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

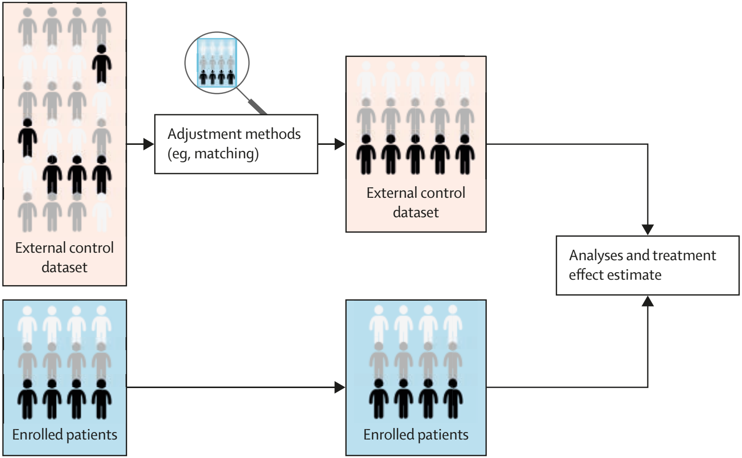
Lorenzo Trippa

Dana-Farber Cancer Institute &
Harvard T.H. Chan School of Public Health

1/16/2026,
ICERM - Brown University

January 16, 2026

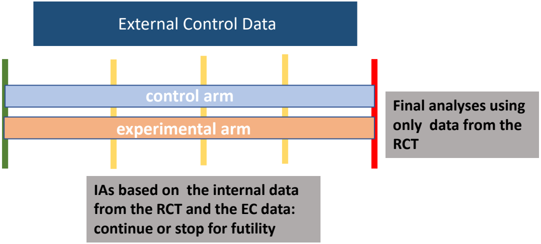
Externally Controlled Trial (ECT)



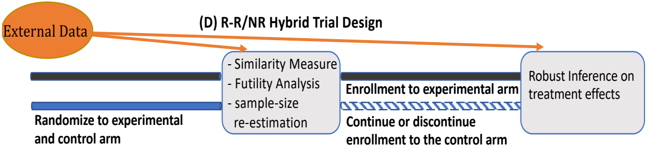
Leveraging external data in the design and analysis of clinical trials in neuro-oncology.
Rahman et al 2021, Lancet Onc

Integration of External Data in Clinical Trials

RCT with Futility stopping



RCT with adaptive randomization



Ventz et a. 2021 Neuro-Oncology
Ventz et a. 2021 Nature-Communications

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

Boyu Ren[1], Federico Ferrari[2], Sandra Fortini[3], Lorenzo Trippa[4,5], and Steffen Ventz[6]

[1]Laboratory for Psychiatric Biostatistics, McLean Hospital, Belmont, MA, USA

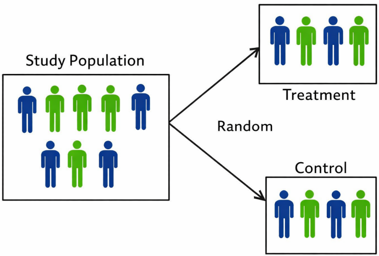
[2]Biostatistics and Research Decision Sciences, Merck & Co., Inc., Rahway, NJ, USA

[3]Department of Decision Sciences, Bocconi University, Milan, Italy

[4]Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA, USA

[5]Department of Data Science, Dana-Farber Cancer Institute, Boston, MA, USA

Biometrika 2026
arxiv.org/abs/2506.04128



Are there treatment effects in some biomarker subgroups?

