Randomization tests to address disruptions in clinical trials

A Report from the NISS Ingram Olkin Forum Series on Unplanned Clinical Trial Disruptions

by Diane Uschner

on February 24, 2026

» Acknowledgements

Thanks to: Nancy Flournoy and the participants of the **NISS Ingram Olkien series on clinical trial disruptions** and the members of the **Randomization WG** for many fruitful discussions.

» Reference

STATISTICS IN BIOPHARMACEUTICAL RESEARCH
2024, VOL. 16, NO. 4, 405–413
https://doi.org/10.1080/19466315.2023.2257894

Taylor & Francis
Taylor & Francis Group

OPEN ACCESS Check for updates

Using Randomization Tests to Address Disruptions in Clinical Trials: A Report from the NISS Ingram Olkin Forum Series on Unplanned Clinical Trial Disruptions

Diane Uschner[a], Oleksandr Sverdlov[b], Kerstine Carter[c], Jonathan Chipman[e,d], Olga Kuznetsova[f], Jone Renteria[a], Adam Lane[g], Chris Barker[h], Nancy Geller[i], Michael Proschan[j], Martin Posch[k], Sergey Tarima[l], Frank Bretz[m,k], and William F. Rosenberger[n]

Background

» Introduction

- Estimands were the statistical concept of the moment when the COVID-19 pandemic started in early 2020, the Suez canal was blocked, and shortly after, the Ukraine war started.
- Numerous disruptions affected clinical trials all around the globe.
- Examples:
  - Population shift
  - Treatment effect drift
  - Change of care
  - Change of data collection
  - Change in availability of study medication

Can we use randomization based inference to address disruptions?

» Motivating example

- Consider a single-blind randomized controlled trial of **two weight loss interventions** on obesity:
  - One intervention is standard of care plus a drug;
  - the other is standard of care plus a lifestyle intervention.
- The primary outcome is **change in BMI** from baseline to 24 months after enrollment.

The primary objective of the trial is to investigate whether one of the two treatments is superior to the other, and to assess the magnitude of the difference of change in BMI between the treatments.

» What are randomization tests??

* Randomization tests test the Fisherian null hypothesis that a patient's response would have been the same, irrespective of the treatment they received:

$$H_0 : y_{i|E} = y_{i|C}$$

Each patient's response would have been the same, if they had been assigned to the other treatment group.

* no assumption about distribution the response,
* identical distribution not necessary,
* based on the counterfactual response.

» Difference in means test statistic

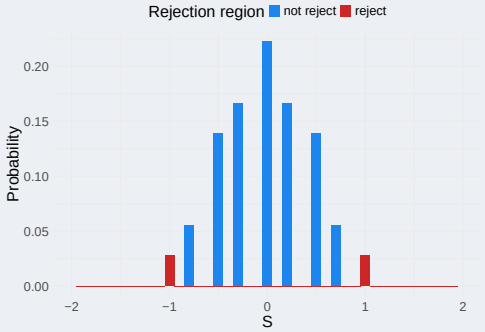
* Consider the difference in means test statistic as a measure for the extremeness of the result

$$S(t, y) = \frac{\sum_i t_i \cdot y_i}{N_E} - \frac{\sum_i (1 - t_i) \cdot y_i}{N_C}$$

* To obtain the distribution of S under H_0, we have to "fix" the null hypothesis, and leave the rest free. $\Rightarrow$ we have to *condition* on a sufficient statistic for H_0:
  $\Rightarrow$ The response $y = (y_1, ..., y_N)$ is sufficient for H_0

» Distribution of the test statistic under the RT H_0

	t	$P(T=t)$	$S(t)$
1	E E E E C C C C	0	-2
2	E E E C E C C C	0	-1.75
9	C E E E C E C C	0	-1.25
10	E E C C E E C C	1/36	-1
17	E C E C E C E C	1/36	-1
34	C C E C E E E C	1/36	-1
35	C C C E E E E C	0	-1.25
70	C C C C E E E E	0	2



