Biometrika (2025), **xx**, x, *p*. 1
Printed in Great Britain

Leveraging External Data for Testing Experimental Therapies with Biomarker Interactions in Randomized Clinical Trials

By B. Ren

Laboratory for Psychiatric Biostatistics, McLean Hospital, Belmont, Massachusetts 02478, U.S.A.

bren@mgb.org

F. Ferrari

Biostatistics and Research Decision Sciences, Merck & Co, Rahway, New Jersey 07065, U.S.A.

federico.ferrari@merck.com

S. Fortini

Department of Decision Sciences, Bocconi University, via Röntgen 1, 20136 Milan, Italy

sandra.fortini@unibocconi.it

S. Ventz

Division of Biostatistics, University of Minnesota, Minneapolis, Minnesota 55414, U.S.A.

ventz001@umn.edu

and L. Trippa

Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, Massachusetts 02115, U.S.A.

ltrippa@jimmy.harvard.edu

Summary

In oncology the efficacy of novel therapeutics often differs across patient subgroups, and these variations are difficult to predict during the initial phases of the drug development process. The relation between the power of randomized clinical trials (RCTs) and heterogeneous treatment effects (HTEs) has been discussed by several authors. In particular, false negative results are likely to occur when the treatment effects concentrate in a subpopulation but the study design did not account for potential HTEs. The use of external data (ED) from completed clinical studies and electronic health records has the potential to improve decision-making throughout the development of new therapeutics, from early-stage trials to registration. Here we discuss the use of ED to evaluate experimental treatments with potential HTEs. We introduce a permutation procedure to test, at the completion of a RCT, the null hypothesis that the experimental therapy does not improve the primary outcomes in any subpopulation. The permutation test leverages the available ED to increase power. Also, the procedure controls the false positive rate at the desired

α-level without restrictive assumptions on the ED, for example, in scenarios with unmeasured confounders, different pre-treatment patient profiles in the RCT population compared to the ED, and other discrepancies between the trial and the ED. We illustrate that the permutation test is optimal according to an interpretable criteria and discuss examples based on asymptotic results and simulations, followed by a retrospective analysis of individual patient-level data from a collection of glioblastoma clinical trials.

Some key words: Bayesian statistics; Decision theory; Glioblastoma; Heterogeneous treatment effect; Permutation test

1. Introduction

During the last decade the characterization of oncogenic alterations and resistance mechanisms have been the basis of a rapid increase in the experimental treatments available for clinical testing in cancer research (Haslam et al., 2021). Novel therapeutics often target patient subpopulations defined by somatic mutations or other biomarkers, and their efficacy can vary across patient subgroups. In the early phases of clinical investigation it is usually unclear and difficult to predict if the experimental therapy improves relevant outcomes, for example survival, and in which subpopulations. Dedicated trial designs (Freidlin et al., 2010, 2012; Ziegler et al., 2012) have been proposed to estimate subgroup-specific treatment effects and to improve decision-making, including go/no-go decisions (Chu & Yuan, 2018) and eligibility of registration studies (Xu et al., 2023) accounting for potential heterogeneous treatment effects (HTEs) across subpopulations. Also, statistical methods for inference on treatment effects and their variations across subgroups have been studied extensively in the literature. We mention the use of treatment effects pattern plots (Bonetti & Gelber, 2004), permutation-based algorithms (Wang et al., 2015), randomization-based procedures (Ding et al., 2016, 2019), Bayesian model averaging to account for multiplicity issues (Berger et al., 2014), and tree-based approaches (Wager & Athey, 2018), among others. Moreover, particular efforts have been made to identify patient subgroups that benefit from the experimental treatment (Rigdon et al., 2018; Wager & Athey, 2018). For example, Morita & Müller (2017) discussed a decision-theoretic solution to identify subgroups, and Rigdon et al. (2018) used regression trees to capture HTEs.

A fundamental requirement for investigating HTEs in randomized clinical trials (RCTs) is the inclusion of more subjects compared to RCTs that ignores HTEs (Yang et al., 2020). However, the sample sizes of most oncology RCTs continue to be chosen for assessing the average treatment effects and are thus inadequate for accurate inference on interactions between treatments and biomarkers. Additionally, in many settings, such as rare diseases, it is not practical or feasible, due to limited resources, to conduct large RCTs that capture treatment effect's variations across subgroups (Nugent et al., 2021). These limitations impact major decisions during the drug development process, including the decision to discontinue the development of the experimental treatment and the choice of the eligibility criteria for registration trials based on previous data from early phase studies.

To mitigate the outlined limitations, we propose a novel method for the analysis of RCTs with potential HTEs, that integrates external data (ED), including individual patient-level information (IPLI) from completed RCTs or *real-world* datasets (RWD; Sherman et al., 2016), such as *electronic health record* (EHR) collected for administrative purposes. This method can improve and accelerate the development of new therapeutics, by leveraging diverse data sources to facilitate the transition from early-phase trials to confirmatory studies, and supporting the critical decision to continue or terminate the investigation of the experimental treatment. The integration of ED in the design and analysis of clinical trials has received substantial attention in oncology.

In particular, recent contributions on the use of external control data in oncology suggest the potential of accelerating the development of new treatments (Rahman et al., 2021; Liau et al., 2023). Moreover, these contributions might facilitate the use of unbalanced randomization ratios (e.g., 1:2 or 1:3 for the control and experimental arms) because the ED, as well as the control arm of the trial, provide information on the control therapy.

We discuss a permutation procedure that incorporates a *Bayesian working model* and augments the data from a RCT with ED. We focus on a single task: assessing if the RCT data provide evidence of positive treatment effects in some of the patients based on hypothesis testing. Our test controls the false positive rate at the desired α-level. Importantly, the control of false positives does not require any assumption on the ED; it covers scenarios with model misspecification, unmeasured confounders, different pre-treatment patient profiles in the trial population compared to the ED, and other discrepancies between the trial and the ED. Our permutation procedure is based on test statistics with straightforward Bayesian interpretations, including popular summaries of the evidence of treatment effects. We will provide a decision-theoretic (Berger, 2013) justification of the permutation procedure, illustrating the optimality of the test according to an easy-to-interpret criteria. Through stylized examples and a data-driven simulation study, we illustrate the properties of our approach, emphasizing its robustness, the control of false positive results, and potential power improvements compared to popular testing procedures the utilize only the RCT data.

2. Method

2.1. *Notation*

We consider a RCT that randomizes n patients with ratio $1 : r$ to the experimental and control arms. We indicate the RCT data (*internal data*, ID) with $\mathcal{D} = (Y, X, A)$, where $Y = (Y_1, \ldots, Y_n)$ are the outcomes, $X = (X_1, \ldots, X_n)$ are pre-treatment patient characteristics ($X_i \in \mathbb{R}^d$), and $A = (A_1, \ldots, A_n)$ are treatment assignment variables ($A_i = 1$ and $A_i = 0$ for the experimental and control treatments). Similarly, $\mathcal{D}_E = (Y_E, X_E, A_E)$ indicates the ED with n_E patients. To simplify the presentation $X_{E,i}$ will include the same variables as X_i, although this assumption can be straightforwardly relaxed. In some cases $\mathcal{D}_E$ includes only patients who received the control therapy (i.e., $A_{E,i} = 0$ for all i), for example $\mathcal{D}_E$ might represent the data from the control arm of a previous RCT that evaluated a different experimental therapy. In other cases $\mathcal{D}_E$ might include both patients treated with the experimental treatment and the control therapy in earlier clinical studies.

We use p and p_E to indicate the unknown distributions of $\mathcal{D}$ and $\mathcal{D}_E$. If the ID and the ED include independent and identically distributed (*iid*) replicates, then $p(y_i|x_i, a_i)$ and $p_E(y_{E,i}|x_{E,i}, a_{E,i})$ are the distributions of the individual outcome $y_i, y_{E,i} \in \mathbb{R}$ conditional on treatment $a_i, a_{E,i} \in \{0, 1\}$ and pre-treatment profile $x_i, x_{E,i} \in \mathbb{R}^d$ in the RCT and the ED. These conditional distributions might be different. We assume that the treatment assignment variable A_i and pre-treatment patient characteristics X_i are independent, and that the random variables $A_1, \ldots, A_n$ are independent or exchangeable.

Our goal is to test the effects of the experimental treatment in the RCT population. We will consider the following null hypothesis,

$$H_0 : p(y, x, a) \text{ is invariant to permutations of } a,\ \forall (x, y) \in \mathbb{R}^{n\times(d+1)} \text{ and } a \in \{0, 1\}^n. \quad (1)$$

In other words, if H_0 holds, then for any configuration of outcomes $y = (y_1, \ldots, y_n) \in \mathbb{R}^n$ and covariates $x = [x_1, \ldots, x_n] \in \mathbb{R}^{n\times d}$ we have $p(y, x, a) = p(y, x, a')$, for every permutation a' of $a \in \{0, 1\}^n$. We use $\mathcal{P}$ and $\mathcal{P}_0 \subset \mathcal{P}$ to indicate the set of potential distributions of $\mathcal{D} \in \mathbb{R}^{n\times(d+1)} \times \{0, 1\}^n$ and the subset of distributions concordant with H_0. The null hypothesis H_0

implies $\mathbb{E}_p(Y_i|X_i = x_i, A_i = 1) = \mathbb{E}_p(Y_i|X_i = x_i, A_i = 0)$ for all $x_i \in \mathbb{R}^d$. Here $\mathbb{E}_p$ indicates the expectation with respect to the unknown distribution $p \in \mathcal{P}$. The alternative hypothesis will be $p \in \mathcal{P} \setminus \mathcal{P}_0$.

To specify the test we define a randomized decision rule (Lehmann & Romano, 2005) $\phi : \mathbb{R}^{n\times(d+1)} \times \{0,1\}^n \to [0,1]$ that rejects H_0 with probability $\phi(\mathcal{D}) \in [0,1]$. We consider randomized decisions ϕ mainly for analytic convenience. We will then discuss a similar non-randomized decision rule $\tilde{\phi}(\mathcal{D}) \in \{0,1\}$ for practical use. In the next paragraphs $\mathbb{E}_p[\phi(\mathcal{D})] = \int \phi(\mathcal{D})dp(\mathcal{D})$ is the probability of rejecting H_0. Also, for any $\alpha \in [0,1]$, the function ϕ is an α-level test if $\mathbb{E}_p[\phi(\mathcal{D})] \leq \alpha$ for every $p \in \mathcal{P}_0$. We summarize the notation in Table 1.

