Adaptive Enrichment Trial Designs: Statistical Methods, Trial Optimization Software, and Case Studies

Michael Rosenblum and Joshua Betz
Department of Biostatistics
Johns Hopkins Bloomberg School of Public Health
mrosen@jhu.edu

Joint work: Aaron Fisher, Jon Steingrimsson, Ivan Diaz,
Adi Gherman, Tianchen Qian, Yu Du
July 13-14, 11:30AM-3PM EDT

To Request Free Account to Use Our Software

Go to: http://rosenblum.jhu.edu

Software runs on Firefox or Chrome

Disclaimer:

The opinions in this presentation are of the author (Michael Rosenblum) and do not necessarily represent Johns Hopkins University, the FDA/HHS, or anyone else.

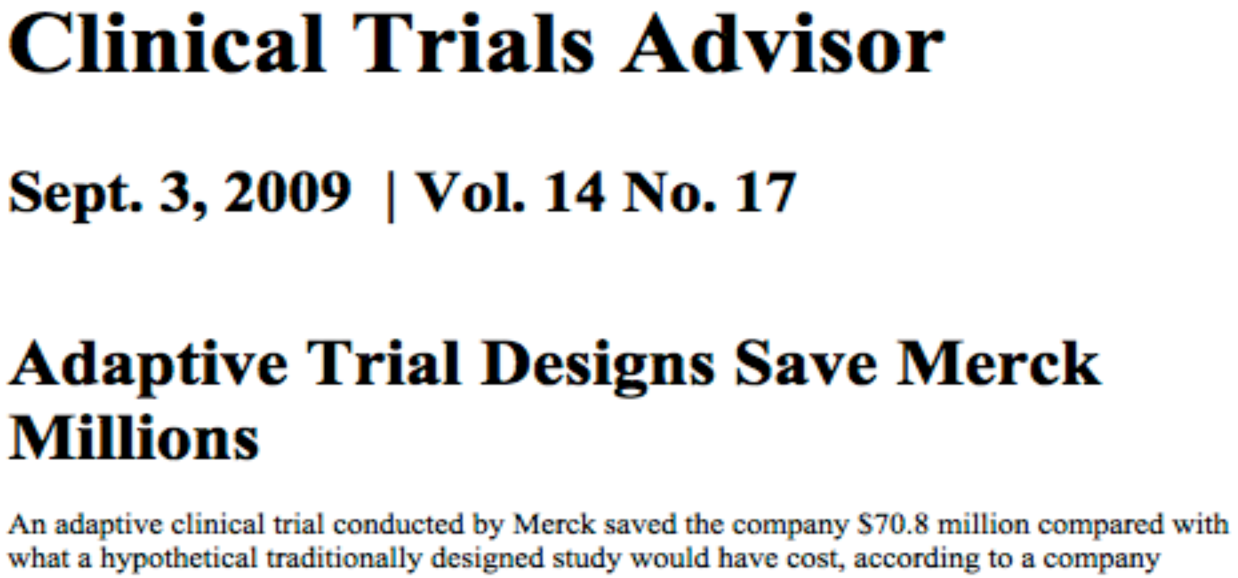
How Can I Help Fight the Climate Crisis?

Do you live in one of these states:
Illinois, Massachusetts, Maryland, Maine, Minnesota, New Jersey, New York, Oregon, Rhode Island.
Then you are likely eligible for Community Solar!

- **Community Solar:** You get environmental benefits of solar without installing solar panels at home.
- Subscribe to "share" of Community Solar project (solar array built on unused land) and get monthly credit on electric bill.
- More information: https://www.solarunitedneighbors.org/
- Email me if you'd like more information.

Adaptive Clinical Trial Designs

Pharmaceutical Companies are Interested:



Clinical Trials Advisor

Sept. 3, 2009 | Vol. 14 No. 17

Adaptive Trial Designs Save Merck Millions

An adaptive clinical trial conducted by Merck saved the company $70.8 million compared with what a hypothetical traditionally designed study would have cost, according to a company

"An adaptive clinical trial conducted by Merck saved the company $70.8 million compared with what a hypothetical traditionally designed study would have cost…"

Why Consider Adaptive Designs?

Potential Benefits:

- Can give More Power to Confirm Effective Treatments/Interventions and Determine Subpopulations who Benefit Most
- Can Reduce Cost, Duration, and Number of Participants
- Caution! adaptive design not always better

Challenge: find the best design tailored to clinical investigator's research question and resource constraints

Adaptive Designs

- Participants Enrolled over Time
- At Interim Analyses, Can Change Sampling in Response to Accrued Data:
  - Adaptive designs could involve changes to:
    - Sample size
    - <u>Enrollment criteria (“enrichment”—my focus)</u>
    - Length of follow-up
    - Randomization probabilities
    - Dose
- SMART designs: If participant fails on initial treatment, randomized to another.

Adaptive Clinical Trials: Overview

- According to guidelines from the European Medicines Agency (EMA) (2007) and US Food and Drug Administration (FDA) (2010, 2016), **adaptations in clinical trials should be protocol based**, i.e. based on a preplanned rule.
- Most common adaptations in phase 2 or 3 clinical trials (our focus) include:

  -Early stopping for efficacy, futility or harm (group sequential designs)

  -Sample size re-estimation

  -Dose selection

  -Biomarker adaptive designs (where biomarker measured at baseline)
- Above from surveys of Morgan et al. (2014), Elsäßer et al. (2014), Hatfield et al. (2016), Lin et al. (2016), Mistry et al. (2017), Bothwell et al. (2018)

Overview: Adaptive Enrichment Designs

Adaptive Enrichment Designs: preplanned rule for modifying enrollment based on accrued data in ongoing trial.

May be useful if suspected that treatment effect differs by subpopulation, e.g., defined by baseline biomarker or disease severity.

Caution: Adaptive design not always better—typically involve tradeoffs.

3 Examples of Adaptive Enrichment Trials

1. **DAWN trial**: mechanical thrombectomy vs. standard medical care for treating acute stroke.
   - Bayesian adaptive enrichment design
   - Inclusion criteria could be modified to restrict enrollment to subpopulations with smaller infarct sizes at baseline.
2. **DEFUSE 3 trial**: “adaptive design will identify, at interim analyses, the group with the best prospect for showing benefit from endovascular treatment, based on baseline core lesion volumes and the times since stroke onset. Interim analyses … at 200 and 340 patients, at which time the study may stop for efficacy/futility, or the inclusion criteria may be adjusted in the case of futility.“
3. **TAPPAS trial**: “adaptive enrichment phase 3 trial of TRC105 and pazopanib versus pazopanib alone in patients with advanced angiosarcoma.” Allows enrichment + sample size increase.

3 Examples of Adaptive Enrichment Trials

1. Tudor G Jovin, Jeffrey L Saver, Marc Ribo, Vitor Pereira, Anthony Furlan, Alain Bonafe, Blaise Baxter, Rishi Gupta, Demetrius Lopes, Olav Jansen, Wade Smith, Daryl Gress, Steven Hetts, Roger J Lewis, Ryan Shields, Scott M Berry, Todd L Graves, Tim Malisch, Ansaar Rai, Kevin N Sheth, David S Liebeskind, Raul G Nogueira. (2017) Diffusion-weighted imaging or computerized tomography perfusion assessment with clinical mismatch in the triage of wake up and late presenting strokes undergoing neurointervention with Trevo (DAWN) trial methods. International Journal of Stroke. Vol 12, Issue 6, pp. 641-652. DOI 10.1177/1747493017710341 [See also 2018 NEJM article: DOI: 10.1056/NEJMoa1706442]
2. Gregory W Albers. Endovascular Therapy Following Imaging Evaluation for Ischemic Stroke 3 (DEFUSE 3) ClinicalTrials.gov Identifier: NCT02586415. https://www.clinicaltrials.gov/ct2/show/NCT02586415?term=DEFUSE&rank=1
3. Robin Lewis Jones, Steven Attia, Cyrus R. Mehta, Lingyun Liu, Kamalesh Kumar Sankhala, Steven Ian Robinson, Vinod Ravi, Nicolas Penel, Silvia Stacchiotti, William D. Tap, Delia Alvarez, Richard Yocum, Charles P. Theuer, and Robert G. Maki. Tappas: An adaptive enrichment phase 3 trial of TRC105 and pazopanib versus pazopanib alone in patients with advanced angiosarcoma (AAS). Journal of Clinical Oncology 2017 35:15_suppl, TPS11081-TPS11081 NCT02979899. http://ascopubs.org/doi/abs/10.1200/JCO.2017.35.15_suppl.TPS11081

My group’s research on adaptive designs

1. New adaptive enrichment designs for time-to-event and other delayed outcomes.

2. User-friendly, free, open-source software to tailor adaptive enrichment design to clinical investigator's research question + compare to standard designs

3. Demonstrate in clinical applications: stroke (Dan Hanley), Alzheimer’s disease (Michela Gallagher), cardiac resynchronization devices (Boston Scientifc), and HIV prevention (Craig Hendrix)

Clinical Applications

Disease	**Data Sources (and Collaborator)**	**Subpop-ulation of Interest**	**Primary Outcome**
Stroke	MISTIE and CLEAR trials	Small vs. large clot volume	180-day disability score
Heart disease	SMART-AV trial	QRS > 150; women, men	Left Ventricular End Systolic Volume at 6 month
Alzheimer’s Disease	ADNI database	ApoE4 genetoype	12 month Clinical Dementia Rating
HIV	PEARLS trial	Women, Men	Time to virologic failure, AIDS, or death

Related Work on Adaptive Enrichment Designs

- **Adapt Treatments and/or Population Sampled:** Thall, Simon, Ellenberg 1988, Schaid, Wieand, Therneau 1990, Wittes and Brittain 1990, Follman 1997, **Russek-Cohen and Simon 1997**, Bauer and Köhne 1994, Bauer and Kieser 1999, Liu, Proschan, Pledger 2002, Stallard and Todd 2003, Sampson and Sill 2005, Bischoff and Miller 2005, Freidlin and Simon 2005, Jennison and Turnbull, 2003, 2006, 2007, **Wang, Hung, O'Neill 2009**, Magirr et al. 2012, Magnusson and Turnbull 2013, Maurer and Bretz (2013), Hampson and Jennison 2015.
- **Related Work on *Optimization* of Adaptive Enrichment Designs:** Götte, Donica, and Mordenti (2015), Graf, Posch, Koenig (2015), Krisam and Kieser (2015).

Related Work on Optimizing; Our Contributions

Our Main Contribution: Simultaneously optimize over large class of designs involving many parameters:

1. Rosenblum, Fang, Liu (2020) : optimize via sparse linear programming.

2. Fisher and Rosenblum (2018) : optimize via simulated annealing.

Our Adaptive Enrichment Design Features

1. Group sequential with prespecified enrollment adaptation rule.

2. Handles many outcome types (e.g., binary, continuous, time-to-event) as long as canonical covariance structure (Jennison and Turnbull 1999).

3. Strong control of familywise Type I error rate.

4. Enrollment adaptation rule and multiple testing procedures only depend on minimal sufficient statistics (Emerson 2006).

5. New multiple testing procedures to optimize power for adaptive enrichment designs (Rosenblum et al. 2016).

Stroke Trial Application

New Surgical Technique to Treat Intracerebral Hemorrhage (MISTIE, PI: Daniel Hanley)

Subpopulations: intraventricular hemorrhage (IVH) < 10ml vs. not.

Projected proportions: 0.33, 0.67.

Primary outcome: 180 day modified Rankin Scale < 4.

Clinically meaningful, minimum treatment effect: 12% risk difference.

Data set used: MISTIE phase 2 trial data.

Alzheimer’s Disease Application

Treatment to slow progression of mild cognitive impairment due to Alzheimer’s disease.

Subpopulations: APOE4 carrier or not.

Primary outcome: 2 year change score in Clinical Dementia Rating Sum of Boxes

Clinically meaningful, minimum treatment effect: 30% reduction in mean change score

Data set used: Alzheimer’s Disease Neuroimaging Initiative (ADNI) cohort study

General Problem Addressed By Our Designs and Software

Two predefined subpopulations that partition overall pop.

Δ_1 = Average treatment effect for subpopulation 1

Δ_2 = Average treatment effect for subpopulation 2

Δ_0 = Average treatment effect for combined population

Goal: construct adaptive enrichment design to test

$$H_{01} : \Delta_1 \leq 0; \quad H_{02} : \Delta_2 \leq 0; \quad H_{00} : \Delta_0 \leq 0$$

that strongly controls familywise Type I error rate,

provides power guarantees, and optimizes expected sample size.

Example of Power and Type I Error Constraints

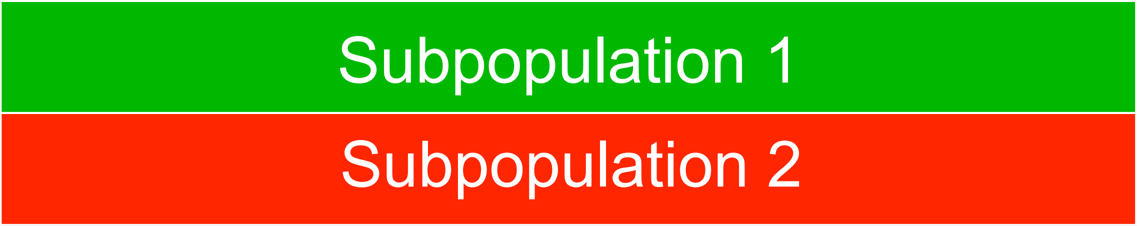
1. If clinically meaningful, minimum effect in both subpops., 80% power combined pop. null.

2. If clinically meaningful, minimum effect in single subpop., 80% power for that null hyp.

3. Strong control of familywise Type I error rate 0.025 (one-sided).

Goal: minimize expected sample size, averaged over scenarios in (1), (2), and global null.

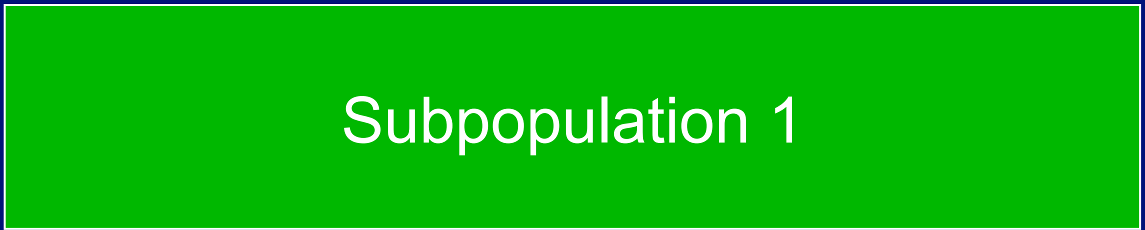
Standard (non-adaptive) Design 1



Subpopulation 1

Subpopulation 2

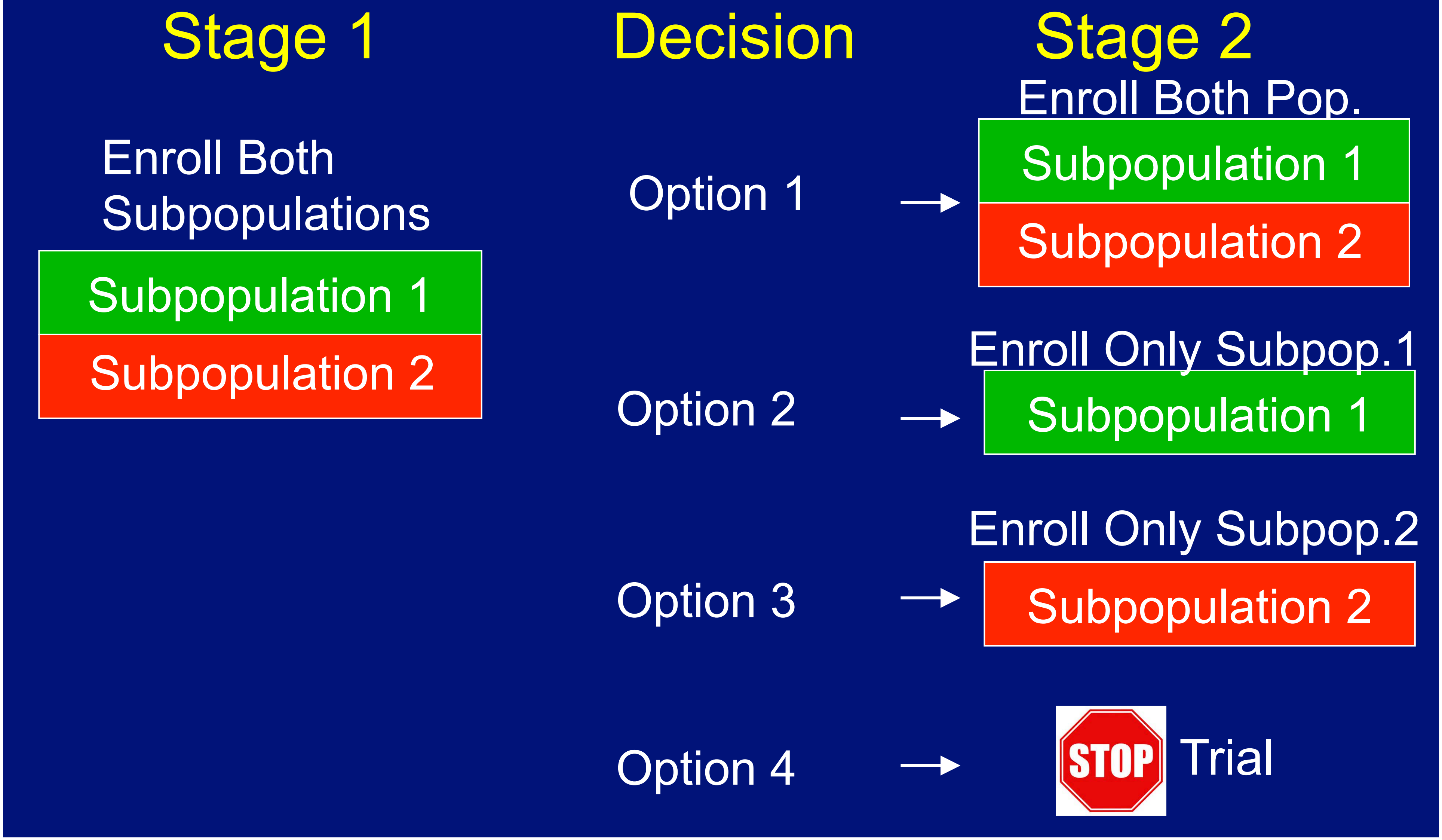
Standard (non-adaptive) Design 2



Subpopulation 1

2 Stage Adaptive Enrichment Design

Flow of Enrollment and Decision



Stage 1

Decision

Stage 2

Enroll Both Subpopulations

🟩 Subpopulation 1

🟥 Subpopulation 2

Option 1 → Enroll Both Pop.