Datasets :**RWD+RCTs**

Table 1. Distribution of pretreatment patient characteristics for the TMZ+RT arm of three clinical studies and three RWE studies

Study **NCT ID** **PubMed ID**	**DFCI-cohort**	**UCLA-cohort**	**NCT01013285** **PM21135282**	**NCT00441142** **PM25910950**	**—** **PM22120301**	**AVAglio** **NCT00943826** **PM24552318**
Data type	RWE	RWE	RWE	Phase II	Phase II	Phase III
Arm	TMZ+RT	TMZ+RT	TMZ+RT	TMZ+RT	TMZ+RT	TMZ+RT
Enrollment period			8/06-11/08	2/09-6/11	8/05-2/11	6/09-3/11
Enrollments to SOC	378	305	110	29	16	460
OS events	269	265	89	24	15	344
Age						
Median	58	57	59	58	59	57
Range	18–91	20–84	20–90	26–73	36–69	18–79
SD	13	13	14	11	11	10
Sex (%)						
Females	0.43	0.36	0.36	0.45	0.5	0.36
Males	0.57	0.64	0.64	0.55	0.5	0.64
KPS (%)						
≤80	0.55	0.39	0.32	0.24	0.44	0.31
>80	0.45	0.61	0.68	0.76	0.56	0.69
Data missing (*n*)	27	17	0	0	0	0
RPA (%)						
3	NA	0.22	0.25	NA	0.12	0.16
4	NA	0.42	0.41	NA	0.75	0.61
5	NA	0.34	0.33	NA	0.13	0.23
6	NA	0.02	0.01	NA	0	0
Data missing (*n*)	378	0	0	29	1	0
Resection (%)						
Biopsy	0.14	0.22	0.21	0.21	0	0.09
Sub total	0.47	0.47	0.36	0.48	0.31	0.49
Gross total	0.39	0.31	0.43	0.31	0.69	0.42
Data missing (*n*)	12	15	0	0	0	0
MGMT (%)						
Unmethylated	0.43	0.71	0.60	0.86	0.43	0.67
Methylated	0.57	0.29	0.40	0.14	0.56	0.32
Data missing (*n*)	194	128	40	7	0	0.23
IDH1 (%)						
Wild-type	0.91	0.91	0.98	0.83	NA	NA
Mutant	0.09	0.09	0.02	0.17	NA	NA
Data missing (*n*)	188	0.46	52	6	16	344

Abbreviations: IDH1, isocitrate dehydrogenase 1; KPS, Karnofsky performance status; MGMT, O6-methylguanine-DNA methyltransferase; RPA, recursive partitioning analysis.

Rahman et al 2023, Clinical Cancer Research

Setting

	RCT, $\mathcal{D} = (Y, X, A)$	ED, $\mathcal{D}_E = (Y_E, X_E, A_E)$
Outcome	$Y = (Y_1, \ldots, Y_n)$	$Y_E = (Y_{E,1}, \ldots, Y_{E,n_E})$
Covariates	$X = (X_1, \ldots, X_n)$	$X_E = (X_{E,1}, \ldots, X_{E,n_E})$
Treatment	$A = (A_1, \ldots, A_n)$	$A_E = (A_{E,1}, \ldots, A_{E,n_E})$
Distribution	$p(y, x, a)$	$p_E(y_E, x_E, a_E)$

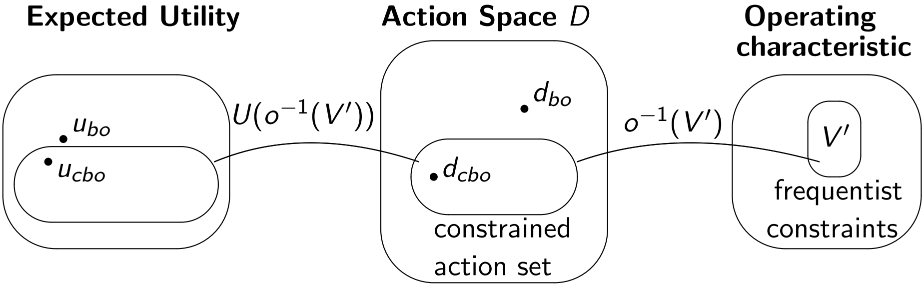
- Permutation-compatible null hypothesis:

$$H_0 : p(y, x, a) \text{ is invariant to any permutation of } a, \forall (y, x, a)$$