Variable	Notation	Definition
Outcome, pre-treatment covariates and treatment in the RCT	Y_i, X_i, A_i	$X_i \in \mathbb{R}^d, A_i \in \{0,1\}$
RCT data (ID) with sample size n	$\mathcal{D} = (Y, X, A)$	$Y = (Y_1, \ldots, Y_n)$, $X = (X_1, \ldots, X_n)$, $A = (A_1, \ldots, A_n)$
Outcome, pre-treatment covariates and treatment in the ED	$Y_{E,i}, X_{E,i}, A_{E,i}$	$X_{E,i} \in \mathbb{R}^{d_E}$
External data (ED) with sample size n_E	$\mathcal{D}_E = (Y_E, X_E, A_E)$	$Y_E = (Y_{E,1}, \ldots, Y_{E,n_E})$, $X_E = (X_{E,1}, \ldots, X_{E,n_E})$, $A_E = (A_{E,1}, \ldots, A_{E,n_E})$
Unknown distributions of $\mathcal{D}$ and $\mathcal{D}_E$	p, p_E	
Randomized and non-random decision rules	$\phi(\mathcal{D}), \tilde{\phi}(\mathcal{D})$	$\phi : \mathbb{R}^n \times \mathbb{R}^{n\times d} \times \{0,1\}^n \to [0,1]$ $\tilde{\phi} : \mathbb{R}^n \times \mathbb{R}^{n\times d} \times \{0,1\}^n \to \{0,1\}$
The working model $\mathcal{M}$ specifies conditional distributions in the RCT and external group with parameters θ	$q_\theta, q_{E,\theta}$	
Prior distribution and the conditional distribution, given the ED, of the parameters θ	$\pi(\cdot), \pi(\theta\|\mathcal{D}_E)$	$\pi(\theta\|\mathcal{D}_E) = q_{E,\theta}(\mathcal{D}_E)\pi(\theta)/\int_\theta q_{E,\theta}(\mathcal{D}_E)\pi(\theta)$
Conditional likelihood of the ID given the ED	$m(\mathcal{D})$	$m(\mathcal{D}) = \int q_\theta(Y\|X, A)\pi(\theta\|\mathcal{D}_E)d\theta$

Table 1: Frequently used notation.

2.2. *An ED Augmented Permutation Test (ED-PT) for RCTs with HTEs*

We developed our testing procedure for the null hypothesis H_0 in (1) with the following goals:

1. The test is tailored to settings with potential HTEs. For instance, several oncology studies enroll patients from various biomarker subgroups, and some groups are more likely to benefit from the experimental treatment than others (Dar et al., 2021). Moreover, there are often biological arguments to expect stronger treatment effects in some subgroups than in others (Lauko et al., 2022).
2. The procedure incorporates information from ED with the aim of increasing the power compared to other testing procedures that do not use ED.
3. The control of the false positive rate at level α is robust with respect to model misspecification and potential discrepancies between the distribution of the ID (p) and the ED (p_E). In particular, the control of the type I error rate is preserved when the conditional distributions $p(\cdot|x_i, a_i)$ and $p_E(\cdot|x_{E,i}, a_{E,i})$ are different.

Working Model. To achieve these aims, we use a working model $\mathcal{M}$ for the ID and the ED, with outcome distributions conditional on pre-treatment profiles and treatments denoted by $q_\theta(y_i|x_i, a_i)$ and $q_{E,\theta}(y_{E,i}|x_{E,i}, a_{E,i})$, respectively. Throughout the manuscript the *true* unknown distributions (p, p_E) and the conditional densities $(q_\theta, q_{E,\theta})$ of the working model $\mathcal{M}$ will repeatedly appear together in the same paragraphs. The model $\mathcal{M}$ is parametrized by $\theta \in \Theta$ and embeds the assumption of conditionally independent outcomes in the RCT and the ED, that is $q_\theta(y|x, a) = \prod_i q_\theta(y_i|x_i, a_i)$ and $q_{E,\theta}(y_E|x_E, a_E) = \prod_i q_{E,\theta}(y_{E,i}|x_{E,i}, a_{E,i})$. The model $\mathcal{M}$ can incorporate HTEs. For example, $\mathcal{M}$ can be a linear regression model including effects of pre-treatment patient characteristics X_i ($X_{E,i}$ for the ED) and treatment A_i ($A_{E,i}$ for the ED) together with interactions. Here the parameters θ include the regression coefficients and the outcome variance. In this example the unknown distributions p and p_E might deviate from the linearity assumptions. We can specify the model $\mathcal{M}$ with identical or distinct regression functions in the ID and ED. Also, we do not require $\mathcal{M}$ to be a simple parametric model, and allow the use of semi-parametric or non-parametric models. Moreover, $\mathcal{M}$ can include any individual pre-treatment information, such as the date of diagnosis or the institution where the patient was enrolled.

We use the Bayesian framework to define the test statistics. We first specify a prior distribution π on the parameter space Θ of $\mathcal{M}$. Then, we summarize the information from the ED through the conditional distribution $\pi(\theta|\mathcal{D}_E) \propto q_{E,\theta}(Y_E|X_E, A_E)\pi(\theta)$. The model $\mathcal{M}$ and the prior distribution π incorporate prior belief on (i) covariate-outcome relationships in the ID and the ED, (ii) potential treatment-biomarker interactions, and (iii) the level of similarity between the regression functions $\mathbb{E}_p(Y_i|X_i, A_i)$ and $\mathbb{E}_{p_E}(Y_{E,i}|X_{E,i}, A_{E,i})$ in the ID and ED. Investigators may use a model $\mathcal{M}$ with identical conditional outcome distributions $q_\theta(\cdot|x, a) = q_{E,\theta}(\cdot|x, a)$ for all $(x, a) \in \mathbb{R}^d \times \{0, 1\}$, or more flexible solutions, such as Bayesian hierarchical models, to allow for different regression functions for the patients treated with the same therapy in the RCT and the ED. Differences between the conditional outcome distributions in these two groups may arise from several factors, such as variations in measurement technologies and treatment schedules of the therapy (Slevin et al., 1989).

Definition of the test statistics. The test statistics is

$$m(\mathcal{D}) = \int q_\theta(Y|X, A)\pi(\theta|\mathcal{D}_E)d\theta. \tag{2}$$

Consider for example a hierarchical Bayesian model $(\mathcal{M}, \pi)$ that includes separate parameters $\theta = (\theta_I, \theta_E)$ for the RCT and the ED, a prior distribution on hyper-parameters ν and conditionally independent group-specific parameters $\theta_I, \theta_E|\nu \overset{iid}{\sim} \pi(\cdot \mid \nu)$. The conditional outcome distributions, given treatments and covariates, in the RCT and the ED are parameterized by θ_I and θ_E respectively. In this case,

$$\begin{aligned} m(\mathcal{D}) &= \frac{\int q_{\theta_I}(Y|X, A)q_{E,\theta_E}(Y_E|X_E, A_E)\pi(\theta_I|\nu)\pi(\theta_E|\nu)\pi(\nu)d\theta_I d\theta_E d\nu}{\int q_{E,\theta_E}(Y_E|X_E, A_E)\pi(\theta_E|\nu)\pi(\nu)d\theta_E d\nu} \\ &= \int q_{\theta_I}(Y|X, A)\pi(\theta_I|\mathcal{D}_E)d\theta_I. \end{aligned} \tag{3}$$

Closed-form expressions for the integrals in (2) and (3) are not necessary, and approximation methods (e.g., importance sampling) to compute these quantity can be used.

The randomized test $\phi(\mathcal{D})$. Let $\tau = (\tau_1, \ldots, \tau_n)$ be a permutation of $(1, 2, \ldots, n)$, $\mathcal{T}$ the set of all permutations, and $A^{(\tau)} = (A_{\tau_1}, \ldots, A_{\tau_n})$. We denote with t_α the $(1 - \alpha)$-quantile of the set $\{m(\mathcal{D}^{(\tau)}); \tau \in \mathcal{T}\}$, where $\mathcal{D}^{(\tau)} = (Y, X, A^{(\tau)})$. We define the randomized test $\phi(\mathcal{D})$ at level α

as

$$\phi(\mathcal{D}) = \begin{cases} 1 & \text{if } m(\mathcal{D}) > t_\alpha, \\ 0 & \text{if } m(\mathcal{D}) < t_\alpha, \\ \frac{\alpha n! - \sum_{\tau\in\mathcal{T}} \mathbb{I}[m(\mathcal{D}^{(\tau)}) > t_\alpha]}{\sum_{\tau\in\mathcal{T}} \mathbb{I}[m(\mathcal{D}^{(\tau)}) = t_\alpha]} & \text{if } m(\mathcal{D}) = t_\alpha, \end{cases} \tag{4}$$

where $\mathbb{I}(\cdot)$ is the indicator function.

A decision-theoretic justification of the permutation test. The Bayesian expected power (BEP; Brown et al. 1987; Liu 2018) of a test ϕ' with respect to $\pi(\theta|\mathcal{D}_E)$ is

$$\text{BEP}(\phi') = \mathbb{E}_{(X,A)\sim p}\left[\int\left(\int \phi'(Y,X,A) q_\theta(Y|X,A)dY\right)\pi(\theta|\mathcal{D}_E)d\theta\right]. \tag{5}$$

The randomized test ϕ in expression (4) has maximal BEP. Proposition 1 states the optimality result. A proof is provided in Section SM1 in the supplementary materials.

Proposition 1. *The permutation test $\phi(\mathcal{D})$ defined in* (4) *has level α and maximizes the BEP in* (5) *among all α-level tests of H_0.*

The non-randomized test $\tilde{\phi}(\mathcal{D})$. Randomized tests are rarely used in practice and enumeration over the set $\mathcal{T}$ of all permutations is typically infeasible even for moderate sample sizes n. We slightly modify $\phi(\mathcal{D})$ in `Algorithm 1` to obtain a practical non-randomized test $\tilde{\phi}(\mathcal{D}) \in \{0,1\}$. We generate J random permutations τ from $\mathcal{T}$, and use the proportion of these permutations with $m(Y,X,A^{(\tau)}) \geq m(\mathcal{D})$ as a p-value approximation.

Algorithm 1. Non-randomized permutation test with level α.

Input: The number of permutations J, the ID $\mathcal{D} = (Y,X,A)$, the working model $\mathcal{M}$, and the conditional distribution $\pi(\theta|\mathcal{D}_E)$
$m(\mathcal{D}) \leftarrow \int_\theta q_\theta(Y|X,A)\pi(\theta|\mathcal{D}_E)d\theta$
For $j = 1$ to $j = J$:
  $\tau \leftarrow$ a random sample from $\mathcal{T}$
  $m_j \leftarrow \int_\theta q_\theta(Y|X,A^{(\tau)})\pi(\theta|\mathcal{D}_E)d\theta$
$\tilde{\phi}(\mathcal{D}) \leftarrow \mathbb{I}\left\{\frac{1+\sum_1^J \mathbb{I}[m_j \geq m(\mathcal{D})]}{1+J} \leq \alpha\right\}$
Output: $\tilde{\phi}(\mathcal{D})$

The proof of Proposition 1 allows us to notice four properties of the ED-PT:

1. False positives are also controlled at the α level in relevant scenarios where the samples of the ID and/or the ED are not independent and identically distributed, for example, when the RCT population varies over time (e.g., Russo et al., 2023; Kennedy et al., 2017).
2. The control of false positive results for ϕ at the α level is maintained if the test statistic $m(\mathcal{D})$ is computed using approximation methods (e.g., importance sampling).
3. The randomized and non-randomized tests (i.e., ϕ and $\tilde{\phi}$) have nearly identical type I and type II error rates when J diverges.
4. The proposed ED-PT is applicable when the external dataset includes both patients treated with the experimental and control therapies.

