The Impact of Randomization Restrictions on Testing Error Rates

Jonathan J. Chipman, PhD

June 24, 2025

University of Utah Intermountain Healthcare
Department of Population Health Sciences
Division of Biostatistics

Research interests and today's talk

Research interests

1. Randomization restrictions
2. Randomization-based inference

Today's talk

1. Chronological restrictions: Impact on Finite Population Type I error
   - Co-authors: Diane Uschner and Alex Sverdlov
2. Covariate restrictions: Improvement on power
   - Co-authors: Lindsay Mayberry and Robert Greevy

Clarifying notes

1. Simple Randomization: coin flip
2. Complete Randomization: coin flip with pre-specified trt allocation
3. Two-sample t-test = ANOVA with 2 groups

Randomization-Based Inference (RBI)

The randomization test minimally assumes

1. The treatment assignment is randomized
2. The potential outcomes values under each arm

Choose a Randomization Model

- Complete Randomization
- Restricted Randomization

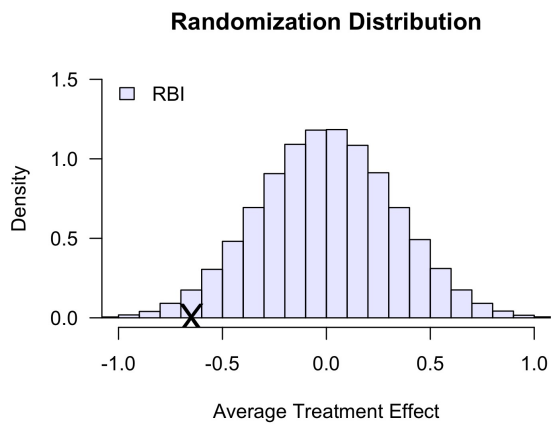
Obtain summary estimate

Compare to null distribution

- Summary measure across all possible randomization sequences

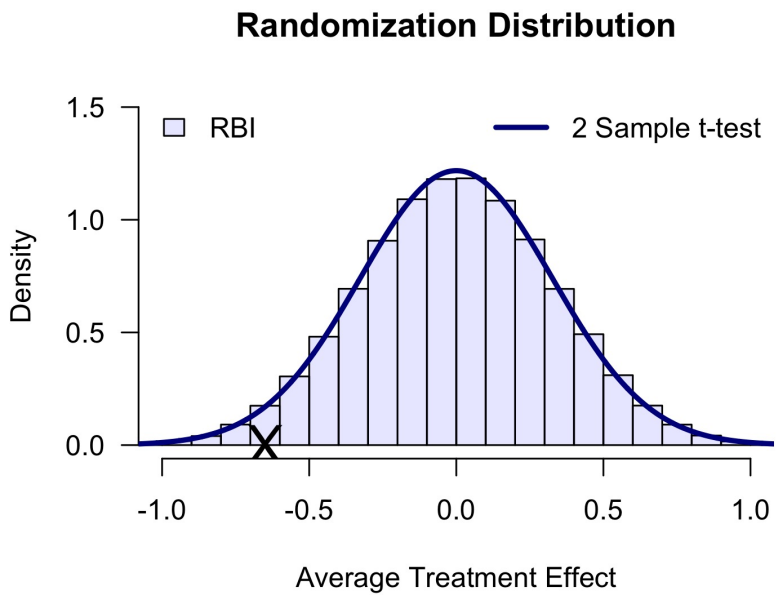
Pedagogical example

- 50 participants, equal allocation
- Observed outcomes ~ N(0,1)
- Sharp null (i.e., $Y_i(0) = Y_i(1)$)



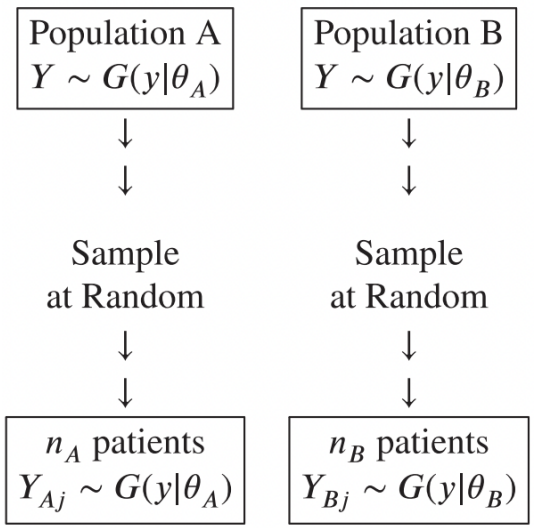
Randomization-Based Inference (RBI)

- 50 participants, equal allocation
- Observed outcomes ~ N(0,1)
- Sharp null (i.e., $Y_i(0) = Y_i(1)$)

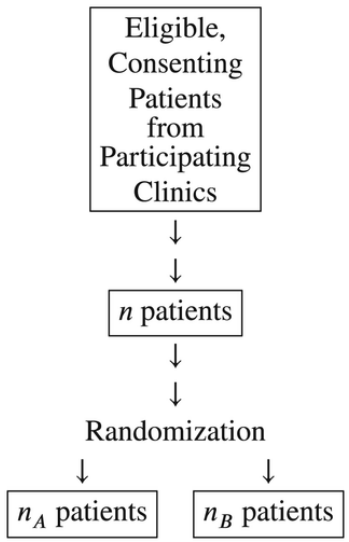


- T-Test: Super-population (no treatment assignment variability)
- RBI: Finite population (no sampling variability)

Clinical Trial Participants and Analysis (Rosenberger et al., 2019)

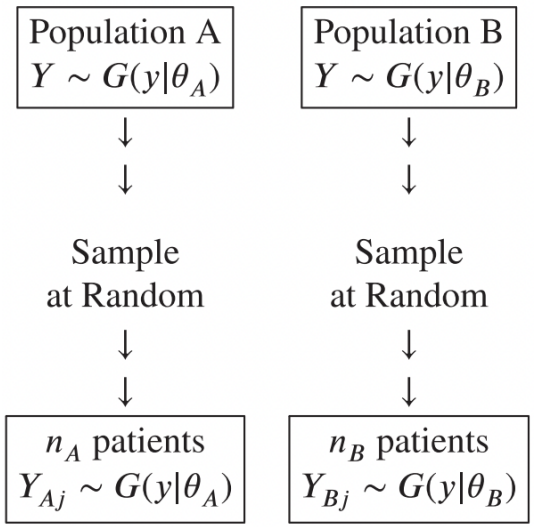


Population model

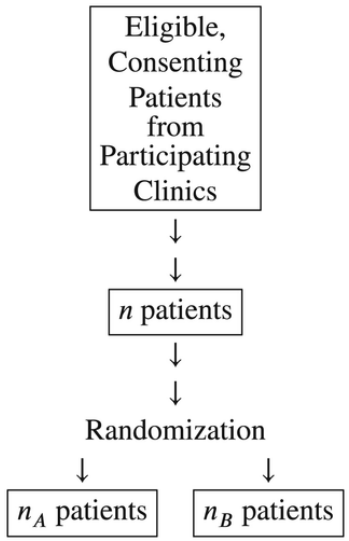


Randomization model

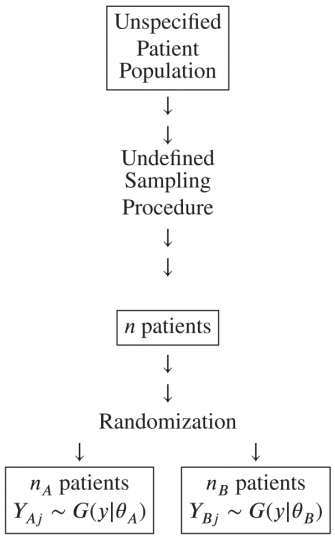
Clinical Trial Participants and Analysis (Rosenberger et al., 2019)