🟩 Subpopulation 1

🟥 Subpopulation 2

Option 2 → Enroll Only Subpop.1

🟩 Subpopulation 1

Option 3 → Enroll Only Subpop.2

🟥 Subpopulation 2

Option 4 → STOP Trial

Adaptive Enrichment Design: Group Sequential, Enrollment Modification Rule

- At each analysis k, compute cumulative statistics $Z_{0,k,}$ $Z_{1,k,}$ $Z_{2,k}$ for combined pop., subpop. 1, and subpop. 2, respectively.

- Decision rule based on these statistics to: stop entire trial, stop single subpopulation accrual but continue other, continue both. (Cannot restart accrual once stopped.)

- No other adaptive features (e.g., randomization ratio fixed).

Multiple Testing Procedure

$$H_{01} : \Delta_1 \leq 0; \quad H_{02} : \Delta_2 \leq 0; \quad H_{00} : \Delta_0 \leq 0$$

At each analysis k:

1. (Test efficacy) For each population s∈{0,1,2}, if $Z_{s,k} > u_{s,k,}$ reject $H_{0s.}$ Also, if both H_{01} and H_{02} are rejected, reject H_{00}.

2. (Modify Enrollment) Stop subpopulation s∈{1,2}, if H_{0s} rejected or $Z_{s,k} < l_{s,k}$ or $Z_{0,k} < l_{0,k.}$

Boundaries $u_{s,k}$, $l_{s,k}$ set by error-spending functions (Maurer and Bretz, 2013; Rosenblum et al. 2016)

Trial Design Optimization Problem

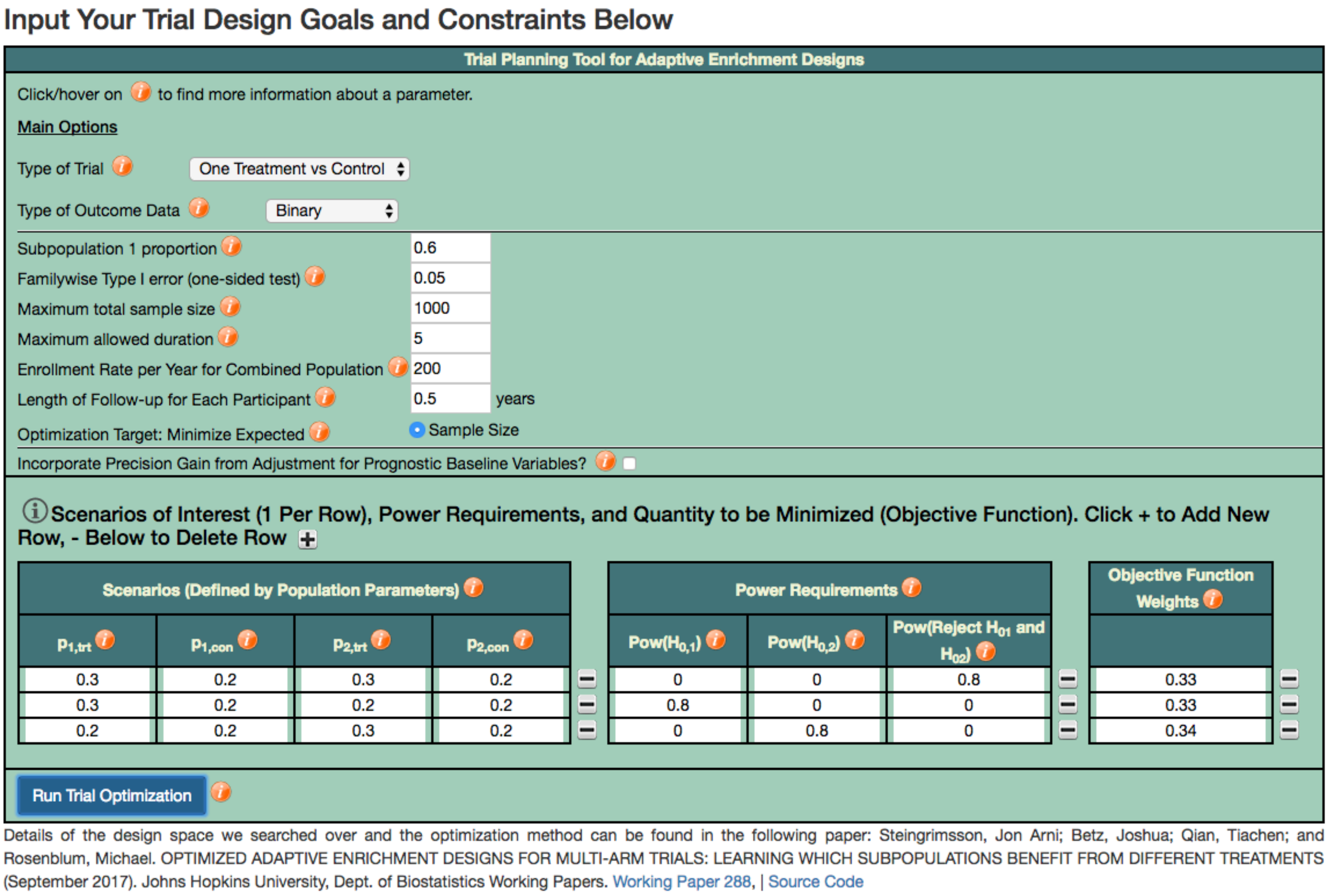
- Challenge: many design parameters to set: number of stages, per-stage sample sizes, efficacy and futility boundaries for each (stage, population) pair.

- We developed software tool (will be available end of summer) to automatically optimize over design parameters; goal is to minimize expected sample size under power and Type I error constraints.

  -User-friendly, graphical user-interface.

  -Produces automated, reproducible reports comparing optimal designs vs. standard designs.

Software User Interface



Input Your Trial Design Goals and Constraints Below

Trial Planning Tool for Adaptive Enrichment Designs

Click/hover on ⓘ to find more information about a parameter.

Main Options

Type of Trial ⓘ	One Treatment vs Control ⬍
Type of Outcome Data ⓘ	Binary ⬍
Subpopulation 1 proportion ⓘ	0.6
Familywise Type I error (one-sided test) ⓘ	0.05
Maximum total sample size ⓘ	1000
Maximum allowed duration ⓘ	5
Enrollment Rate per Year for Combined Population ⓘ	200
Length of Follow-up for Each Participant ⓘ	0.5 years
Optimization Target: Minimize Expected ⓘ	● Sample Size

Incorporate Precision Gain from Adjustment for Prognostic Baseline Variables? ⓘ □

ⓘ **Scenarios of Interest (1 Per Row), Power Requirements, and Quantity to be Minimized (Objective Function). Click + to Add New Row, - Below to Delete Row** ⊞

Scenarios (Defined by Population Parameters) ⓘ					Power Requirements ⓘ				Objective Function Weights ⓘ	
$p_{1,trt}$ ⓘ	$p_{1,con}$ ⓘ	$p_{2,trt}$ ⓘ	$p_{2,con}$ ⓘ		Pow($H_{0,1}$) ⓘ	Pow($H_{0,2}$) ⓘ	Pow(Reject H_{01} and H_{02}) ⓘ			
0.3	0.2	0.3	0.2	⊟	0	0	0.8	⊟	0.33	⊟
0.3	0.2	0.2	0.2	⊟	0.8	0	0	⊟	0.33	⊟
0.2	0.2	0.3	0.2	⊟	0	0.8	0	⊟	0.34	⊟

Details of the design space we searched over and the optimization method can be found in the following paper: Steingrimsson, Jon Arni; Betz, Joshua; Qian, Tiachen; and Rosenblum, Michael. OPTIMIZED ADAPTIVE ENRICHMENT DESIGNS FOR MULTI-ARM TRIALS: LEARNING WHICH SUBPOPULATIONS BENEFIT FROM DIFFERENT TREATMENTS (September 2017). Johns Hopkins University, Dept. of Biostatistics Working Papers. Working Paper 288, | Source Code

Design Optimizer Outputs

1. Optimized adaptive and standard designs that satisfy all power and Type I error constraints.

2. Performance comparisons: sample size, duration, power, Type I error, Bias, Variance, Mean Squared Error, Confidence Interval Cov.

3. Highlight key tradeoffs.

4. Plots of efficacy and futility boundaries.

Example of Optimization: Stroke Trial Application

From Fisher and Rosenblum (2017):

Search over 4 classes designs:

1. Separate error spending functions for efficacy and futility boundaries using power family, unequal per-stage sample sizes, up to 10 stages

2. O'Brien-Fleming boundaries, 5 stages, equal per-stage sample sizes

3. Pocock boundaries, 5 stages, equal per-stage sample sizes

4. Single stage designs

Example of Power and Type I Error Constraints

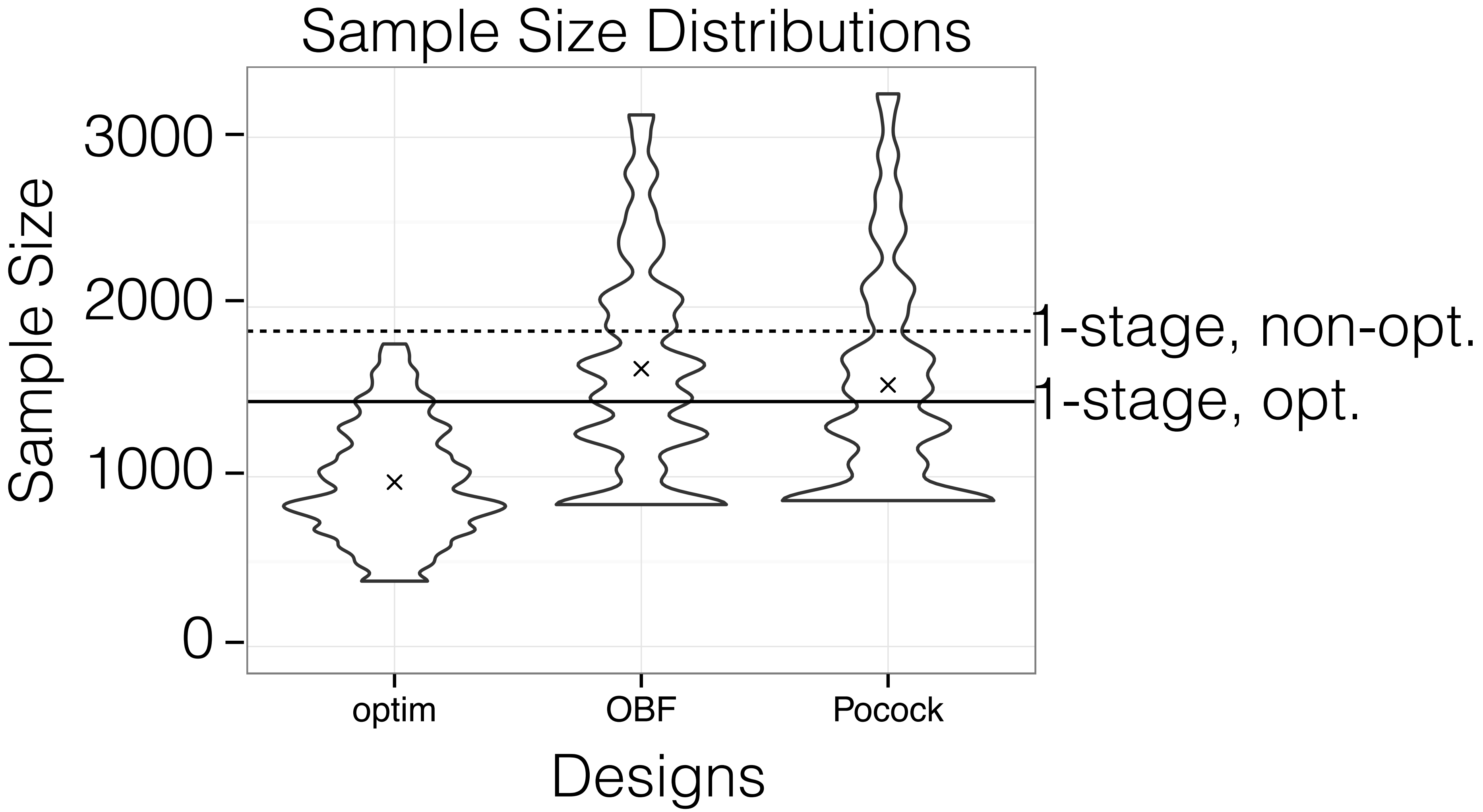
1. If clinically meaningful, minimum effect in both subpops., 80% power combined pop. null.

2. If clinically meaningful, minimum effect in single subpop., 80% power for that null hyp.

3. Strong control of familywise Type I error rate 0.025 (one-sided).

Goal: minimize expected sample size, averaged over scenarios in (1), (2), and global null.