We refer to Section SM2 in the supplementary materials for a discussion of these properties.

2.3. *One-sided Testing*

The testing procedure that we introduced in the previous subsection does not distinguish between positive and negative treatment effects. In different words, the test rejects H_0 with high probability when the experimental therapy is inferior compared to the control. In many settings there are strong arguments to exclude the possibility of negative effects of the experimental intervention. We can mention for example an experimental plan, with frequent text and email reminders, to improve adherence to the recommendations of a cancer prevention program. However, in some trials the experimental treatment may have negative effects. For instance the treatment may reduce survival due to treatment-related toxicities. In such cases our ED-PT may reject H_0 due to negative effects.

Several variations of the permutation procedure (`Algorithm 1`) allow the user to (i) reject H_0 when some of the patients benefit from the experimental treatment and (ii) control the likelihood of rejecting H_0 when the effects are absent or negative. These variations are based on test statistics with simple Bayesian interpretations. We describe two modified versions of the testing procedure:

(i) The first one replaces $m(\mathcal{D})$ in (2) with

$$\tilde{m}_1(\mathcal{D}) = \int_{\tilde{\Theta}} \pi(\theta|\mathcal{D}, \mathcal{D}_E) d\theta. \tag{6}$$

Here we restrict the integral to a subset $\tilde{\Theta} \subset \Theta$ of the parameter space. For example we can restrict integration to a subset $\tilde{\Theta}$ with positive and clinically relevant effects for at least one subgroup of patients. We can modify the permutation procedure in `Algorithm 1`, using the statistics $\tilde{m}_1(\mathcal{D})$ instead of $m(\mathcal{D})$.

(ii) Alternatively, we can define $\tilde{m}_2(\mathcal{D})$ as the expected regret (i.e., the difference in expected utility) between (i) the optimal policy (Murphy, 2003) that treats every patient i with the best available therapy $\arg\max_{a\in\{0,1\}} \mathbb{E}_p(Y_i|X_i, A_i = a)$, and (ii) the policy that assigns every patient to the control therapy. The Bayesian working model can be used for inference on the optimal policy. Let $\tilde{a}_i(\theta) = \arg\max_{a\in\{0,1\}} \mathbb{E}_{q_\theta}(Y_i|X_i, A_i = a)$ and define the utility $\frac{1}{n}\sum_{i=1}^n \mathbb{E}_{q_\theta}[Y_i|X_i, A_i = \tilde{a}_i(\theta)]$, an interpretable function of θ and $X = (X_1, \ldots, X_n)$, which we integrate with respect to the posterior of θ, conditional on the RCT and EC data. The resulting integral is a summary of the efficacy of the optimal policy in the trial population, and it can be compared to the policy that treats every patient with the control therapy using the following statistics:

$$\tilde{m}_2(\mathcal{D}) = \int \frac{1}{n}\sum_{i=1}^n \left\{\mathbb{E}_{q_\theta}[Y_i|X_i, A_i = \tilde{a}_i(\theta)] - \mathbb{E}_{q_\theta}(Y_i|X_i, A_i = 0)\right\} d\pi(\theta \mid \mathcal{D}, \mathcal{D}_E). \tag{7}$$

2.4. *ED-PT with binary outcomes*

We consider a RCT with binary outcomes and 1:r randomization (experimental vs. control). For simplicity we do not include covariates in this example. Additional examples with potential treatment-biomarker interactions and negative treatment effects are discussed in Section SM6 in the supplementary materials. The ED include patients treated with the control therapy, i.e. $A_{E,i} = 0$ for all i. We first describe the unknown data distributions, p and p_E (ID and ED):

$$\begin{gathered} A_i \overset{iid}{\sim} \text{Bernoulli}[1/(1+r)], \quad Y_i|A_i \overset{ind}{\sim} \text{Bernoulli}(w_0 + \gamma A_i), \\ Y_{E,i} \overset{iid}{\sim} \text{Bernoulli}(w_0 + \beta_0). \end{gathered} \tag{8}$$

Here w_0 is the response rate in the control arm, $\gamma \in [-w_0, 1 - w_0]$ and $\beta_0 \in [-w_0, 1 - w_0]$ indicate the treatment effect and the mean difference between the outcomes in the internal (i.e.,

within the RCT) and external control groups. We set $n = 100$, $n_E = 500$, $r = 0.5$ and $w_0 = 0.5$. To investigate the type I error rate and power of the test we consider $\gamma = 0$ and $\gamma = 0.25$. We vary β_0 from -0.1 to 0.1 to examine the robustness of the test procedure with respect to discrepancies between the control arm in the RCT and the ED.

Working model $\mathcal{M}$. We use a Beta-Bernoulli working model:

$$Y_i|A_i = a, \theta_1, \theta_0 \overset{ind}{\sim} \text{Bernoulli}(\theta_a), \quad Y_{E,i}|\theta_1, \theta_0 \overset{ind}{\sim} \text{Bernoulli}(\theta_0), \\ \theta_a \overset{iid}{\sim} U[0,1], \quad a \in \{0,1\}. \tag{9}$$

The model parameters are $\theta = (\theta_0, \theta_1)$, the response rates in the control and experimental arms. Here $U[0,1]$ indicates the uniform distribution on $[0,1]$, and conditional on θ, the outcomes are independent. The model $\mathcal{M}$ assumes identical response rates in the internal and external control groups. $\mathcal{M}$ is misspecified when $\beta_0 \neq 0$. Let $n_1 = \sum_i A_i$, $n_0 = n - n_1$, $s_1 = \sum_i A_i Y_i$, $s_0 = \sum_i (1 - A_i) Y_i$, $s_E = \sum_i Y_{E,i}$ and $s = s_0 + s_1$. The conditional distribution of θ_0 given $\mathcal{D}_E$ is a $\text{Beta}(s_E + 1, n_E - s_E + 1)$ distribution, while the conditional distribution of θ_1 given $\mathcal{D}_E$ remains $U[0,1]$. The conditional likelihood $m(\mathcal{D})$, based on standard results on the Beta-Bernoulli model, is

$$m(\mathcal{D}) = \frac{s_1!(n_1 - s_1)!(s_0 + s_E)!(n_0 + n_E - s_0 - s_E)!(n_E + 1)!}{(n_1 + 1)!(n_E + n_0 + 1)! s_E!(n_E - s_E)!}. \tag{10}$$

Other testing procedures. In our comparisons we considered:

Test-A. A permutation test without ED, identical to `Algorithm 1`. We use the same working model for the ID as described above and we do not incorporate ED in the analyses. The test statistic is $m'(\mathcal{D}) = s_1!(n_1 - s_1)! s_0!(n_0 - s_0)!/[(n_1 + 1)!(n_0 + 1)!]$, which is proportional across permutations to the Bayes factor contrasting the hypotheses $\theta_0 \neq \theta_1$ and $\theta_0 = \theta_1$.

Test-B. A Wald test based on the ID with test statistic $Z = (s_1/n_1 - s_0/n_0)/(s_1(n_1 - s_1)/n_1^3 + s_0(n_0 - s_0)/n_0^3)^{1/2}$ and α-level rejection region $|Z| > \Phi^{-1}(1 - \alpha/2)$, where Φ is the cumulative distribution function of the standard normal distribution.

Test-C. A Wald test that merges ID and ED with statistic $Z = (s_1/n_1 - (s_0 + s_E)/(n_0 + n_E))/(s_1(n_1 - s_1)/n_1^3 + (s_0 + s_E)(n_0 + n_E - s_0 - s_E)/(n_0 + n_E)^3)^{1/2}$, and rejection region $|Z| > \Phi^{-1}(1 - \alpha/2)$.

Test-D. An oracle procedure. The oracle knows the response probability w_0 under the control treatment in the RCT. The test statistic is $Z = (s_1/n_1 - w_0)/[s_1(n_1 - s_1)/n_1^3]^{1/2}$ and the rejection region is $|Z| > \Phi^{-1}(1 - \alpha/2)$.

Simulation results. We considered scenarios with different β_0 values, and for each scenario repeated 10,000 simulations. We estimate the type I error rate for two significance levels $\alpha = 0.01$ and 0.05 (Figure 1, Panels a and b). We also estimate the power for $\alpha = 0.05$ (Figure 1, Panel c). The graphs show that when $p(y_i|a_i = 0) = p_E(y_{E,i}|a_{E,i} = 0)$, i.e. $\beta_0 = 0$, both our permutation test and the Wald test (Test-B and Test-C, illustrated in blue) benefit from the use of the ED, which leads to substantial improvements of the power. Importantly, when $p(\cdot|a_i = 0) \neq p_E(\cdot|a_{E,i} = 0)$ the ED-PT (solid green line) controls the type I error rate, while the Wald test that incorporates ED (Test-C; solid blue line) has an inflated type I error.

Asymptotic Analysis. We discuss the asymptotic behavior of our ED-PT. We consider a sequence of $(\mathcal{D}, \mathcal{D}_E)$ pairs with increasing sample sizes n and n_E, and assume the following:

(A1) $r = n_0/n_1 > 0$ and $r_E = n_E/n_1 > 0$ are fixed and $n_1 \to \infty$,

(A2) $\gamma = a/n_1^{1/2}$, for $a > 0$, and $\beta_0 = b/n_1^{1/2}$, with a and b fixed.

We focus on the test $\tilde{\phi}$ (`Algorithm 1`) when J is large (i.e., $J/n! \to \infty$). In other words, we consider $\tilde{\phi}$ and the exact p-value, equal to the proportion of permutations $\tau \in \mathcal{T}$ that satisfy $m(\mathcal{D}^{(\tau)}) \geq m(\mathcal{D})$. We are interested in obtaining the limiting ($n \to \infty$) power of $\tilde{\phi}$ under the Assumptions (A1) and (A2). To evaluate the power we utilize the following proposition, which provides a large-sample approximation $\hat{\mathtt{pv}}_{\tilde{\phi}}$ of the exact p-value $\mathtt{pv}_{\tilde{\phi}}$.