Population model



Randomization model

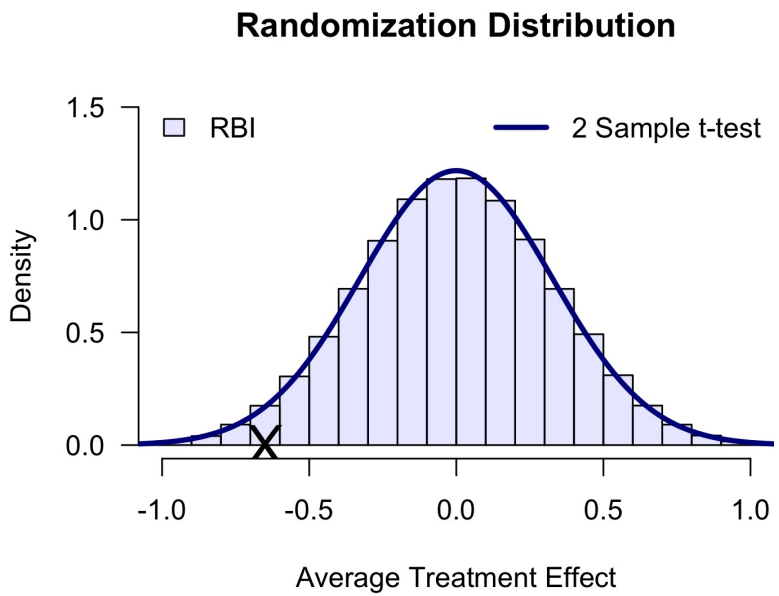


Invoked population model

ANOVA on finite or hypothetical population

Randomization-Based Inference (RBI)

- 50 participants, equal allocation
- Observed outcomes ~ N(0,1)
- Sharp null (i.e., $Y_i(0) = Y_i(1)$)



- T-Test: Super-population (no treatment assignment variability)
- RBI: Finite population (no sampling variability)

Randomization restrictions

No restrictions: Finite Central Limit Theorems (Li and Ding (2017))

- Derived assuming Complete Randomization (pre-specified n's)

Reducing chronological imbalance (with sequential enrollment)

- Permuted Block Design
- Maximum Tolerable Imbalance Designs
  - Less predictable than block designs (Berger et al., 2021)
  - Example: Big Stick Design

Reducing covariate imbalance*

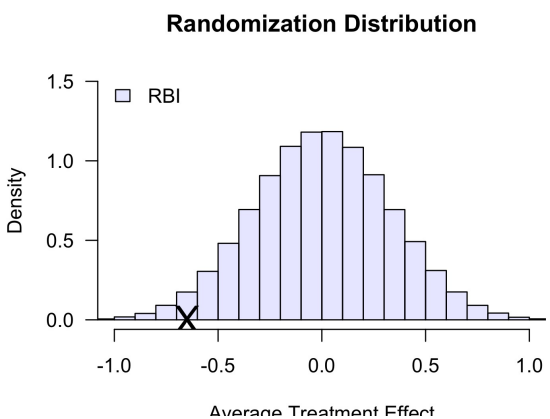
- Stratification
- Rerandomization
- Matching
- Minimization with a biased coin

* Covariate Adjusted Randomization (CAR)

Covariate-adjusted randomization

- 50 participants, equal allocation
- Observed outcomes ~ N(0,1)
- Sharp null (i.e., $Y_i(0) = Y_i(1)$)

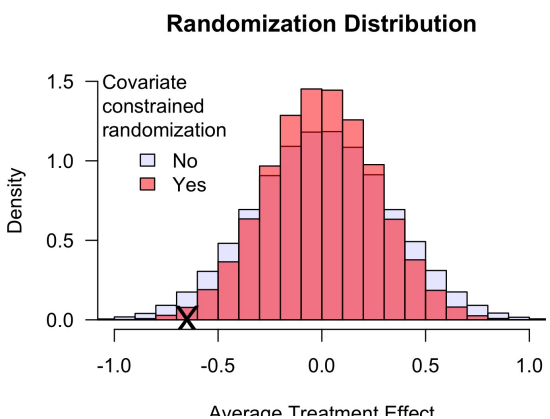
- $X \sim N(0,1)$ baseline health status, $\rho_{X,Y} = 0.67$
- Constrain allowable randomization sequences
  - Standardized Mean Difference (SMD): $|(\overline{X}_T - \overline{X}_C)/SD(X)|$
  - Exclude randomizations resulting in SMD > 0.2 SD
  - Re-randomization (Morgan and Rubin, 2012)



Covariate adjusted randomization

- 50 participants, equal allocation
- Observed outcomes ~ N(0,1)
- Sharp null (i.e., $Y_i(0) = Y_i(1)$)

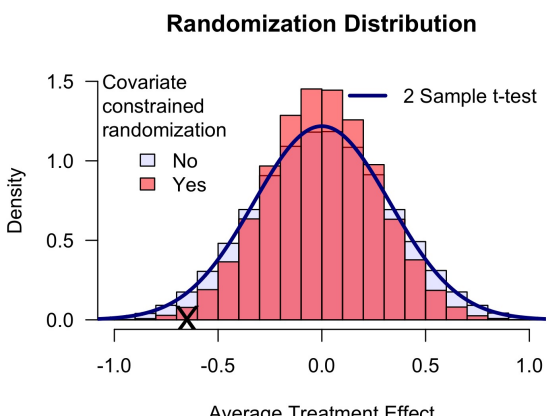
- $X \sim N(0,1)$ baseline health status, $\rho_{X,Y} = 0.67$
- Constrain allowable randomization sequences
  - Standardized Mean Difference (SMD): $|(\overline{X}_T - \overline{X}_C)/SD(X)|$
  - Exclude randomizations resulting in SMD > 0.2 SD
  - Re-randomization (Morgan and Rubin, 2012)



CAR+RBI

- 50 participants, equal allocation
- Observed outcomes ~ N(0,1)
- Sharp null (i.e., $Y_i(0) = Y_i(1)$)

- $X \sim N(0,1)$ baseline health status, $\rho_{X,Y} = 0.67$
- Constrain allowable randomization sequences
  - Standardized Mean Difference (SMD): $|(\overline{X}_T - \overline{X}_C)/SD(X)|$
  - Exclude randomizations resulting in SMD $>$ 0.2 SD
  - Re-randomization (Morgan and Rubin, 2012)



After CAR

- ANOVA is conservative
- Inference requires correction/adjustment (e.g., ANCOVA)
- Or ... use RBI

Restrictions to reduce chronological imbalance

Real-world example: Rare and ultra-rare diseases

In the US

- Rare: $\leq$ 200K individuals
- Ultra-rare: $\leq$ 6600 individuals (per England and Scotland definition)

Suppose you perform a trial

- 200 participants in finite population of 200K individuals
- Target population: 200K patient horizon
- Sample population: Eligible, Consenting, and Available individuals
- Finite sample population: Yes
- Random sample of finite population: Arguably so with assumptions
  1. Eligibility: Limits generalization
  2. Willingness: Possible high level of willingness
  3. Availability (Sites and timing): Questionable impact

RBI and ANOVA inference

Sampling Frame	Inference: Randomization-Based Inference (RBI)	Inference: Analysis of Variance (ANOVA)
Finite ($n = N$)	✓ Consistent with sampling frame Estimation: Unbiased Uncertainty: 1. Exactly controlled Type I error 2. Generalizes to observed sample	✗ Not consistent with sampling frame Estimation: Unbiased Uncertainty: 1. At least asymptotically correct Type I error* 2. Extrapolates to super-population reflective of sample
Super Population ($n < N < \infty$)	✗ Not consistent with sampling frame Estimation: Unbiased Uncertainty: 1. Exactly controlled Type I error 2. Generalizes to observed sample only	✓ Consistent with sampling frame Estimation: Unbiased Uncertainty: 1. At least asymptotically correct Type I error* 2. Generalizes to super-population reflective of sample

Contribution of paper, in finite populations with/without sampling:

1. Chronological restrictions slow the convergence of ANOVA Type I error
2. Adjusting for restrictions can improve convergence

Type I error convergence in a finite population?