Comparison of 5 Optimized Designs: Stroke Trial Application



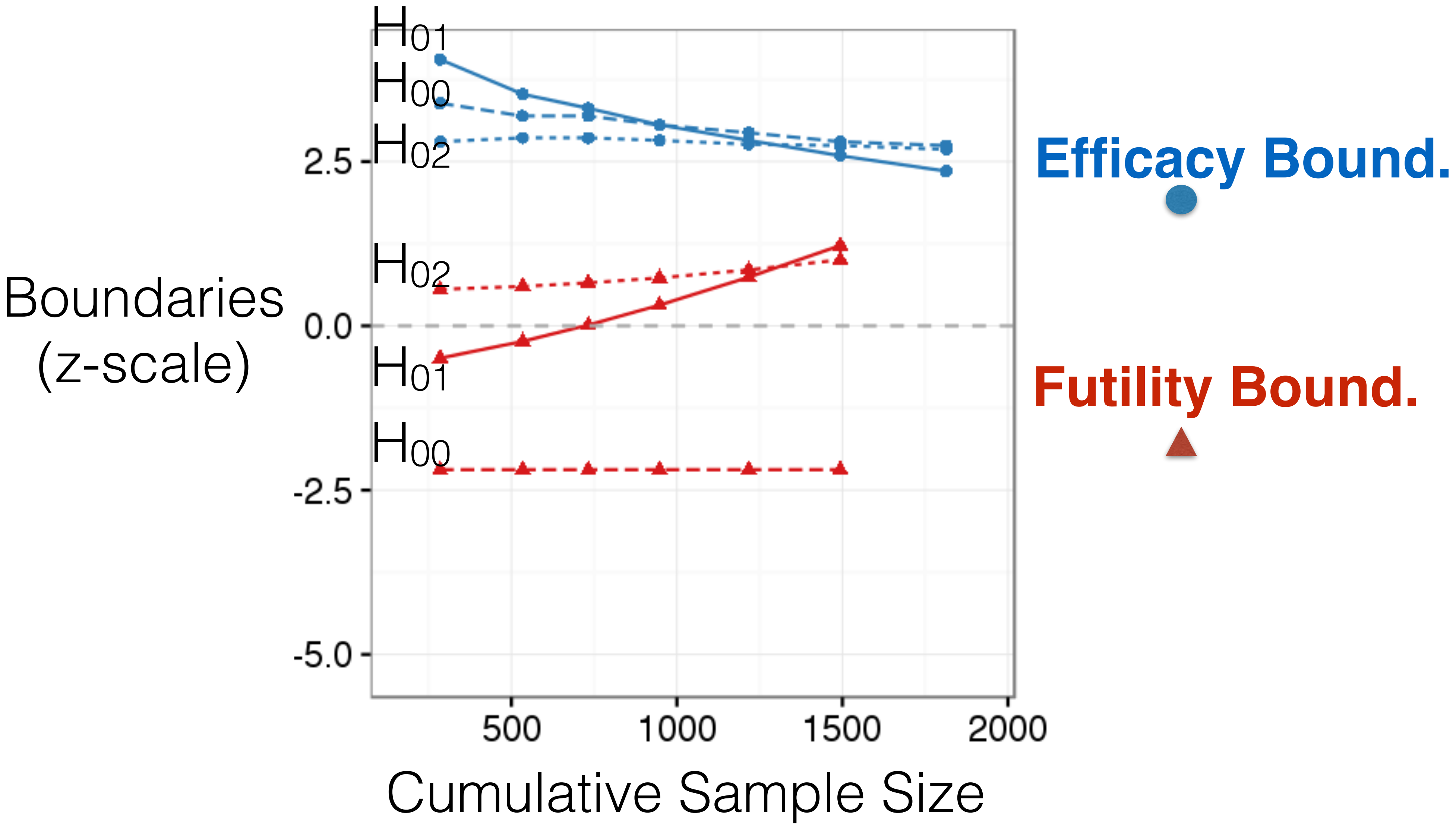
x=Expected Sample Size

Comparison of Optimized Designs: Stroke Trial Application

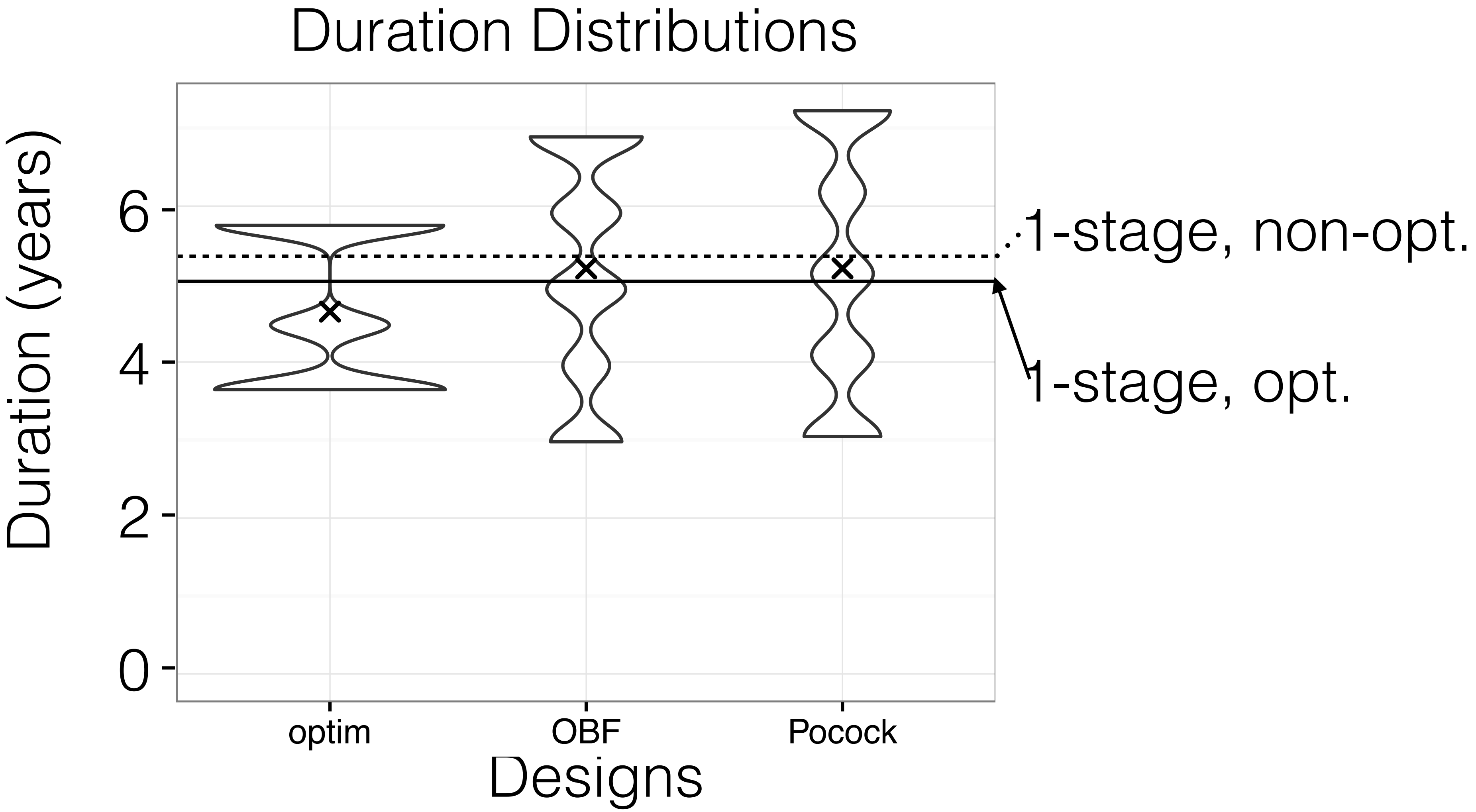
Performance Tradeoff Summary among Best Designs

	Optimized Adaptive Enrichment Design	Optimized 1-Stage
Expected Sample Size	968	1430
Maximum Sample Size	1787	1430

Optimized Adaptive Design Boundaries: Stroke Trial Application



Alzheimer’s Disease Application: Comparison of Optimized Designs



1-stage, non-opt.

1-stage, opt.

x=Expected Duration (years)

Performance Tradeoff Summary among Best Designs

	Optimized Adaptive Enrichment Design	Optimized 1-Stage
Expected Duration (Years)	4.65	5.02
Maximum Duration (Years)	5.75	5.02

HIV Treatment Application

Motivation: Prospective Evaluation of Antiretrovirals in Resource Limited Settings (PEARLS) study of the AIDS Clinical Trials Group (ACTG Trial A5175).

Subpopulations: women/men

Projected proportions: 0.47 women

Primary outcome (composite): time to virologic failure, AIDS, or death.

Goal: Test whether treatment arm A is **non-inferior** to arm B, i.e., whether the hazard ratio comparing arm A to B is at most 1.35. We aimed to conduct this non-inferiority test for each of the two subpopulations (defined by sex).

Scenarios

	Subpop 1 HR	Subpop 2 HR	Weight
Scenario 1	**1**	**1**	0.25
Scenario 2	**1**	1.35	0.25
Scenario 3	**1**	2.14	0.25
Scenario 4	1.35	1.35	0.25

Power constraints: 0.8 to reject each **false null hypothesis** (where Hazard Ratio < 1.35).
Goal: Minimize Expected Sample Size (equal weight on each scenario) under power and familywise Type I error constraints.

We compared 3 Types of Designs

(i) a single stage design;

(ii) a two-stage adaptive enrichment design (denoted $D_{\text{Adapt-Both}}$) that starts enrolling both subpopulations during stage 1;

(iii) a two-stage adaptive enrichment design (denoted $D_{\text{Adapt-Only-Subpop.1}}$) that enrolls only subpopulation 1 (women) in stage 1, but may expand the enrollment criteria to also include men after the first interim analysis.

We optimized the design parameters: alpha allocation, duration of enrollment, timing of the interim analysis, futility boundaries, rule for deciding whether to enroll men during stage 2.

Results (HIV application)

For enrollment rate 724 per year, optimized adaptive designs did not improve on the optimized single stage design.

For enrollment rate 362 per year: the tradeoff compared to the optimized single stage design was a reduction in the expected sample size of 2% at the cost of an increase in maximum sample size of 6%.

Results (HIV application)

For enrollment rate 362 per year: the tradeoff compared to the optimized single stage design was a reduction in the expected sample size of 2% at the cost of an increase in maximum sample size of 6%.

(i) Optimized single stage design enrolls 4.7 years and allocated 88% of alpha to the hypothesis test for subpopulation 1 (women). Sample size = 1702.

(ii) Optimized $D_{\text{Adapt-Both}}$ design, analysis times: 3.4 and 8 years. For each subpopulation that was not stopped at the interim analysis, enrollment continued to 5.0 years (and follow-up continued to 8 years).

Futility boundaries used at the interim analysis were -2.1 for women and -0.74 for men (on the z-scale). The expected sample size for $D_{\text{Adapt-Both}}$ was 1660 and the maximum sample size was 1799.

(iii) Did not improve on single stage design.

References

This research was supported by the U.S. Food and Drug Administration (HHSF223201400113C) and the Patient-Centered Outcomes Research Institute (ME-1306-03198). This work is solely the responsibility of the authors and does not represent the views of these agencies.

Selected References:

Betz, Josh; Steingrimsson, Jon Arni; Qian, Tianchen; and Rosenblum, Michael, "COMPARISON OF ADAPTIVE RANDOMIZED TRIAL DESIGNS FOR TIME-TO-EVENT OUTCOMES THAT EXPAND VERSUS RESTRICT ENROLLMENT CRITERIA, TO TEST NON-INFERIORITY" (September 2017). *Johns Hopkins University, Dept. of Biostatistics Working Papers.* http://biostats.bepress.com/jhubiostat/paper289

Fisher, A. and Rosenblum, M. (2017), Stochastic Optimization of Adaptive Enrichment Designs for Two Subpopulations. Johns Hopkins University, Dept. of Biostatistics Working Papers. http://goo.gl/wcQAxP

Qian, T., Colantuoni, E., Fisher, A., and Rosenblum, M., Sensitivity of Trial Performance to Delay Outcomes, Accrual Rates, and Prognostic Variables Based on a Simulated Randomized Trial with Adaptive Enrichment. Johns Hopkins University, Dept. of Biostatistics Working Papers. https://goo.gl/ty61Bk

Rosenblum, M., and Hanley, D.F. (2017) Topical Review: Adaptive Enrichment Designs for Stroke Clinical Trials. Stroke. 48(6). https://doi.org/10.1161/STROKEAHA.116.015342

Rosenblum, M., Qian, T., Du, Y., and Qiu, H., Fisher, A. (2016a) Multiple Testing Procedures for Adaptive Enrichment Designs: Combining Group Sequential and Reallocation Approaches. Biostatistics. (17)4, 650-662. http://goo.gl/extFAI

Rosenblum, M., Fang, X., and Liu, H. (2017) Optimal, Two Stage, Adaptive Enrichment Designs for Randomized Trials Using Sparse Linear Programming. Johns Hopkins University, Dept. of Biostatistics Working Papers: https://goo.gl/67lBga

Rosenblum, M., Thompson, R., Luber, B., Hanley, D. (2016b) Group Sequential Designs with Prospectively Planned Rules for Subpopulation Enrichment. Statistics in Medicine. 35(21), 3776-3791. http://goo.gl/7nHAVn

Steingrimsson, Jon Arni; Betz, Joshua; Qian, Tiachen; and Rosenblum, Michael, "OPTIMIZED ADAPTIVE ENRICHMENT DESIGNS FOR MULTI-ARM TRIALS: LEARNING WHICH SUBPOPULATIONS BENEFIT FROM DIFFERENT TREATMENTS" (September 2017). *Johns Hopkins University, Dept. of Biostatistics Working Papers.* http://biostats.bepress.com/jhubiostat/paper288

Stochastic Optimization of Adaptive Enrichment Designs for Two Subpopulations: Two Treatments vs. Control

Josh Betz, Michael Rosenblum, Jon Steingrimsson, Adi Gherman

Johns Hopkins Bloomberg School of Public Health

Aaron Fisher, Harvard TH Chan School of Public Health

Designs and Results from the following:

Jon Arni Steingrimsson, Joshua Betz, Tianchen Qian, Michael Rosenblum
Optimized adaptive enrichment designs for three-arm trials: learning which subpopulations benefit from different treatments, *Biostatistics*, Volume 22, Issue 2, April 2021, Pages 283–297, https://doi.org/10.1093/biostatistics/kxz030

Application: SMART-AV Trial

- Cardiac Resynchronization Therapy + Defibrillator (CRT+D): patients with medically-refractive heart failure (HF) with severe left ventricular systolic dysfunction (LVSD).
- Timing of atrioventricular (AV) delay may improve prognosis and quality of life
- SMART-AV (Stein et al. 2010): Multi-center RCT comparing 3 types of AV Optimization:
  - Doppler Echo Guided (DEO)
  - SmartDelay™ algorithmic optimization (SDO)
  - Fixed Delay – No Optimization (Control, Standard of Care)
- Suspected treatment effect varied by disease severity
  - Short QRS (≤150 ms: healthier, greater chance to benefit) vs. Long QRS (>150 ms: more severe HF, less chance to benefit)

Optimization of Adaptive Enrichment Designs

- Most prior research on two arm trials – Notable exception is Wason and Jaki (2012):
  - 6 parameterized designs over 3 treatment scenarios
  - Binding futility rules
- Our Software: Platform for Optimizing Designs
  - One or two treatments vs. control; Continuous, binary, and survival outcomes
  - Familywise Type I Error Rate (FWER) control by design;
  - Non-binding Futility Stopping
  - Compare performance across user-specified scenarios;
- Optimization is over rather large parameter space

Cardiac Resynchronization Therapy Application

Compare 2 new algorithms (treatments **A** and **B**) vs standard algorithm for setting Atrioventricular Delay in Cardiac Resynchronization Therapy Defibrillator for heart failure patients.

Subpopulations: Long vs. Short QRS duration

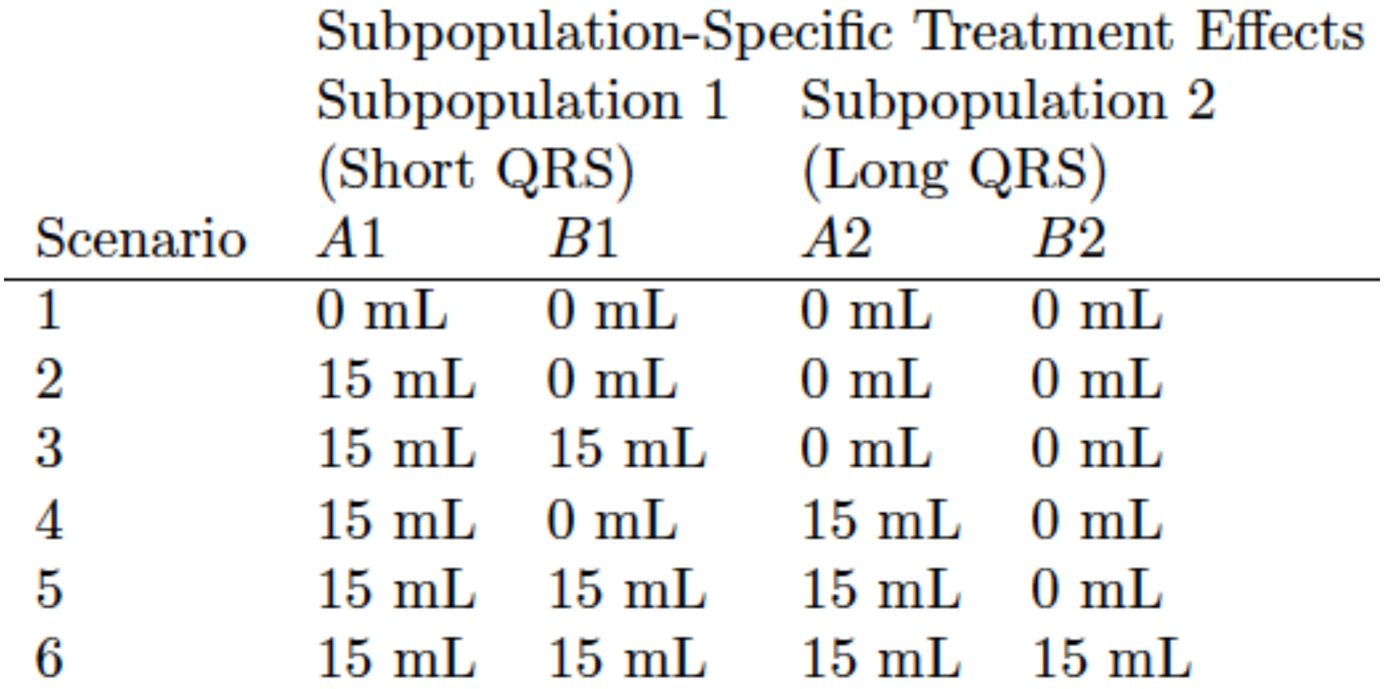
Primary Outcome: 6-month change Left Ventricular End Systolic Volume

Clinically meaningful, minimum treatment effect: 15ml

Data set used: SMART-AV phase 4 clinical trial

4 Null Hypotheses: One for each subpopulation by treatment combination.

6 Scenarios Used in Trial Planning:



	Subpopulation-Specific Treatment Effects			
	Subpopulation 1 (Short QRS)		Subpopulation 2 (Long QRS)	
Scenario	*A*1	*B*1	*A*2	*B*2
1	0 mL	0 mL	0 mL	0 mL
2	15 mL	0 mL	0 mL	0 mL
3	15 mL	15 mL	0 mL	0 mL
4	15 mL	0 mL	15 mL	0 mL
5	15 mL	15 mL	15 mL	0 mL
6	15 mL	15 mL	15 mL	15 mL

Require: 80% power to reject each null hyp. for which treatment effect at least $\Delta_{MIN.}$

Goal: Minimize Expected Sample Size under Power Constraints and Strong Control Familywise Type I Error Rate

Multiple Testing Procedure and Enrollment Modification Rule

At each analysis k, treatment $a \in \{A,B\}$, subpop $s \in \{1,2\}$:

1. **Test efficacy**: if $Z_{a,s,k} > u_{s,k,}$ then reject $H_{a,s,}$ i.e., declare treatment a efficacious for subpopulation s.
2. **Test futility**: if $Z_{a,s,k} < f_{s,k,}$ declare futility for treatment a for subpopulation s [denoted (a,s)]

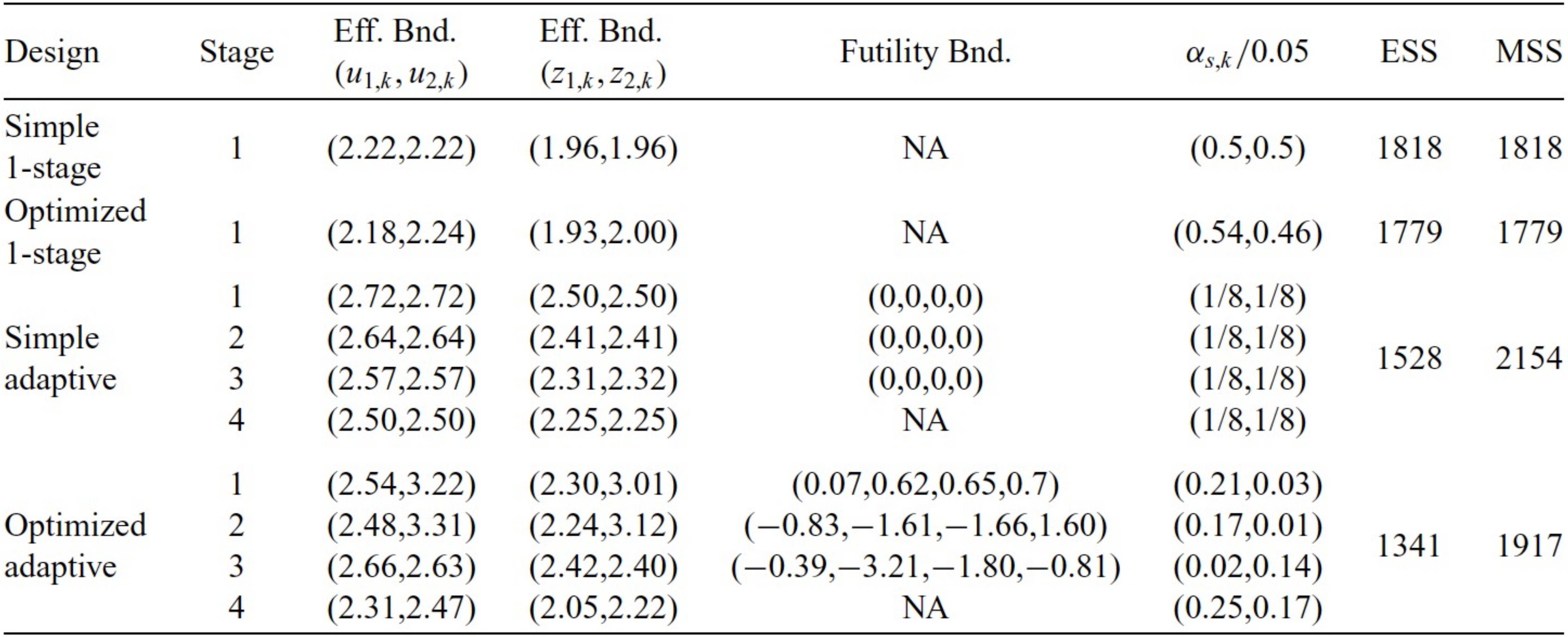
If declare efficacy or futility for (a,s), the no more patients from subpopulation s assigned to arm a.