Proposition 2. *Under assumptions (A1) and (A2),*

$$\frac{\mathtt{pv}_{\tilde{\phi}}}{\hat{\mathtt{pv}}_{\tilde{\phi}}} \xrightarrow{p} 1,$$

when the sample sizes diverge. Here the convergence is in probability, $\mathtt{pv}_{\tilde{\phi}}$ *is the exact p-value of the test* $\tilde{\phi}$ *and*

$$\hat{\mathtt{pv}}_{\tilde{\phi}} = 1 - \Phi\left(\frac{\max\left[s_1, \frac{2(s+s_E)}{r+r_E+1} - s_1\right] - \frac{s}{r+1}}{\left\{\frac{sr[(r+1)n_1-s]}{[(r+1)n_1-1](r+1)^2}\right\}^{1/2}}\right) + \Phi\left(\frac{\min\left[s_1, \frac{2(s+s_E)}{r+r_E+1} - s_1\right] - \frac{s}{r+1}}{\left\{\frac{sr[(r+1)n_1-s]}{[(r+1)n_1-1](r+1)^2}\right\}^{1/2}}\right). \tag{11}$$

We evaluated the accuracy of the approximation numerically (see Figure 1, Panel d). In this panel $n_1 = 10,000$, $r = 0.5$, and $r_E = 5$. We varied w_0, a and b using a grid, with $w_0 \in [0.2, 0.8]$, $a \in [0, 2]$, $b \in [-3, 3]$. Figure 1(d) illustrates 1000 $(\mathcal{D}, \mathcal{D}_E)$ simulations, each corresponding to a different combination of w_0, a and b. It illustrates the accuracy of the approximation $\hat{\mathtt{pv}}_{\tilde{\phi}}$ in Proposition 2.

We can now derive the limiting ($n \to \infty$) power function of $\tilde{\phi}$.

Proposition 3. *Under assumptions (A1) and (A2), the limiting power function of the* α*-level test* $\tilde{\phi}$ *is*

$$g(r, r_E, a, b, w_0) = pr\left\{\Phi[\max(U_1, U_0)] - \Phi[\min(U_1, U_0)] > 1 - \alpha\right\}, \tag{12}$$

where $U = (U_1, U_0) \sim N(\mu, \Sigma)$ *has bivariate normal distribution, with parameters*

$$\mu = \begin{bmatrix} \frac{ar^{1/2}}{[(r+1)w_0(1-w_0)]^{1/2}} \\ \frac{-[r(r+r_E+1)+2r_E]a+2(r+1)r_E b}{(r+r_E+1)[r(r+1)w_0(1-w_0)]^{1/2}} \end{bmatrix}, \quad \Sigma = \begin{bmatrix} 1, & -1 \\ -1, & 1 + \frac{4r_E}{r(r+r_E+1)} \end{bmatrix}. \tag{13}$$

Computing (12) requires numerical integration, for example Monte Carlo simulations. When $a > 0$ is large (above 1.5) and $b = 0$, we can approximate (12) with a simpler closed-form expression (see Section SM4 in the supplementary materials for the derivation):

$$g(r, r_E, a, b, w_0) \approx \Phi\left[\frac{ar^{1/2}}{[(r+1)w_0(1-w_0)]^{1/2}} - \Phi^{-1}(1-\alpha)\right]. \tag{14}$$

Figure 1(e) illustrates the limiting power function in (12) with dots, computed with 100,000 Monte Carlo simulations, the approximation in (14) (solid lines) and Test-B with dashed lines, when a varies between 0 and 3. Here $r \in \{1/2, 1, 2\}$, $w_0 = 0.5$, $n_1 = 10,000$, $r_E = 5$ and $b = 0$. In this example the ED-PT test is asymptotically more powerful than the two-sided Wald test (Test-B) that utilizes only ID. Also, Figure SM5 illustrates large sample rejection regions for the two tests; here $r = 1/2$, $n_1 = 10,000$, $r_E = 5$ and $(n_E, s_E) = (50,000,\ 24,995)$.

The power function in Proposition 3 allows us to discuss the behavior of the ED-PT test when the user has access to a large external dataset representative of the control therapy (i.e., $r_E \to \infty$). The following corollary provides the limit of $g(r, r_E, a, b, w_0)$ when r_E diverges.

COROLLARY 1. *We consider the α-level test $\tilde{\phi}$ with access to large ED. The limit $\lim_{r_E\to\infty} g(r, r_E, a, b, w_0)$ is equal to the right-hand side of (12) with*

$$\Sigma = \begin{bmatrix} 1 & -1 \\ -1 & 1+4/r \end{bmatrix}, \quad \mu^\intercal = \left[\frac{ar^{1/2}}{[(r+1)w_0(1-w_0)]^{1/2}}, \frac{-a(r+2)+2(r+1)b}{[r(r+1)w_0(1-w_0)]^{1/2}} \right].$$

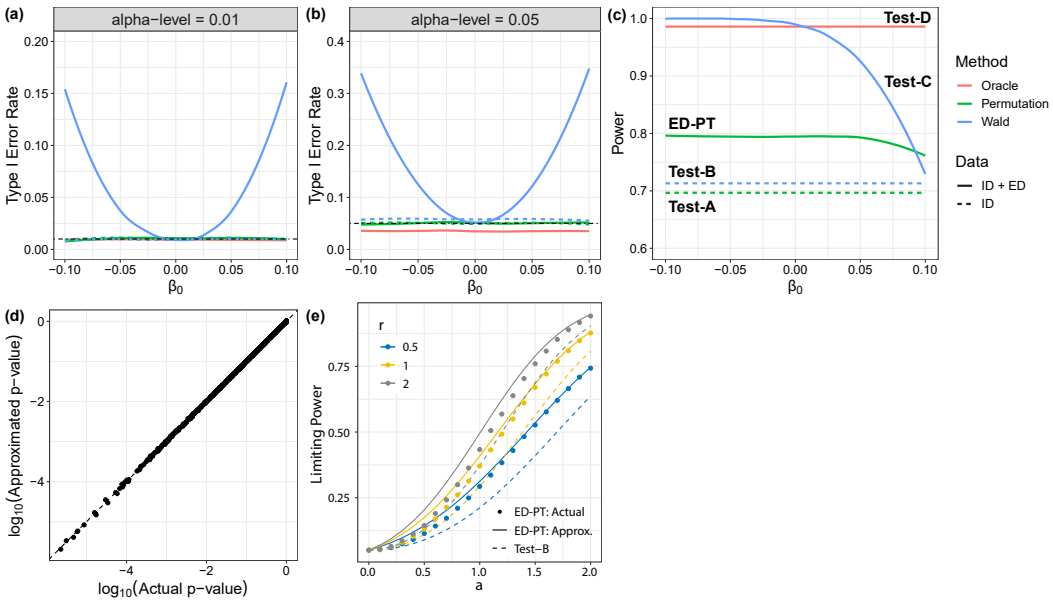

Fig. 1: ED-PT, an example with binary outcomes. Panels (a) and (b) illustrate type I error rates at two α levels, while Panel (c) shows the power. All panels (a-c) compare different tests (ED-PT and Test-A to Test-D) with ED (solid lines) and without ED (dashed lines). Here, β_0, i.e., x-axis in Panels (a) to (c), indicates the difference between the response rates of the control groups in the ED and the ID (see expression 8). Panel (d) illustrates the comparison between approximate p-values in expression (11) and the exact p-values of $\tilde{\phi}$. In these simulations, w_0 varies between 0.2 and 0.8, a varies between 0 and 2, and b varies between -3 and 3. Panel (e) shows, with dots, for a between 0 and 2 (x-axis), the limiting power function $g(r, r_E, a, b, w_0)$ of ED-PT in (12) when $w_0 = 0.5$, $r_E = 5$, $b = 0$ and $r \in \{0.5, 1, 2\}$. It also includes the approximation in (14) with solid lines, and the limiting power function for Test-B (dashed lines).

2.5. *ED-PT with normally distributed outcomes*

We now consider normally distributed outcomes and include pre-treatment covariates. We assume again $1 : r$ randomization (experimental and control arms) for the RCT, and all patients in the ED are treated with the control therapy, as in Section 2.4. The ID and the ED have the following distributions,

$$\begin{aligned} A_i &\overset{iid}{\sim} \text{Bernoulli}[1/(1+r)], \quad Y_i|A_i, X_i \overset{ind}{\sim} N(\eta_0 + \gamma A_i + \beta_1^\intercal X_i + A_i\gamma_1^\intercal X_i, 1), \\ Y_{E,i}|X_{E,i} &\overset{ind}{\sim} N[\eta_0 + \beta_{0,0} + (\beta_1 + \beta_{0,1})^\intercal X_{E,i}, 1]. \end{aligned} \tag{15}$$

Recall that $X_i, X_{E,i} \in \mathbb{R}^d$. Here γ is the treatment effect for patients with pre-treatment covariates X_i equal to 0_d, where 0_d is a vector of zeros with length d. The terms $A_i\gamma_1^\intercal X_i$

determine variations of the treatment effects across pre-treatment profiles, and the parameters $\beta_0 = (\beta_{0,0}, \beta_{0,1}) \in \mathbb{R}^{d+1}$ quantify the discrepancies between ED and the controls within the RCT.

Working Model $\mathcal{M}$. We use a linear model,

$$Y_i|A_i, X_i \overset{ind}{\sim} N(\theta_0 + \theta_1^\intercal X_i + \theta_2 A_i + A_i \theta_3^\intercal X_i, 1), \quad Y_{E,i}|X_{E,i} \overset{ind}{\sim} N(\theta_0 + \theta_1^\intercal X_{E,i}, 1), \tag{16}$$

where $\theta_1, \theta_3 \in \mathbb{R}^d$. The Bayesian model uses independent $N(0, \sigma^2)$ prior distributions for the components of $\theta = (\theta_0, \theta_1, \theta_2, \theta_3)$. Based on standard conjugacy results, $\theta|\mathcal{D}_E \sim N(\mu_E, V_E)$, where $V_E = \text{diag}\{(\tilde{X}_E^\intercal \tilde{X}_E + \sigma^{-2} I_{d+1})^{-1}, \sigma^2 I_{d+1}\}$, $\mu_E = [(\tilde{X}_E^\intercal \tilde{X}_E + \sigma^{-2} I_{d+1})^{-1} \tilde{X}_E^\intercal Y_E, 0_{d+1}]$ and $\tilde{X}_E = (1_{n_E}, X_E)$. Here 1_{n_E} is a vector of ones of length n_E, and I_{d+1} is the $(d+1) \times (d+1)$ identity matrix. The conditional likelihood of the ID (expression 2) is

$$m(\mathcal{D}) \propto \exp\left(\frac{1}{2}\mu^\intercal V^{-1}\mu\right)|V|^{1/2}, \tag{17}$$

where $V = (\tilde{X}^\intercal \tilde{X} + V_E^{-1})^{-1}$, $\mu = V(\tilde{X}^\intercal Y + V_E^{-1}\mu_E)$, and $\tilde{X} = [1_n, X, A, (A1_n^\intercal)X]$.

Other testing procedures. Our comparisons include:

Test-A. A permutation test without ED identical to the ED-PT; the test statistic is $m'(\mathcal{D}) \propto \exp\left(\frac{1}{2}\mu'^\intercal V'^{-1}\mu'\right)|V'|^{1/2}$, where $V' = (\tilde{X}^\intercal \tilde{X} + \sigma^{-2} I_{2(d+1)})^{-1}$ and $\mu' = V'\tilde{X}^\intercal Y$.