- H_0 implies no treatment effect in **any** patient subpopulation

$$E_p(Y_i | X_i = x_i, A_i = 1) = E_p(Y_i | X_i = x_i, A_i = 0), \forall x_i$$

Bayes Optimum (ob) vs Constrained Bayes Optimum (cob)

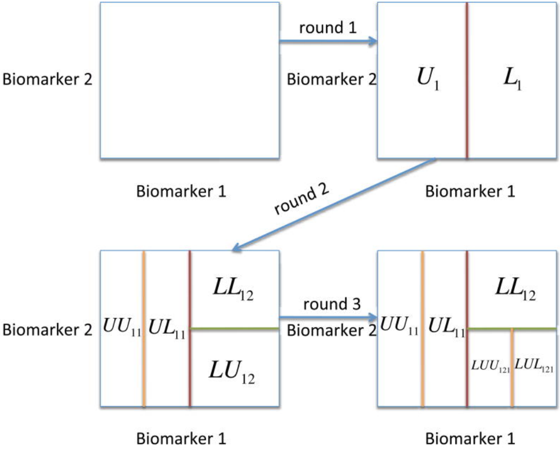


Action space: candidate testing functions.
Utility criteria: expected power.
Operating characteristic : Type I error $< \alpha$ (robust control)

Prior + Utility Criteria + Regulator Constraint → Test

Step1: select a joint prior model for $\mathcal{D}$ and $\mathcal{D}_E$

Example:



Xu et. al, Stat Biosci 2016

Integration of ED through a Bayesian model

- A working model $\mathcal{M}$ facilitating integration of ED

$$\mathcal{M} = \left\{ \begin{array}{ll} \text{RCT:} & q_\theta(y|x,a) = \prod_i q_\theta(y_i|x_i,a_i) \\ \text{ED:} & q_{E,\theta}(y_E|x_E,a_E) = \prod_i q_{E,\theta}(y_{E,i}|x_{E,i},a_{E,i}) \\ \text{Prior:} & \pi(\theta), \theta \in \Theta \end{array} \right\}$$

- Specification of q_θ and $q_{E,\theta}$ allows for HTE (e.g., interaction terms)
- Reflects prior belief on the discrepancy between RCT and ED
- Conditional distribution $\pi(\theta|\mathcal{D}_E)$ summarizes the information in ED

$$\pi(\theta|\mathcal{D}_E) \propto \pi(\theta) q_{E,\theta}(Y_E|X_E,A_E)$$

- A test statistic for RCT data incorporating $\pi(\theta|\mathcal{D}_E)$

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

ED-augmented Permutation Test

$\tau = (\tau_1, \ldots, \tau_n)$ a permutation of $(1, \ldots, n)$, $\tau \in \mathcal{T}$.

Algorithm *permutation test*

1: **Input:** The number of permutations J, ID $\mathcal{D} = (Y, X, A)$, working model $\mathcal{M}$, conditional distribution $\pi(\theta|\mathcal{D}_E)$

2: $m(\mathcal{D}) = \int_\theta q_\theta(Y|X, A)\pi(\theta|\mathcal{D}_E)d\theta$;

3: **for** $j \leftarrow 1$ to J **do**

4: $\quad\tau \leftarrow$ a random sample from $\mathcal{T}$;

5: $\quad m_j = \int_\theta q_\theta(Y|X, A^{(\tau)})\pi(\theta|\mathcal{D}_E)d\theta$;

6: **Output:** $\widetilde{\phi}(\mathcal{D}) = \mathbb{I}\left[\frac{1+\sum_1^J \mathbb{I}(m_j > m(\mathcal{D}))}{1+J} \leq \alpha\right]$