» Rejection of the null

* A large absolute value of the test statistic is evidence against the null hypothesis.
* We compare the observed test statistic to the potential values that would have resulted if a different randomization sequence had been used in a trial.
* H_0 is rejected if the probability to observe a larger absolute value of the test statistic is lower than 5%:

$$p = \sum_{t \in \Omega} 1(|S(t)| \geq |S(t_{obs})|) \cdot \mathbb{P}(T = t) < 0.05.$$

* A consistent estimator for p can be obtained by Monte-Carlo resampling:

$$\hat{p} = \frac{1}{L} \sum_{l=1}^{L} 1(|S(t_l)| \geq |S(t_{obs})|).$$

Estimand

» Defining an estimand in the Fisherian framework

- Target population: defined by the inclusion and exclusion criteria of the trial, **conditional on the enrolled participants.**
- Primary outcome: Change in BMI from enrollment to 24 months after follow-up **in the group of enrolled participants**.
- Treatments: Randomized treatments
- Intercurrent events and their strategies:
  1. Withdrawal from treatment due to reasons associated with the treatment: treatment policy strategy
  2. Withdrawal from treatment unrelated to treatment: hypothetical strategy
- Population level summary measure: difference in means *of the observed outcomes*.

» Intercurrent events due to the pandemic

During the course of the trial, the trial protocol needed to be amended for the impact of the pandemic, and the following intercurrent events were added due to the pandemic:

3. Measurement of the primary outcome in clinic is replaced with at-home measurement of the primary outcome: Treatment policy strategy
4. The study drug is replaced with a different drug because of supply chain-issues: Hypothetical strategy

» Estimands

1. the effect of the treatment for the overall population Θ, but assuming that all participants can come to the clinic to get their BMI measured, and the study drug is still available:

$$\theta_1 = E_{\Theta}(Y|T = 1, C^3 = C^4 = 0) - E_{\Theta}(Y|T = 0, C^3 = C^4 = 0)$$

2. the effect in post-pandemic world $\Theta_{\texttt{after}}$, but assuming that all BMI measurements are conducted at home

$$\theta_1 = E_{\Theta_{\texttt{after}}}(Y|T = 1, C^3 = 0, C^4 = 1) - E_{\Theta_{\texttt{after}}}(Y|T = 0, C^3 = 0, C^4 = 1)$$

» Implementation in the randomization testing framework

- Treatment policy strategy: Randomization test will provide unbiased results because the null hypothesis $H_0 : P(Y \geq y|T) = P(Y \geq y)$ continues to be true
- Hypothetical strategy:
  - Conditional randomization tests: restrict the reference set to sequences where the treatment assignment for a participant i is the same as the assignment i in the observed randomization sequence (Heussen et al. 2024)
  - Multiple imputation: baseline and longitudinal covariates as well as longitudinal outcome values of earlier time points could be used to impute missing outcomes at the time point of interest. (Ivanova et al. 2022)

» Computational considerations for the conditional randomization test

- The conditional reference set consists of all sequences of the full reference set that fulfill certain conditions.
- The sampling approach can either be naive (sampling from the full set and discarding sequences that do not meet the condition) or complex (sampling directly from the conditional set).
- Currently, sampling directly from the conditional set is only possible in simple cases

Simulation Study

» Model for the effect of trial disruptions

We model the outcome using a normal distribution $Y_i \sim N(\delta \cdot T_i + \vartheta_i, 1)$, where δ is the treatment effect and ϑ_i represents a mean shift due to a trial disruption.

1. Increasing trend over time: We model the effect ϑ_i based on a step trend (Tamm and Hilgers 2014), reflecting a change in the patient population at time n_0.

$$\vartheta_i = \begin{cases} \theta & i \geq n_0 \\ 0 & i < n_0 \end{cases}$$

2. Heterogeneity of the treatment effect: Assume that there is an interaction between treatment and the disruption, i.e., $\hat{\vartheta}_i = T_i \cdot \vartheta_i$.