Type 1 Error convergence when $n = N$ and $n < N$

Step 1: Generate a population of size N with independent draws from $\mathcal{N}(0,1)$

Sample (or not):
A) If $n < N$, sample n without replacement. Skip step if $n = N$.

Treatment assignment uncertainty:
B) Generate a randomization sequence from the randomization scheme.
C) Perform ANOVA and ANCOVA tests and record whether the test would reject ($p_{test} < \alpha$):
D) Repeat B) and C) `nrands` times to estimate rejection rate within sample.

Sampling uncertainty (random sampling uncertainty):
E) If $n < N$, Repeat A) through D) `nsamps` times to estimate rejection rate across samples.

Overall rejection rate:
F) Estimate Type I error, per test, for the population as the proportion of total rejections.

Step 2: Repeat Step 1 `npops` times

Empirical investigations

Randomization Schemes:

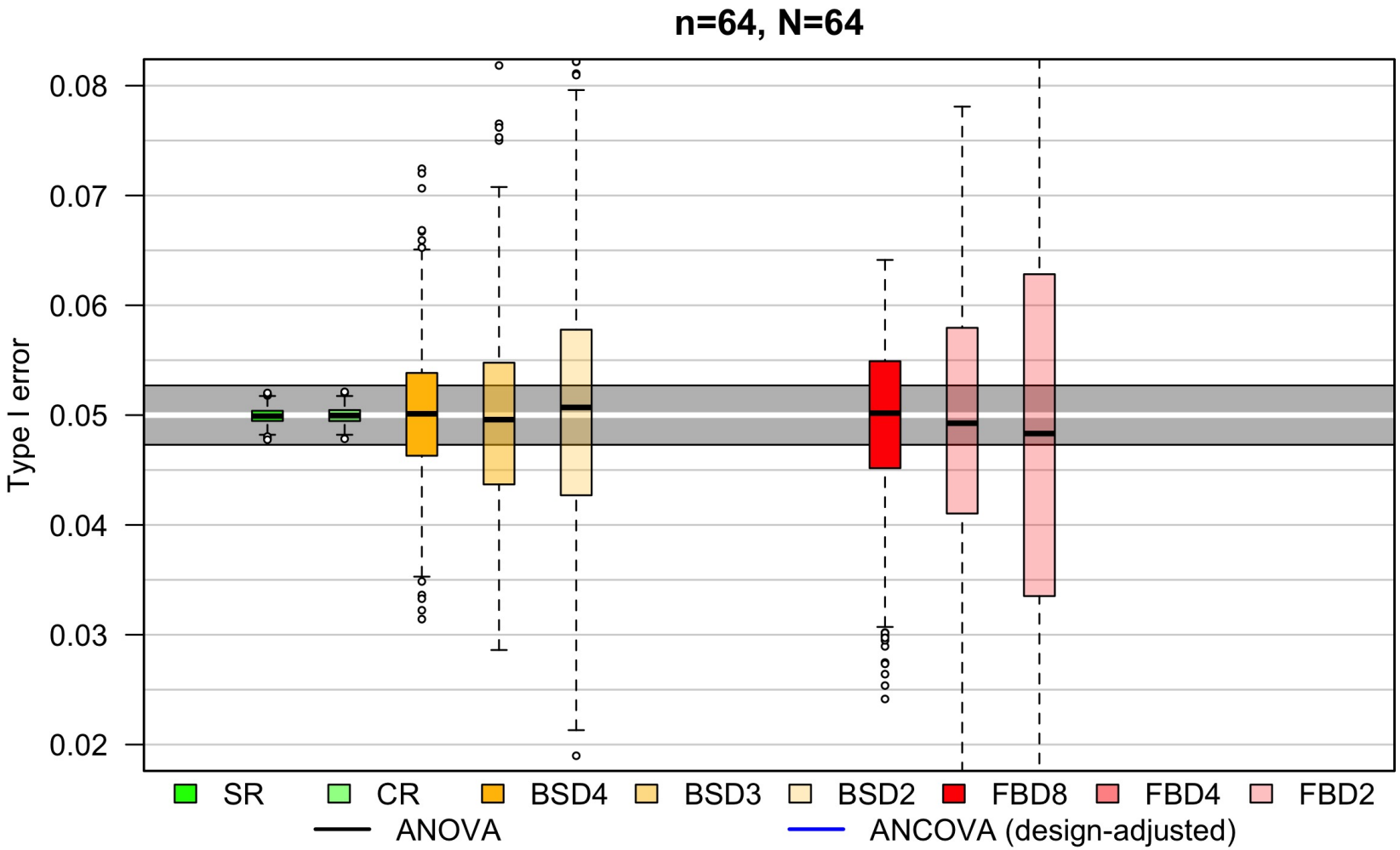
- Simple Randomization
- Complete/Equal Randomization
- Fixed Block Randomization (block sizes: 2, 4, 6, 8)
- Big Stick Design (MTI: 2, 3, 4)

Analysis:

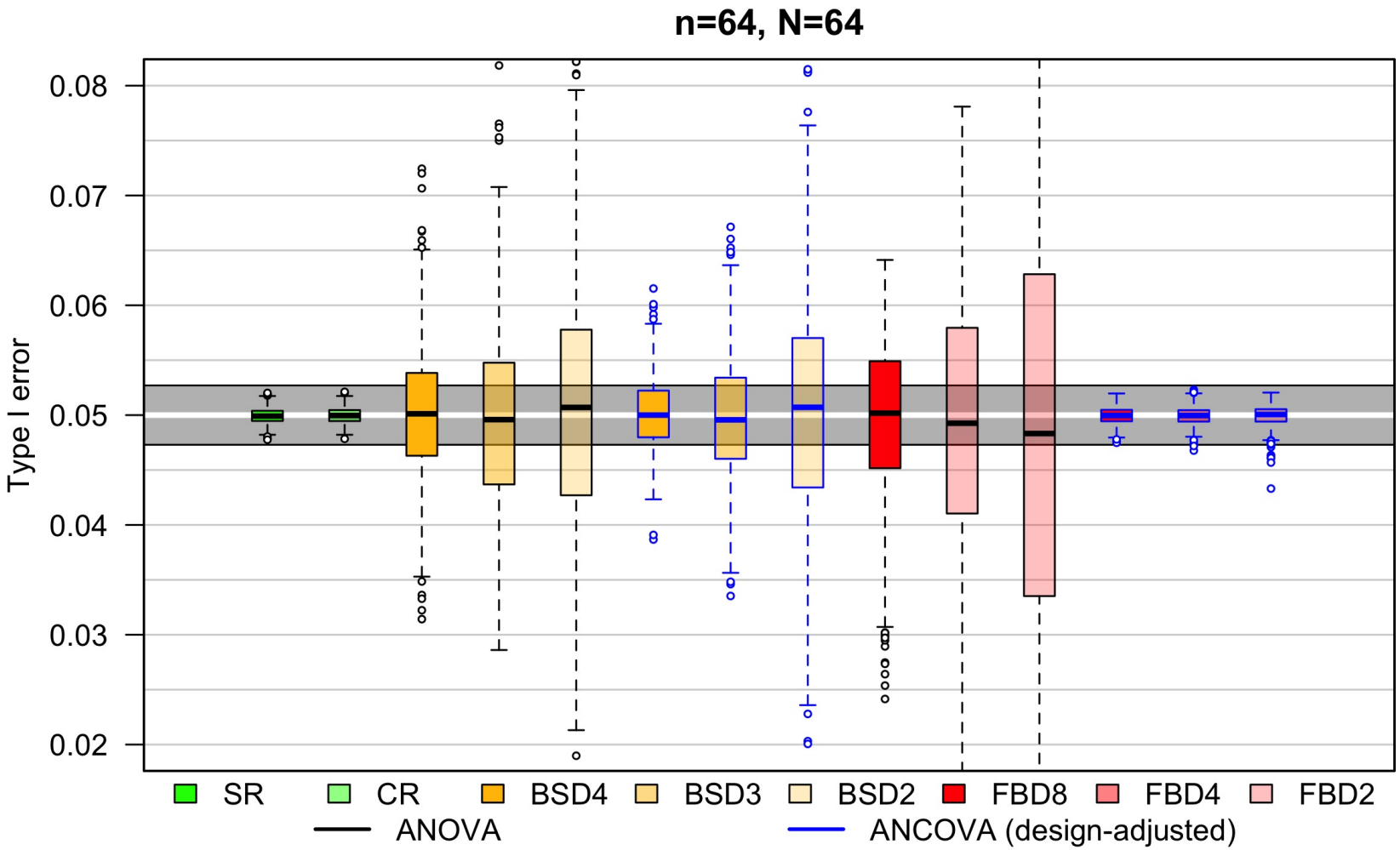
- ANOVA
- ANCOVA adjusting for randomization restrictions
  - Fixed Block: Block ID
  - Big Stick Design: Indicator of being at MTI threshold
  - Motivation: Subset to Simple/Complete Randomization