Proposition (Optimality of ED-PT)

$\phi(D)$ has maximal Bayesian expected power (BEP) among all level α tests, where BEP of a test ϕ' is defined as

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

recall:

π prior model

ϕ is the testing procedure$D \to \{0,1\}$

A simple example

- Compare the operating characteristics of ED-PT and other testing procedure, especially the robustness against **discrepancy between RCT and ED**
- $X_i, X_{E,i} \in \{0, 1\}$ are subpopulation indicators
- Data generating mechanism

$$A_i \overset{iid}{\sim} \text{Bernoulli}(2/3), A_{E,i} = 0,$$
$$Y_i | A_i, X_i \sim N(\gamma A_i + \beta_1 X_i + A_i \gamma_1 X_i, 1),$$
$$Y_{E,i} | X_{E,i} \sim N(\boldsymbol{\beta_0} + \beta_1 X_{E,i}, 1).$$

- $\beta_1 = 0.5, \gamma = 0.5, \gamma_1 = -0.3$
- HTE: 0.5 and 0.2 in $X_i = 0$ and $X_i = 1$ respectively
- β_0 quantifies the discrepancy between RCT and ED

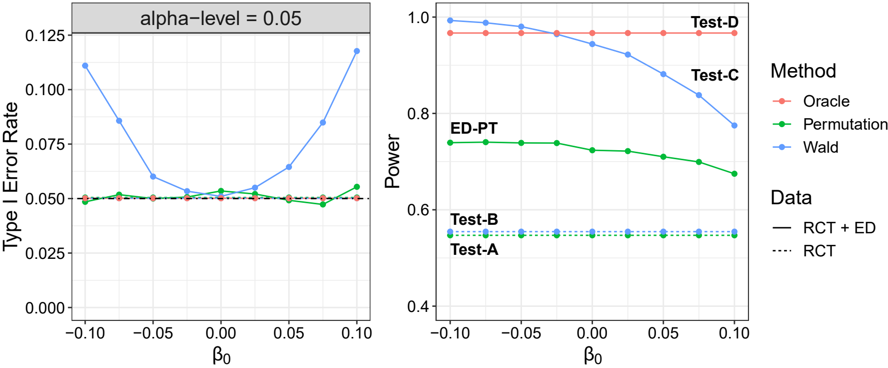
A simple example: testing procedures

- All methods are based on the working model

$$
\begin{gathered}
Y_i | A_i, X_i \sim N(\theta_0 + \theta_1 X_i + \theta_2 A_i + A_i \theta_3 X_i, 1), \\
Y_{E,i} | X_{E,i} \sim N(\theta_0 + \theta_1 X_{E,i}, 1),
\end{gathered}
$$

1. **ED-PT**: the proposed testing procedure in Algorithm 1
2. **Test-A**: the same algorithm as ED-PT but without using ED
3. **Test-B**: a Wald test for (θ_2, θ_3) based on the RCT only
4. **Test-C**: a Wald test for (θ_2, θ_3) based on RCT + ED
5. **Test-D**: an *oracle* Wald test that knows the outcome model of the control in the RCT

An example: simulation results



one-sided testing

- H_0 does not distinguish between positive and negative treatment effects
- the experimental treatments could perform worse than the control (e.g., toxicities)
- with negative effects we don't want to reject the null
- ED-PT has to be modified
- **Main idea**: we modify the test statistic

Modified ED-PT for one-sided alternatives

We propose two types of modifications:

1. Posterior probability:

$$\widetilde{m}_1(\mathcal{D}) = \int_{\widetilde{\Theta}} \pi(\theta|\mathcal{D}, \mathcal{D}_E) d\theta,$$

   where $\widetilde{\Theta} \subset \Theta$ indicates the parameter space corresponding to relevant treatment effects. In the illustrating exampling, we can set $\widetilde{\Theta} = \{\theta_2 > 0 \text{ or } \theta_2 + \theta_3 > 0\}$