Boundaries $u_{s,k}$, $l_{s,k}$ set by error-spending functions using Dunnett intersection tests with alpha-reallocation.

4 Design Types that We Compare:

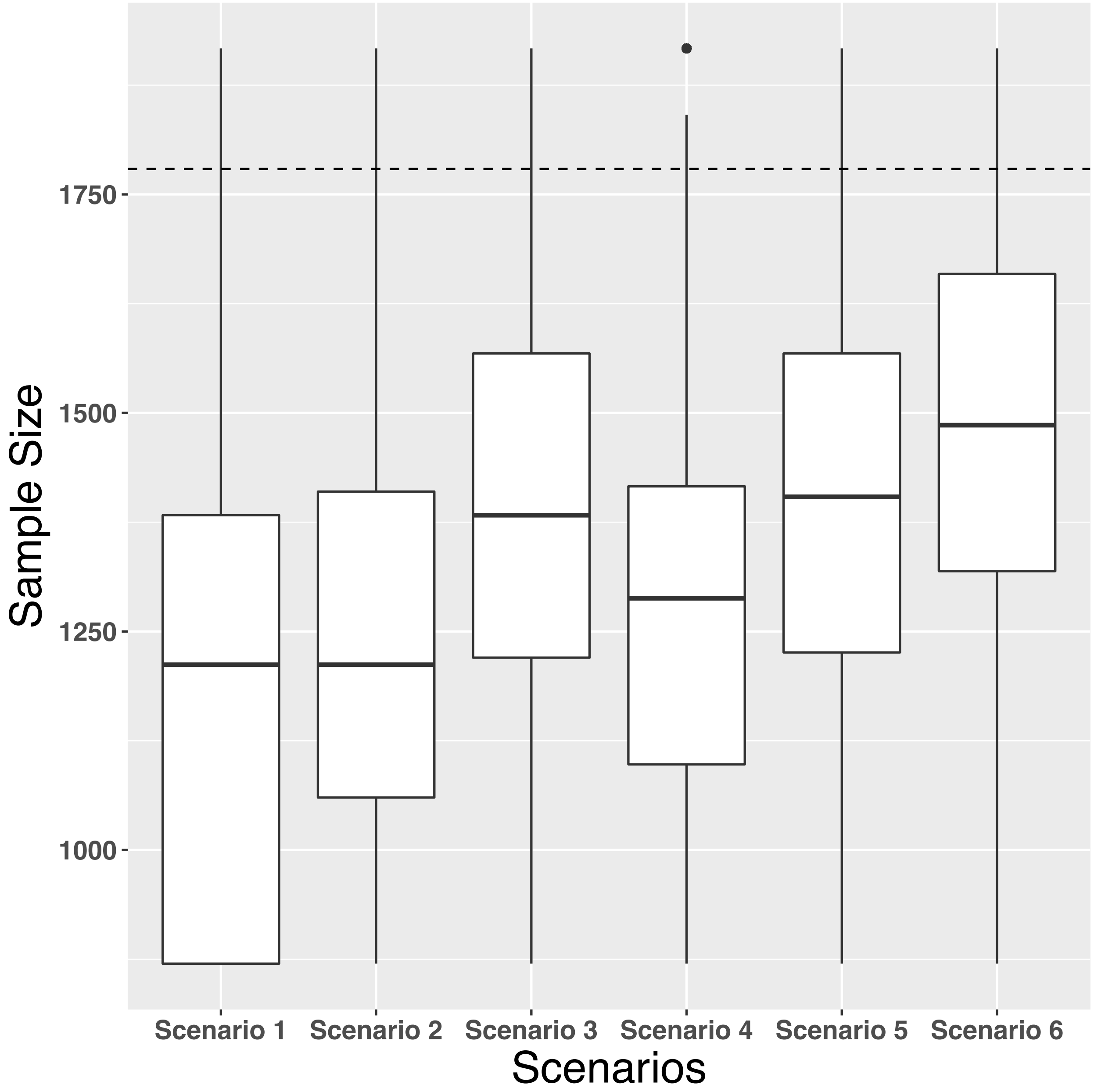
- Standard Fixed Design: 1 Stage (No Interim)
  - 1 Parameter: Feasible Sample Size;
- Optimized Fixed Design: 1 Stage (No Interim)
  - 3 Parameters: α-allocation chosen by optimization
- Standard Adaptive Design: 2 Stages (1 Interim)
  - Equal α-allocation similar to Pocock Boundaries
  - 5 Parameters: futility chosen by optimization
  - Interim analysis at 50% of observed primary outcomes
- Optimized Adaptive Design: 2 Stages (1 Interim)
  - 10 parameters; α-allocation, futility, and interim analysis time chosen by optimization

Optimized Design from Each of Four Classes



Design	Stage	Eff. Bnd. $(u_{1,k}, u_{2,k})$	Eff. Bnd. $(z_{1,k}, z_{2,k})$	Futility Bnd.	$\alpha_{s,k}/0.05$	ESS	MSS
Simple 1-stage	1	(2.22,2.22)	(1.96,1.96)	NA	(0.5,0.5)	1818	1818
Optimized 1-stage	1	(2.18,2.24)	(1.93,2.00)	NA	(0.54,0.46)	1779	1779
Simple adaptive	1	(2.72,2.72)	(2.50,2.50)	(0,0,0,0)	(1/8,1/8)	1528	2154
	2	(2.64,2.64)	(2.41,2.41)	(0,0,0,0)	(1/8,1/8)		
	3	(2.57,2.57)	(2.31,2.32)	(0,0,0,0)	(1/8,1/8)		
	4	(2.50,2.50)	(2.25,2.25)	NA	(1/8,1/8)		
Optimized adaptive	1	(2.54,3.22)	(2.30,3.01)	(0.07,0.62,0.65,0.7)	(0.21,0.03)	1341	1917
	2	(2.48,3.31)	(2.24,3.12)	(−0.83,−1.61,−1.66,1.60)	(0.17,0.01)		
	3	(2.66,2.63)	(2.42,2.40)	(−0.39,−3.21,−1.80,−0.81)	(0.02,0.14)		
	4	(2.31,2.47)	(2.05,2.22)	NA	(0.25,0.17)		

Sample Size Distributions



- Boxplots give distributions for optimized adaptive enrichment design

- Dashed line is for optimized 1-stage design.

Conclusions

- Optimization and Adaptation can be beneficial
- Optimization platform is needed for comparing different classes of designs:
    - Flexible, modular design
    - Find right benefit/complexity tradeoff
    - Specification of objective function is important
- Future Directions
    - Implementing additional design modules
    - Improving optimization

Acknowledgements

- Funded by Patient-Centered Outcomes Research Institute (ME-1306-03198) and the US Food and Drug Administration (HHSF223201400113C).
- We would like to thank Boston Scientific for access to SMART-AV data
- We would like to thank Estelle Russek-Cohen, Shaila Shaheed, York Tomita, Elizabeth Colantuoni, Marina Kondratovich, Greg Levin, Amal Manseur, Sean Wybenga, and Jingyee Kou for their helpful comments.
- The views expressed in this talk are solely those of the speaker and do not represent the organizations above
- Conflicts of Interest: none

References

- Liu, Q, and KM Anderson. 2008. On Adaptive Extensions of Group Sequential Trials for Clinical Investigations. *JASA* 103 (484): 1621–30.
- Spiessens, B, and Debois M. 2010. Adjusted Significance Levels for Subgroup Analyses in Clinical Trials. *Contemp Clin Trials* 31 (6): 647–56.
- Stein, KM, KA Ellenbogen, MR Gold, B Lemke, IF Lozano, S Mittal, FG Spinale, JE Van Eyk, AD Waggoner, and TE Meyer. 2010. SmartDelay Determined AV Optimization: A Comparison of AV Delay Methods Used in Cardiac Resynchronization Therapy (SMART-AV): Rationale and Design. *J Pacing Clin Electrophysiol* 33 (1): 54–63.
- Wason, JMS, and T Jaki. 2012. Optimal Design of Multi-Arm Multi-Stage Trials. *Stat Med* 31 (30): 4269–79.

Multiple testing procedures for adaptive enrichment designs: combining group sequential and reallocation approaches

Michael Rosenblum
Biostatistics Department
Johns Hopkins Bloomberg School of Public Health (JHBSPH)

Joint work with:
Tianchen Qian, Yu Du, Huitong Qiu, Aaron Fisher

From 2016 paper in *Biostatistics*: http://goo.gl/extFAI

November 20, 2017

Adaptive Enrichment Designs

- Involve preplanned rules for modifying enrollment criteria based on accrued data
- Multiple populations of interest, each with corresponding null hypothesis to test
- Challenge: construct group sequential multiple testing procedure with all of the following properties:
    1. Strong control of familywise Type I error rate (probability of rejecting one or more true null hypotheses)
    2. Leverages correlation among statistics over time and for overlapping populations
    3. Provides strictly greater power than several known methods
    4. Does not require knowing covariance matrix in advance

Related Work

- Methods that leverage covariance among statistics: Tang and Geller (1999), Stallard (2011), Magirr and others (2012), Magnusson and Turnbull (2013)
- Methods that lower rejection thresholds for the remaining null hypotheses after others have been rejected, by reallocating alpha across hypotheses (populations): Holm (1979), Bretz et al. (2009), Maurer and Bretz (2013).
  These don't leverage covariance among statistics.

We combine features from these two types of approaches.

Note: Bretz et al. (2011, Section 3.2) do this, but unlike our method require covariance of future statistics known in advance.

Interleaved Error Spending Functions

- Error spending function approach of Slud and Wei (1982) and Lan and DeMets (1983). Advantage: information accrual rates don't need to be known in advance.
- We use separate error-spending function for each composite population of interest.
- Tests for different populations are interleaved to take advantage of correlations among statistics for different but overlapping populations and statistics for the same population at different times.

Hypotheses and Statistics

- Null Hypotheses: $H_{0j} : \Delta_j \leq 0$, for each $j \in \{0, \ldots, J\}$. Global null hypothesis $H_0 : \Delta_j = 0$ for all $j \leq J$.
- A sequence of analyses $1, \ldots, K$ are preplanned, where analysis k takes place at the end of stage k.
- At analysis k, observe cumulative, Wald statistics $Z_{0,k}, Z_{1,k}, \ldots, Z_{J,k}$.
- Assume $EZ_{j,k} \leq 0$ under H_{0j} for all stages k.
- Covariance matrix of $Z_{j,k}$ fixed but unknown.
- Alpha increments $\alpha_{j,k}$ determined by error spending functions at each stage.

Simple Version Without Interleaved Error Spending Functions

Alpha increments $\alpha_{j,k} \geq 0$ and $\sum_{j \geq 0, k \geq 0} \alpha_{j,k} = 0.05$.

At stage k, consider each null hypothesis $H_{0j}, j = 0, 1, \ldots, J$ and reject H_{0j} if $Z_{j,k} > \mathbf{u_{j,k}}$.

Each efficacy boundary $\mathbf{u_{j,k}}$ is set to be solution to:

$$\alpha_{j,k} = P_{H_0}\left\{Z_{j,k} > \mathbf{u_{j,k}}; Z_{j,k'} \leq u_{j,k'} \text{ for all } k' \leq k\right\}.$$

This uses covariances for same population across stages, but ignores covariance among populations.
Can improve by interleaving error-spending functions.

Interleaved error spending functions

Alpha increments $\alpha_{j,k} \geq 0$ and $\sum_{j\geq 0, k\geq 0} \alpha_{j,k} = 0.05$.

At stage k, consider each null hypothesis $H_{0j}, j = 0, 1, \ldots, J$ and reject H_{0j} if $Z_{j,k} > \mathbf{u_{j,k}}$.

Each efficacy boundary $\mathbf{u_{j,k}}$ is set to be solution to:

$$\alpha_{j,k} = P_{H_0}\left\{Z_{j,k} > \mathbf{u_{j,k}}; Z_{j',k'} \leq u_{j',k'} \text{ for all } (j',k') \text{ preceding } (j,k)\right\},$$

where (j',k') preceding (j,k) if $k' < k$ or if $(k' = k$ and $j' \leq j)$;
and where $\alpha_{j,k} \geq 0$ and $\sum_{j\geq 0, k\geq 0} \alpha_{j,k} = \alpha$ (e.g., 0.05).

This leverages covariances, but does not use alpha reallocation.

Example of Efficacy Boundary Improvements

Efficacy Boundaries (z-scale) for 3 Stage, 2 Hypothesis Trial

Analysis (k)	1	2	3
	Without Interleaving ($\mathcal{M}^{MB}$)		
H_{00} boundaries $u_{0,k}$	2.57	2.32	2.10
H_{01} boundaries $u_{1,k}$	3.45	3.29	3.14
	With Interleaving ($\mathcal{M}^{NEW}$)		
H_{00} boundaries $u_{0,k}$	2.57	2.32	2.09
H_{01} boundaries $u_{1,k}$	3.16	2.95	2.74

$\mathcal{M}^{NEW}$: Our new multiple testing procedure.
$\mathcal{M}^{MB}$: Maurer and Bretz (2013) procedure.

Interleaved error spending functions with alpha reallocation

Closure principle: For each $F \subseteq \{0, \ldots, J\}$, define local test of intersection null hyp. $H_F = \cap_{j \in F} H_{0j}$ with level α. Reject elementary null H_{0j} if every H_F with $j \in F$ rejects.

Local test of H_F: reject if $Z_{j,k} > \mathbf{u^F_{j,k}}$ for any $j \in F$, $k \geq 0$, where at analysis k, for each $j \in F$, set $\mathbf{u^F_{j,k}}$ to be solution to:

$$\begin{aligned} c_j^F \alpha_{j,k} \quad = \quad & P_{H_0} \Big\{ Z_{j,k} > \mathbf{u^F_{j,k}}; \\ & \qquad Z_{j',k'} \leq u^F_{j',k'} \text{ for all } (j', k') \text{ preceding } (j, k) \text{ and } j' \in F \Big\}, \end{aligned}$$

where $c_j^F \geq 1$ and $\sum_{j \in F, k \geq 0} c_j^F \alpha_{j,k} = \alpha$.
Intuitively, c_j^F is alpha inflation factor, reallocating alpha from hypotheses $j \notin F$. It can be set, e.g., using graphical approach of Bretz et al. (2009), but not restricted to this.

Example of Efficacy Boundary Improvements

Efficacy Boundaries (z-scale) for 3 Stage, 2 Hypothesis Trial

Analysis (k)	1	2	3
	Without Interleaving ($\mathcal{M}^{MB}$)		
H_{00} boundaries $u_{0,k}$	2.57	2.32	2.10
H_{01} boundaries $u_{1,k}$	3.45	3.29	3.14
	With Interleaving ($\mathcal{M}^{NEW}$)		
H_{00} boundaries $u_{0,k}$	2.57	2.32	2.09
H_{01} boundaries $u_{1,k}$	3.16	2.95	2.74
	After Alpha Reallocation (both)		
H_{00} boundaries after reject H_{01}	2.55	2.30	2.07
H_{01} boundaries after reject H_{00}	2.55	2.30	2.07

$\mathcal{M}^{NEW}$: Our new multiple testing procedure.

$\mathcal{M}^{MB}$: Maurer and Bretz (2013) procedure.

Main Result

Our new multiple testing procedure: $\mathcal{M}^{NEW}$.
Maurer and Bretz (2013) procedure: $\mathcal{M}^{MB}$.