» Simulation study to investigate the effect of trial disruptions

We assess the effect of disruptions on the test decision of the randomization test using a simulation study with the following parameters:

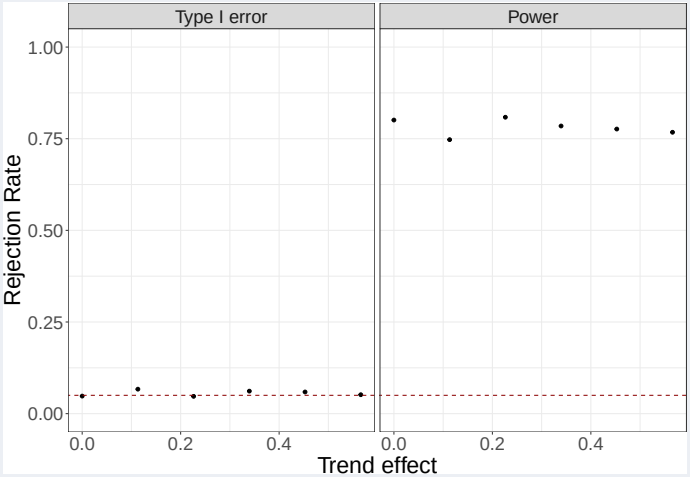
Parameter	Value
Sample size	N = 100
Randomization Procedures	PBD(N), PBD(N), EBCD(0.667), BSD(3), CR
Treatment effect	$\delta \in \{0, 0.556\}$
Trend effect	$\theta \in \{0, 0.0556\}$, $\theta = 0.556 \cdot j, j = 0, .2, .4, \dots, 1$
Disruption time	$n_0 = \frac{N}{2}$
Size of the reference set	$L = 2,500$
Repetitions	$r = 10,000$

Results

» Increasing time trend

Figures 1 shows type I error probability and power of the randomization test in case of trial for increasing effects of the trial disruption, $\theta = 0.556 \cdot i, i = 0, .2, .4, \ldots, 1$ using Random Allocation Rule. For all investigated scenarios, the type I error probability and power meet the nominal significance level (up to some simulation variability).

» Error rates under increasing time trend

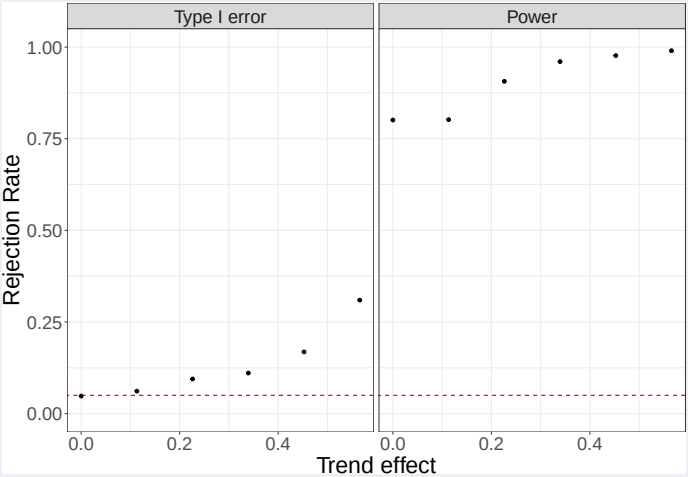


Rejection probability under the null hypothesis ($\delta = 0$) for increasing trend effect.

» Treatment effect heterogeneity

Figures 2 shows type I error probability and power of the randomization test in case of a heterogeneous treatment effect introduced by the trial disruption, $\theta = 0.556 \cdot i, i = 0, .2, .4, \dots, 1$ using Random Allocation Rule. We can see that the type I error probability and power are inflated when the heterogeneity is large.

