r)}\right) &= \phi^2 \left(M_0^\top M_0\right)_{2,2}^{-1}, \end{aligned} \qquad (26)$$

$$\text{Cov}\left(\hat{\theta}_{1:K}^{(r+e)}, \hat{\theta}^{(r)}\right) = \phi^2 \begin{bmatrix} 0_{K\times K}, I_K, 0_{K\times d} \end{bmatrix} \left(M_1^\top M_1\right)^{-1} M_1^\top \begin{bmatrix} I_{n^{(r)}}, 0_{n^{(e)}\times n^{(r)}} \end{bmatrix}^\top M_0 \left(M_0^\top M_0\right)^{-1} \begin{bmatrix} 0, 1, 0_{1\times d} \end{bmatrix}^\top, \qquad (27)$$

where M_0 is the $n^{(r)} \times (2 + d)$ design matrix for model (21).

We use these equations and the next three propositions to discuss $\hat{\theta}_{1:K}^b$. The propositions can be useful to examine harmonized estimators in settings beyond the linear model, as we will see in later subsections. We use the notation $S = \text{Var}(\hat{\theta}_{1:K}^{(r+e)}, \hat{\theta}^{(r)})$ and $u = c(\lambda/\lambda + c)\Sigma\pi$. The harmonized estimator in Equation (5) can be rewritten as

$$\hat{\theta}_{1:K}^b = \hat{\theta}_{1:K}^{(r+e)} + \left(\hat{\theta}^{(r)} - \pi^\top \hat{\theta}_{1:K}^{(r+e)}\right)u.$$

Proposition 2   If the following two assumptions

> (A1) $E(\hat{\theta}^{(r)}) - \pi^\top \theta_{1:K} = 0$, and
> (A2) $E(\hat{\theta}_{1:K}^{(r+e)}) - \theta_{1:K} = B\gamma_{1:K}$, where B is a $K \times K$ matrix,
>   hold, then

$$\text{Bias}\left(\hat{\theta}_{1:K}^b, \theta_{1:K}\right) = \left(I_K - u\pi^\top\right)B\gamma_{1:K} \qquad (28)$$

and

$$\text{Var}\left(\hat{\theta}_{1:K}^b\right) = PSP^\top, \qquad (29)$$

where $P = [I_K - u\pi^\top, u]$.

For the setting that we are discussing, with $\hat{\theta}_{1:K}^{(r+e)}$ and $\hat{\theta}^{(r)}$ based on models (22) and (21), respectively, and the data-generating model summarized by expression (23), this proposition allows us to derive the bias and variance of $\hat{\theta}_{1:K}^{b}$. Assumption (A1) holds under minimal design assumptions and (A2) is concordant with Equation (25). Thus, we only need to plug the values of ϕ, M_0, M_1, M_2, and $\gamma_{1:k}$ into Equations (25)–(29). Moreover, analytic expressions of the conditional bias and variance, analogous to Equations (28) and (29), can be easily obtained when $\hat{\theta}^{(r)} = \bar{Y}_{\cdot,1}^{(r)} - \bar{Y}_{\cdot,0}^{(r)}$ and/or for parameterizations of the data-generating model (23) and design matrices that violate the assumption (A1), because the harmonized estimator remains a linear transformation of the outcome data $Y_i^{(r)}$ and $Y_i^{(e)}$. The same comment holds if $\hat{\theta}^{(r)} = \pi^T \hat{\theta}_{1:K}^{(r)}$, with group-specific estimates of the treatment effects $\hat{\theta}_{1:K}^{(r)}$ obtained through OLS by restricting model (22) only to the RCT data.

The next result indicates for which fixed choice of λ and Σ the harmonized estimator is unbiased under the SDM assumption (i.e. the differences between the EC and RCT control groups are constant across subpopulations). We use the notation $b = (b_1, \ldots, b_K) = B1_{K \times 1}$.

> **Proposition 3** If assumptions (A1) and (A2) are satisfied, $\gamma_{1:K} = \gamma 1_{K \times 1}$ for some $\gamma \in \mathbb{R}$, and $b \neq 0_{K \times 1}$, then $\hat{\theta}_{1:K}^{b}$ is unbiased (i.e. $E(\hat{\theta}_{1:K}^{b}) = \theta_{1:K}$) if and only if $\lambda = \infty$ and $\Sigma \pi = \kappa b$ for some $\kappa \neq 0$.

This proposition can be used to construct a BD harmonized estimator that is unbiased under the SDM assumption (24). The main requirement is to solve the equation $\Sigma \pi = \kappa b$. We only need the design matrices and it is not necessary to know the regression coefficients of model (23). For example, if the signs of the elements of $b = (b_1, \ldots, b_K)$ are all the same, then taking $\Sigma = \text{diag}(|b_k|/\pi_k)$ solves the equation. Recall that b is proportional to the bias of $\hat{\theta}_{1:K}^{(r+e)}$ when $\gamma_{1:K} \propto 1_{K \times 1}$; therefore, one can expect the individual entries b_k to have identical signs. The next proposition indicates that we can choose Σ that makes $\hat{\theta}_{1:K}^{b}$ unbiased when $\gamma_{1:K} \propto 1_{K \times 1}$.

> **Proposition 4** There exists a positive-definite matrix $\Sigma \in \mathbb{R}^{K \times K}$ such that $\Sigma \pi = \kappa b$ for some $\kappa \neq 0$ if and only if $\pi^\top b \neq 0$.

If many positive matrices $\Sigma \in \mathbb{R}^{K \times K}$ solve the equation $\Sigma \pi = \kappa b$, with $\lambda = \infty$, they all define the same harmonized estimator. In other words, the BD harmonized estimator is unique. Indeed, when $\lambda = \infty$, the estimator $\hat{\theta}_{1:K}^{b}$ depends on Σ only through $(\pi^\top \Sigma \pi)^{-1} \Sigma \pi$ (see Equation (5)), which becomes $(\pi^\top b)^{-1} b$ if $\Sigma \pi = \kappa b$. In other words, the harmonized estimator that satisfies the equation $E(\hat{\theta}_{1:K}^{b}) = \theta_{1:K}$ is unique. We also note that the condition $\pi^\top b \neq 0$ in Proposition 4 is not particularly restrictive. For example, if the covariates $X_i^{(r)}$ and $X_i^{(e)}$ have a continuous distribution, then the event $\pi^\top b \neq 0$ has probability 1.

Figure 4 illustrates some simulations. We simulated data from the linear model (23), with $\gamma_{1:K}$ defined as in scenarios 1–3 in Figure 1. We generated a single covariate (i.e. $d = 1$), $X_i^{(r)} \overset{\text{ind}}{\sim} N(0, 1)$ for $i = 1, \ldots, n^{(r)}$ in the RCT and $X_i^{(e)} \overset{\text{ind}}{\sim} N(2, 1)$ for $i = 1, \ldots, n^{(e)}$ in the EC. The vector of covariate values was the same across simulation replicates. The regression coefficient is $\beta = 0.5$. Beyond the addition of the covariate X_i all simulation details (RCT design, EC design, and treatment effects) are the same as in Figure 1. We compare three estimators: (i) the pooled OLS estimator $\bar{\theta}_{1:K}^{(r+e)}$ based on model (22), (ii) the BD harmonized estimator that we described, using $\lambda = \infty$ and $\Sigma = \text{diag}(|b_k|/\pi_k)$ ('BD harmonized'), and (iii) the RCT-only OLS estimator $\hat{\theta}_{1:K}^{(r)}$ based on the subgroup-specific model (22) excluding the EC data. The harmonized estimator has variance comparable to the pooled estimator and it is more accurate than the RCT-only analysis in scenarios 1 and 2. In scenario 3, with $\gamma_{1:K} \not\propto 1_{K \times 1}$, as expected the harmonized estimator inherits the bias of the pooled estimator, although moderately reduced.