Test-B. A Wald test for (θ_2, θ_3) based only on the ID, using the model (16). The test statistic is $Z = (\hat{\theta}_2, \hat{\theta}_3)^\intercal \Sigma^{-1} (\hat{\theta}_2, \hat{\theta}_3)$, where $(\hat{\theta}_2, \hat{\theta}_3)$ is the MLE of (θ_2, θ_3) under the working model and Σ is the submatrix of $\text{Cov}(\hat{\theta}) = (\tilde{X}^\intercal \tilde{X})^{-1}$ that corresponds to the estimates $(\hat{\theta}_2, \hat{\theta}_3)$. The rejection region is $Z > \chi^2_{1-\alpha,d+1}$. Here $\chi^2_{1-\alpha,d+1}$ is the $1-\alpha$ quantile of a Chi-square distribution.

Test-C. A Wald test for (θ_2, θ_3) based on merged ID and ED, using the model in (16). The test statistic is $Z = (\hat{\theta}_2, \hat{\theta}_3)^\intercal \Sigma^{-1} (\hat{\theta}_2, \hat{\theta}_3)$, as in Test-B, but in this case $(\hat{\theta}_2, \hat{\theta}_3)$ and Σ are the MLEs and their covariance estimate computed after merging ID and ED. The rejection region is $Z > \chi^2_{1-\alpha,d+1}$.

Test-D. An oracle procedure. The oracle knows the regression function of the internal control group. The test is based on the RCT residuals $R_i = Y_i - \eta_0 - \beta_1^\intercal X_i$ for all individuals i that received the experimental treatment (i.e. $A_i = 1$). In particular, it is a Wald test for (θ_2, θ_3) based on the regression model $R_i|X_i \sim N(\theta_2 + \theta_3^\intercal X_i, 1)$. We compute the MLEs $(\hat{\theta}_2, \hat{\theta}_3)$ and their covariance matrix Σ, including only patients that received the experimental treatment $(A_i = 1)$. Then, similar to Test-B and Test-C, we use the test statistic Z and the rejection region $Z > \chi^2_{1-\alpha,d+1}$.

S1, a scenario with discrete covariates (X_i and $X_{E,i}$) indicating the patient subgroups. Pre-treatment profiles are often summarized by binary and categorical variables. We consider a trial with patients partitioned into K subgroups. In particular X_i and $X_{E,i}$ include $K-1$ binary indicators (i.e., $X_{E,i}, X_i \in \{0,1\}^{K-1}$) that point to the individual subgroup. That is, X_i and $X_{E,i}$ can take only K different values, and the $(k-1)$-th element of X_i (or $X_{E,i}$) is equal to one if patient i belongs to group k. In Figure 2(a) and (b) we show simulation results for $K = 2$ patient subgroups. Here the proportion of patients in subgroup 1 is 0.5 for the RCT and the ED. Also, $\eta_0 = 0, \beta_1 = 0.5$ and $r = 1/2$. With $K = 2$ the treatment effects in subgroups 1 and 2 are γ and $\gamma + \gamma_1$. We considered $\gamma = \gamma_1 = 0$ to assess the type I error rate of the test, and $(\gamma = 0.5, \gamma_1 = -0.2)$ to evaluate the power. We set $\beta_{0,1} = 0$ and varied $\beta_{0,0}$ from -0.1 to

0.1 to evaluate the robustness of the ED-PT procedure with respect to discrepancies between $p_E(\cdot|x_{E,i}, a_{E,i} = 0)$ and $p(\cdot|x_i, a_i = 0)$.

The results in Figure 2(a) and (b) were computed with $J = 1,000$ permutations, $10,000$ simulations per scenario, $n = 150$, $n_E = 750$, and $\sigma^2 = 10$. If $p_E(\cdot|x_{E,i}, a_{E,i} = 0) = p(\cdot|x_i, a_i = 0)$ for every pair of pre-treatment profiles $x_{E,i} = x_i$ (i.e., $\beta_{0,0} = 0$), then the use of ED increases the power of all testing procedures. Moreover, when $p_E(\cdot|x_{E,i}, a_{E,i} = 0) \neq p(\cdot|x_i, a_i = 0)$ for some values of $x_{E,i} = x_i$, our ED-PT controls the type I error rate, while the control of false positives deteriorates for Test-C.

Asymptotic analysis. We investigate the asymptotic behavior of our ED-PT. We focus on the scenarios that we described, with $X_i, X_{E,i} \in \{0,1\}^{K-1}$ indicating subgroups. The unknown outcome distributions (p and p_E) are summarized by model (15), and the working model by expression (16). We derive the limiting power function when the population includes $K > 1$ subgroups. Similar to Section 2.4, we consider a sequence of $(\mathcal{D}, \mathcal{D}_E)$ pairs with increasing samples sizes n and n_E. See Section SM5 in the supplementary materials for details. In Figure 2(c) we computed the limiting power (see expression S.2 in Section SM5 in the supplementary materials) using 100,000 Monte Carlo simulations (dots). We then compared these results to estimates of the power based on trial simulations (solid lines), with $K = 2$, $\rho = (0.5, 0.5)$, $a_1 \in [0, 10]$, $\eta_0 = \beta_{0,1} = a_2 = b_1 = b_2 = 0$, $\beta_1 = 0.5$, $r \in \{0.5, 1, 2\}$, $r_E = 7.5$, $n_1 = 10,000$, $J = 1,000$ and 10,000 simulation replicates per scenario. The panel illustrates that in this scenario the limiting power of the ED-PT is larger than the power of Test-B (dashed lines).

Modified S1, negative treatment effects. We modified Scenario *S1* above to showcase the excessive rejection rate of the null hypothesis by the original ED-PT when treatment effects are negative, and illustrate how the modified ED-PT, based on test statistics $\tilde{m}_1(\mathcal{D})$ and $\tilde{m}_2(\mathcal{D})$, effectively control the rejection rate in such scenarios.

We simulated 10,000 times the pair $(\mathcal{D}, \mathcal{D}_E)$ using model (15) with $K = 2$ patient subgroups. We set $\eta_0 = 0, \beta_1 = 0.5, \gamma = 0$ (the treatment effect in group 1) and $\gamma_1 = -1$ (the treatment effect in group 2) in (15). Also, $n_1 = 100, r = 0.5, r_E = 7.5, \beta_{0,0} = \beta_{0,1} = 0$, and $\alpha = 0.05$. In this modified scenario `Algorithm 1` rejected the null hypothesis with a frequency equal to 0.85.

To use $\tilde{m}_1$, we specify $\tilde{\Theta}$ as $\tilde{\Theta} = \{\theta \in \Theta : \theta_2 > \tilde{\theta} \text{ or } \theta_2 + \theta_3 > \tilde{\theta}\}$, where $\tilde{\theta}$ is a threshold that defines clinically relevant treatment effects. The modified permutation procedure in `Algorithm 1`, using the statistics $\tilde{m}_1(\mathcal{D})$ instead of $m(\mathcal{D})$, rejected H_0 with a frequency equal to 0.04 when we set $\tilde{\theta} = 0$. If we use $\tilde{m}_2(\mathcal{D})$ in `Algorithm 1` to replace $m(\mathcal{D})$ in this example, we have

$$\tilde{m}_2(\mathcal{D}) = \mathbb{E}[\rho_1\theta_2\mathbb{I}(\theta_2 > 0) + (1 - \rho_1)(\theta_2 + \theta_3)\mathbb{I}(\theta_2 + \theta_3 > 0) \mid \mathcal{D}, \mathcal{D}_E].$$

Recall that ρ_1 is the prevalence of the first subgroup of patients. In this case, the rejection rate decreased from 0.85 to 0.05.

Figure SM3 in the supplementary materials shows the frequency of rejections of the ED-PT with the test statistic $m(\mathcal{D})$ (dashed line) and the modified versions of the testing procedure with statistics $\tilde{m}_1(\mathcal{D})$ and $\tilde{m}_2(\mathcal{D})$ (solid blue and red lines, respectively) in Scenario *S1* when $\gamma = 0$ and $\gamma_1 \in [-1, 1]$. It illustrates that the modified versions of our ED-PT in the presence of negative treatment effects (i.e., $\gamma_1 < 0$) control the frequency of false positive results. Moreover, with positive treatment effects these modified versions of the ED-PT procedure have power similar to the ED-PT with test statistics $m(\mathcal{D})$.

S2, a scenario with two subgroups and continuous pre-treatment covariates. We conclude this subsection by adding continuous pre-treatment covariates to the Scenario *S1*. In particular, we specify pre-treatment profiles $X_i, X_{E,i} \in \{0,1\} \times R^{d-1}$, where the first entries of X_i and $X_{E,i}$ are

iid Bernoulli$(1/2)$ as in Scenario *S1* and the remaining $d-1$ components of X_i and $X_{E,i}$ have $N(0_{d-1}, I_{d-1})$ distributions. We set $n_1 = 100$, $r = 0.5$, $r_E = 7.5$, $\eta_0 = \beta_{0,0} = 0$ and $\beta_{0,1} = 0_d$ in the outcome model (15). In other words, there are no discrepancies between the outcome distributions $p_E(\cdot|x_{E,i}, a_{E,i} = 0)$ and $p(\cdot|x_i, a_i = 0)$ of the ED and the RCT control arm. In Figure 2(d-e) the dimension d of X_i grows from 2 to 41. The ED-PT is compared to Test-B, and a two-sample Z-test (ID only) that ignores the covariates. In Panel (d) both patient subgroups benefit from the experimental treatment: $\gamma = 0.6$ and $\gamma_1 = (-0.2, 0_{d-1})$. In Panel (e) only one group of patients benefits from the experimental treatment: $\gamma = 0.75$ and $\gamma_1 = (-0.75, 0_{d-1})$. We set $\beta_1 = [0.5, 1_{d-1}/(d-1)^{-1/2}]$, therefore the marginal variability of the outcome Y_i does not vary with the number of continuous covariates $(d-1)$. In both panels, the power of ED-PT remains nearly the same as d increases, while for Test-B, which uses only the RCT data, the power decreases. This result suggests that ED-PT successfully leverages the information about β_1 provided by the ED. Also, the Z-test has lower power compared to the ED-PT and Test-B.

3. Glioblastoma Clinical Trials

We report the results of retrospective analyses of a collection of Glioblastoma (GBM) datasets. The analyses are based on a resampling schema described in Ventz et al. (2022a). The goal is to assess power and type I error rates of our ED-PT and other testing procedures.

Datasets. The data include IPLI of newly diagnosed GBM patients that were treated with temozolomide and radiation therapy (TMZ+RT), the current standard of care in GBM (Stupp et al., 2005). Pre-treatment patient variables (X_i and $X_{E,i}$) include age, sex, Karnofsky performance status (KPS), MGMT methylation status, and extent of tumor resection (EOR). We use IPLI (i.e., pre-treatment variables and outcomes) of patients treated with TMZ+RT from the AVAGLIO RCT (Chinot et al., 2014) and the DFCI EHRs database. We refer to Rahman et al. (2023) for further details on these datasets.