Sampling frame: Trial sample size $n = 16 - 1024$, Population $N \geq n$ trial sample size

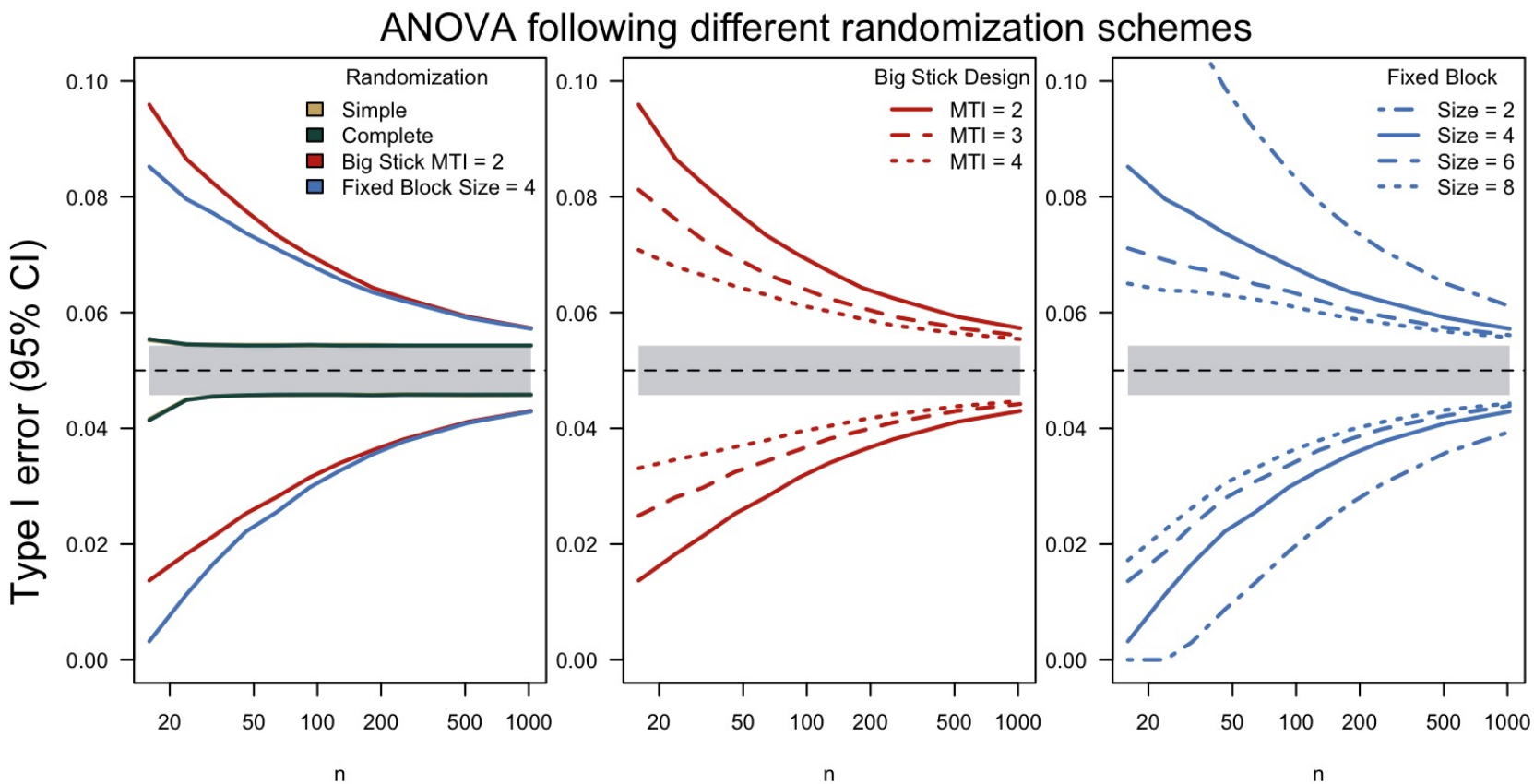
Empirical investigations, $n = N$



Empirical investigations, $n = N$

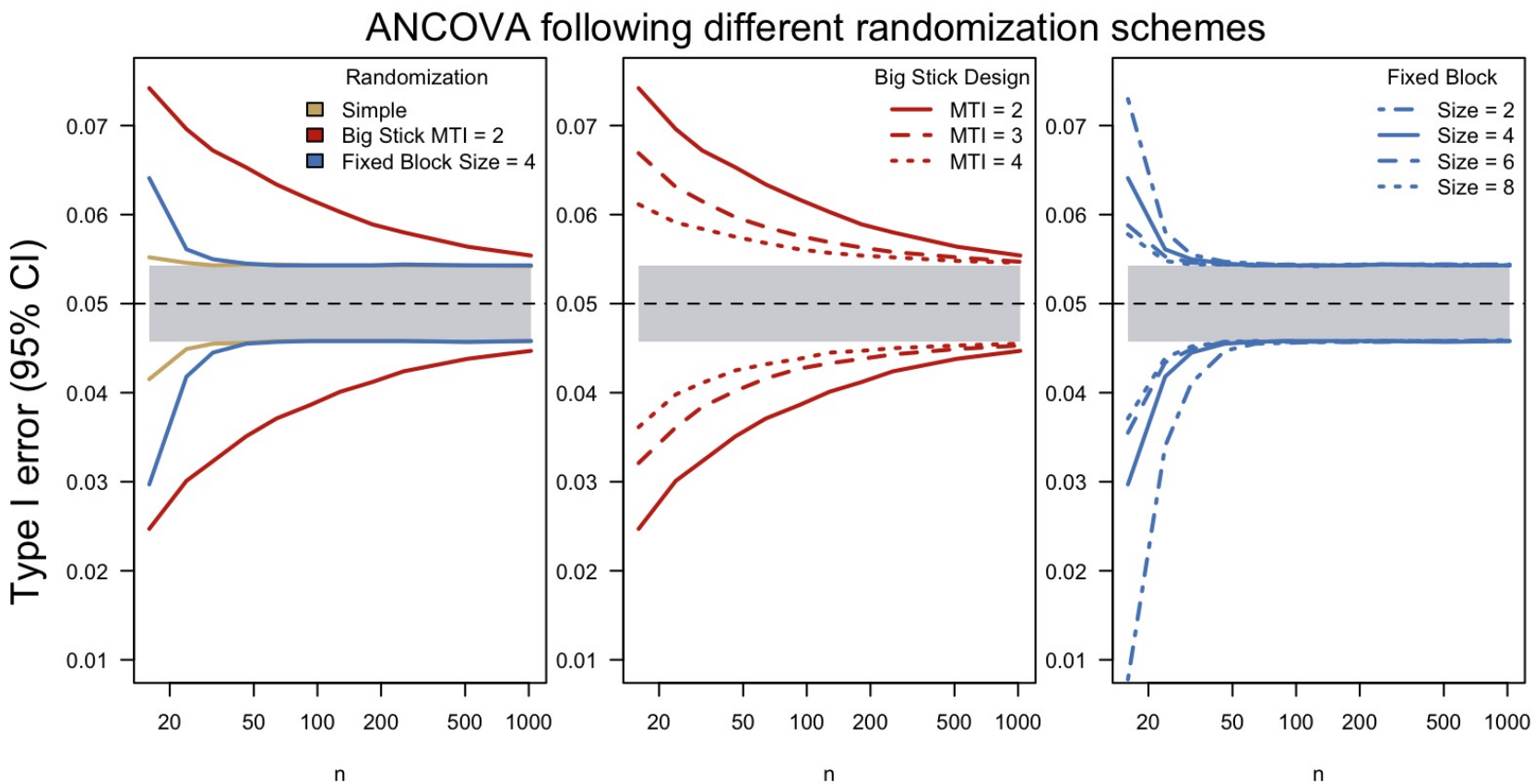


Empirical investigations, $n = N$



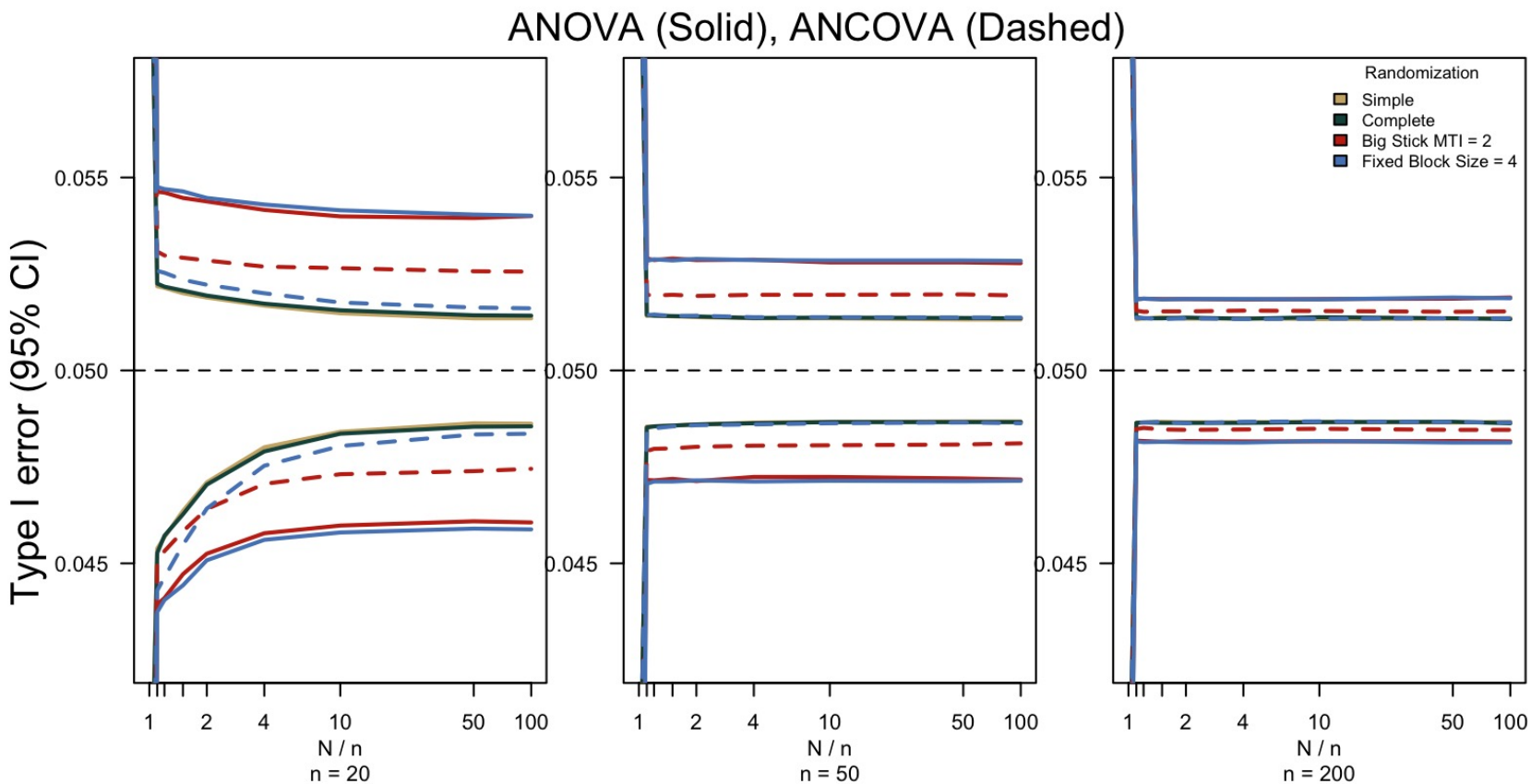
npops = 30K and nrands = 10K

Empirical investigations, $n = N$



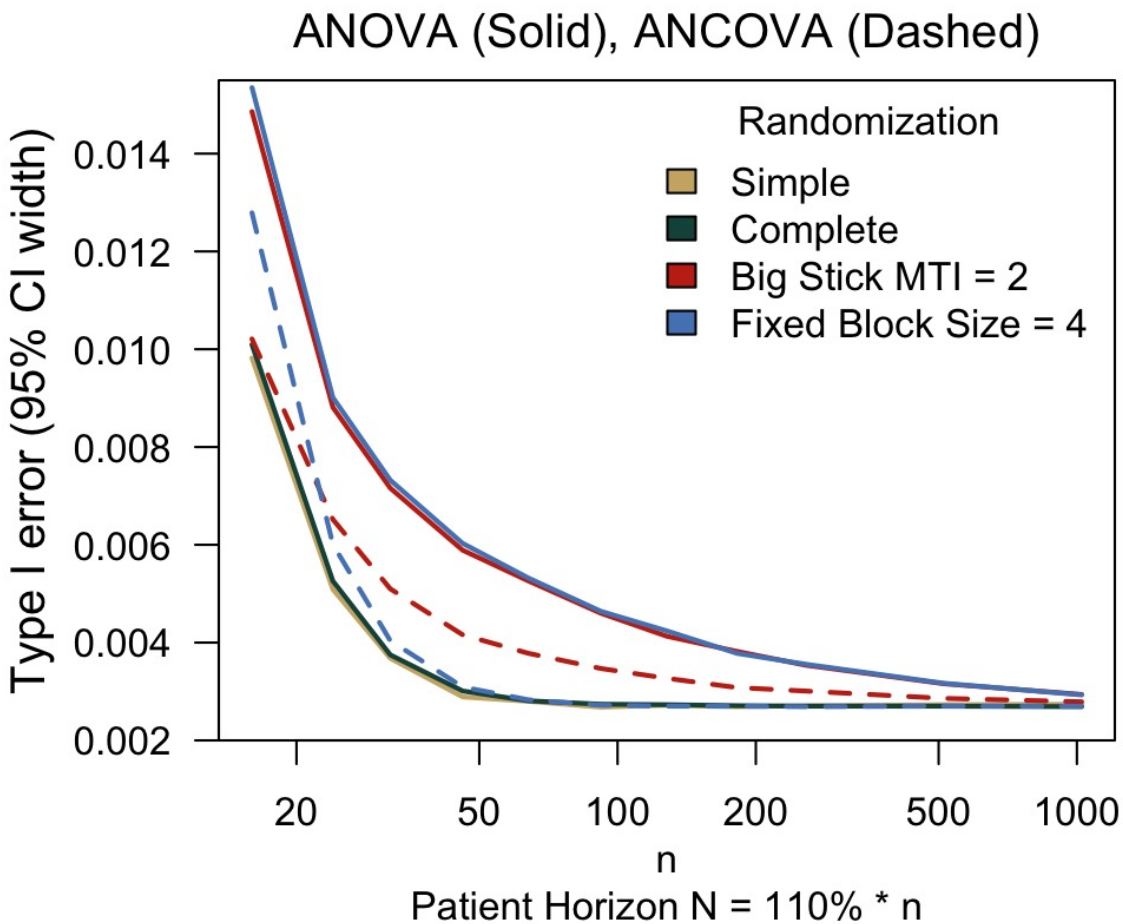
`npops` $= 30K$ and `nrands` $= 10K$

Empirical investigations, $n < N$



npops $= 30$K, nrands $= 1$K, and nrands $= 100$

Empirical investigations, $n < N$



`npops` $=$ 30K, `nrands` $=$ 1K, and `nrands` $=$ 100

Conclusions

For finite populations, empirical results suggest that:

1. Randomization restrictions impact ANOVA Type I error convergence
2. ANCOVA adjustment for restrictions can improve convergence
3. Seemingly, the ANCOVA adjustment is best when it reduces to subsets of Simple or Complete Randomization
4. Random sampling accelerates Type I error convergence, though ANCOVA was still preferable

Echoes of: "As ye randomize, so shall you analyze"

Viewpoints on RBI and clinical trial estimands

I lean toward the finite population interpretation of trial participants.

Yet, RBI is not as thoroughly developed as other inference strategies.

Supposing RBI can sufficiently address the analysis, then use RBI. Otherwise, use model-based strategies as needed for the analysis.

I (we, the co-authors) would like to see the estimand framework to better distinguish between the target and trial population and ability to extrapolate from the trial to target population.

- Eligibility
- Willingness
- Availability (Sites and timing)

End of Part I ... Questions

Paper near submission: Randomization Restrictions: Their Impact on Type I Error When Experimenting with Finite Populations

Restrictions to reduce covariate imbalance

FDA 2023 Guidance for covariate-adjustment in RCTs (FDA, 2023)

FDA 2023 Guidance for covariate-adjustment in RCTs (FDA, 2023)

Estimand of interest: Population-level treatment effect

FDA 2023 Guidance for covariate-adjustment in RCTs (FDA, 2023)