See online supplementary material Figure S3 (Section S2.8) for similar comparisons of these estimators that focus on the bias, variance, and RMSE when the trial size $n^{(r)} = 100$ is fixed (with $n_{\cdot,0}^{(r)} = 33$ and $n_{\cdot,1}^{(r)} = 67$), the EC sample size $n^{(e)}$ increases from 33 to 627 (i.e. $q = n_{\cdot,0}^{(e)}(n_{\cdot,0}^{(e)} + n_{\cdot,0}^{(r)})^{-1}$ increases from 0.5 to 0.95) and the SDM assumption is violated. The bias and RMSEs of
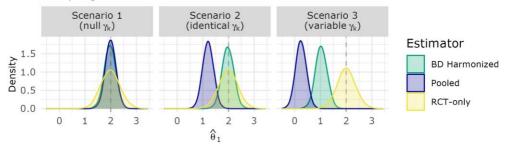
Sampling distributions of the estimators

Figure 4. Sampling distributions of three estimators of θ_1. The densities shown are kernel estimates based on 2,000 simulated trials per scenario. The dashed lines indicate the true value of θ_1. The scenarios present different configurations of $\gamma_{1:K}$, the subgroup-specific differences between the EC and RCT control mean outcomes conditional on X. These configurations are identical to those in Figure 1. Other than the addition of X_i all simulation details are the same as in Figure 1.

both the pooled and harmonized estimators increase with $n^{(e)}$. As expected this trend is less marked for the harmonized estimator compared to the pooled estimator. Importantly, when $n^{(e)}$ is large other definitions of $\hat{\theta}_{1:K}^{(r+e)}$, beyond pooling (e.g. Bayesian estimators based on hierarchical borrowing or power priors), can be used to reduce the outlined risks associated with distortion mechanisms that violate the SDM assumption.

In these simulations, we also consider the *cut* estimator (Equation (16)). Here, Analyst 1 uses model (21) and flat priors for μ, θ, and β, and Analyst 2 uses model (22) with flat priors for $\mu_{1:K}$, $\theta_{1:K}$, and β. As before, the mean of the resulting *cut distribution* is a VD harmonized estimator with $\lambda = \infty$ and Σ equal to Analyst 2's posterior variance of $\theta_{1:K}$. With flat priors $\hat{\theta}^{(r)} = E^{(A1)}(\theta \mid D^{(r)})$ and $\hat{\theta}_{1:K}^{(r+e)} = E^{(A2)}(\theta_{1:K} \mid D^{(r)}, D^{(e)})$ are identical to the OLS estimates discussed throughout this subsection. In our simulations the cut estimates are nearly identical ($R^2 > 0.99$, see online supplementary material) to the BD harmonized estimates in Figure 4.

3.2 Logistic regression

We now consider harmonized estimators for binary outcomes. The estimator $\hat{\theta}_{1:K}^{(r+e)}$ is based on the *pooled* logistic regression model:

$$
\begin{aligned}
p\left(Y_i^{(r)} = 1 \mid W_i^{(r)}, T_i^{(r)}, X_i^{(r)}\right) &= g\left(v_{W_i^{(r)}} + \eta_{W_i^{(r)}} T_i^{(r)} + \beta^\top X_i^{(r)}\right), \\
p\left(Y_i^{(e)} = 1 \mid W_i^{(e)}, X_i^{(e)}\right) &= g\left(v_{W_i^{(e)}} + \beta^\top X_i^{(e)}\right),
\end{aligned}
\tag{30}
$$

where $g(x) = 1/(1 + e^{-x})$, and the outcomes are independent. The maximum likelihood estimates (MLEs) are $(\hat{v}_{1:K}^{(r+e)}, \hat{\eta}_{1:K}^{(r+e)}, \hat{\beta}^{(r+e)})$. We consider the treatment effects

$$
\theta_k = E\left(Y_i^{(r)} \mid T_i^{(r)} = 1, W_i^{(r)} = k\right) - E\left(Y_i^{(r)} \mid T_i^{(r)} = 0, W_i^{(r)} = k\right),
\tag{31}
$$

and the estimates

$$
\hat{\theta}_k^{(r+e)} = \frac{1}{n_{k,\cdot}^{(r)}} \sum_{i:\, W_i^{(r)} = k} \left[g\left(\hat{v}_k^{(r+e)} + \hat{\eta}_k^{(r+e)} + \hat{\beta}^{(r+e)\top} X_i^{(r)}\right) - g\left(\hat{v}_k^{(r+e)} + \hat{\beta}^{(r+e)\top} X_i^{(r)}\right) \right].
\tag{32}
$$

We also specify a consistent estimator $\hat{\theta}^{(r)}$ of $\theta = \pi^\top \theta_{1:K}$. Then, we examine the resulting harmonized estimator (Equation (5)) when the individual records in $D^{(r)}$ and $D^{(e)}$ are independent replicates, with

$$p\left(Y_i^{(r)} = 1 \mid W_i^{(r)}, T_i^{(r)}, X_i^{(r)}\right) = g\left(v_{W_i^{(r)}} + \eta_{W_i^{(r)}} T_i^{(r)} + \beta^\top X_i^{(r)}\right), \quad \text{and}$$
$$p\left(Y_i^{(e)} = 1 \mid W_i^{(e)}, X_i^{(e)}\right) = g\left(v_{W_i^{(e)}} + \delta_{W_i^{(e)}} + \beta^\top X_i^{(e)}\right). \tag{33}$$

Expression (33) summarizes the true outcome distributions and expression (30) describes the working model used to compute $\hat{\theta}_{1:K}^{(r+e)}$. If the parameters δ_k are different from zero, then the RCT and EC controls have different subgroup-specific response rates after conditioning on the covariates X. Here, the SDM assumption is formalized by the equality

$$\delta_k = g^{-1}\left(E\left(Y_i^{(e)} \mid X_i^{(e)} = x, \, W_i^{(e)} = k\right)\right) - g^{-1}\left(E\left(Y_i^{(r)} \mid X_i^{(r)} = x, \, W_i^{(r)} = k, \, T_i^{(r)} = 0\right)\right) = \delta, \tag{34}$$

for all subgroups $k = 1, \ldots, K$ and some $\delta \in \mathbb{R}$.

We provide results similar to those discussed for the linear model in Section 3.1. Here, we consider the asymptotic properties of a sequence of harmonized estimates, when the sample sizes increase:

(B1) Both $n^{(r)} \to \infty$ and $n^{(e)} \to \infty$. The ratio $n^{(e)}/n^{(r+e)}$ converges to a value in $(0, 1)$. Also, $n_{k,1}^{(r)}/n_{\cdot,1}^{(r)}$, $n_{k,0}^{(r)}/n_{\cdot,0}^{(r)}$, and $n_{k,0}^{(e)}/n^{(e)}$ converge to values in $(0, 1)$ for each $k = 1, \ldots, K$.