Theorem

- *For each null hypothesis H_{0j}, $\mathcal{M}^{NEW}$ rejects it by analysis k whenever $\mathcal{M}^{MB}$ does.*
- *If H_{0J} is false, $\mathcal{M}^{NEW}$ has strictly greater power than $\mathcal{M}^{MB}$ to reject H_{0J} by analysis k, under the condition that covariance matrix full rank and each $\alpha_{j,k} > 0$.*

Limitations

- Computation of multivariate normal distribution function restricts total number of stages and hypotheses. For confirmatory trials with 2-4 elementary null hypotheses and 3-5 analysis times, can use Genz et al. (2014).
- Advantages from leveraging covariance only useful if substantial overlap in populations, e.g., a subpopulation that makes up 2/3 of overall population.

References

Rosenblum, M., Qian, T., Du, Y., and Qiu, H., Fisher, A. (2016) Multiple Testing Procedures for Adaptive Enrichment Designs: Combining Group Sequential and Reallocation Approaches. *Biostatistics*. http://goo.gl/extFAI

Bretz, F., W. Maurer, W. Brannath, and M. Posch (2009). A graphical approach to sequentially rejective multiple test procedures. *Statistics in medicine 28*(4), 586–604.

Bretz, F., M. Posch, E. Glimm, F. Klinglmueller, W. Maurer, and K. Rohmeyer (2011). Graphical approaches for multiple comparison procedures using weighted bonferroni, simes, or parametric tests. *Biometrical Journal 53*(6), 894–913.

Genz, A., F. Bretz, T. Miwa, X. Mi, F. Leisch, F. Scheipl, and T. Hothorn (2014). mvtnorm: Multivariate Normal and t Distributions. R package version 1.0-0. URL http://CRAN.R-project.org/package=mvtnorm.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scand. J. Statist. 6*, 65–70.

Optimal Tests of Treatment Effects for the Overall Population and Two Subpopulations in Randomized Trials, using Sparse Linear Programming

Michael Rosenblum
Johns Hopkins Bloomberg School of Public Health

Joint work with:
Han Liu and Xingyuan (Ethan) Fang at Princeton University,
En-Hsu Yen at University of Texas, Austin

Outline

Part I: **Standard (Non-adaptive) Randomized Trial**; Goal is to Optimize Multiple Testing Procedure

Outline

Part I: **Standard (Non-adaptive) Randomized Trial**; Goal is to Optimize Multiple Testing Procedure

Part II: **Adaptive Enrichment Designs**; Goal is to Simultaneously Optimize Decision Rule and Multiple Testing Procedure

Problem Motivation

Goal: Testing Treatment Effects in Two Subpopulations and the Overall Population in Randomized Trials

Example:

- Treating resistant HIV. Recent HIV drugs (maraviroc, raltegravir) have shown stronger benefit in those with lower phenotypic sensitivity to background therapy.

We assume two, predefined, subpopulations that partition the overall population.
We optimize analysis at end of randomized trial.

Multiple Testing Problem: Null Hypotheses Definition

Define three treatment effects of interest:

- Δ_1: Mean Treatment Effect for Subpopulation 1 (i.e., difference between population mean of the primary outcome under treatment and under control)
- Δ_2: Mean Treatment Effect for Subpopulation 2
- $\Delta_C = p_1\Delta_1 + (1 - p_1)\Delta_2$: Mean Treatment Effect for Combined Population

Multiple Testing Problem: Null Hypotheses Definition

Define three treatment effects of interest:

- Δ_1: Mean Treatment Effect for Subpopulation 1 (i.e., difference between population mean of the primary outcome under treatment and under control)
- Δ_2: Mean Treatment Effect for Subpopulation 2
- $\Delta_C = p_1\Delta_1 + (1 - p_1)\Delta_2$: Mean Treatment Effect for Combined Population

Goal: construct multiple testing procedure M for null hypotheses:

- $H_{01} : \Delta_1 \leq 0,$
- $H_{02} : \Delta_2 \leq 0,$
- $H_{0C} : p_1\Delta_1 + (1 - p_1)\Delta_2 \leq 0,$

that strongly controls familywise Type I error rate,
and optimizes power in sense described below.

Multiple Testing Problem: Optimization Criteria

Goal: multiple testing procedure for:

- $H_{01} : \Delta_1 \leq 0$,
- $H_{02} : \Delta_2 \leq 0$,
- $H_{0C} : p_1\Delta_1 + (1 - p_1)\Delta_2 \leq 0$.

Assume known variances σ_1^2, σ_2^2 and normally distributed outcomes; subpopulation z-statistics Z_1, Z_2 are then sufficient statistics.

Let $\Delta^{\text{min}} > 0$ denote minimum, clinically meaningful treatment benefit. Let L denote loss function and Λ denote prior on (Δ_1, Δ_2).

Constrained Bayes Optimization Problem

For loss function L and prior Λ on parameters (Δ_1, Δ_2),

find multiple testing procedure M minimizing Bayes criterion:

$$\int E_{\Delta_1,\Delta_2} L[M(Z_1, Z_2); \Delta_1, \Delta_2] d\Lambda(\Delta_1, \Delta_2),$$

Constrained Bayes Optimization Problem

For loss function L and prior Λ on parameters (Δ_1, Δ_2),

find multiple testing procedure M minimizing Bayes criterion:

$$\int E_{\Delta_1,\Delta_2} L[M(Z_1, Z_2); \Delta_1, \Delta_2] d\Lambda(\Delta_1, \Delta_2),$$

under **familywise Type I error constraints**:

$$\sup_{(\Delta_1,\Delta_2)\in\mathbb{R}^2} P_{\Delta_1,\Delta_2}[M \text{ rejects any true null hypothesis}] \leq \alpha,$$

Constrained Bayes Optimization Problem

For loss function L and prior Λ on parameters (Δ_1, Δ_2),

find multiple testing procedure M minimizing Bayes criterion:

$$\int E_{\Delta_1,\Delta_2} L[M(Z_1, Z_2); \Delta_1, \Delta_2] d\Lambda(\Delta_1, \Delta_2),$$

under **familywise Type I error constraints**:

$$\sup_{(\Delta_1,\Delta_2)\in\mathbb{R}^2} P_{\Delta_1,\Delta_2}[M \text{ rejects any true null hypothesis}] \leq \alpha,$$

and **power constraint for combined population**:

$$P_{\Delta^{\min},\Delta^{\min}}[M \text{ rejects } H_{0C}] \geq 1 - \beta.$$

Our Method to Solve Optimization Problem

Computational challenge: continuum of Type I error constraints. (It's not enough to satisfy only at global null $(\Delta_1, \Delta_2) = (0, 0)$.)

Our Method to Solve Optimization Problem

Computational challenge: continuum of Type I error constraints. (It's not enough to satisfy only at global null $(\Delta_1, \Delta_2) = (0, 0)$.)

Our solution:

1. Discretize decision region $\mathbb{R}^2$ into small rectangles $\mathcal{R}$; for any $r \in \mathcal{R}$, enforce test procedure M rejects same set of hypotheses for any $(Z_1, Z_2) \in r$.
2. Discretize constraints into fine grid on boundaries of null spaces.

Our Method to Solve Optimization Problem

Computational challenge: continuum of Type I error constraints. (It's not enough to satisfy only at global null $(\Delta_1, \Delta_2) = (0, 0)$.)

Our solution:

1. Discretize decision region $\mathbb{R}^2$ into small rectangles $\mathcal{R}$; for any $r \in \mathcal{R}$, enforce test procedure M rejects same set of hypotheses for any $(Z_1, Z_2) \in r$.
2. Discretize constraints into fine grid on boundaries of null spaces.

Discretized, constrained Bayes opt. problem can be represented as sparse, linear program:

$$\max_{x} \quad \mathbf{c}^{\mathbf{T}}\mathbf{x} \qquad \text{s.t.} \quad \mathbf{A}\mathbf{x} \leq \mathbf{b}.$$

Even in simple example below, $\mathbf{A}$ is 1,757,113 × 1,506,006 matrix. We solve this by tailoring advanced optimization methods to the structure of this problem (and leveraging sparseness of $\mathbf{A}$).

Optimal Procedures at $\alpha = 0.05$; $1 - \beta = 0.9$ and 0.88

Example: $p_1 = 1/2$; $\sigma_1^2 = \sigma_2^2$. **Loss function** penalizes 1 unit for failure to reject H_{0k} at $\Delta_k \geq \Delta^{\min}$. **Prior**: equally weighted pt. masses at $(0,0), (\Delta^{\min}, 0), (0, \Delta^{\min}), (\Delta^{\min}, \Delta^{\min})$. **Sample size** is min s.t. UMP test of H_{0C} has 90% power at $(\Delta^{\min}, \Delta^{\min})$.

Optimal Procedures at $\alpha = 0.05$; $1 - \beta = 0.9$ and 0.88

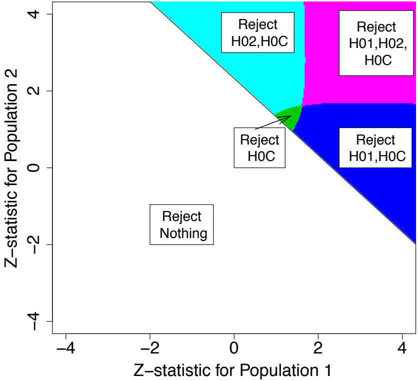
Example: $p_1 = 1/2$; $\sigma_1^2 = \sigma_2^2$. **Loss function** penalizes 1 unit for failure to reject H_{0k} at $\Delta_k \geq \Delta^{\min}$. **Prior**: equally weighted pt. masses at $(0,0), (\Delta^{\min}, 0), (0, \Delta^{\min}), (\Delta^{\min}, \Delta^{\min})$. **Sample size** is min s.t. UMP test of H_{0C} has 90% power at $(\Delta^{\min}, \Delta^{\min})$.

H_{0C} Power Constr. $1 - \beta = 0.90$

Optimal Procedures at $\alpha = 0.05$; $1 - \beta = 0.9$ and 0.88

Example: $p_1 = 1/2$; $\sigma_1^2 = \sigma_2^2$. **Loss function** penalizes 1 unit for failure to reject H_{0k} at $\Delta_k \geq \Delta^{\min}$. **Prior**: equally weighted pt. masses at $(0,0), (\Delta^{\min}, 0), (0, \Delta^{\min}), (\Delta^{\min}, \Delta^{\min})$. **Sample size** is min s.t. UMP test of H_{0C} has 90% power at $(\Delta^{\min}, \Delta^{\min})$.

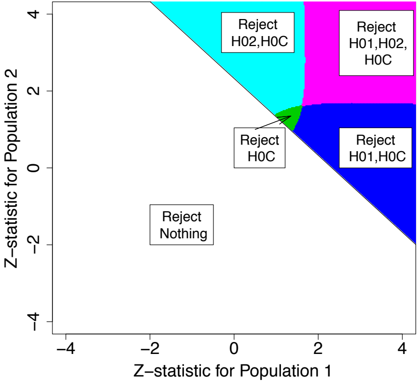
H_{0C} Power Constr. $1 - \beta = 0.90$



Optimal Procedures at $\alpha = 0.05$; $1 - \beta = 0.9$ and 0.88

Example: $p_1 = 1/2$; $\sigma_1^2 = \sigma_2^2$. **Loss function** penalizes 1 unit for failure to reject H_{0k} at $\Delta_k \geq \Delta^{\min}$. **Prior**: equally weighted pt. masses at $(0,0), (\Delta^{\min}, 0), (0, \Delta^{\min}), (\Delta^{\min}, \Delta^{\min})$. **Sample size** is min s.t. UMP test of H_{0C} has 90% power at $(\Delta^{\min}, \Delta^{\min})$.

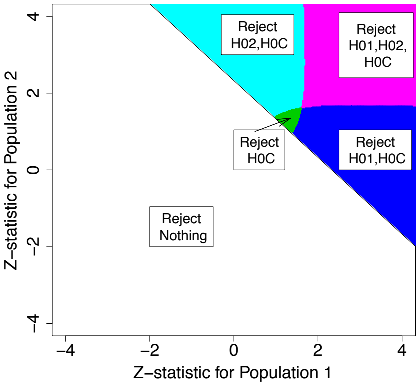
H_{0C} Power Constr. $1-\beta = 0.90$ H_{0C} Power Constr. $1-\beta = 0.88$



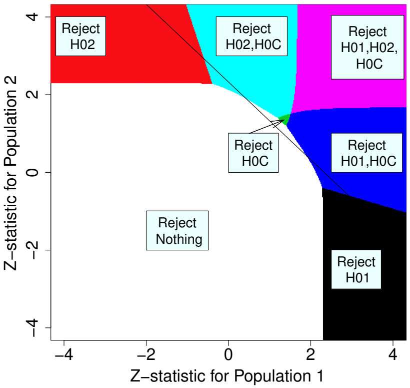
Optimal Procedures at $\alpha = 0.05$; $1-\beta = 0.9$ and 0.88

Example: $p_1 = 1/2$; $\sigma_1^2 = \sigma_2^2$. **Loss function** penalizes 1 unit for failure to reject H_{0k} at $\Delta_k \geq \Delta^{\min}$. **Prior**: equally weighted pt. masses at $(0,0), (\Delta^{\min}, 0), (0, \Delta^{\min}), (\Delta^{\min}, \Delta^{\min})$. **Sample size** is min s.t. UMP test of H_{0C} has 90% power at $(\Delta^{\min}, \Delta^{\min})$.

H_{0C} Power Constr. $1-\beta = 0.90$



H_{0C} Power Constr. $1-\beta = 0.88$



Power for optimal multiple testing procedures at $1-\beta = 0.9$ and $1-\beta = 0.88$.

Let $m^*(1-\beta)$ denote optimal procedure having power $1-\beta$ for H_{0C} at alternative $(\Delta^{\min}, \Delta^{\min})$.

Power Comparison of Two Optimal Procedures

	Procedure:	
	$m^*(0.9)$	$m^*(0.88)$
Power for H_{01} at $(\Delta^{\min}, 0)$:	0.39	0.51
Power for H_{02} at $(0, \Delta^{\min})$:	0.39	0.51
Power for H_{0C} at $(\Delta^{\min}, \Delta^{\min})$:	0.90	0.88

Part II: **Adaptive Enrichment Designs**:

Part II: **Adaptive Enrichment Designs**; Goal is to Simultaneously Optimize Decision Rule and Multiple Testing Procedure

Part II: **Adaptive Enrichment Designs**:

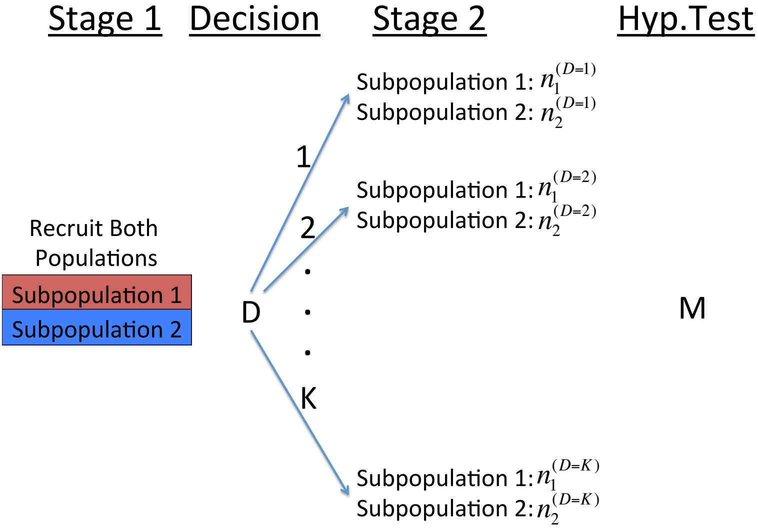
Part II: **Adaptive Enrichment Designs**; Goal is to Simultaneously Optimize Decision Rule and Multiple Testing Procedure

Goal: construct adaptive enrichment design D and multiple testing procedure M for:

- $H_{01} : \Delta_1 \leq 0$,
- $H_{02} : \Delta_2 \leq 0$,
- $H_{0C} : p_1\Delta_1 + (1 - p_1)\Delta_2 \leq 0$,

that strongly controls familywise Type I error rate, and is optimal in sense defined below.

General Two-Stage Adaptive Enrichment Design



Stage 1 Decision Stage 2 Hyp.Test

Recruit Both Populations

Subpopulation 1

Subpopulation 2

D

1

Subpopulation 1: $n_1^{(D=1)}$

Subpopulation 2: $n_2^{(D=1)}$

2

Subpopulation 1: $n_1^{(D=2)}$

Subpopulation 2: $n_2^{(D=2)}$

. . .

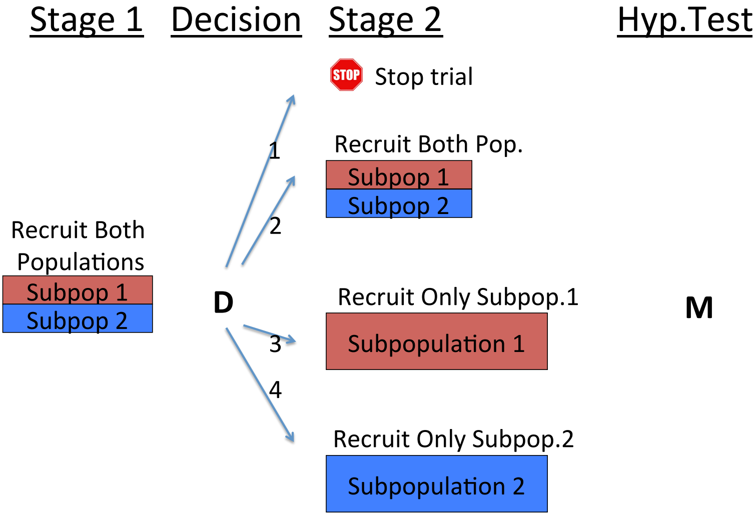
K

Subpopulation 1: $n_1^{(D=K)}$

Subpopulation 2: $n_2^{(D=K)}$

M

Example Two-Stage Adaptive Enrichment Design



Stage 1

Decision

Stage 2

Hyp.Test

Stop trial

Recruit Both Pop.