Estimand of interest: Population-level treatment effect

An unadjusted estimator is internally valid

- Unbiased
- Correct Type I error when modeling assumptions are met
- Could be more efficient by adjusting for covariates

FDA 2023 Guidance for covariate-adjustment in RCTs (FDA, 2023)

Estimand of interest: Population-level treatment effect

An unadjusted estimator is internally valid

- Unbiased
- Correct Type I error when modeling assumptions are met
- Could be more efficient by adjusting for covariates

Podcast Q and A: What are a couple of key items that you especially want listeners to remember?

I really want listeners to remember that **the FDA encourages covariate adjustment** because we believe that it is a **low hanging fruit** that can be used **to improve the efficiency** of a clinical trial analysis without creating additional burdens for sponsors. We encourage sponsors to discuss covariate adjustment with the FDA during the development of the protocol, particularly for situations not explicitly covered in the guidance. - Dan Rubin

FDA 2023 Guidance for covariate-adjustment in RCTs (FDA, 2023)

Estimand of interest: Population-level treatment effect

An unadjusted estimator is internally valid

- Unbiased
- Correct Type I error when modeling assumptions are met
- Could be more efficient by adjusting for covariates

Podcast Q and A: What are a couple of key items that you especially want listeners to remember?

I really want listeners to remember that **the FDA encourages covariate adjustment** because we believe that it is a **low hanging fruit** that can be used **to improve the efficiency** of a clinical trial analysis without creating additional burdens for sponsors. We encourage sponsors to discuss covariate adjustment with the FDA during the development of the protocol, particularly for situations not explicitly covered in the guidance. - Dan Rubin

CAR+RBI

Proactive covariate adjustment

Benefits

- Could increase efficiency
- Could reduce covariate imbalances
- Non-parametric, exact test
- Unadjusted estimator