(B2) We also assume that $(\hat{v}_{1:K}^{(r+e)}, \hat{\eta}_{1:K}^{(r+e)}, \hat{\beta}^{(r+e)}, \hat{\theta}_{1:K}^{(r+e)})$ converges in probability to some point $(v_{1:K}^\circ, \eta_{1:K}^\circ, \beta^\circ, \theta_{1:K}^\circ)$, which might differ from the true $(v_{1:K}, \eta_{1:K}, \beta, \theta_{1:K})$. See the online supplementary material for mild conditions that ensure convergence.

Given $(X_i^{(r)}, W_i^{(r)}, T_i^{(r)})_{i \geq 1}$, $(X_i^{(e)}, W_i^{(e)})_{i \geq 1}$ and the limit of $n^{(e)}/n^{(r+e)}$, the vector $\theta_{1:K}^\circ$ can be represented as a function of the parameters $(v_{1:K}, \eta_{1:K}, \beta, \delta_{1:K})$ in model (33). For instance, when $\delta_{1:K} = 0_{K \times 1}$, then the working model (30) is correctly specified and $\theta_{1:K}^\circ = \theta_{1:K}$ by standard maximum likelihood arguments. For other values of $\delta_{1:K}$ close to $0_{K \times 1}$, we have the Taylor approximation

$$\theta_{1:K}^\circ = \theta_{1:K} + B\delta_{1:K} + r_1(\delta_{1:K}), \tag{35}$$

where $B = \frac{\partial \theta_{1:K}^\circ}{\partial \delta_{1:K}}$ is the Jacobian matrix, and the remainder r_1 satisfies

$$\lim_{\delta_{1:K} \to 0_{K \times 1}} \frac{r_1(\delta_{1:K})}{\|\delta_{1:K}\|_1} = 0_{K \times 1}. \tag{36}$$

We approximate the difference $\theta_{1:K}^\circ - \theta_{1:K}$ with a linear transformation of $\delta_{1:K}$. The approximation is analogous to Equation (25). To select Σ (in Equation (5)) and compute the harmonized estimator $\hat{\theta}_{1:K}^b$, we will use the factorization

$$B = \frac{\partial \theta_{1:K}^\circ}{\partial(v_{1:K}^\circ, \eta_{1:K}^\circ, \beta^\circ)} \frac{\partial(v_{1:K}^\circ, \eta_{1:K}^\circ, \beta^\circ)}{\partial \delta_{1:K}}. \tag{37}$$

Analytic expressions for computing the derivatives are included in the online supplementary material.

The next result, similar to Proposition 2, focuses on $\hat{\theta}_{1:K}^b$ when the sample sizes diverge. Here, $\hat{\theta}_{1:K}^b$ is computed using a data-dependent matrix $\hat{\Sigma}$, the value assigned to the regularization parameter Σ in Equation (5). As before, $u = c(\lambda/\lambda + c)\Sigma\pi$, with $c = (\pi^\top \Sigma \pi)^{-1}$.

Proposition 5 If assumption (B1) holds, and the sequence of designs ensures that for any value of $\delta_{1:k}$ in model (33)

(B3) $\hat{\theta}^{(r)} \xrightarrow{p} \pi^{\top}\theta_{1:K}$,

(B4) $\hat{\theta}^{(r+e)}_{1:K} \xrightarrow{p} \theta_{1:K} + B\delta_{1:K} + r_1(\delta_{1:K})$, where $r_1(\delta_{1:K})$ satisfies Equation (36), and

(B5) $\hat{\Sigma} \xrightarrow{p} \Sigma$, where Σ is a positive-definite $K \times K$ matrix, then

$$\hat{\theta}^b_{1:K} \xrightarrow{p} \theta_{1:K} + (I_K - u\pi^{\top})B\delta_{1:K} + r_2(\delta_{1:K}),$$

where the approximation error $r_2(\delta_{1:K})$ satisfies

$$\lim_{\delta_{1:K} \to 0_{K\times 1}} \frac{r_2(\delta_{1:K})}{\|\delta_{1:K}\|_1} = 0_{K\times 1}.$$

In other words, the bias of $\hat{\theta}^b_{1:K}$ is approximately equal to $(I_K - u\pi^{\top})B\delta_{1:K}$. This allows us to select λ and Σ so that the harmonized estimator $\hat{\theta}^b_{1:K}$ is nearly unbiased under the SDM assumption (34). In particular, when $B1_{K\times 1} \neq 0_{K\times 1}$, the estimator $\hat{\theta}^b_{1:K}$ with parameters (λ, Σ) is approximately unbiased if (i) $\lambda = \infty$ and (ii) $\Sigma\pi = \kappa B1_{K\times 1}$ for some $\kappa \neq 0$. In practice, we can compute a consistent estimate $\hat{B}$ of B and select $\hat{\Sigma}$ such that

$$\hat{\Sigma}\pi = \kappa\hat{B}1_{K\times 1}. \tag{38}$$

It is only necessary that $\pi^{\top}\hat{B}1_{K\times 1} \neq 0$ (see Proposition 4). With $\lambda = \infty$ all the positive matrices $\hat{\Sigma}$ that solve Equation (38) produce the same harmonized estimator. Note that if we fix κ, then the difference between the convex set of solutions of Equation (38) and the solutions of $\Sigma\pi = \kappa B1_{K\times 1}$ vanishes when the sample sizes increase. To estimate B we first compute the MLEs $(\hat{\nu}^{(r)}_{1:K}, \hat{\eta}^{(r)}_{1:K}, \hat{\beta}^{(r)})$ of model (30) without using EC data and then plug these estimates into the right-hand side of expression (37). This does not require Monte Carlo simulations or other approximation methods (see the online supplementary material). The analytic computation of $\hat{B}$ utilizes the joint empirical distributions of covariates X and biomarkers W in the RCT and EC datasets.

Although here we focused on logistic regression, the approach to choose $\hat{\Sigma}$ (see Equations (5) and (38)) that we described can also be used in other settings. Consider, for example, count outcome data Y. We can proceed as follows:

(a) We specify a working model—say, a Poisson model with assumptions and covariate-outcome relationships similar to the GLM in expression (30)—that leverages the EC and RCT data, to compute $\hat{\theta}^{(r+e)}_{1:K}$. We also compute a consistent estimator $\hat{\theta}^{(r)}$ of the overall treatment effect θ.

(b) We specify a data-generating model, say an extended Poisson model with subgroup-specific bias terms $\delta_{1:K}$ having the same interpretation as in expression (33). Under mild assumptions (e.g. *iid* replicates of biomarkers, covariates, and treatment in the RCT and external group), the asymptotic difference $\theta^\circ_{1:K} - \theta_{1:K}$ is well defined. This difference can be viewed as a function of $\delta_{1:K}$ and is null if $\delta_{1:K} = 0_{K\times 1}$.

(c) Therefore, as before, we can use the Taylor approximation $\theta^\circ_{1:K} - \theta_{1:K} \approx B\delta_{1:K}$, compute an estimate $\hat{B}$ of B, find a positive definite matrix $\hat{\Sigma}$ that solves $\hat{\Sigma} = \hat{B}1_{K\times 1}$, and specify $\Sigma = \hat{\Sigma}$ in Equation (5).