Subpop 1

Subpop 2

Recruit Both Populations

Subpop 1

Subpop 2

D

1

2

3

4

Recruit Only Subpop.1

Subpopulation 1

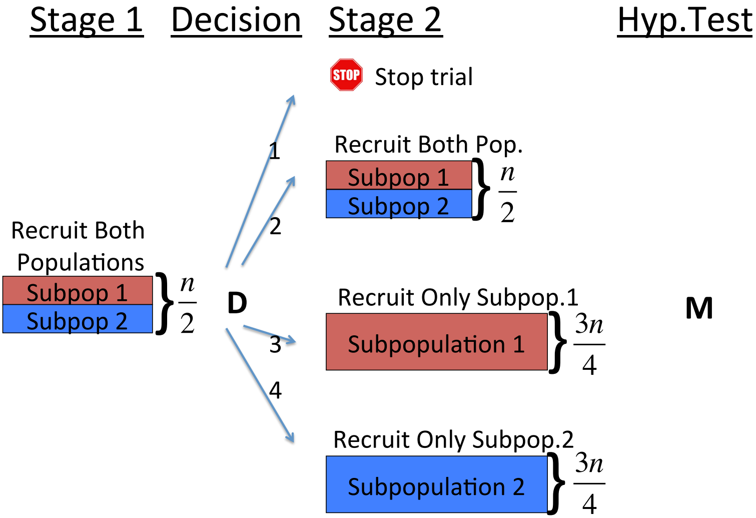
Recruit Only Subpop.2

Subpopulation 2

M

Example Two-Stage Adaptive Enrichment Design

n = total sample size if both subpopulations enrolled in stage 2.



Stage 1

Decision

Stage 2

Hyp.Test

Recruit Both Populations

Subpop 1

Subpop 2

$\frac{n}{2}$

D

1 Stop trial

2 Recruit Both Pop.

Subpop 1

Subpop 2

$\frac{n}{2}$

3 Recruit Only Subpop.1

Subpopulation 1

$\frac{3n}{4}$

4 Recruit Only Subpop.2

Subpopulation 2

$\frac{3n}{4}$

M

Two-Stage Adaptive Enrichment Design

Assume known variances and normally distributed outcomes; subpopulation cumulative sample sizes and z-statistics and are then sufficient statistics for $\Delta_1, \Delta_2, \Delta_C$.

Two-Stage Adaptive Enrichment Design

Assume known variances and normally distributed outcomes; subpopulation cumulative sample sizes and z-statistics and are then sufficient statistics for $\Delta_1, \Delta_2, \Delta_C$.

- $Z_s^{(1)}$: z-statistic for subpopulation s at end of stage 1;
- $Z_s^{(F)}$: Final, cumulative z-statistic for subpop. s at end of stage 2.

Two-Stage Adaptive Enrichment Design

Assume known variances and normally distributed outcomes; subpopulation cumulative sample sizes and z-statistics and are then sufficient statistics for $\Delta_1, \Delta_2, \Delta_C$.

- $Z_s^{(1)}$: z-statistic for subpopulation s at end of stage 1;
- $Z_s^{(F)}$: Final, cumulative z-statistic for subpop. s at end of stage 2.

Decision rule D is map from $\mathbf{Z}^{(1)} = (Z_1^{(1)}, Z_2^{(1)})$ to possible enrollment decisions $\mathcal{D}$.

Multiple testing procedure M is map from $\mathbf{Z}^{(F)} = (Z_1^{(F)}, Z_2^{(F)})$ and decision D to set of null hypotheses rejected (if any).

Two-Stage Adaptive Enrichment Design

Assume known variances and normally distributed outcomes; subpopulation cumulative sample sizes and z-statistics and are then sufficient statistics for $\Delta_1, \Delta_2, \Delta_C$.

- $Z_s^{(1)}$: z-statistic for subpopulation s at end of stage 1;
- $Z_s^{(F)}$: Final, cumulative z-statistic for subpop. s at end of stage 2.

Decision rule D is map from $\mathbf{Z}^{(1)} = (Z_1^{(1)}, Z_2^{(1)})$ to possible enrollment decisions $\mathcal{D}$.

Multiple testing procedure M is map from $\mathbf{Z}^{(F)} = (Z_1^{(F)}, Z_2^{(F)})$ and decision D to set of null hypotheses rejected (if any).

User specifies: (i) loss function $L(D, M; \Delta_1, \Delta_2)$, e.g., total sample size; and (ii) distribution Λ on alternatives (Δ_1, Δ_2).

Constrained Bayes Optimization Problem

Problem inputs: p_1; set of possible stage 2 decisions; σ_1^2, σ_2^2; clinically meaningful min. treatment effect Δ^{min}; loss function L; distribution Λ on alternatives (Δ_1, Δ_2); $\alpha, \beta_1, \beta_2, \beta_C$.

Recall $D = D(\mathbf{Z}^{(1)})$ and $M = M(\mathbf{Z}^{(F)}, D(\mathbf{Z}^{(1)}))$.

Constrained Bayes Optimization Problem

Problem inputs: p_1; set of possible stage 2 decisions; σ_1^2, σ_2^2; clinically meaningful min. treatment effect $\Delta^{\min}$; loss function L; distribution Λ on alternatives (Δ_1, Δ_2); $\alpha, \beta_1, \beta_2, \beta_C$.

Recall $D = D(\mathbf{Z}^{(1)})$ and $M = M(\mathbf{Z}^{(F)}, D(\mathbf{Z}^{(1)}))$.

Constrained Bayes Opt. Problem: Find pair (D, M) minimizing:

$$\int E_{\Delta_1,\Delta_2}[L(D, M; \Delta_1, \Delta_2)]d\Lambda(\Delta_1, \Delta_2),$$

under **familywise Type I error constraints**:

$$\sup_{(\Delta_1,\Delta_2)\in\mathbb{R}^2} \mathrm{Pr}_{\Delta_1,\Delta_2}[M \text{ rejects any true null hypothesis}] \leq \alpha,$$

Constrained Bayes Optimization Problem

Problem inputs: p_1; set of possible stage 2 decisions; σ_1^2, σ_2^2; clinically meaningful min. treatment effect $\Delta^{\min}$; loss function L; distribution Λ on alternatives (Δ_1, Δ_2); $\alpha, \beta_1, \beta_2, \beta_C$.

Recall $D = D(\mathbf{Z}^{(1)})$ and $M = M(\mathbf{Z}^{(F)}, D(\mathbf{Z}^{(1)}))$.

Constrained Bayes Opt. Problem: Find pair (D, M) minimizing:

$$\int E_{\Delta_1,\Delta_2}[L(D, M; \Delta_1, \Delta_2)] d\Lambda(\Delta_1, \Delta_2),$$

under **familywise Type I error constraints**:

$$\sup_{(\Delta_1,\Delta_2)\in\mathbb{R}^2} \mathrm{Pr}_{\Delta_1,\Delta_2}[M \text{ rejects any true null hypothesis}] \leq \alpha,$$

and **power constraints**:

$$\begin{aligned} \mathrm{Pr}_{\Delta^{\min},0}[M \text{ rejects } H_{01}] &\geq 1-\beta_1. \\ \mathrm{Pr}_{0,\Delta^{\min}}[M \text{ rejects } H_{02}] &\geq 1-\beta_2. \\ \mathrm{Pr}_{\Delta^{\min},\Delta^{\min}}[M \text{ rejects } H_{0C}] &\geq 1-\beta_C. \end{aligned}$$

Our Method to Solve Optimization Problem

Computational challenge: continuum of Type I error constraints. (It's not enough to satisfy only at global null $(\Delta_1, \Delta_2) = (0, 0)$.)

Our Method to Solve Optimization Problem

Computational challenge: continuum of Type I error constraints. (It's not enough to satisfy only at global null $(\Delta_1, \Delta_2) = (0, 0)$.)

Our solution:

1. Discretize decision region $\mathbb{R}^2$ into small rectangles $\mathcal{R}$; for any $r \in \mathcal{R}$, enforce decision rule D makes same decision for any $(Z_1^{(1)}, Z_2^{(1)}) \in r$.
2. For each decision $d \in \{1, \ldots, K\}$, discretize rejection regions $\mathbb{R}^2$ into small rectangles $\mathcal{R}'_d$; for any $r' \in \mathcal{R}'_d$, enforce that if $D = d$, multiple testing procedure M rejects same set of hypotheses for any $(Z_1^{(F)}, Z_2^{(F)}) \in r'$.
3. Discretize Type I error constraints into fine grid on boundaries of null spaces.

Our Method to Solve Optimization Problem

Computational challenge: continuum of Type I error constraints. (It's not enough to satisfy only at global null $(\Delta_1, \Delta_2) = (0, 0)$.)

Our solution:

1. Discretize decision region $\mathbb{R}^2$ into small rectangles $\mathcal{R}$; for any $r \in \mathcal{R}$, enforce decision rule D makes same decision for any $(Z_1^{(1)}, Z_2^{(1)}) \in r$.
2. For each decision $d \in \{1, \dots, K\}$, discretize rejection regions $\mathbb{R}^2$ into small rectangles $\mathcal{R}'_d$; for any $r' \in \mathcal{R}'_d$, enforce that if $D = d$, multiple testing procedure M rejects same set of hypotheses for any $(Z_1^{(F)}, Z_2^{(F)}) \in r'$.
3. Discretize Type I error constraints into fine grid on boundaries of null spaces.

Discretized opt. problem is not convex. However, we construct reparametrization that is sparse, linear program:

$$\max_x \quad \mathbf{c}^\mathbf{T}\mathbf{x} \qquad \text{s.t.} \qquad \mathbf{A}\mathbf{x} \leq \mathbf{b}.$$

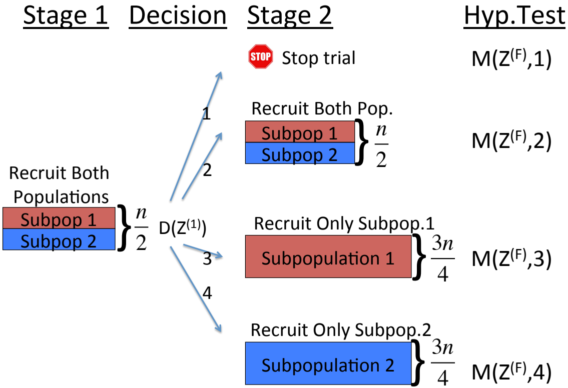
We apply advanced optimization methods to solve this.

Example

$p_1 = 1/2$, $\alpha = 0.05$, $\sigma_1^2 = \sigma_2^2$. L =total sample size. Prior Λ equally weighted pt. masses at (Δ_1, Δ_2) equal to $(0,0), (\Delta^{\min}, 0)$, $(0, \Delta^{\min})$, $(\Delta^{\min}, \Delta^{\min})$.

Example

$p_1 = 1/2$, $\alpha = 0.05$, $\sigma_1^2 = \sigma_2^2$. $L =$total sample size. Prior Λ equally weighted pt. masses at (Δ_1, Δ_2) equal to $(0,0), (\Delta^{\min}, 0)$, $(0, \Delta^{\min})$, $(\Delta^{\min}, \Delta^{\min})$. Sample size n is min s.t. standard design satisfies each power constraint at $1 - \beta_j = 0.64$ for $j \in \{1, 2, C\}$. Set each $1 - \beta_j = 0.82$.



Stage 1 | Decision | Stage 2 | Hyp.Test

Recruit Both Populations: Subpop 1, Subpop 2 $\}\ \frac{n}{2}$ → $D(Z^{(1)})$

1: Stop trial — $M(Z^{(F)},1)$

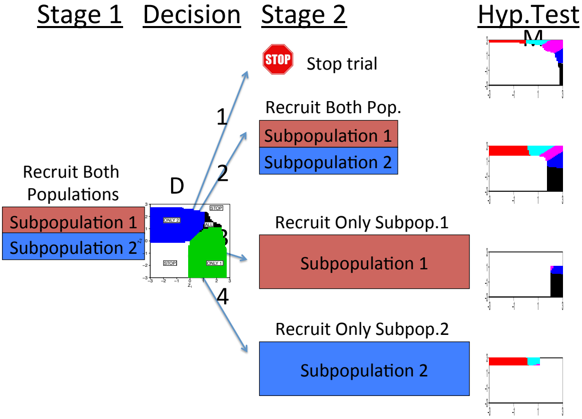
2: Recruit Both Pop.: Subpop 1, Subpop 2 $\}\ \frac{n}{2}$ — $M(Z^{(F)},2)$

3: Recruit Only Subpop.1: Subpopulation 1 $\}\ \frac{3n}{4}$ — $M(Z^{(F)},3)$

4: Recruit Only Subpop.2: Subpopulation 2 $\}\ \frac{3n}{4}$ — $M(Z^{(F)},4)$

Example

$p_1 = 1/2$, $\alpha = 0.05$, $\sigma_1^2 = \sigma_2^2$. L =total sample size. Prior Λ equally weighted pt. masses at (Δ_1, Δ_2) equal to $(0,0), (\Delta^{\min}, 0)$, $(0, \Delta^{\min})$, $(\Delta^{\min}, \Delta^{\min})$. Sample size n is min s.t. standard design satisfies each power constraint at $1 - \beta_j = 0.64$ for $j \in \{1, 2, C\}$. Set each $1 - \beta_j = 0.82$.



Stage 1

Decision

Stage 2

Hyp.Test

Recruit Both Populations

Subpopulation 1

Subpopulation 2

D

1

2

3

4

STOP Stop trial

Recruit Both Pop.

Subpopulation 1

Subpopulation 2

Recruit Only Subpop.1

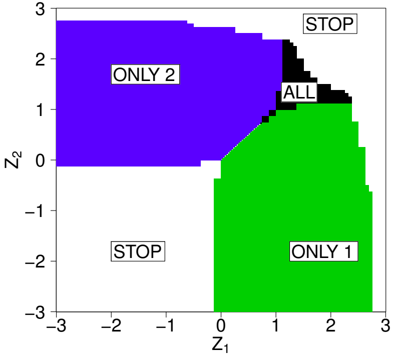
Subpopulation 1

Recruit Only Subpop.2

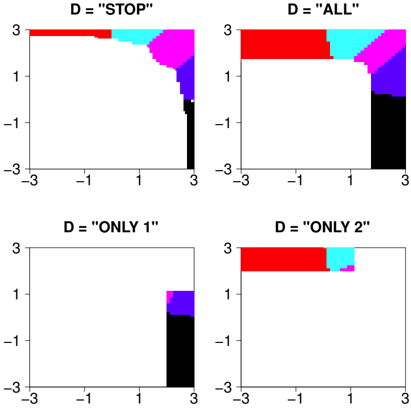
Subpopulation 2

Approximately Optimal Solution

Decision Rule for Stage 2 Enrollment:

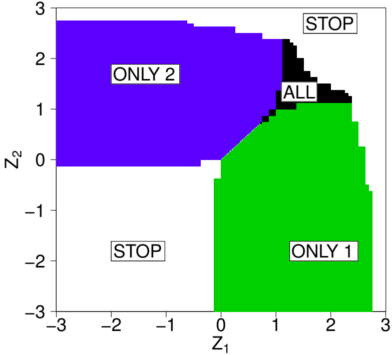


Rejection Regions under Each Decision: (in terms of $Z_1^{(F)}, Z_2^{(F)}$)

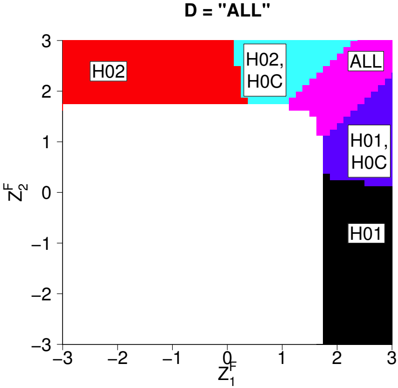


Approximately Optimal Solution

Decision Rule to Enroll Stage 2:

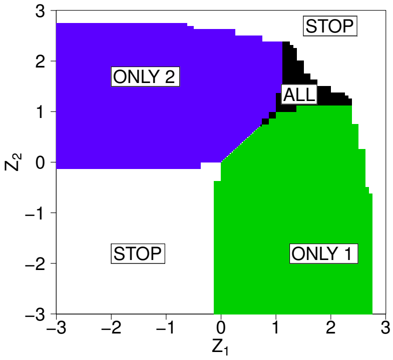


Rejection Regions for Decision:

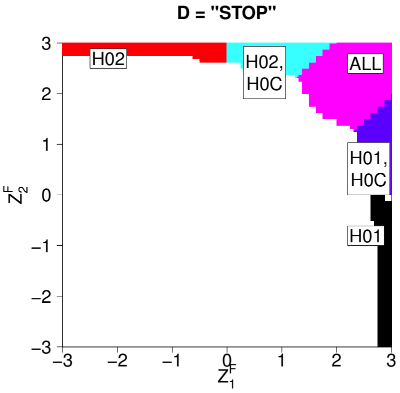


Approximately Optimal Solution

Decision Rule to Enroll Stage 2:

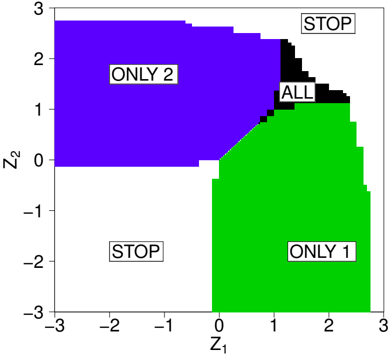


Rejection Regions for Decision:

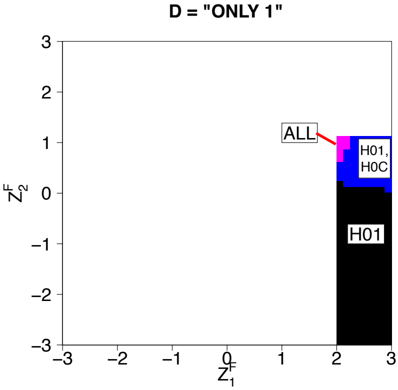


Approximately Optimal Solution

Decision Rule to Enroll Stage 2:

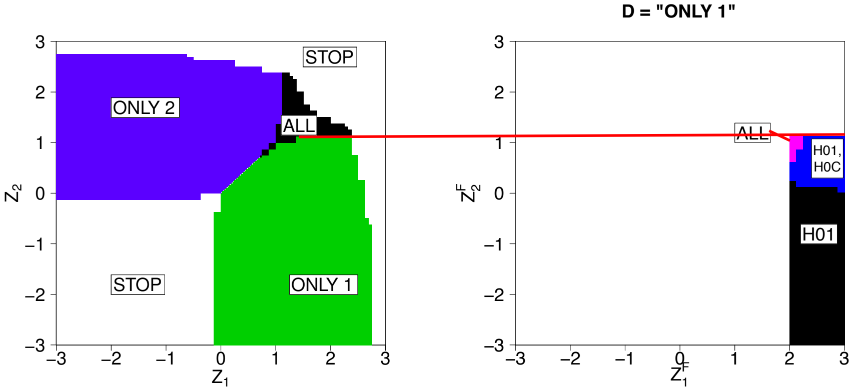


Rejection Regions for Decision:



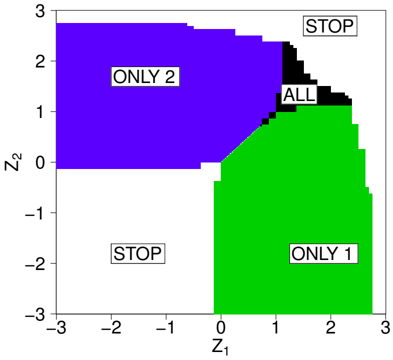
Approximately Optimal Solution

Decision Rule to Enroll Stage 2: Rejection Regions for Decision:

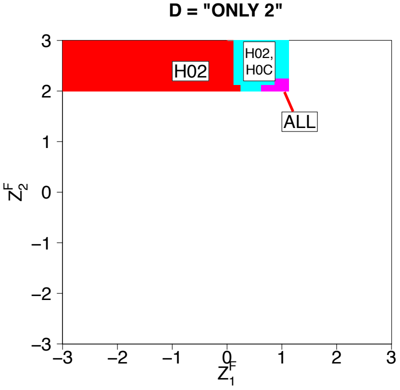


Approximately Optimal Solution

Decision Rule to Enroll Stage 2:



Rejection Regions for Decision:



Power Comparison

Compare to adaptive enrichment design using p-value combination approach (Bauer and Köhne, 1994), with Dunnett intersection test and inverse-normal combination function. Early stopping is incorporated using O'Brien-Fleming boundaries for each intersection null hypothesis. Decision rule for stage 2:

- if combined population statistic $(Z_1^{(1)} + Z_2^{(1)})/\sqrt{2} > t_c$, enroll both subpop.
- else, enroll from each subpopulation s for which $Z_s^{(1)} > t$.

Consider $\beta = \beta_1 = \beta_2 = \beta_C$. For each power threshold $1 - \beta$, we optimized over t, t_c to minimize expected sample size under the power constraints. $n =$ total sample size if both enrolled stage 2.

Table: Minimum of $\int ESS \, d\Lambda$, as power constraint $1 - \beta$ varied.

Required Power $1 - \beta$:	70%	74%	78%	82%
Comparator	$0.97n$	$1.01n$	infeasible	infeasible
Optimal	$0.79n$	$0.84n$	$0.92n$	$1.03n$

References

Rosenblum, M., Fang, X., and Liu, H. (2020) Optimal, Two Stage, Adaptive Enrichment Designs for Randomized Trials Using Sparse Linear Programming. *Journal of the Royal Statistical Society, Series B.* 82, 749-772. http://doi.org/10.1111/rssb.12366

Rosenblum, M., Liu, H., and Yen, E.-H. (2014), Optimal Tests of Treatment Effects for the Overall Population and Two Subpopulations in Randomized Trials, using Sparse Linear Programming, *Journal of American Statistical Association, Theory and Methods Section*, Volume 109. Issue 507. 1216-1228.

Learning Objective Recap: To understand the potential benefits and limitations of adaptive enrichment designs for confirmatory randomized trials.

Software for Planning Adaptive Enrichment Designs

Open Source (free) Software

Our software: http://rosenblum.jhu.edu

Optimizes Multistage enrichment designs for two subpopulations; open-source; graphical user-interface.

asd (by Parsons et al.): Two-stage enrichment designs for two subpopulations; early and final outcomes.

Commercial Software

ADDPLAN PE (ICON PLC): Multistage designs for multiple subpopulations; uses combination-tests; graphical user-interface.; must prespecify stage where enrichment can occur.

FACTS (Berry Consultants): Multistage designs for multiple subpopulations. Graphical user-interface; Bayesian hierarchical model for subpopulation treatment effects.

asd (Adaptive Seamless Design)

For planning confirmatory trials (Phase II/III)

Two-stage enrichment designs for two subpopulations; allows early and final outcomes (e.g., survival times).

Multiple choices for testing procedure, based on combination testing and closure principle (but doesn’t generally use sufficient statistics).

Allows time-to-event endpoints

Only two-stages

ADDPLAN PE

For planning confirmatory trials (Phase II/III)

Can handle more than 2 subpopulations;

Graphical User-Interface;

Multiple choices for testing procedure, based on combination testing and closure principle (but doesn’t generally use sufficient statistics).

Must a priori designate a particular stage at which change to enrollment can be made.

FACTS (Berry Consultants)

For planning Phase II trials

Can handle more than 2 subpopulations;

Graphical User-Interface;

Bayesian hierarchical model for subpopulation treatment effects.

Does not guarantee strong control of familywise Type I error rate (FWER), but gives simulated FWER at different distributions.

Assumptions Required to Estimate Different Parameters from Randomized Trial Data in the Presence of Intercurrent Events

Michael Rosenblum
Department of Biostatistics
Johns Hopkins Bloomberg School of Public Health

July 31, 2019

JOHNS HOPKINS
BLOOMBERG SCHOOL
of PUBLIC HEALTH

Disclaimer

The views presented here are my own and do not necessarily represent those of the FDA, John Hopkins University, or anyone else.

My Main Points Regarding Draft ICH E9 Revision

1. Estimands vary substantially in the assumptions needed to identify and estimate them. [Mallinckrodt et al, 2019]
2. Stronger assumptions = lower level of evidence.
3. Level of evidence ranges from ideal RCT (with perfect compliance, no missing data) to observational study.
4. Different estimands/analyses have different power. Though it may initially seem like power can be substantially improved (compared to ITT) by using some alternative estimands, probably not the case in practice.
5. Need strong reasons to justify use of non-ITT estimand.

Example: MIRA Trial (Padian et al. 2007)

- Research Question: What is effectiveness of providing diaphragms and gel in preventing HIV infection in susceptible women?
- Two arm, randomized, controlled trial (n=4948)
- Primary intervention: diaphragm and gel provision to diaphragm arm (not to control arm). Trial not blinded.
- Secondary Intervention: Intensive condom provision and counseling given to both arms.

Results of MIRA Trial (Padian et al. 2007)

- Intention-to-Treat (ITT) Analysis:
  - 158 new HIV infections in Diaphragm Arm
  - 151 new HIV infections in Control Arm
- But Avg. Reported Condom use (at last sex)
  - 53.5% in Diaphragm Arm
  - 85.1% in Control Arm
- Challenge: Remove Impact of Differential Condom use. Condom use = intercurrent event.

Intercurrent Events (ICH E9 R1 addendum)

- Could be anything measured after randomization that potentially distorts clinical interpretation of ITT effect.
- E.g.,
  - Non-compliance with study protocol (discontinuation of treatment, taking alternative treatment, switching treatments)
  - Use of rescue therapy if assigned therapy is not efficacious
  - Death (if primary outcome is something else)

Estimands in ICH E9 R1 addendum (renamed by me)

1. **Intention-To-Treat (ITT):** Effect of assigning population to treatment arm vs. control arm
2. **ITT with outcome redefined** to incorporate intercurrent events (e.g., HIV infection in time periods with no condom use)
3. **Set-Treatment-Regime (Hypothetical):** Effect of making all in population follow a treatment policy that specifies intercurrent events (e.g., setting all to no condom use)
4. **Principal strata:** ITT effect in the subpopulation where intercurrent events not impacted by arm assignment (e.g., those who would never use condoms in any case)

Estimands in ICH E9 R1 addendum (renamed by me)

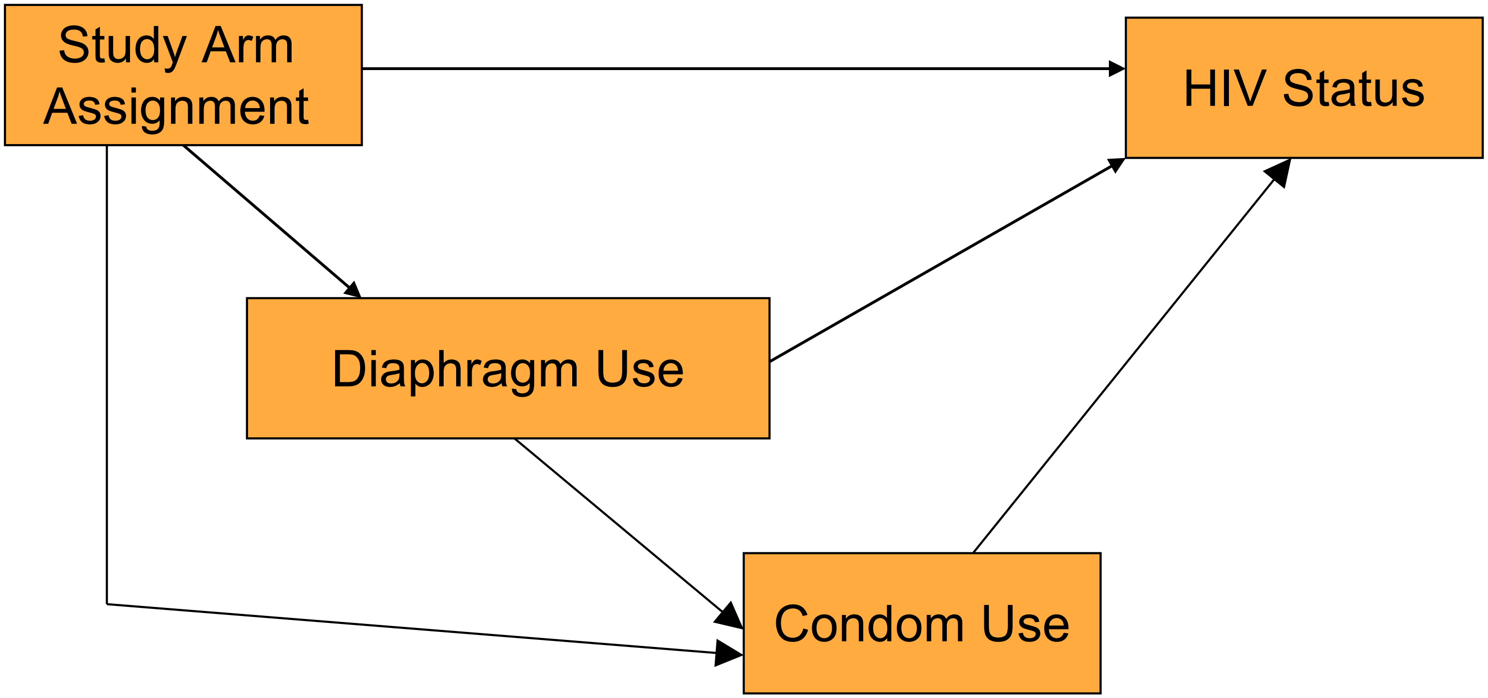
1. **Intention-To-Treat (ITT):** ignore intercurrent events
2. **ITT with outcome redefined:** absorb intercurrent events into the outcome definition
3. **Set-Treatment-Regime (Hypothetical):** what if intercurrent events experimentally set to be same for all
4. **Principal strata:** restrict to subset of population where intercurrent events can't be impacted by arm assignment.

According to Mallinckrodt et al (2019):

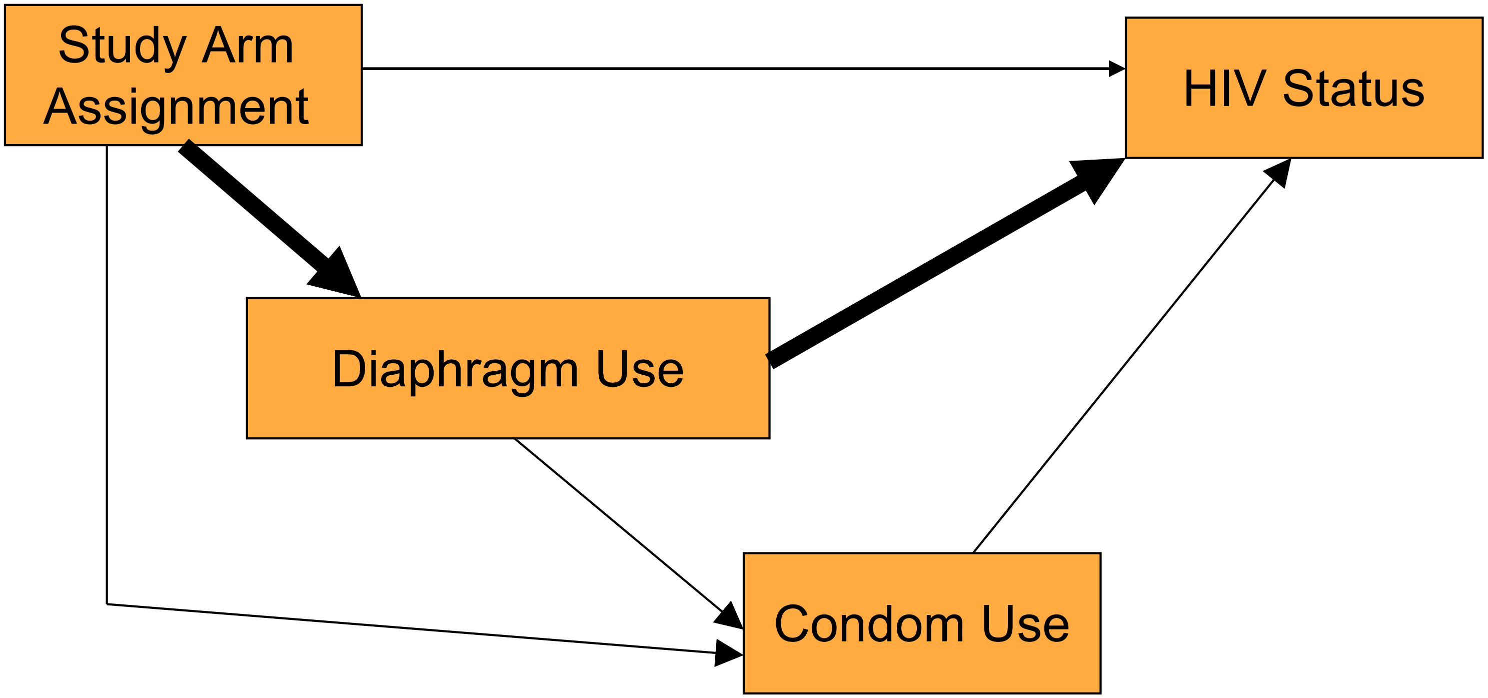
1 and **2** easy to estimate but hard to interpret. (Causal effect of treatment entangled with intercurrent events.)

3 and **4** hard to identify but better isolate treatment efficacy.

Intercurrent Events and Causal Diagrams: Isolating Causal Pathways of Clinical Interest



Intercurrent Events and Causal Diagrams: Isolating Causal Pathways of Clinical Interest



Assumptions Needed to Identify and Estimate Estimands, Beyond Missing Data Assumptions

1+2. ITT: no assumptions

3. Set-Treatment-Regime assumptions:

a. no unmeasured confounders (untestable from trial data)

b. confounders and compliance accurately measured

c. possible to follow regime given any observed history

d. can correctly model (i) compliance given observed history or (ii) outcome given observed history

4. Principal Strata: (a) exclusion restriction (untestable from trial data); (b) monotonicity. (Frangakis et al. 2004)

Power as function of compliance rate (where intercurrent event = compliance with assigned arm)

- Thought experiment: blinded drug trial with no side effects.

Assume: compliance completely at random, only those who take treatment are impacted by it, constant individual treatment effect, and constant variance. Then to achieve a desired power the required sample size (n) is:

- **ITT:** n is proportional to $1/p^2$ (for p=probability of comply)
- **Set-Treatment-Regime:** n is proportional to 1/p

E.g., at p=0.5, using **Set-Treatment-Regime** requires ½ the sample size as **ITT.** Same true for **Principal Strata**.