Outcome and Subgroups. The primary outcome is a binary variable that captures survival after 12 months of treatment (OS-12). We consider treatment effects that vary across subgroups, defined by patients' KPS (> 90 and ≤ 90) and MGMT status (positive and negative). For each patient treated with TMZ+RT we define $S_i = S(X_i) \in \{1, 2, 3, 4\}$ if the patients' KPS and MGMT variables are $[> 90, \text{positive}]$, $[\leq 90, \text{positive}]$, $[> 90, \text{negative}]$, or $[\leq 90, \text{negative}]$. Hypotheses of treatment effects variations modulated by MGMT and KPS in neuro-oncology have been the subject of extensive literature, see for example Chen et al. (2018). The proportions of patients in these groups and the corresponding OS-12 rates in the AVAGLIO and DFCI datasets are provided in Table 2.

	KPS≥90, MGMT+	KPS<90, MGMT+	KPS≥90, MGMT-	KPS<90, MGMT-
AVAGLIO (ID) N (%)	78 (23%)	30 (9%)	161 (48%)	68 (20%)
OS-12	0.83	0.70	0.68	0.48
DFCI (ED) N (%)	95 (29.60%)	59 (18.38%)	92 (28.66%)	75 (23.36%)
OS-12	0.84	0.83	0.81	0.63

Table 2: Subgroup sizes and the corresponding OS-12 rates in the AVAGLIO and DFCI datasets.

In silico RCTs *and* ED, (p, p_E). We use a resampling schema, similar to the one in Ventz et al. (2022a), which allows us to generate *in silico* datasets. In particular, we generate ED and *in silico* RCT with pre-treatment characteristics $(X_{E,i}, X_i \in \mathbb{R}^5)$ distributions identical to the empirical

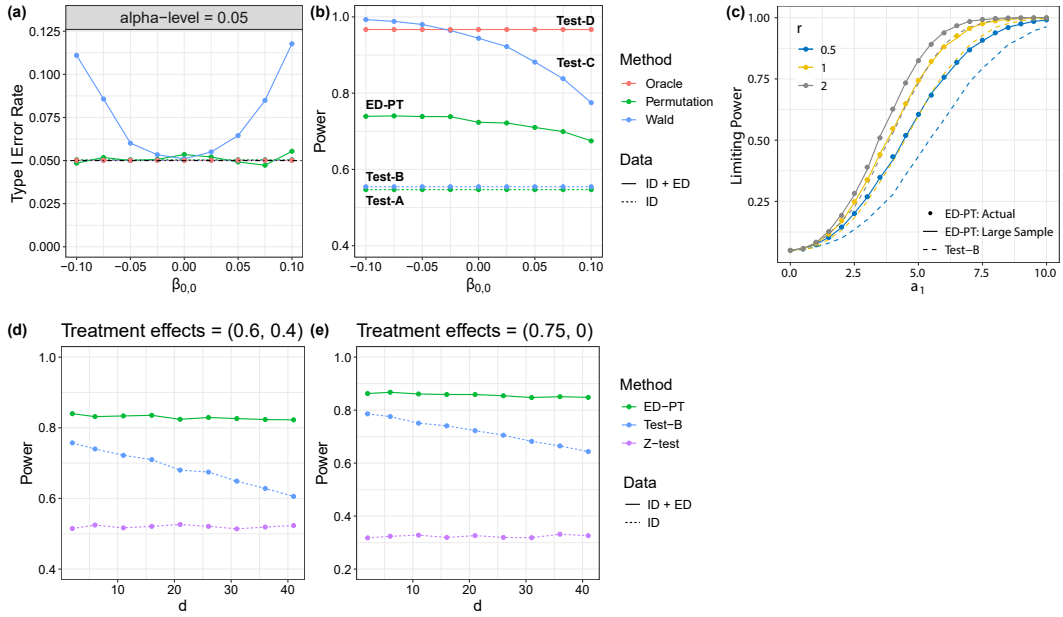

Fig. 2: Power and type I error rates of the ED-PT, an example with continuous outcomes. Panels (a-c) show results when covariates are discrete subgroup indicators. Here $K = 2$ with $\rho = (0.5, 0.5)$, $\eta_0 = \beta_{0,1} = 0$, $\beta_1 = 0.5$, $r_E = 7.5$ and $\sigma^2 = 10$. We used $J = 1{,}000$ permutations and 10,000 simulation replicates per scenario. Panels (a) and (b): Type I error rates ($\alpha = 0.05$) and power for the ED-PT and alternative tests (Test-A to Test-D) with (solid lines) and without (dashed lines) ED. The type I error rate was evaluated with $\gamma = \gamma_1 = 0$ and the power was assessed with $\gamma = 0.5$ and $\gamma_1 = -0.2$. The $\beta_{0,0}$ parameter (x-axis) summarizes the discrepancy between the ED and the control arm of the RCT (see also expression 15). The results were computed with $r = 0.5$ and $n = 150$. Panel (c): Limiting power of the ED-PT, evaluated using expression (S.2) in Section SM5 in the supplementary material with 100,000 Monte Carlo replicates (dots). The results are compared to estimates of the power based on trial simulations, using `Algorithm 1` and $n_1 = 10{,}000$ (solid lines). We considered $r \in \{0.5, 1, 2\}$ and specified γ, γ_1, $\beta_{0,0}$, and $\beta_{0,1}$ as in Assumption (B3), with $a_1 \in [0, 10]$, $a_2 = 0$, and $b = (0, 0)$. The panel also illustrates the power of Test-B (dashed lines). Panels (d) and (e) show the power of three testing procedures when the covariates include subgroup indicators and continuous pre-treatment variables. The dimension of X_i varied between 2 and 41, with 1 to 40 continuous pre-treatment variables. The details of the simulation are described in the manuscript.

distributions in the DFCI and AVAGLIO datasets. To generate the *in silico* RCT and ED, we followed these three steps:

(i.a) *In silico ID*. We sample n patients (pre-treatment profiles and OS-12 outcomes) with replacement from the TZM+RT arm of the AVAGLIO study.

(i.b) *In silico ID, treatment assignment*. We randomly assign $n_1 = n/(1 + r)$ of the n patients to the *in silico* experimental arm, and the remaining $n_0 = n - n_1$ to the control arm. The experimental and control arms of the *in silico* RCT include n_0 and n_1 *iid* replicates with the same joint distribution of pre-treatment variables and outcomes. This joint distribution

is identical to the empirical distribution of the TZM+RT group of the AVAGLIO study. In different words, in this *in silico* RCT the treatment effects are null.

(i.c) *Treatment effects.* To introduce subgroup-specific positive treatment effects in the *in silico* RCT, we randomly relabel with probability $g(S_i) \geq 0$ each negative outcome ($Y_i = 0$) in the experimental arm (Step i.b) into a positive outcome ($Y_i = 1$). Similarly, to specify scenarios with negative treatment effects we randomly relabel each positive outcome ($Y_i = 1$) into a negative one ($Y_i = 0$) with probabilities $g(S_i)$ that vary across subgroups.

(ii) *In silico ED.* We sample n_E patients with replacement from either the TZM+RT arm of the AVAGLIO study or the DFCI dataset (pre-treatment profiles and OS-12 outcomes) that constitute the *in silico* ED.

(iii) *Hypothesis testing.* We apply our ED-PT in `Algorithm 1` to the *in silico* RCT (Step i) and the ED (Step ii).

We set $n = 150$, $r = 1/2$, $n_E = 50, 100, \ldots, 250$, and consider six scenarios with distinct configurations of the treatment effects (Table 3). In each scenario the relabeling probabilities $g(s)$, $s \in \{1, 2, 3, 4\}$, which varies across patient subgroups, match the desired group-specific log-odds ratio LOR_s (treatment vs control). In particular, if $\text{LOR}_s > 0$, then $g(s) = 1/\{1 + [h_s(e^{\text{LOR}_s} - 1)]^{-1}\}$, where h_s is the response rate in subgroup s of the AVAGLIO dataset. A similar map is used with negative effects ($\text{LOR}_s < 0$).

In scenarios 1-5 the treatment effects are null or positive. In contrast, in scenario 6 the experimental treatment has negative effects. We include this scenario to compare the three versions of ED-PT (based on m, $\tilde{m}_1$ and $\tilde{m}_2$) introduced in Section 2.1 and 2.2 and to examine the control of false positives.

Scenario	LOR_s values	Description
1	$(0, 0, 0, 0)$	No effects
2	$(0.5, 1, 1.5, 2)$	Positive effects all groups
3	$(0, 0, 3, 0)$	Positive effect for patients with MGMT+ and KPS < 90
4	$(5, 5, 0, 0)$	Positive effects for patients with MGMT+
5	$(2, 0, 2, 0)$	Positive effects for patients with KPS ≥ 90
6	$(-1, 0, 0, 0)$	Negative effects for patients with MGMT+ and KPS ≥ 90

Table 3: Six scenarios in which the log-odds ratios LOR_s, $s = 1, 2, 3, 4$ vary across patient subgroups.

Working model $\mathcal{M}$. We specify the working model, with conditional distributions q_θ and $q_{E,\theta}$ parameterized by $\theta = (\theta_{ID}, \theta_{ED})$. We use a Bayesian logistic regression model, with six pre-treatment covariates $x_i^{1:6}$: age (x_i^1), sex (x_i^2), EOR (x_i^3), MGMT (x_i^4), KPS (x_i^5), and the interaction term MGMT $\times$ KPS (x_i^6). The model includes additional coefficients to capture the treatment effect and interactions between the treatment and MGMT, KPS, or MGMT $\times$ KPS. To summarize, $(x_i, a_i) \to \text{logit}[q_\theta(y_i = 1|x_i, a_i)]$ is a linear map of $(1, x_i^{1:6}, a_i, a_i \times x_i^{4:6})$ with coefficients $\theta_{ID} = (\theta_0, \theta_x, \theta_a, \theta_I) \in \mathbb{R}^{11}$, where $\theta_x = (\theta_x^1, \ldots, \theta_x^6)$. Similarly for patients treated with the control therapy in the external group, $x_{E,i} \to \text{logit}[q_{E,\theta}(y_{E,i} = 1|x_{E,i})]$ is a linear map with coefficients $\theta_{ED} = (\theta_{E,0}, \theta_{E,x}) \in \mathbb{R}^7$.

Prior model π. Based on our previous analyses of GBM RCTs and EHRs (Ventz et al., 2019, 2022a; Rahman et al., 2023; Ventz et al., 2022b), we assume that $\theta_0 = \theta_{E,0}$, $\theta_{E,x}^j = \theta_x^j$ for $j \in \{1, 2, 4, 5\}$ and $\theta_{E,x}^j = \theta_x^j + \theta_B^j$ for $j \in \{3, 6\}$. We use a normal prior with large variances $\theta = (\theta_0, \theta_x, \theta_a, \theta_I, \theta_B) \sim N(0_{13}, 100 I_{13})$.