In this Poisson example, and in other GLMs, $\hat{B}$ can be computed analytically using a nearly identical formula to the one that we use for logistic regression (see the online supplementary material).

There are other strategies to select Σ. For example, as suggested by the Bayesian interpretation in Section 2.3, one may choose Σ to be an estimate of the variance of $\hat{\theta}^{(r+e)}_{1:K}$, giving a VD harmonized
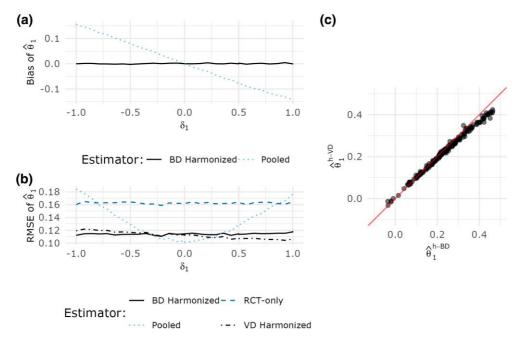
Figure 5. Logistic regression and harmonized estimates. The results are based on 2,000 replicates per scenario. (a) Bias when we estimate θ_1 using the pooled estimator and the BD harmonized estimator as a function of δ_1. (b) The root mean squared error of four estimators. (c) Scatter plot comparing the BD harmonized estimator (denoted $\hat{\theta}_1^{(h-BD)}$) and the VD harmonized estimator (denoted $\hat{\theta}_1^{(h-VD)}$) across a random subsample of 200 replicates with $\delta = 1_{K \times 1}$ and $\theta_1 = 0.23$.

estimator. In our implementations, we estimate this variance using the Fisher information of $(v_{1:K}, \eta_{1:K}, \beta)$ under model (30).

Figure 5 illustrates a simulation study. We generate outcome data from model (33), with $\delta_{1:K} = (\delta, \ldots, \delta)$ for $\delta \in [-1, 1]$. This is a scenario with systematic bias across subgroups (SDM assumption holds); the online supplementary material includes a scenario with bias terms $\delta_{1:K}$ that vary across subgroups. We consider $K = 5$ subgroups. We generate a single covariate $X_i^{(r) \text{ind}} \sim N(0, 1)$ in the RCT and $X_i^{(e) \text{ind}} \sim N(2, 1)$ in the EC, with the regression coefficient $\beta = 0.2$. The other details of the simulations are similar to the scenarios used for Figure 4. We compare four estimators of $\theta_{1:K}$: (i) the pooled estimator $\hat{\theta}_{1:K}^{(r+e)}$, using MLE based on model (30); (ii) an RCT-only estimator denoted $\hat{\theta}_{1:K}^{(r)}$, using MLEs based on model (30) without the EC data; (iii) the BD harmonized estimator denoted $\hat{\theta}_{1:K}^{h-B}$, with $\lambda = \infty$ and Σ selected using Equation (38); and (iv) the VD harmonized estimator denoted $\hat{\theta}_{1:K}^{h-V}$, with $\lambda = \infty$ and Σ equal to the estimate of the variance of $\hat{\theta}_{1:K}^{(r+e)}$. Figure 5a shows that the bias of $\hat{\theta}_1^{(r+e)}$ is approximately linear in δ_1, and based on this fact harmonization can reduce the bias of $\hat{\theta}_1^{(r+e)}$ under the SDM assumption (see Equation (34)). Figure 5b shows that the relative performance of the estimators depends on δ_1. Figure 5c shows that harmonization with Σ chosen to reduce bias (BD harmonization) or Σ chosen as the estimated variance of $\hat{\theta}_{1:K}^{(r+e)}$ (VD harmonization) produce similar estimates in the simulation scenarios that we considered. Online supplementary material Figure S5 illustrates scenarios in which the SDM assumption is violated ($\delta_{1:K} = (\delta - 0.5, \delta + 0.5, \delta - 0.5, \delta + 0.5, \delta - 0.5)$ with $\delta \in [-1, 1]$) and the harmonized estimator presents lower RMSE than the RCT-only estimator despite postharmonization bias.

3.3 Propensity score weighting and harmonization
Statistical methods from the causal inference literature (Li et al., 2022) can be useful for combining RCT and EC data. We discuss the harmonization (Equation (1)) of subgroup-specific effect

estimates obtained through propensity score weighting to leverage RCT and EC data. In Section 4, we use this approach to analyse GBM data.

Here, we consider an initial estimate $\hat{\theta}_{1:K}^{(r+e)}$ based on an inverse probability weighting scheme (Lunceford & Davidian, 2004; Tipton, 2013). Other approaches, such as marginal structural models (Cole & Hernan, 2008) and direct standardization (Robins, 1986; Snowden et al., 2011), can be used. We refer to Li et al. (2022) and other reviews for discussions of the principles, assumptions, and sensitivity of relevant causal inference methods. We consider again binary outcomes, with the subgroup-specific treatment effects defined in Equation (31) and the overall treatment effect defined as $\theta = \pi^{\top}\theta_{1:K}$.

To compute $\hat{\theta}_{1:K}^{(r+e)}$, we fit the logistic regression model (30), using weighted maximum likelihood, and apply Equation (32) as in Section 3.2 to estimate treatment effects. The individual log-likelihood contribution is multiplied by a weight $w_i^{(r)}$ or $w_i^{(e)}$, for RCT and EC patients, respectively. In our implementation, the weights are $w_i^{(r)} = 1$ for RCT patients and

$$w_i^{(e)} = \zeta \frac{\hat{\rho}\left(W_i^{(e)}, X_i^{(e)}\right)}{1 - \hat{\rho}\left(W_i^{(e)}, X_i^{(e)}\right)} \tag{39}$$

for EC patients, where $\hat{\rho}(W_i^{(e)}, X_i^{(e)})$ is an estimate of the propensity score that we compute through standard logistic regression as in Tipton (2013). The constant ζ is selected to have $\max_i w_i^{(e)} \leq 1$, i.e. the RCT patients are weighted more than the the patients in the EC group. The ratio $\hat{\rho}(W, X)/(1 - \hat{\rho}(W, X))$ in Equation (39) has a straightforward interpretation. Indeed, if we multiply it by $n^{(e)}/n^{(r)}$, we obtain an estimate of the ratio of the densities at potential (W, X) values in the RCT and EC populations.

To harmonize $\hat{\theta}_{1:K}^{(r+e)}$, we compute the treatment effect estimate $\hat{\theta}^{(r)}$, which is based on RCT data only. We use $\lambda = \infty$, and Σ is selected with minor changes to the procedure discussed in Section 3.2:

(a) The working model to compute $\hat{\theta}_{1:K}^{(r+e)}$ remains (30). Recall that we assign weights (39) to the EC patients to compute the estimates $(\hat{v}_{1:K}^{(r+e)}, \hat{\eta}_{1:K}^{(r+e)}, \hat{\beta}^{(r+e)})$.

(b) We build on the same assumptions as in Section 3.2, including the data-generating model (33). We focus on the parameterization $(\hat{v}_{1:K}^{(r)}, \hat{\eta}_{1:K}^{(r)}, \hat{\beta}^{(r)})$, with unknown bias components $\delta_{1:K}$. The estimates $(\hat{v}_{1:K}^{(r)}, \hat{\eta}_{1:K}^{(r)}, \hat{\beta}^{(r)})$ are based on *RCT-only* regression. Under this data-generating model $\theta_{1:K}^{\circ}$, the limit of $\hat{\theta}_{1:K}^{(r+e)}$, is a function of $\delta_{1:K}$.