  (Simple summary measure)

Considerations

- Finite population estimand
- Real-time implementation
- See also slides 33-36

CAR+RBI

Proactive covariate adjustment

Benefits

- Could increase efficiency
- Could reduce covariate imbalances
- Non-parametric, exact test
- Unadjusted estimator
  (Simple summary measure)

Considerations

- Finite population estimand
- Real-time implementation
- See also slides 33-36

Can CAR+RBI be as powerful as regression adjustment?

Can CAR+RBI be as powerful as regression adjustment?

Context of linear models, power increases when ...

- Model specification: Closely approximating the relationship between the outcomes, treatment, and covariates
  - Y ~ TX + COV + TX*COV (Lin, 2013)
  - Can be complex to specify

- Treatment assignment: TX $\perp$ COV (no covariate imbalance) (Atkinson, 1982; Senn et al., 2010)

Conjecture: A good CAR with RBI can competitively capture regression adjustment efficiency

1. Can reduce chance imbalances
2. Averts model complexity

Selected CAR strategies

When X is not known for all participants before randomization

1. Stratified randomization
2. Minimization with a biased coin (Pocock, 1977)
3. Sequential Matched (and Rematched) randomization (Kapelner and Krieger, 2014; Chipman et al., 2023)
4. Sequential Re-randomization (Zhou et al., 2018)

Stratified Randomization

- Most commonly implemented covariate-adjusted randomization scheme (Sverdlov et al., 2023; McPherson et al., 2012)

- Randomize within categorized patient profiles
  - Continuous covariates must be categorized
  - Quickly limited by the number of adjusting covariates

Minimization with a biased coin

Weighted randomization to arm that reduces imbalance (Pocock and Simon, 1975)

Examples:

1. D_A Biased Coin Design (D_A-BCD) (Atkinson, 1982)

   Minimize standard error of ATE from pre-specified model

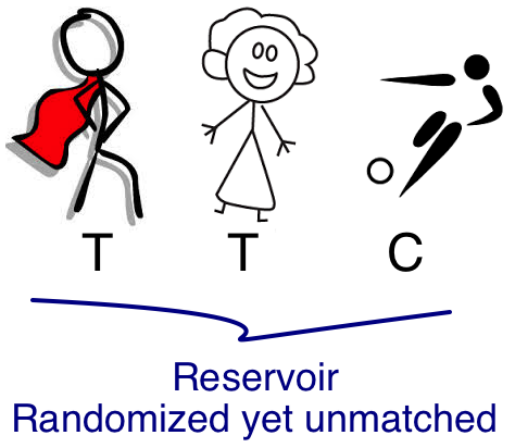
2. Pairwise Sequential Randomization (PSR) (Qin et al., 2016; Ma et al., 2020):

$$M = (\overline{x}_T - \overline{x}_C)' cov(\overline{x}_T - \overline{x}_C)^{-1} (\overline{x}_T - \overline{x}_C)$$