Approximation of the test statistics $m(\mathcal{D})$. In our ED-PT the conditional likelihood $m(\mathcal{D}) = \int q_\theta(Y|X,A) \times \pi(\theta|\mathcal{D}_E)d\theta$ is computed using a Laplace approximation (De Bruijn, 1981; Tierney & Kadane, 1986),

$$m(\mathcal{D}) \approx (2\pi)^{13/2}|-\hat{H}|^{-1/2} q_{\hat{\theta}}(\mathcal{D}) q_{E,\hat{\theta}}(\mathcal{D}_E)\pi(\hat{\theta}), \tag{18}$$

where $\hat{\theta}$ is the maximum a posteriori (MAP) estimate and $\hat{H}$ is the Hessian of the log posterior at $\hat{\theta}$.

In `Algorithm 1` we compute the MAP estimate and the Hessian matrix for J permutations $\mathcal{D}^{(\tau)}$. Similarly, to compute the modified statistics $\tilde{m}_1(\mathcal{D})$ and $\tilde{m}_2(\mathcal{D})$ in Section 2.3 we used Laplace approximations. For $\tilde{m}_1(\mathcal{D})$, we restrict the integration to $\tilde{\Theta} \subset \Theta$, the parameter configurations with positive effects. Also, to compute $\tilde{m}_2(\mathcal{D})$, we iteratively sample from $N[\hat{\theta}, (-\hat{H})^{-1}]$ to approximate the integral in (7).

Alternative testing procedures. Our comparisons include several other testing procedures:

Test-A-m and $\tilde{m}$. Permutation tests for ID only using the statistics m (Test-A-m) and $\tilde{m}_j$ (Test-A-$\tilde{m}_j$), $j = 1, 2$, respectively.

Test-B and C. Wald test for proportions, as in Section 2.4, using only the ID (Test-B) or merging the ID and the ED (Test-C), without accounting for pre-treatment covariates.

Test-LR and LR-ED. Likelihood ratio test based on our working model (null hypothesis: $(\theta_a, \theta_I) = 0_4$) using only the ID (Test-LR) or merging the ID and the ED (Test-LR-ED).

Test-Matching. A matching-based testing procedure. We first apply a matching algorithm to estimate the average effect in the treated group. Specifically, we match each patient in the RCT's experimental arm to one patient in either the control arm of the RCT or the ED based on propensity scores. We use the R package `MatchIt` to perform matching. We then use g-computation implemented in the R package `marginaleffects` to estimate the treatment effects.

Test-IPW. An inverse probability weighting (IPW) procedure. We first estimate $\hat{e}(x)$, which is the conditional probability that a randomly selected patient (i.e., from the RCT or the ED) with pre-treatment characteristics x was enrolled into the RCT, using a logistic regression model. We then assign weights equal to one to the patients in the ID, while for the ED the individual wights are $w_E\hat{e}(x_{E,i})/[1-\hat{e}(x_{E,i})]$. Here $w_E \in \mathbb{R}^+$ determines the relative weights of the external dataset with respect to the RCT data. We followed the approach in Li et al. (2018) and Wang et al. (2023) to obtain average treatment effect estimates in the RCT population. We use the R package `ipw` to compute weights and weighted linear regression to estimate the treatment effects. The R package `sandwich` is used to obtain the robust standard errors (White, 1980) of the estimates.

Comparative analyses. Our analyses focus on newly diagnosed GBM patients, and are based on two main groups of $(\mathcal{D}, \mathcal{D}_E)$ *in silico* replicates. In the first group the distributions $p(y_i, x_i|a_i = 0)$ and $p_E(y_{E,i}, x_{E,i}|a_{E,i} = 0)$ are identical and match the empirical joint distribution of pre-treatment profiles and the outcomes in the AVAGLIO trial (Step i and ii of our schema). In the second group $p(y_i, x_i|a_i = 0)$ and $p_E(y_{E,i}, x_{E,i}|a_{E,i} = 0)$ are different because the ED are generated using a different dataset, the DFCI EHRs. Treatment effects are included in some of the *in silico* RCTs (see step i.c and Table 3). Then we focus on testing, using a variety of approaches that differ substantially in several aspects, including the use or exclusion of ED and the potential lack of control of false positives due to unmeasured confounders or other distortion mechanisms. The aim of the comparative analyses is to identify testing procedures suitable for future GBM trials.

Results. Figure 3(a) shows the type I error rates of all testing procedures in Scenario 1 with ED generated using the DFCI EHRs. We find that Test-C and Test-LR-ED, which use the ID and the ED, have inflated type I error rates (Test-C: 0.10 for $n_E = 50$, and up to 0.36 when $n_E = 250$; Test-LR-ED: 0.12 for $n_E = 50$, and up to 0.32 when $n_E = 250$). Also Test-Matching and Test-IPW present inflated type I error rates that are likely due to unmeasured confounding or other types of unadjusted discrepancies between the conditional distributions $p(\cdot|x_i, a_i = 0)$ and $p_E(\cdot|x_{E,i}, a_{E,i} = 0)$. In contrast the permutation tests with (ED-PT-m, ED-PT-$\tilde{m}_1$, and ED-PT-$\tilde{m}_2$) and without (Test-A) ED control the type I error rate close to the nominal $\alpha = 0.05$ level.

When negative treatment effects are present (Scenario 6 in Table 3), all testing procedures except ED-PT-$\tilde{m}_1(\mathcal{D})$ and ED-PT-$\tilde{m}_2(\mathcal{D})$ have high rejection probabilities (see Figure SM6). This is expected because the results (reject H_0 or not) of several testing procedures (Test-LR, Test-LR-ED, Test-B and Test-C) do not depend on the sign of the estimated treatment effects. On the other hand, in Scenario 6 the ED-PTs based on the modified statistics $\tilde{m}_1(\cdot)$ and $\tilde{m}_2(\cdot)$ have a rejection rate close to 0.05. Also, for Test-B and Test-C, the rejection rate becomes smaller than 2.5% when we modify the rejection region to implement one-sided testing.

Based on the results on the control of false positives, we focused on ED-PT-m, Test-A, Test-LR and Test-B for power comparisons. Figure 3 illustrates the power of these tests in Scenario 2-5. The top row illustrates results when we use the AVAGLIO trial to generate both the *in silico* RCTs and the ED while the bottom row reports results when we use the DFCI EHRs to generate the ED (see point ii of the schema used to generate RCT and external data). The top panels show an ideal setting, $p(\cdot|x_i, a_i = 0) = p_E(\cdot|x_{E,i}, a_{E,i} = 0)$ when $x_i = x_{E,i}$, whereas the bottom panels provide more realistic evaluations of the ED-PT procedure when $p(\cdot|x_i, a_i = 0) \neq p_E(\cdot|x_{E,i}, a_{E,i} = 0)$ for some $x_i = x_{E,i}$. In this second row of panels, with the ED generated using the DFCI EHRs, we have potential distortion mechanisms (Rahman et al., 2023) such as unmeasured confounding or subtle differences in the definition of the outcomes.

We vary the size of the ED (n_E) as indicated by the x-axis to examine its impact on the power of our ED-PT. The solid curve indicates ED-PT-m and the dashed curves correspond to three alternative tests (Test-A in red, Test-LR in green, and Test-B in blue). The dash-dotted red curve in Figure 3 indicates the power of the ED-PT test when the sample size n_E of the ED diverges ($n_E \to +\infty$; see Section SM10 in the supplementary materials for details). We did not include the tests that failed to control the type I error rate at the nominal α level.

For all HTE configurations (Scenarios 2 to 5 in Table 3) we observe a gain in power for the ED-PT-m compared to Test-A. As expected the power increase of ED-PT-m compared to Test-A is larger when the ED are generated from the AVAGLIO study and smaller when the ED are generated from the DFCI EHRs. Except for Scenario 2, where all four subgroups have positive treatment effects, ED-PT-m with a moderate n_E ($n_E > 100$) outperforms the other procedures. The improvements in power of ED-PT-m compared to the best performing ID-only approaches in Panel (b) vary across scenarios from 3.7% (Scenario 3, DFCI EHRs used to generate ED) to 29.3% (Scenario 4, AVAGLIO trial used to generate ED). These increments in power can be attained by the best ID-only testing procedures if the size n of the ID increases between 10% (Scenario 3) and 33% in (Scenario 4). In Scenario 2 the ED-PT is the most powerful test when the ED are generated using the AVAGLIO trial, but it is 9.6% less powerful than Test-B when the ED are generated using the DFCI EHRs. These results suggest the importance of selecting adequate ED (e.g., previous RCTs or EHR), avoiding obsolete data repository and potential distortion mechanisms (Ventz et al., 2019). For all scenarios, we observe that the power of ED-PT-m with $n_E = 250$ is close to the power of ED-PT-m with $n_E \to \infty$ (ED-PT-Inf).

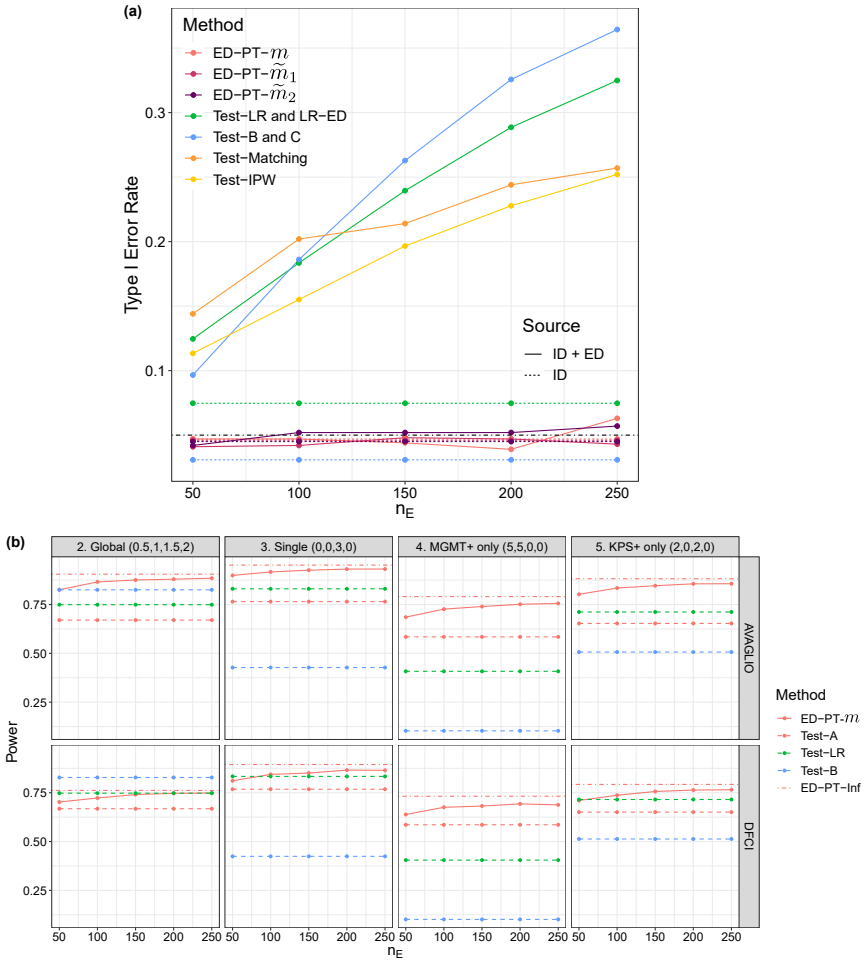

Fig. 3: Comparative analyses with *in silico* RCTs and EDs. Type I error rates (a) and power (b) of our ED-PT and alternative testing procedures as n_E increases. In Panel (a), we illustrate the type I error rates of all testing procedures when the ED are generated using the DFCI EHRs. The dash-dotted black line indicates the nominal level $\alpha = 0.05$. In Panel (b), the top row illustrates results when the ED are generated using the AVAGLIO trial, while for the bottom row the ED were generated using the DFCI EHRs.