The limit $\theta_{1:K}^{\circ}$ can be computed for any $(v_{1:K}^{(r)}, \eta_{1:K}^{(r)}, \beta^{(r)}, \delta_{1:K})$ configuration by maximizing the weighted log-likelihood function with $4 \times n^{(r)} + 2 \times n^{(e)}$ terms: 2 terms ($Y_i^{(e)} = 0$ or 1) for each patient in the EC group (weights: the individual $w_i^{(e)}$ weights and the conditional probability that $Y_i^{(e)} = 0$ or 1 are multiplied) and 4 ($T_i^{(r)} = 0$ or 1, and $Y_i^{(r)} = 0$ or 1) for each patient in the RCT group (weights: $w_i^{(r)} = 1$, the conditional probabilities that $Y_i^{(r)} = 0$ or 1, and the probability of $T_i^{(r)} = 0$ or 1 are multiplied). This approach assumes that the distributions of biomarkers and covariates (X, W) in the RCT and EC groups match the available empirical estimates.

(c) Lastly, we use the Taylor approximation $\theta_{1:K}^{\circ} - \theta_{1:K} \approx B\delta_{1:K}$, where $B = \partial\theta_{1:K}^{\circ}/\partial\delta_{1:K}$. We compute $\hat{B}$ through a weighted version of the analytic expressions that we used in Section 3.2 (see online supplementary material). As before, we select a matrix Σ that solves the equation $\Sigma\pi = \kappa\hat{B}1_{K\times 1}$. Recall that with $\lambda = \infty$ any Σ that solves the equation (i.e. BD harmonization) defines the same harmonized estimator.

Alternatively, the analyst can choose Σ equal to the estimate of the variance of $\hat{\theta}_{1:K}^{(r+e)}$, using popular sandwich estimators (Lunceford & Davidian, 2004) or bootstrap procedures. See online supplementary material Figure S3 (Section S2.8) for similar comparisons of these estimators that focus on the bias, variance, and RMSE when the trial size $n^{(r)} = 100$ is fixed (with $n_{\cdot,0}^{(r)} = 33$

and $n_{.1}^{(r)} = 67$), the EC sample size $n^{(e)}$ increases from 33 to 627 (i.e. $q = n_{.0}^{(e)}(n_{.0}^{(e)} + n_{.0}^{(r)})^{-1}$ increases from 0.5 to 0.95) and the SDM assumption is violated. The bias and RMSEs of both the pooled and harmonized estimators increase with $n^{(e)}$. As expected this trend is less marked for the harmonized estimator compared to the pooled estimator. Importantly, when $n^{(e)}$ is large other definitions of $\hat{\theta}_{1:K}^{(r+e)}$, beyond pooling (e.g. Bayesian estimators based on hierarchical borrowing or power priors), can be used to reduce the outlined risks associated with distortion mechanisms that violate the SDM assumption.

4 Analysis of GBM data

We apply the harmonized estimator described in Section 3.3 to a GBM trial. We include *real-world* data in our analyses (Ventz et al., 2019). The data collection that we use has been previously described (Rahman et al., 2023). We evaluate the harmonized estimator through a model-free resampling scheme (Avalos-Pacheco et al., 2023; Ventz et al., 2022). The scheme is similar to bootstrapping and it allows us to generate *in silico* RCTs and EC datasets. Using this resampling scheme we can explore the operating characteristics of different methods and subgroup analyses.

Subgroup analysis is an important task in GBM trials. Preclinical data often suggest that the effects of experimental treatments might vary substantially across patient subgroups. Also, experimental treatments might benefit only some groups. See Trippa and Alexander (2017) and references therein for an extensive discussion. As an early example, we mention the Stupp et al. (2005) phase III trial. The final analyses of this large trial provided evidence that temozolomide combined with radiation therapy (TMZ + RT), which is the current standard of care in GBM, improves survival compared to radiation therapy (RT) alone. Subsequent subgroup analyses (Stupp et al., 2009) supported the hypothesis that the treatment benefit of TMZ + RT compared to RT was limited to patients with MGMT-negative methylation status.

Data collection: In our analyses we used data from the phase III AVAGLIO trial (profiles and outcomes of 352 newly diagnosed GBM patients treated with TMZ + RT) and larger real world data set from Dana-Farber Cancer Institute (including 532 newly diagnosed GBM patients treated with TMZ + RT). We previously described these data collections in Rahman et al. (2023) and Ventz et al. (2022). We use individual-level patient data. We consider as the primary outcome 12-month survival, i.e. $Y_i = 1$ if the patient is alive 12 months after treatment initiation and $Y_i = 0$ otherwise. For patients with a follow-up of less than 12 months we imputed their 12-month survival status using a Cox model (see online supplementary material Section S4.1 for details). In our subgroup analyses we have $K = 4$ nonoverlapping subpopulations based on the patient's MGMT methylation status and their Karnofsky performance status (KPS) before treatment. The four subgroups are as follows: ($W_i = 1$) MGMT-negative and $KPS < 90$, ($W_i = 2$) MGMT-negative and $KPS \geq 90$, ($W_i = 3$) MGMT-positive and $KPS < 90$, and ($W_i = 4$) MGMT-positive and $KPS \geq 90$. The four subgroups have proportions (0.20, 0.47, 0.10, 0.23) in the trial and (0.27, 0.30, 0.19, 0.24) in the real-world dataset, respectively. Relevant pretreatment covariates used in our analyses include age, sex, and extent of tumor resection prior to the treatment. See Ventz et al. (2019) and references therein for a discussion on relevant prognostic variables in GBM. Based on the likelihood ratio test described in online supplementary material Section S3.8, we did not find evidence of potential violations of the SDM assumption (*p*-value of 0.336).

Resampling: We use a resampling procedure (Avalos-Pacheco et al., 2023; Ventz et al., 2022) similar to bootstrapping to generate *in silico* GBM RCTs. Specifically, we generate 1, 000 *in silico* trials repeating the following steps:

(a) *Control arm*. Sample $n_{.0}^{(r)} = 100$ patient records ($Y_i^{(r)}$, $W_i^{(r)}$, $X_i^{(r)}$) with replacement from the actual control arm of the clinical trial.

(b) *Experimental arm*. Sample $n_{.1}^{(r)} = 200$ patient records ($Y_i^{(r)}$, $W_i^{(r)}$, $X_i^{(r)}$) with replacement again from the actual control arm of the clinical trial.

(c) *Treatment effects*. Since we sampled the same group of patients treated with TMZ + RT to generate the *in silico* control and experimental arms, in our computer experiments the treatment effects are $\theta_{1:K} = (0, 0, 0, 0)$. Positive treatment effects can be spiked in by extending the survival in the *in silico* experimental arm (Ventz et al., 2019).
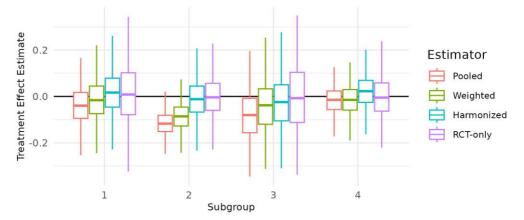
Figure 6. Boxplots illustrating the distribution of the $\theta_{1:K}$ estimates across the 1,000 *in silico* GBM trials simulated as we described.