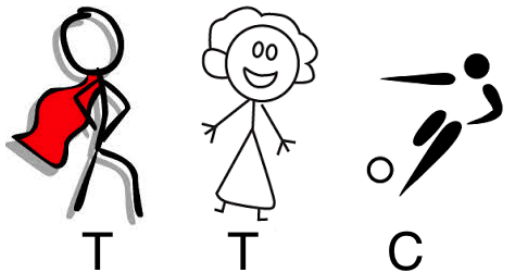
Sequential Matching (Kapelner and Krieger, 2014)

Sequential Matching (Kapelner and Krieger, 2014)

Sequential Matching (Kapelner and Krieger, 2014)



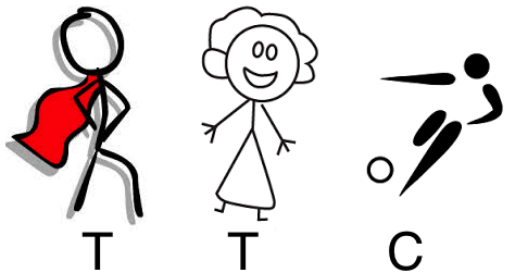
Sequential Matching (Kapelner and Krieger, 2014)



If $\text{Distance}_2 < \text{Threshold}$

Reservoir
Randomized yet unmatched

Sequential Matching (Kapelner and Krieger, 2014)



If $\text{Distance}_2 > \text{Threshold}$

Sequential Matching, additional notes (Kapelner and Krieger, 2014)

Implementation details ... See also (Chipman et al., 2023).

CAR + RBI in real-world case study

REACH Trial: Causal questions, estimands, and design

Rapid **E**ducation/Encouragement **A**nd **C**ommunications for **H**ealth

Population	Adults with Type 2 Diabetes (DM)
Purpose	Increase glycemic control and adherence to medications
Main Intervention	Text message-delivered diabetes support for 12 months
Outcome	12 month glycemic control (A1c) compared to control
Multi-site enrollment	512 patients from Vanderbilt and Non-Vanderbilt Clinics

Key Baseline Covariates

Biological Factors

- Baseline A1c*
- Age at baseline
- Time since DM dx*
- DM type*
- Race / Ethnicity

Socio-economic Factors

- Yrs of education
- Income level
- Insurance type

* Greater priority for balancing ($R^2 = 0.26$ vs 0.32 for all covariates)

Questions of interest

1. Can CAR+RBI be as powerful as regression adjustment?

2. How sensitive are CAR schemes to the number of covariates?

Randomization models:

- Complete Randomization (CR)
- Covariate-Adjusted Randomization (CAR)
  - Stratified randomization
  - D_A-BCD (3/4 biased coin)
  - PSR (3/4 biased coin)
  - Sequential Matched Randomization and extensions

OLS-Adjustment: Linear adjustment for covariates (no interactions)

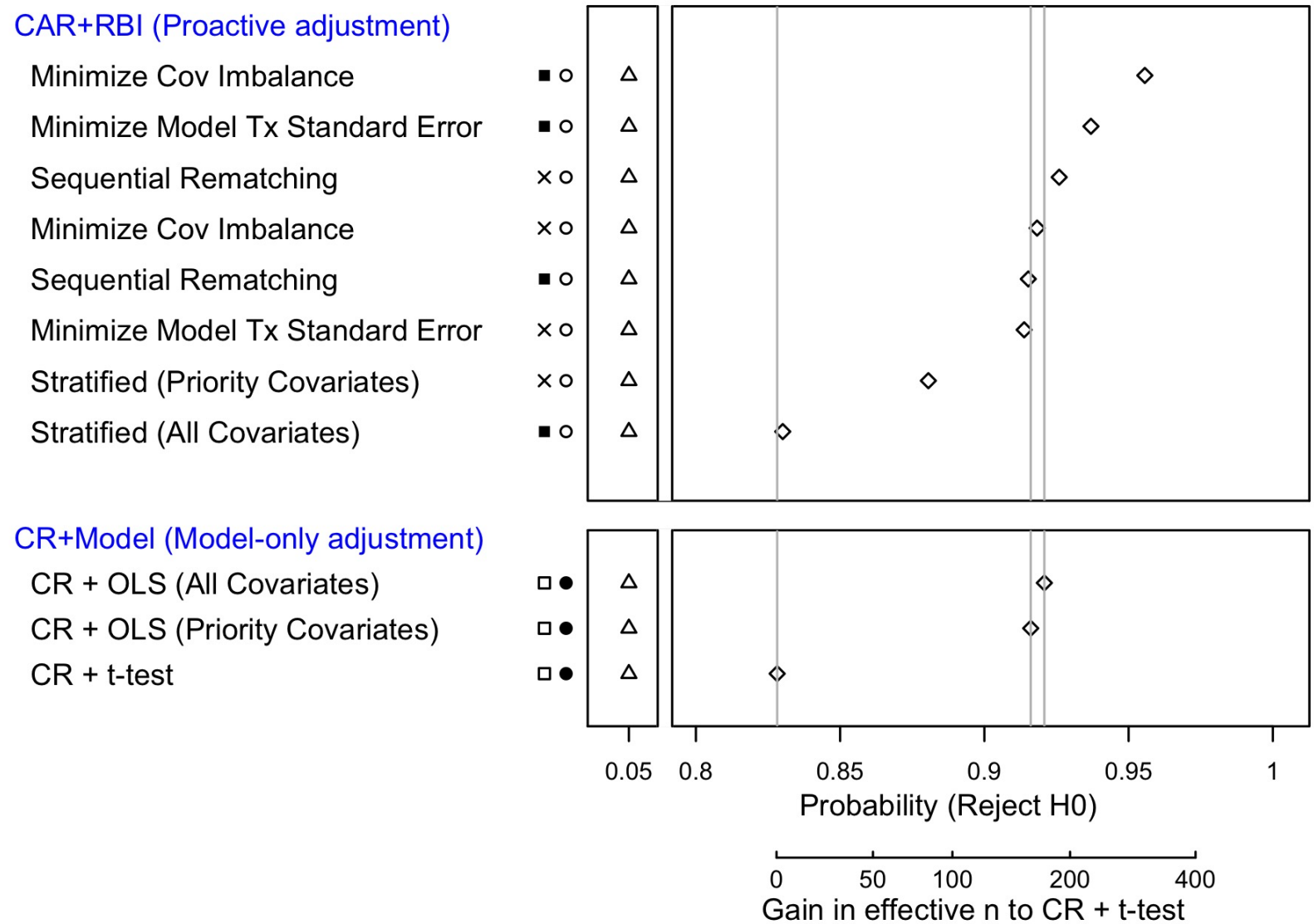
Simulation of REACH outcomes

For sharp treatment effects of 0 (null) and -0.5 (beneficial)

1. Conditioning on enrollment (i.e., randomness via trt assignment)
2. 100K Randomization sequences for each randomization procedure
3. For each sequence, record
   - Covariate balance: Average SMD across covariates
   - Study efficiency: Whether to reject H_0: ATE $= 0$
   - Match quality: Distance of matched pairs

Case-study caveats:

- Assumes variability due to observed outcomes
- Results may differ for each trial



Covariate-Adjustment of Randomization
- None
- Site and Priority Covariates
- All Covariates

Inference
- Randomization-Based
- Parametric

Treatment Effect
- 0 (H0)
- -0.50

REACH Trial: Case study

1. Can CAR+RBI be as powerful as regression adjustment?
   - SRR, PSR, and D_A-BCD were more powerful
   - SRR, PSR, and D_A-BCD with RBI increased effective n by ≥ 170 compared to CR with unadjusted test
2. How sensitive are CAR schemes to the number of covariates?
   - Stratification: Worsened when over-adjusted
   - Sequential Matching: Mixed impact
   - PSR and DA-BCD: improved with more covariates

Generalized code to reassess questions for any trial (Chipman et al., 2023)

Questions to consider

Covariate-adjustment with RBI:

- Can increase efficiency
- Can reduce covariate imbalances
- Can use an unadjusted estimator (simple summary measure)
- Is a non-parametric, exact test

What barriers/considerations remain for adopting this analytic strategy?