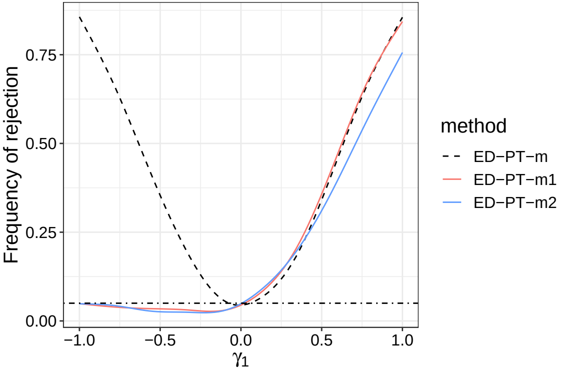
2. Expected regret:

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

   where $a_i(\theta) = \texttt{arg-max}_a \mathbb{E}_{q_\theta}(Y_i|X_i, A_i = a)$ is the optimal treatment for subject i based on the working model $\mathcal{M}$

example (continued): negative treatment effects

- Inflated rejection probability of the original ED-PT under negative treatment effects
- Both $\widetilde{m}_1$ and $\widetilde{m}_2$ can resolve this issue
- The same data-generating model as before with $\gamma = 0$ and $\gamma_1 \in [-1, 1]$



Glioblastoma (GBM) datasets

- Collections of multiple GBM trials and EHR data (Rahman et. al 2023)
  - Patients treated with temozolomide and radiation therapy (TMZ+RT)
  - Focus on the AVAGLIO study and DFCI EHR
- **Outcome**: 12-month survival (binary)
- **Covariates**: age, sex, Karnofsky performance status (KPS), MGMT methylation status and extent of tumor resection (EOR)
- Four subgroups defined by KPS (≤ 90 vs. > 90) and MGMT (positive vs. negative) status
  - Two biomarkers that might modulate treatment effects (Chen et al. 2018)

Generating *in silico* RCTs and EDs

- A resampling schema as in **?** to create synthetic RCTs and EDs
- Accurate evaluation of operating characteristics
- The simulation follows four steps:
  1. *In silico RCT*: sample with replacement n patients from TZM+RT arm of the AVAGLIO study
  2. *Treatment assignment*: randomly assign $n_1 = n/(1+r)$ to the *in silico* experimental arm and the rest to the control
  3. *Introduce treatment effects*: randomly flipped negative outcome in the experimental arm into positive with a pre-specified probability
  4. *In silico ED*: sample with replace n_E patients from either the TZM+RT arm of the AVAGLIO study or the DFCI EHR data

Methods in comparison

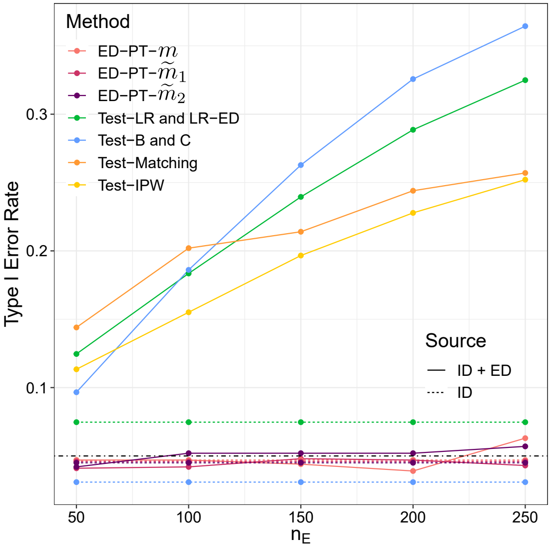
- We consider the following working model:

$$\text{logit}[q_\theta(y=1|x,a)] = \theta_0 + \theta_x^\intercal x + \theta_a a + \theta_I^\intercal x_{4:6} a,$$
$$\text{logit}[q_{E,\theta}(y=1|x_E)] = \theta_0 + (\theta_x + \theta_{E,x})^\intercal x,$$