(d) *External data.* Sample $n_{.,0}^{(e)} = 600$ patient records $(Y_i^{(e)}, W_i^{(e)}, X_i^{(e)})$ from the actual EC dataset.

(e) *Estimation.* Compute the subgroup-specific treatment effect estimates. We compare four estimators:

   (i) The *pooled* logistic regression estimator (Section 3.2, Equation (32)).

  (ii) The propensity score *weighted* logistic regression estimator from Section 3.3, with weights defined in Equation (39).

 (iii) The *harmonized* estimator. We harmonize the weighted estimator as proposed in Section 3.3. We use $\lambda = \infty$ and specify Σ equal to the identity matrix (see Equation (5)).

 (iv) An *RCT-only* estimator based on maximum likelihood under model (30), without EC data.

We chose $n_{.,0}^{(r)} = 100$, $n_{.,1}^{(r)} = 200$, and $n^{(e)} = 600$, realistic sample sizes close to the sizes of the source data sets.

Figure 6 shows the distribution of the $\theta_{1:K}$ estimates across the *in silico* RCTs. In all four subgroups, the pooled estimator is the most biased. The propensity score weighted estimator slightly reduces this bias, and harmonization further reduces the bias. We repeated the analyses implementing the data-dependent selection of Σ to reduce bias as discussed in Section 3.3. The bias, compared to harmonization with fixed Σ equal to the identity matrix, was reduced in three of the four subgroups, with reductions up to 90%, while it remained nearly identical in one of the subgroups. In addition, the propensity score weighted and harmonized estimators have similar variability. In particular, the standard deviations of the weighted and harmonized estimators are, respectively, 0.092 and 0.097 in subgroup 1, 0.060 and 0.081 in subgroup 2, 0.125 and 0.122 in subgroup 3, and 0.065 and 0.073 in subgroup 4. Harmonization in our computer experiments produces estimates with lower variability compared to the RCT-only estimator, which is unbiased. For further details on bias, variance, and RMSE in these simulations, see online supplementary material Table S1 (in Section S4.2).

5 Discussion

Subgroup analyses have become an essential component of testing the efficacy and safety of new experimental treatments in precision medicine. These analyses allow clinicians to identify subpopulations that respond to novel therapies. However, conducting clinical trials with sample sizes and rigorous statistical plans for accurate subgroup analyses is a challenging and costly task. Compared to RCTs designed to estimate the average treatment effect in the overall study population, trials designed to estimate subgroup-specific treatment effects require substantially larger sample sizes.

These challenges may be partially addressed by allowing subgroup analyses to incorporate external control data from past trials and electronic health records. This can be particularly valuable for rare diseases and rare patient subpopulations. However, variations in the definitions of prognostic variables across studies, unmeasured confounders, different measurement standards across institutions, and other mechanisms can bias the research findings when external datasets augment RCT data.

In this work, we introduced harmonized estimators for subgroup analysis that balance potential efficiency gains from the integration of EC data and the risk of bias due to confounding and other distortion mechanisms. The harmonized estimation procedure has two components: (i) an initial estimator $\hat{\theta}_{1:K}^{(r+e)}$ for the subgroup-specific treatment effects that leverages EC data and (ii) an estimator $\hat{\theta}^{(r)}$ for the average treatment effect in the RCT across subgroups that does not involve the use of external information. The proposed harmonization shrinks the initial subgroup-specific estimates $\hat{\theta}_{1:K}^{(r+e)}$ towards a linear subspace of potential subgroup-specific treatment effects with a prevalence-weighted average concordant with the estimate $\hat{\theta}^{(r)}$ obtained using only the RCT data. Any initial estimator $\hat{\theta}_{1:K}^{(r+e)}$ can be used and we provide strategies for selecting the parameters λ and Σ. For example, throughout this article, we discuss parameterizations chosen with the goal of obtaining harmonized estimates $\hat{\theta}_{1:K}^{b}$ that are unbiased or approximately unbiased when the SDM assumption holds. In Section 2.1, we described how under some regularity conditions this BD harmonized estimator can be viewed as a modified version of the estimator $\hat{\theta}_{1:K}^{(r+e)}$ obtained by estimating and subtracting the bias $\psi_{1:K} = E(\hat{\theta}_{1:K}^{(r+e)}) - \theta_{1:K}$ under the SDM assumption. Similarly, BD harmonization of estimators $\hat{\theta}_{1:K}^{(r+e)}$ based on GLMs (Section 3.2) subtracts from $\hat{\theta}_{1:K}^{(r+e)}$ an estimate of $\psi_{1:K}$. Compared to other strategies to reduce or remove bias due to SDMs from subgroup analyses (see online supplementary material Section S3.9 for an example), harmonization ensures coherence between subgroup analyses and the primary analysis (i.e. $\pi^{\top}\hat{\theta}_{1:K}^{b} = \hat{\theta}^{(r)}$).

We discussed harmonized estimators based on various outcome models, including MLE for the linear and logistic regression models and procedures from the causal inference literature. We provided sufficient conditions to obtain BD harmonized treatment effect estimates. Using patient-level data from brain cancer studies, we showed that, under realistic scenarios, the harmonized estimator can reduce the mean squared error of the subgroup-specific treatment effect estimates by up to 50% compared to estimates that do not leverage external data. Moreover, our comparisons based on brain cancer trial data showed a reduction in bias via harmonization of up to 86% compared to treatment effects estimates that use propensity score weighting to combine external and RCT data.

In our work, we focused on harmonization with $\lambda = \infty$ to ensure that the harmonized estimates $\hat{\theta}_{1:K}^{b}$ satisfy $\pi^{\top}\hat{\theta}_{1:K}^{b} = \hat{\theta}^{(r)}$. Moreover, taking $\lambda = \infty$ and selection of Σ based on Proposition 3 or Equation (38), allows analysts to obtain a harmonized estimator that is unbiased (or approximately unbiased) when the SDM assumption holds. Alternatively, taking Σ as an estimate of the variance of $\hat{\theta}_{1:K}^{(r+e)}$ has a Bayesian justification.

Harmonization is a procedure applicable to any subgroup analysis that involves RCT data and ECs, including estimators $\hat{\theta}_{1:K}^{(r+e)}$ beyond those discussed in this article. In online supplementary material Section S3.7, we describe additional simulations in which harmonization improves MSE and reduces bias of subgroup-specific estimates $\hat{\theta}_{1:K}^{(r+e)}$ obtained with (i) a Bayesian causal forest model (Hahn et al., 2020), (ii) a hierarchical model with borrowing of information across subgroups (Jones et al., 2011), and (iii) a propensity score adjusted power prior procedure (Lin et al., 2019). In these simulations, we considered the same scenarios as in Sections 3.1 and 3.2 and computed $\hat{\theta}_{1:K}^{(r+e)}$ (the posterior mean of $\theta_{1:K}$), which was then modified following the proposed harmonization approach. The comparisons of the operating characteristics pre- and postharmonization remained similar to what we illustrated throughout Section 3. In these cases, it is challenging to analytically derive the bias of $\hat{\theta}_{1:K}^{(r+e)}$, but we can compute BD harmonized estimators. This can be achieved by estimating the bias of $\hat{\theta}_{1:K}^{(r+e)}$ with resampling procedures (see online supplementary material Section S3.10). We also point interested readers to online supplementary material Section S5, where we discuss the bias of $\hat{\theta}_{1:K}^{(r+e)}$ and $\hat{\theta}_{1:K}^{b}$ in the presence of unmeasured

predictors. These results show that various estimators can be harmonized, and that harmonization can improve their operating characteristics. In addition, harmonization can reduce the bias of $\hat{\theta}_{1:K}^{(r+e)}$ in scenarios in which the model used for estimating the treatment effects $\theta_{1:K}$ is misspecified.