Power as function of compliance rate (where intercurrent event = compliance with assigned arm)

- If compliance confounded, need to adjust for baseline and post-randomization variables when estimating **Set-Treatment-Regime**
- Recommend to use longitudinal double robust estimator such as targeted maximum likelihood estimator or augmented inverse probability weighted estimator. Inverse weighting can lead to larger variance and decrease power.
- E.g., in MIRA trial the estimated relative risk of HIV infection:

--ITT estimate: 1.05, 95% CI (0.84,1.30)

--Set-Treatment-Regime (No Condom Use): 0.59, 95% CI (0.26, 4.56)

(Estimated here with IPW—gives general idea of potential variance impact. See Rosenblum et al., 2009)

Implications for Trial Design/Implementation

1. Invest more resources to ensure high compliance and follow-up, e.g., run-ins, directly observed therapy, home-visit if missed visit. (Hernán and Scharfstein, 2018; Scharfstein, 2019)
2. Lower compliance and/or follow-up = lower level of evidence. Alternative analyses cannot make up for this.
3. Other estimands can remove some distortion due to inter-current events, but strong assumptions required, similar as those needed to draw cause-effect conclusions from observational (e.g., cohort) study data.
4. Need strong reasons to justify use of non-ITT estimand.

Other Possible Estimands for Supplementary Analysis

1. ITT restricted to subpopulation with high probability of compliance given baseline variables, e.g., Pr(compliance | baseline) > 0.8. Population defined in terms of baseline variables. (“Moving the goal posts” idea of Crump, Hotz, Imbens, Mitnik, 2006)
2. ITT restricted to subset predicted to benefit in terms of baseline variables (van der Laan and Luedtke, 2016)

References

1. Mallinckrodt, C. H., J. Bell, G. Liu, B. Ratitch, M. O'Kelly, I. Lipkovich, P. Singh, L. Xu, and G. Molenberghs. "Aligning Estimators With Estimands in Clinical Trials: Putting the ICH E9 (R1) Guidelines Into Practice." Therapeutic innovation & regulatory science (2019): 2168479019836979.
2. Ratitch, Bohdana, Niti Goel, Craig Mallinckrodt, James Bell, Jonathan W. Bartlett, Geert Molenberghs, Pritibha Singh, Ilya Lipkovich, and Michael O'Kelly. "Defining efficacy estimands in clinical trials: examples illustrating ICH E9 (R1) Guidelines." Therapeutic innovation & regulatory science (2019): 2168479019841316.
3. ICH E9 (R1)Addendum. Estimands and sensitivity analysis in clinical trials. EMA/CHMP/ICH/436221/2017 2. http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2017/08/WC500233916.pdf. Accessed December 15, 2017.
4. OCEBM Levels of Evidence Working Group*. "The Oxford 2011 Levels of Evidence". Oxford Centre for Evidence-Based Medicine. http://www.cebm.net/index.aspx?o=5653 OCEBM Table of Evidence Working Group = Jeremy Howick, Iain Chalmers (James Lind Library), Paul Glasziou, Trish Greenhalgh, Carl Heneghan, Alessandro Liberati, Ivan Moschetti, Bob Phillips, Hazel Thornton, Olive Goddard and Mary Hodgkinson

References

5. Crump, R., Hotz, V. J., Imbens, G., & Mitnik, O. (2006). Moving the goalposts: Addressing limited overlap in the estimation of average treatment effects by changing the estimand.

6. Luedtke, Alexander R.; van der Laan, Mark J. Statistical inference for the mean outcome under a possibly non-unique optimal treatment strategy. Ann. Statist. 44 (2016), no. 2, 713--742. doi:10.1214/15-AOS1384. https://projecteuclid.org/euclid.aos/1458245733

7. Hernán, M. A., & Scharfstein, D. (2018). Cautions as regulators move to end exclusive reliance on intention to treat. *Annals of internal medicine*, *168*(7), 515-516.

8. Constantine E Frangakis, Ronald S Brookmeyer, Ravi Varadhan, Mahboobeh Safaeian, David Vlahov & Steffanie A Strathdee (2004) Methodology for Evaluating a Partially Controlled Longitudinal Treatment Using Principal Stratification, With Application to a Needle Exchange Program. Journal of the American Statistical Association, Volume 99, 2004 - Issue 465

9. Padian N., van der Straten A., Ramjee G., Chipato T., de Bruyn D., Blanchard K., Shiboski S., Montgomery E., Fancher H., Cheng H., Rosenblum M., van der Laan M., Jewell N., McIntyre J., and the MIRA team (2007), Diaphragm and Lubricant Gel for Prevention of HIV Acquisition in Southern African Women: a Randomised Controlled Trial. The Lancet, 370(9583), 251-261.

10. Rosenblum M., Jewell N. P., van der Laan M. J., Shiboski S., van der Straten A., Padian N. (2009), Analyzing Direct Effects in Randomized Trials with Secondary Interventions: An Application to HIV Prevention Trials. Journal of the Royal Statistical Society. Series A, 172(2), 443-465.

11. Hernán, Miguel A., and James M. Robins. "Instruments for causal inference: an epidemiologist's dream?." Epidemiology (2006): 360-372

Adaptive design in surveys and clinical trials: similarities, differences and opportunities for cross-fertilization

Michael Rosenblum
Department of Biostatistics
Johns Hopkins Bloomberg School of Public Health (JHBSPH)
Joint work with: Peter Miller and Michael Thieme (US Census Bureau),
Benjamin Reist (US Department of Agriculture),
Elizabeth A. Stuart and Thomas A. Louis (JHBSPH)

JOHNS HOPKINS
BLOOMBERG SCHOOL
of PUBLIC HEALTH

Disclaimer

The views presented here are my own and do not necessarily represent those of anyone else. Funding sources are listed on last slide.

Presentation is based on:

Rosenblum, M., Miller, P., Reist, B., Stuart, E., Thieme, M., and Louis, T. (2019) Adaptive Design in Surveys and Clinical Trials: Similarities, Differences, and Opportunities for Cross-Fertilization. *Journal of the Royal Statistical Society, Series A (Statistics in Society)*. 182, 963-982. https://doi.org/10.1111/rssa.12438

Types of Adaptation

In Trials:

(a) Early stopping for efficacy, futility or harm (group sequential designs)

(b) Modifying enrolment criteria, dose, sample size, follow-up time, randomization probabilities or end points

(c) Rerandomizing participants with poor outcomes to another treatment "Sequential, Multiple Assignment Randomized Trials" (SMART designs)

In Surveys:

(a) Changing modes of contact or data collection (mail, phone, online, visit)

(b) Modifying timing or frequency of contact attempts

(c) Changing incentives for responding

(d) Deciding when to stop data collection

Adaptive Survey Samples: Monitoring Representativeness

- We use terms ‘balance’ and ‘representativeness’ interchangeably to denote similarity between distribution of auxiliary variables in the sample and in the (subset of) respondents.
- These measured by **balance indicators and R-indicators** (Särndal, 2008, 2011; Särndal and Lundström, 2010; Lundquist and Särndal, 2013) and R-indicators (Schouten et al., 2009, 2011).
- If auxiliary variables highly correlated with outcomes, increased sample balance has potential to reduce non-response bias under missing at random.
- R-indicators identify (i) which variables contribute most to sample imbalance and (ii) direction of misrepresentation (Schouten et al., 2011).

Ideas for Cross-fertilization: Overview

Adaptive Trials and Surveys share goals of minimizing costs while making valid statistical inferences in the presence of non-response or missed visits.

We present 5 potential opportunities where methods from one domain could be applied to the other.

Common themes:

1. Adaptive methods developed for surveys may be applied to improve external validity (and sometimes also internal validity) in trials.
2. Adaptive methods from trials may be used to improve internal validity and efficiency in surveys.

1. Survey ➜ clinical trial: systematically monitor representativeness and improve by targeted enrolment or follow-up

- To improve internal validity: compare baseline variables of respondents to those of overall sample, and target intensive follow-up (double-sampling) of non-responders to increase balance/representativeness
- To improve external validity: monitor how representative the enrolled participants are of the target population and selectively increase efforts to enrol underrepresented groups.
- In both above, could use R-indicators as formal measure of balance/representativeness, and to determine which baseline variables contributing most to imbalance.

2. Survey➜clinical trial: collect and use paradata to improve retention and protocol compliance

- Paradata = contextual information collected alongside or before outcome data
- Paradata examples (collected at or before clinic visits):

  -number of attempts needed to schedule visit

  -arrival time (late or early)

  -number of questions answered and time on each question in interviews

  -clinician observations on participant (dis)satisfaction with study experience
- Paradata could be used to predict participant retention and protocol compliance. Can then identify whom to target with interventions that encourage participation and/or protocol compliance.

3. Clinical trial ➜ survey: when to stop

- Survey conducted in waves of data collection.
- Preplanned rules for when to stop data collection based on accrued data.
- **Survey stopping rules** of Rao et al. (2008); Wagner and Raghunathan (2010): focus on reducing non-response bias; predict probability that next data collection wave will impact key survey estimates by minimum threshold
- **Clinical trial stopping rules**: (Scharfstein et al., 1997; Jennison and Turnbull, 1999) focus on precision and power; use information monitoring
- Proposal: compare both types of rules in simulated adaptive surveys.
- Performance criteria: number of contact attempts, duration, and cost of survey; bias, variance, mean-squared error, and confidence interval coverage

4. Clinical trial ➜ survey: sequential multiple-assignment randomized trial (SMART) designs

- In each wave, nonrespondents randomized to different contact modes, intensities or incentives to respond.
- Goal: learn which sequences most effective in achieving sample balance, decreasing cost and decreasing length of data collection.
- E.g., Dworak and Chang (2015) randomized non-respondents in Health and Retirement Survey to different sequences of $$ and persuasive messages.
- Proposed Idea: **Estimate optimal sequential treatment rule within strata of auxiliary variables** using methods of Murphy (2003); Robins (2004); van der Laan and Luedtke (2015). Use to efficiently target non-respondents most likely to increase sample representativeness (**R–**indicator) at lowest cost.
- Limitation: Likely to require model assumptions; vulnerable to model misspec.

5. Clinical trial ➜ survey: formal adaptation protocol and a Data Monitoring Committee

Adaptation protocol: prospectively planned adaptation rules (including stopping rule).

Data Monitoring Committee (DMC):

- An arms-length committee (not including survey director) of experts who bring disinterested perspective to bear on judgements about the survey conduct.
- E.g., members of a survey organization in which adaptive survey is being planned, but who have no stake in the adaptive survey themselves.
- Monitor survey quality measures, implementation of adaptive decisions about timing or frequency or mode of contact for non-respondents, and respondent burden (from multiple contacts)

Cautions/Warnings

- Important to ensure that methods are robust to model misspecification and other violations of assumptions!
- We recommend conducting simulation studies to assess the effect of measurement error and model misspecification before conducting an adaptive design or survey (recommended by U.S. Food and Drug Administration (2010, 2016)).
- Important to do more research examining trade-offs between different types of error (e.g. non-response and measurement) and between, for example, mean-squared error and cost.
- Valid analyses of data generated by adaptive methods require more care and sophistication than those generated from a fixed plan approach.

References

- Dworak, P. and Chang, W. (2015) SMART on health and retirement study. American Association for Public Opinion Research A. Conf., Hollywood.
- Elsäßer, A., Regnstrom, J., Vetter, T., Koenig, F., Hemmings, R. J., Greco, M., Papaluca-Amati, M. and Posch, M. (2014) Adaptive clinical trial designs for European marketing authorization: a survey of scientific advice letters from the European Medicines Agency. Trials, 15, no. 1, article 383.
- European Medicines Agency (2007) Reflection paper on methodological issues in confirmatory clinical trials
- planned with an adaptive design. Technical Report. Committee for Medicinal Products for Human Use, European Medicines Agency, London.
- Food and Drug Administration (2010) Draft guidance for industry: adaptive design clinical trials for drugs and biologics. Technical Report. US Food and Drug Administration, Silver Spring.
- Food and Drug Administration (2016) Adaptive designs for medical device clinical studies: guidance for industry and Food and Drug Administration staff. Technical Report. US Food and Drug Administration, Silver Spring.
- Hatfield, I., Allison, A., Flight, L., Julious, S. A. and Dimairo, M. (2016) Adaptive designs undertaken in clinical research: a review of registered clinical trials. Trials, 17, article 150.
- Jennison, C. and Turnbull, B. W. (1999) Group Sequential Methods with Applications to Clinical Trials. London: Chapman and Hall.
- Lin, M., Lee, S., Zhen, B., Scott, J., Horne, A., Solomon, G. and Russek-Cohen, E. (2016) CBERS experience with adaptive design clinical trials. Therp. Innovn Reglatry Sci., 50, 195–203.
- Lundquist, P. and Särndal, C.-E. (2013) Aspects of responsive design with applications to the Swedish living conditions survey. J. Off. Statist., 29, 557–582.
- Mistry, P., Dunn, J. A. and Marshall, A. (2017) A literature review of applied adaptive design methodology within the field of oncology in randomised controlled trials and a proposed extension to the CONSORT guidelines. BMC Med. Res. Methodol., 17, article 108.

References

- Morgan, C. C., Huyck, S., Jenkins, M., Chen, L., Bedding, A., Coffey, C. S., Gaydos, B. and Wathen, J. K. (2014) Adaptive design: results of a 2012 survey on perception and use. Therp. Innovn Reglatry Sci., 48, 473–481.
- Murphy, S. A. (2003) Optimal treatment regimens. J. R. Statist. Soc. B, 65, 331–355.
- Rao, R. S., Glickman, M. E. and Glynn, R. J. (2008) Stopping rules for surveys with multiple waves of nonrespondent follow-up. Statist. Med., 27, 2196–2213.
- Robins, J. M. (2004) Optimal structural nested models for optimal sequential decisions. In Proc. 2nd Seattle Symp. Biostatistics. New York: Springer.
- Rosenblum, M., Miller, P., Reist, B., Stuart, E., Thieme, M., and Louis, T. (2019) Adaptive Design in Surveys and Clinical Trials: Similarities, Differences, and Opportunities for Cross-Fertilization. *Journal of the Royal Statistical Society, Series A (Statistics in Society)*. 182, 963-982. https://doi.org/10.1111/rssa.12438
- Särndal, C.-E. (2008) Assessing auxiliary vectors for control of nonresponse bias in the calibration estimator. J. Off. Statist., 24, no. 2, article 167.
- Särndal, C.-E. (2011) Dealing with survey nonresponse in data collection, in estimation. J. Off. Statist., 27, no. 1, article 1.
- **Särndal,C.-E. and Lundström, S. (2010) Design for estimation: identifying auxiliary vectors to reduce nonresponse bias.** Surv. Methodol.**,** 36**, 131–144.**
- Scharfstein, D. O., Tsiatis, A. A. and Robins, J. M. (1997) Semiparametric efficiency and its implication on the design and analysis of group-sequential studies. J. Am. Statist. Ass., 92, 1342–1350.
- Schouten, B., Cobben, F. and Bethlehem, J. (2009) Indicators for the representativeness of survey response. Surv. Methodol., 35, 101–113.
- Schouten, B., Shlomo, N. and Skinner, C. (2011) Indicators for monitoring and improving survey response. J. Off. Statist., 27, 231–253.
- van der Laan, M. J. and Luedtke, A. R. (2015) Targeted learning of the mean outcome under an optimal dynamic treatment rule. J. Causl Inf., 3, 61–95.
- Wagner, J. and Raghunathan, T. E. (2010) A new stopping rule for surveys. Statist. Med., 29, 1014–1024.

Acknowledgments/Funding

Presentation is based on: Rosenblum, M., Miller, P., Reist, B., Stuart, E., Thieme, M., and Louis, T. (2019) Adaptive Design in Surveys and Clinical Trials: Similarities, Differences, and Opportunities for Cross-Fertilization. *Journal of the Royal Statistical Society, Series A (Statistics in Society)*. 182, 963-982. https://doi.org/10.1111/rssa.12438

Acknowledgments: T. A. Louis was supported by an interagency personnel agreement with Johns Hopkins University while serving as Associate Director for Research and Methodology at the US Census Bureau, M. Rosenblum by the Patient-Centered Outcomes Research Institute (grant ME-1306-03198) and the US Food and Drug Administration (grant HHSF223201400113C) and E. A. Stuart by the National Science Foundation (grant DRL-1335843) and the Institute of Education Sciences (grant R305D150001). This research was supported in part by the Intramural Research Program of the US Department of Agriculture National Agricultural Statistics Service. This work is solely the responsibility of the authors and does not represent the views of these agencies; the findings and conclusions in this publication have not been formally disseminated by any of the above agencies and should not be construed to represent any agency determination or policy.

Overview of My Research: Developing Improved Methods and Software for Design and Analysis of Clinical Trials

1. Reduce sample size by **optimal adjustment for prognostic variables**

2. **New adaptive enrichment designs** that maximize power under cost constraints; open-source software tool for trial planning/optimization; Clinical applications in Stroke, Cardiac Resynchronization Devices, Alzheimer’s Disease, and HIV Treatment

3. **Stress-testing trial designs:** software tool that systematically probes trial designs to find and fix weaknesses before trial conducted.

FDA Guidance Documents on Adaptive Designs

FDA Adaptive Design Clinical Trials for Drugs and Biologics Guidance for Industry, 2019

https://www.fda.gov/regulatory-information/search-fda-guidance-documents/adaptive-design-clinical-trials-drugs-and-biologics-guidance-industry

FDA Adaptive Designs for Medical Device Clinical Studies Guidance for Industry, 2016

https://www.fda.gov/regulatory-information/search-fda-guidance-documents/adaptive-designs-medical-device-clinical-studies