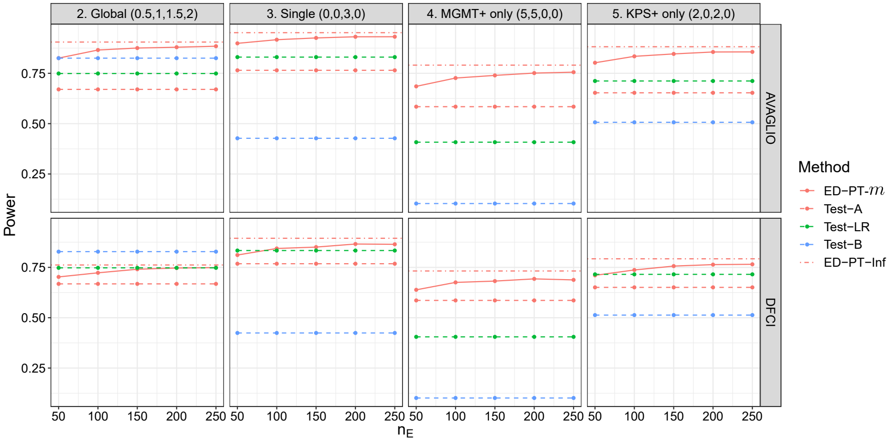
where $x = (\text{age, sex, EOR, MGMT, KPS}, \text{MGMT} \times \text{KPS})$

- Laplace approximation to compute $m(\mathcal{D})$ and its variants
- We consider four classes of methods:
  1. ED-PT, ED-PT-$\widetilde{m}_1$, ED-PT-$\widetilde{m}_2$ and a permutation test without using ED (Test-A)
  2. Wald test without accounting for covariates (Test-B, C)
  3. Likelihood ratio test for (θ_a, θ_I) using RCT and RCT + ED
  4. Causal inference based methods for external control integration: a matching approach and an IPW approach

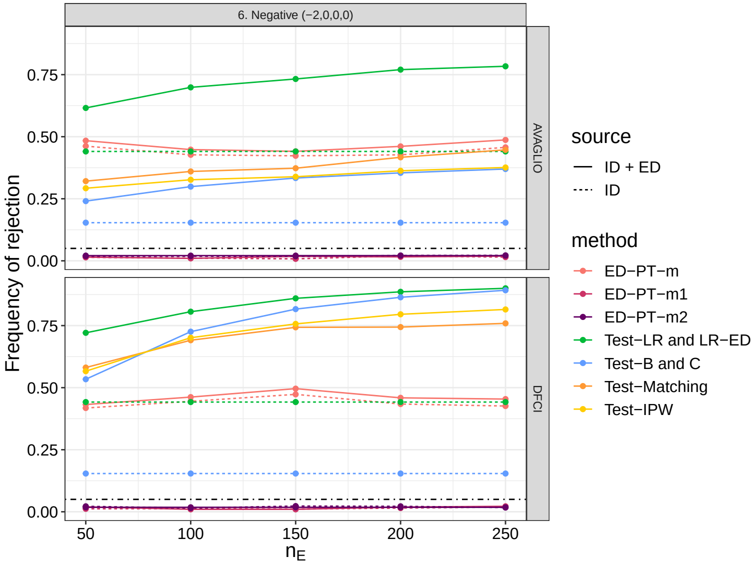
Type I error rates and power



Type I error rates and power



Type I error rates and power



Conclusion

- We investigate the use of ED in the analysis of RCTs where HTEs are plausible
- We propose a permutation test that leverage information from external data through a Bayesian model with the aim of enhancing power
- We illustrate the strength of our permutation procedure with a simulated example and a retrospective analysis of GBM data collections