Conflicts of interest: None declared.

Funding

Data availability

Data used for the case study can be analysed upon request via the GBM-X Data Platform (https://rconnect.dfci.harvard.edu/gbmdata). In the supplementary code folder, we include a synthetic version of the case study data which can be used to replicate the analyses in Section 4.

Supplementary material

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

References

Avalos-Pacheco A., Ventz S., Arfè A., Alexander B. M., Rahman R., Wen P. Y., & Trippa L. (2023). Validation of predictive analyses for interim decisions in clinical trials. *JCO Precision Oncology*, 7(7), e2200606. https://ascopubs.org/doi/10.1200/PO.22.00606. https://doi.org/10.1200/PO.22.00606

Bayarri M. J., Berger J. O., & Liu F. (2009). Modularization in Bayesian analysis, with emphasis on analysis of computer models. *Bayesian Analysis*, 4(1), 119–150. https://projecteuclid.org/journals/bayesian-analysis/volume-4/issue-1/Modularization-in-Bayesian-analysis-with-emphasis-on-analysis-of-computer/10.1214/09-BA404.full. https://doi.org/10.1214/09-BA404

Chamberlain G., & Leamer E. E. (1976). Matrix weighted averages and posterior bounds. *Journal of the Royal Statistical Society: Series B (Methodological)*, 38(1), 73–84. https://onlinelibrary.wiley.com/doi/10.1111/j.2517-6161.1976.tb01569.x. https://doi.org/10.1111/j.2517-6161.1976.tb01569.x

Cole S. R., & Hernan M. A. (2008). Constructing inverse probability weights for marginal structural models. *American Journal of Epidemiology*, 168(6), 656–664. https://academic.oup.com/aje/article-lookup/doi/10.1093/aje/kwn164. https://doi.org/10.1093/aje/kwn164

Corrigan-Curay J., Sacks L., & Woodcock J. (2018). Real-world evidence and real-world data for evaluating drug safety and effectiveness. *JAMA*, 320(9), 867. http://jama.jamanetwork.com/article.aspx?doi=10.1001/jama.2018.10136. https://doi.org/10.1001/jama.2018.10136

Dabush D. R., Shepshelovich D., Shochat T., Tibau A., Amir E., & Goldvaser H. (2021). The impact of radiological assessment schedules on progression-free survival in metastatic breast cancer: A systemic review and meta-analysis. *Cancer Treatment Reviews*, 100(Suppl 2), 102293. https://linkinghub.elsevier.com/retrieve/pii/S0305737221001419. https://doi.org/10.1016/j.ctrv.2021.102293

Di Maio M., Perrone F., & Conte P. (2020). Real-world evidence in oncology: Opportunities and limitations. *The Oncologist*, 25(5), e746–e752. https://academic.oup.com/oncolo/article/25/5/e746/6444361. https://doi.org/10.1634/theoncologist.2019-0647

Dijkman B., Kooistra B., & Bhandari M. (2009). How to work with a subgroup analysis. *Canadian Journal of Surgery*, 52(6), 515–522. https://www.canjsurg.ca/content/52/6/515

Efron B., & Tibshirani R. (1986). Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Statistical Science*, 1(1), 54–75. https://projecteuclid.org/journals/statistical-science/volume-1/issue-1/Bootstrap-Methods-for-Standard-Errors-Confidence-Intervals-and-Other-Measures/10.1214/ss/1177013815.full. https://doi.org/10.1214/ss/1177013815

Ghosal S., & van der Vaart A. (2017). *Fundamentals of nonparametric Bayesian inference* (1st ed.). Cambridge University Press. https://www.cambridge.org/core/product/identifier/9781139029834/type/book

Hahn P. R., Murray J. S., & Carvalho C. M. (2020). Bayesian regression tree models for causal inference: Regularization, confounding, and heterogeneous effects (with discussion). *Bayesian Analysis*, 15(3), 965–1056. https://projecteuclid.org/journals/bayesian-analysis/volume-15/issue-3/Bayesian-Regression-Tree-Models-for-Causal-Inference--Regularization-Confounding/10.1214/19-BA1195.full. https://doi.org/10.1214/19-BA1195

Hobbs B. P., Carlin B. P., & Sargent D. J. (2013). Adaptive adjustment of the randomization ratio using historical control data. *Clinical Trials*, 10(3), 430–440. http://journals.sagepub.com/doi/10.1177/1740774513483934. https://doi.org/10.1177/1740774513483934

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://projecteuclid.org/journals/bayesian-analysis/volume-7/issue-3/Commensurate-Priors-for-Incorporating-Historical-Information-in-Clinical-Trials-Using/10.1214/12-BA722.full. https://doi.org/10.1214/12-BA722

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://onlinelibrary.wiley.com/doi/10.1002/sim.6728. https://doi.org/10.1002/sim.6728

Jacob P. E., Murray L. M., Holmes C. C., & Robert C. P. (2017). 'Better together? Statistical learning in models made of modules', arXiv, arXiv:1708.08719 [stat], preprint: not peer reviewed. https://doi.org/10.48550/arXiv.1708.08719

Jones H. E., Ohlssen D. I., Neuenschwander B., Racine A., & Branson M. (2011). Bayesian models for subgroup analysis in clinical trials. *Clinical Trials*, 8(2), 129–143. http://journals.sagepub.com/doi/10.1177/1740774510396933. https://doi.org/10.1177/1740774510396933

Li F., Buchanan A. L., & Cole S. R. (2022). Generalizing trial evidence to target populations in non-nested designs: Applications to AIDS clinical trials. *Journal of the Royal Statistical Society: Series C, Applied Statistics*, 71(3), 669–697. https://academic.oup.com/jrsssc/article/71/3/669/7067595. https://doi.org/10.1111/rssc.12550

Lim J., Walley R., Yuan J., Liu J., Dabral A., Best N., Grieve A., Hampson L., Wolfram J., Woodward P., Yong F., Zhang X., & Bowen E. (2018). Minimizing patient burden through the use of historical subject-level data in innovative confirmatory clinical trials: Review of methods and opportunities. *Therapeutic Innovation & Regulatory Science*, 52(5), 546–559. http://link.springer.com/10.1177/2168479018778282. https://doi.org/10.1177/2168479018778282

Lin J., Gamalo-Siebers M., & Tiwari R. (2019). Propensity-score-based priors for Bayesian augmented control design. *Pharmaceutical Statistics*, 18(2), 223–238. https://onlinelibrary.wiley.com/doi/10.1002/pst.1918. https://doi.org/10.1002/pst.1918

Lunceford J. K., & Davidian M. (2004). Stratification and weighting via the propensity score in estimation of causal treatment effects: A comparative study. *Statistics in Medicine*, 23(19), 2937–2960. https://onlinelibrary.wiley.com/doi/10.1002/sim.1903. https://doi.org/10.1002/sim.1903