1. Reconciling with finite population estimand
2. Assumption of a 'sharp'-null hypothesis
3. Adoption into databases for real-time randomization
4. Modelling may be desirable for more complex questions
5. Data may have missing outcomes or covariates

Reconciling with finite population estimands

FDA Estimand of interest: Population-level treatment effect

RCT's carry internal validity but are generally limited in external validity:

- Participants are not a random sample of broader population

Viewpoint: A finite population estimand is consistent with the generalizability of the trial's participants.

Additional considerations

- FDA supports RBI and has approved treatments based upon RBI(FDA, 2023; FDA, 2017)
- A super-population model can be used Imbens and Menzel (2021)

Questions to consider

What barriers/considerations remain for adopting this analytic strategy?

1. Reconciling with finite population estimands
   - FDA supports RBI and has approved treatments based upon RBI(FDA, 2023; FDA, 2017)
   - A super-population model can be used Imbens and Menzel (2021)
   - Trial participants are often unlike the population of interest
2. Assumption of a 'sharp'-null hypothesis
   - A sharp-null can be relaxed (Imbens and Menzel, 2021)
3. Adoption into databases for real-time randomization
   - Proof of principle in ECOG (Lange and MacIntyre, 1985)
4. Modelling may be desirable for more complex questions
   - (Shao et al., 2010; Kapelner and Krieger, 2014; Ma et al., 2020; Bannick et al., 2023)
5. Data may have missing outcomes or covariates
   - (Rubin, 1998; Ivanova et al., 2022; Heussen et al., 2023)

End of part II

Chipman, J. J., Mayberry, L., and Greevy, R. A. J. (2023). Rematching on-the-fly: Sequential matched randomization and a case for covariate-adjusted randomization.
Statistics in medicine, 42(22):3981–3995

Overview:

1. Unadjusted chronological restrictions can impact Type I error
2. Adjusting for covariates in randomization can be more powerful than adjusting in an ANCOVA model

References I

Atkinson, A. C. (1982). Optimum Biased Coin Designs for Sequential Clinical Trials with Prognostic factors. *Biometrika*, 69(1):61.

Bannick, M. S., Shao, J., Liu, J., Du, Y., Yi, Y., and Ye, T. (2023). A General Form of Covariate Adjustment in Randomized Clinical Trials. *arXiv preprint arXiv:2306.10213*.

Berger, V. W., Bour, L. J., Carter, K., Chipman, J. J., Everett, C. C., Heussen, N., Hewitt, C., Hilgers, R. D., Luo, Y. A., Renteria, J., Ryeznik, Y., Sverdlov, O., Uschner, D., and Beckman, R. A. (2021). A roadmap to using randomization in clinical trials. *BMC Medical Research Methodology*, 21(1).

Chipman, J. J., Mayberry, L., and Greevy, R. A. J. (2023). Rematching on-the-fly: Sequential matched randomization and a case for covariate-adjusted randomization. *Statistics in medicine*, 42(22):3981–3995.

References II

Heussen, N., Hilgers, R.-D., Rosenberger, W. F., Tan, X., and Uschner, D. (2023). Randomization-based inference for clinical trials with missing outcome data. *Statistics in Biopharmaceutical Research*, pages 1–12.

Imbens, G. and Menzel, K. (2021). A causal bootstrap. *Annals of Statistics*, 49(3):1460–1488.

Ivanova, A., Lederman, S., Stark, P. B., Sullivan, G., and Vaughn, B. (2022). Randomization tests in clinical trials with multiple imputation for handling missing data. *Journal of Biopharmaceutical Statistics*, 32(3):441–449.

Kapelner, A. and Krieger, A. (2014). Matching on-the-fly: Sequential allocation with higher power and efficiency. *Biometrics*, 70(2):378–388.

Lange, N. and MacIntyre, J. (1985). A computerized patient registration and treatment randomization system for multi-institutional clinical trials. *Controlled Clinical Trials*, 6(1):38–50.

References III

Li, X. and Ding, P. (2017). General Forms of Finite Population Central Limit Theorems with Applications to Causal Inference. *Journal of the American Statistical Association*, 112(520):1759–1769.

Lin, W. (2013). Agnostic notes on regression adjustments to experimental data: Reexamining Freedman' s critique. *The Annals of Applied Statistics*, 7(1):295–318.

Ma, W., Qin, Y., Li, Y., and Hu, F. (2020). Statistical Inference for Covariate-Adaptive Randomization Procedures. *Journal of the American Statistical Association*, 115(531):1488–1497.

McPherson, G. C., Campbell, M. K., and Elbourne, D. R. (2012). Use of randomisation in clinical trials: a survey of UK practice. *Trials*, 13(1):1–7.

Morgan, K. L. and Rubin, D. B. (2012). Rerandomization to improve covariate balance in experiments. *The Annals of Statistics*.

References IV

Pocock, S. J. (1977). Group sequential methods in the design and analysis of clinical trials. *Biometrika*, 64(2):191–199.

Pocock, S. J. and Simon, R. (1975). Sequential Treatment Assignment with Balancing for Prognostic Factors in the Controlled Clinical Trial. *Biometrics*, 31(1):103–115.

Qin, Y., Li, Y., Ma, W., and Hu, F. (2016). Pairwise Sequential Randomization and Its Properties. *arXiv:1611.02802*.

Rosenberger, W. F., Uschner, D., and Wang, Y. (2019). Randomization: The forgotten component of the randomized clinical trial. *Statistics in Medicine*, 38(1):1–12.

Rubin, D. B. (1998). More powerful randomization-based p-values in double-blind trials with non-compliance. *Statistics in Medicine*, 17(3):371–385.

References V

Senn, S., Anisimov, V. V., and Fedorov, V. V. (2010). Comparisons of minimization and Atkinson's algorithm. *Statistics in medicine*, 29(7-8):721–730.

Shao, J., Yu, X., and Zhong, B. (2010). A theory for testing hypotheses under covariate-adaptive randomization. *Biometrika*, 97(2):347–360.

Sverdlov, O., Carter, K., Hilgers, R.-D., Everett, C. C., Berger, V. W., Luo, Y. A., Chipman, J. J., Ryeznik, Y., Ross, J., Knight, R., and Yamada, K. (2023). Which Randomization Methods Are Used Most Frequently in Clinical Trials? Results of a Survey by the Randomization Working Group. *Statistics in Biopharmaceutical Research*, 0(ja):1–21.

U.S. Department of Health and Human Services Food and Drug Administration (2023). Adjusting for Covariates in Randomized Clinical Trials for Drugs and Biological Products. *Guidance Document*, (May).

References VI

U.S. Food and Drug Administration (2017). Statistical Review and Evaluation of NDA/Serial Number: 125261 / 138. https://www.fda.gov/media/108953/download.

Zhou, Q., Ernst, P. A., Morgan, K. L., Rubin, D. B., and Zhang, A. (2018). Sequential rerandomization. *Biometrika*, 105(3):745–752.

BSD ANCOVA adjustments