» Error rates under treatment effect heterogeneity



Rejection probability under the null hypothesis ($\delta = 0$) for heterogeneous treatment effect after the disruption.

Conclusions and outlook

» Conclusions

We showed that randomization tests are robust against clinical trial disruptions in certain scenarios, namely if the disruption can be considered an ancillary statistic to the treatment effect. As a consequence, randomization tests maintain type I error probability and power at their nominal levels.

Randomization tests can provide a useful sensitivity analysis in clinical trials that are affected by clinical trial disruptions.

Thank you for your attention!

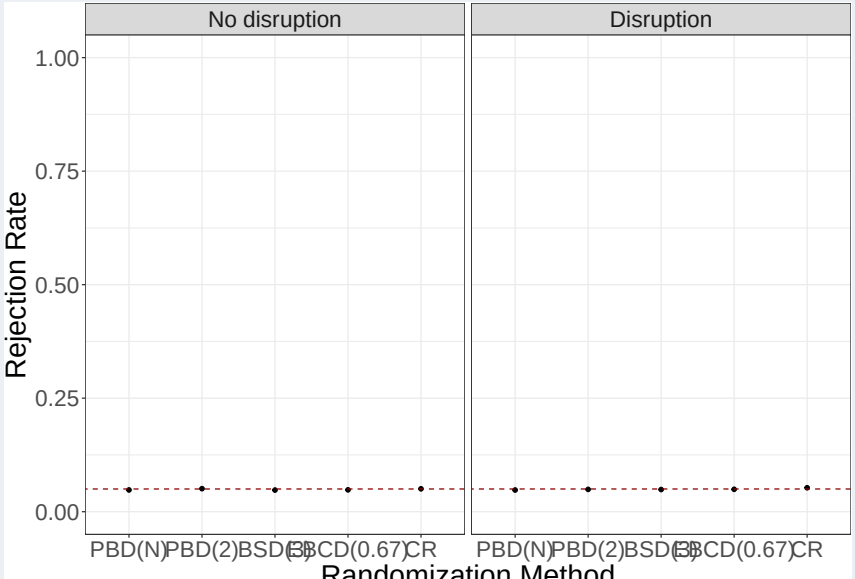
Stay in touch: diane.uschner@roche.com

Randomization WG https://randomization-wg.org/

» Time trend

Figures 3 and 4 show type I error probability and power of the randomization test in case of trial disruption ($\theta = 0$) and no trial disruption ($\theta = 0.0565$) for different randomization procedures. For all investigated scenarios, the type I error probability and power meet the nominal significance level. We can conclude that the rejection probability of the randomization test does not depend on the randomization procedure in this scenario.

» Type I error probability under time trend



» Power under time trend

References

- Uschner, D., Sverdlov, O., Carter, K., Chipman, J., Kuznetsova, O., Renteria, J., ..., Rosenberger, W. F. (2023). Using Randomization Tests to Address Disruptions in Clinical Trials: A Report from the NISS Ingram Olkin Forum Series on Unplanned Clinical Trial Disruptions. Statistics in Biopharmaceutical Research, 16(4), 405-413. https://doi.org/10.1080/19466315.2023.2257894
- Ivanova, A., Lederman, S., Stark, P. B., Sullivan, G., & Vaughn, B. (2022). Randomization tests in clinical trials with multiple imputation for handling missing data. Journal of Biopharmaceutical Statistics, 32(3), 441-449. https://doi.org/10.1080/10543406.2022.2080695
- Heussen, N., Hilgers, R. D., Rosenberger, W. F., Tan, X., & Uschner, D. (2023). Randomization-Based Inference for Clinical Trials with Missing Outcome Data. Statistics in Biopharmaceutical Research, 16(4), 456-467. https://doi.org/10.1080/19466315.2023.2250119

» References

Tamm, M., and R.-D. Hilgers. 2014. "Chronolgical Bias in Randomized Clinical Trials Under Different Ttype of Unobserved Time Trends." *Meth. Inf. Med.*