Morita S., & Müller P. (2017). Bayesian population finding with biomarkers in a randomized clinical trial. *Biometrics*, 73(4), 1355–1365. https://onlinelibrary.wiley.com/doi/10.1111/biom.12677. https://doi.org/10.1111/biom.12677

Pletcher M. J., & McCulloch C. E. (2017). The challenges of generating evidence to support precision medicine. *JAMA Internal Medicine*, 177(4), 561. http://archinte.jamanetwork.com/article.aspx?doi=10.1001/jamainternmed.2016.9138. https://doi.org/10.1001/jamainternmed.2016.9138

Rahman R., Ventz S., Redd R., Cloughesy T., Alexander B. M., Wen P. Y., & Trippa L. (2023). Accessible data collections for improved decision making in neuro-oncology clinical trials. *Clinical Cancer Research*, 29(12), 2194–2198. https://aacrjournals.org/clincancerres/article/29/12/2194/727078/Accessible-Data-Collections-for-Improved-Decision. https://doi.org/10.1158/1078-0432.CCR-22-3524

Robins J. (1986). A new approach to causal inference in mortality studies with a sustained exposure period-application to control of the healthy worker survivor effect. *Mathematical Modelling*, 7(9-12), 1393–1512. https://linkinghub.elsevier.com/retrieve/pii/0270025586900886. https://doi.org/10.1016/0270-0255(86)90088-6

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: Robust meta-analytic-predictive priors. *Biometrics*, 70(4), 1023–1032. https://onlinelibrary.wiley.com/doi/10.1111/biom.12242. https://doi.org/10.1111/biom.v70.4

Sivaganesan S., Müller P., & Huang B. (2017). Subgroup finding via Bayesian additive regression trees. *Statistics in Medicine*, 36(15), 2391–2403. https://onlinelibrary.wiley.com/doi/10.1002/sim.7276. https://doi.org/10.1002/sim.7276

Snowden J. M., Rose S., & Mortimer K. M. (2011). Implementation of G-computation on a simulated data set: Demonstration of a causal inference technique. *American Journal of Epidemiology*, 173(7), 731–738. https://academic.oup.com/aje/article-lookup/doi/10.1093/aje/kwq472. https://doi.org/10.1093/aje/kwq472

Stupp R., Hegi M. E., Mason W. P., van den Bent M. J., Taphoorn M. J., Janzer R. C., Ludwin S. K., Allgeier A., Fisher B., Belanger K., Hau P., Brandes A. A., Gijtenbeek J., Marosi C., Vecht C. J., Mokhtari K., Wesseling P., Villa S., Eisenhauer E., … Mirimanoff R.-O. (2009). Effects of radiotherapy with concomitant and adjuvant temozolomide versus radiotherapy alone on survival in glioblastoma in a randomised phase III study: 5-year analysis of the EORTC-NCIC trial. *The Lancet: Oncology*, 10(5), 459–466. https://linkinghub.elsevier.com/retrieve/pii/S1470204509700257. https://doi.org/10.1016/S1470-2045(09)70025-7

Stupp R., Mason W. P., van den Bent M. J., Weller M., Fisher B., Taphoorn M. J., Belanger K., Brandes A. A., Marosi C., Bogdahn U., Curschmann J., Janzer R. C., Ludwin S. K., Gorlia T., Allgeier A., Lacombe D., Cairncross J. G., Eisenhauer E., & Mirimanoff R. O. (2005). Radiotherapy plus concomitant and adjuvant temozolomide for glioblastoma. *New England Journal of Medicine*, 352(10), 987–996. http://www.nejm.org/doi/abs/10.1056/NEJMoa043330. https://doi.org/10.1056/NEJMoa043330

Sully B. G. O., Julious S. A., & Nicholl J. (2013). A reinvestigation of recruitment to randomised, controlled, multicenter trials: A review of trials funded by two UK funding agencies. *Trials*, 14(1), 166. https://trialsjournal.biomedcentral.com/articles/10.1186/1745-6215-14-166. https://doi.org/10.1186/1745-6215-14-166

Sun X., Briel M., Busse J. W., You J. J., Akl E. A., Mejza F., Bala M. M., Bassler D., Mertz D., Diaz-Granados N., Vandvik P. O., Malaga G., Srinathan S. K., Dahm P., Johnston B. C., Alonso-Coello P., Hassouneh B., Walter S. D., Heels-Ansdell D., … Guyatt G. H. (2012). Credibility of claims of subgroup effects in randomised controlled trials: Systematic review. *BMJ*, 344(mar15 1), e1553–e1553. https://www.bmj.com/lookup/doi/10.1136/bmj.e1553. https://doi.org/10.1136/bmj.e1553

Tipton E. (2013). Improving generalizations from experiments using propensity score subclassification: Assumptions, properties, and contexts. *Journal of Educational and Behavioral Statistics*, 38(3), 239–266. http://journals.sagepub.com/doi/10.3102/1076998612441947. https://doi.org/10.3102/1076998612441947

Trippa L., & Alexander B. M. (2017). Bayesian baskets: A novel design for biomarker-based clinical trials. *Journal of Clinical Oncology: Official Journal of the American Society of Clinical Oncology*, 35(6), 681–687. https://doi.org/10.1200/JCO.2016.68.2864

Tung N. M., Robson M. E., Ventz S., Santa-Maria C. A., Nanda R., Marcom P. K., Shah P. D., Ballinger T. J., Yang E. S., Vinayak S., Melisko M., Brufsky A., DeMeo M., Jenkins C., Domchek S., D'Andrea A., Lin N. U., Hughes M. E., Carey L. A., … Garber J. E. (2020). TBCRC 048: Phase II study of olaparib for metastatic breast cancer and mutations in homologous recombination-related genes. *Journal of Clinical Oncology*, 38(36), 4274–4282. https://ascopubs.org/doi/10.1200/JCO.20.02151. https://doi.org/10.1200/JCO.20.02151

Ventz S., Comment L., Louv B., Rahman R., Wen P. Y., Alexander B. M., & Trippa L. (2022). The use of external control data for predictions and futility interim analyses in clinical trials. *Neuro-Oncology*, 24(2), 247–256. https://academic.oup.com/neuro-oncology/article/24/2/247/6295505. https://doi.org/10.1093/neuonc/noab141

Ventz S., Lai A., Cloughesy T. F., Wen P. Y., Trippa L., & Alexander B. M. (2019). Design and evaluation of an external control arm using prior clinical trials and real-world data. *Clinical Cancer Research*, 25(16), 4993–5001. https://aacrjournals.org/clincancerres/article/25/16/4993/125037/Design-and-Evaluation-of-an-External-Control-Arm. https://doi.org/10.1158/1078-0432.CCR-19-0820

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://onlinelibrary.wiley.com/doi/10.1002/pst.1589. https://doi.org/10.1002/pst.1589

Wang R., Lagakos S. W., Ware J. H., Hunter D. J., & Drazen J. M. (2007). Statistics in medicine—reporting of subgroup analyses in clinical trials. *New England Journal of Medicine*, 357(21), 2189–2194. http://www.nejm.org/doi/abs/10.1056/NEJMsr077003. https://doi.org/10.1056/NEJMsr077003