4. Discussion

The efficacy of novel therapeutics can vary substantially across patient subgroups, and in early phase clinical studies it is often unclear whether and which patient subgroups benefit from the treatment. RCTs are an essential component of the development of new experimental therapies. These trials are crucial to demonstrate causal effects on clinical outcomes. However, as previously discussed in the literature, RCTs require large sample sizes to investigate HTEs, hence long execution times and high costs. The use of patient-level ED from completed clinical trials and electronic health records has been proposed to address these challenges and accelerate drug development processes.

In this manuscript, we defined three test statistics, m, $\tilde{m}_1$ and $\tilde{m}_2$, to evaluate the treatment effects. The first, m, is used to detect non-zero effects regardless of their signs, while the other

two are specified to detect positive effects. We notice that the same approach can be used to determine whether the RCT data provide evidence of negative effects in some of the enrolled patients. In particular, we can use modified versions of the test statistics $\tilde{m}_1$ and $\tilde{m}_2$ for detecting negative effects. We refer to Sections SM9.1 and SM9.2 in the supplementary materials for additional examples and scenarios with negative effects of the experimental therapy in subgroups of patients.

An experimental therapy may have positive effects on some patients and negative on others. With such variations of the treatment effects across subgroups, challenging decisions arise throughout the drug development process. As examples of relevant decisions we mention recommending a phase 3 confirmatory RCT based on phase 2 results, and selecting a phase 3 design with appropriate eligibility criteria and adequate sample size. A comprehensive methodological toolbox is essential to support these decisions. This includes procedures for testing—the primary focus of this manuscript—as well as approaches for estimation and prediction. We also emphasize that context-specific knowledge, such as previous results from preclinical models that characterized the experimental therapy's mechanism of action, can be fundamental to support key decisions during clinical trials.

Acknowledgement

The authors are grateful to Sergio Bacallado for helpful discussions. Funding was provided by the National Institutes of Health (NIH) grant R01 LM013352 and Project Data Sphere. Steffen Ventz was supported by the National Cancer Institute (5P30CA077598-23), a DSI Grant from the Minnesota Supercomputing Institute, and a Medtronic Faculty Fellowship.

Supplementary materials

Supplementary materials include proofs for all propositions, additional simulations, and additional considerations on the method that we introduced.

References

Berger, J. O. (2013). *Statistical decision theory and Bayesian analysis*. Springer.

Berger, J. O., Wang, X. & Shen, L. (2014). A bayesian approach to subgroup identification. *J Biopharm Stat* **24**, 110–129.

Bonetti, M. & Gelber, R. D. (2004). Patterns of treatment effects in subsets of patients in clinical trials. *Biostatistics* **5**, 465–481.

Brown, B. W., Herson, J., Atkinson, E. N. & Rozell, M. E. (1987). Projection from previous studies: a bayesian and frequentist compromise. *Controlled clinical trials* **8**, 29–44.

Chen, X., Zhang, M., Gan, H., Wang, H., Lee, J.-H. et al. (2018). A novel enhancer regulates mgmt expression and promotes temozolomide resistance in glioblastoma. *Nat Commun* **9**, 2949.

Chinot, O. L., Wick, W., Mason, W., Henriksson, R., Saran, F. et al. (2014). Bevacizumab plus radiotherapy–temozolomide for newly diagnosed glioblastoma. *NEJM* **370**, 709–722. PMID: 24552318.

Chu, Y. & Yuan, Y. (2018). Blast: Bayesian latent subgroup design for basket trials accounting for patient heterogeneity. *JRSS-C: Applied Statistics* **67**, 723–740.

Dar, H., Johansson, A., Nordenskjöld, A., Iftimi, A., Yau, C. et al. (2021). Assessment of 25-year survival of women with estrogen receptor–positive/erbb2-negative breast cancer treated with and without tamoxifen therapy. *JAMA Network Open* **4**, e2114904–e2114904.

De Bruijn, N. G. (1981). *Asymptotic methods in analysis*, vol. 4. Courier Corporation.

Ding, P., Feller, A. & Miratrix, L. (2016). Randomization inference for treatment effect variation. *Journal of the Royal Statistical Society Series B: Statistical Methodology* **78**, 655–671.

Ding, P., Feller, A. & Miratrix, L. (2019). Decomposing treatment effect variation. *Journal of the American Statistical Association* **114**, 304–317.

Freidlin, B., McShane, L. M. & Korn, E. L. (2010). Randomized clinical trials with biomarkers: design issues. *JNCI* **102**, 152–160.

Freidlin, B., McShane, L. M., Polley, M.-Y. C. & Korn, E. L. (2012). Randomized phase ii trial designs with biomarkers. *JCO* **30**, 3304.

Haslam, A., Kim, M. & Prasad, V. (2021). Updated estimates of eligibility for and response to genome-targeted oncology drugs among us cancer patients, 2006-2020. *Annals of Oncology* **32**, 926–932.

Kennedy, A. D., Torgerson, D. J., Campbell, M. K. & Grant, A. M. (2017). Subversion of allocation concealment in a randomised controlled trial: a historical case study. *Trials* **18**, 1–6.

Lauko, A., Lo, A., Ahluwalia, M. S. & Lathia, J. D. (2022). Cancer cell heterogeneity & plasticity in glioblastoma and brain tumors. In *Seminars in Cancer Biology*, vol. 82. Elsevier.

Lehmann, E. & Romano, J. (2005). *Testing statistical hypotheses*, vol. 3. Springer.

Li, F., Morgan, K. L. & Zaslavsky, A. M. (2018). Balancing covariates via propensity score weighting. *JASA* **113**, 390–400.

Liau, L. M., Ashkan, K., Brem, S., Campian, J. L. & Trusheim, J. E. a. o. (2023). Association of autologous tumor lysate-loaded dendritic cell vaccination with extension of survival among patients with newly diagnosed and recurrent glioblastoma. *JAMA Onc* **9**, 112–121.

Liu, F. (2018). Assessment of bayesian expected power via bayesian bootstrap. *Stat Med* **37**, 3471–3485.

Morita, S. & Müller, P. (2017). Bayesian population finding with biomarkers in a randomized clinical trial. *Biometrics* **73**, 1355–1365.

Murphy, S. A. (2003). Optimal dynamic treatment regimes. *JRSS-B: Statistical Methodology* **65**, 331–355.

Nugent, B. M., Madabushi, R., Buch, B., Peiris, V., Crentsil, V. et al. (2021). Heterogeneity in treatment effects across diverse populations. *Pharmaceutical Statistics* **20**, 929–938.

Rahman, R., Ventz, S., McDunn, J., Louv, B., Reyes-Rivera, I. et al. (2021). Leveraging external data in the design and analysis of clinical trials in neuro-oncology. *Lancet Onc* **22**, e456–e465.

Rahman, R., Ventz, S., Redd, R., Cloughesy, T., Alexander, B. M. et al. (2023). Accessible Data Collections for Improved Decision Making in Neuro-Oncology Clinical Trials. *CCR* **29**, 2194–2198.

Rigdon, J., Baiocchi, M. & Basu, S. (2018). Preventing false discovery of heterogeneous treatment effect subgroups in randomized trials. *Trials* **19**, 1–15.

Russo, M., Ventz, S., Wang, V. & Trippa, L. (2023). Inference in response-adaptive clinical trials when the enrolled population varies over time. *Biometrics* **79**, 381–393.

Sherman, R. E., Anderson, S. A., Dal Pan, G. J., Gray, G. W., Gross, T. et al. (2016). Real-world evidence—what is it and what can it tell us. *NEJM* **375**, 2293–2297.

Slevin, M., Clark, P., Joel, S., Malik, S., Osborne, R. et al. (1989). A randomized trial to evaluate the effect of schedule on the activity of etoposide in small-cell lung cancer. *JCO* **7**, 1333–1340.

Stupp, R., Mason, W. P., Van Den Bent, M. J., Weller, M., Fisher, B. et al. (2005). Radiotherapy plus concomitant and adjuvant temozolomide for glioblastoma. *NEJM* **352**, 987–996.

Tierney, L. & Kadane, J. B. (1986). Accurate approximations for posterior moments and marginal densities. *JASA* , 82–86.

Ventz, S., Comment, L., Louv, B., Rahman, R., Wen, P. Y. et al. (2022a). The use of external control data for predictions and futility interim analyses in clinical trials. *Neuro Onc* **24**, 247–256.

Ventz, S., Khozin, S., Louv, B., Sands, J., Wen, P. Y., Rahman, R., Comment, L., Alexander, B. M. & Trippa, L. (2022b). The design and evaluation of hybrid controlled trials that leverage external data and randomization. *Nat Commun* **13**, 5783.

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. *CCR* **25**, 4993–5001.

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

Wang, J., Zhang, H. & Tiwari, R. (2023). A propensity-score integrated approach to bayesian dynamic power prior borrowing. *Stat Biopharm Research* , 1–23.

Wang, R., Schoenfeld, D. A., Hoeppner, B. & Evins, A. E. (2015). Detecting treatment-covariate interactions using permutation methods. *Stat Med* **34**, 2035–2047.

White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica* , 817–838.

Xu, J., Zhang, H., Zhang, H., Bian, J. & Wang, F. (2023). Machine learning enabled subgroup analysis with real-world data to inform clinical trial eligibility criteria design. *Scientific Reports* **13**, 613.

Yang, S., Li, F., Starks, M. A., Hernandez, A. F., Mentz, R. J. et al. (2020). Sample size requirements for detecting treatment effect heterogeneity in cluster randomized trials. *Stat Med* **39**, 4218–4237.

Ziegler, A., Koch, A., Krockenberger, K. & Großhennig, A. (2012). Personalized medicine using dna biomarkers: a review. *Human genetics* **131**, 1627–1638.