

Design and Analysis of Clinical Trials

Pamela Shaw & Michael Proschan

UW Summer Institute in Statistics for Clinical &
Epidemiological Research (SISCER)

Virtual Course, July 8-10, 2024

Introductions

Pamela Shaw

- ▶ Kaiser Permanente Washington Health Research Institute
- ▶ Biostatistics Division
- ▶ `pamela.a.shaw@kp.org`

Michael Proschan

- ▶ National Institute of Allergy and Infectious Diseases (NIAID)
- ▶ Office of Biostatistics Research
- ▶ `proschan@niaid.nih.gov`

Course Outline

All times are Pacific Daylight Savings Time (PDT)

Day 1

1. 8:30-8:40 Introductions
2. 8:40-9:30 Choice of primary outcome and analysis
9:30-9:45 Break
3. 9:45-10:30 Randomization
10:30-10:45 Break
4. 10:45-12:00 Sample size/ Power

Day 2

6. 8:30-10:15 Interim monitoring
7. 10:15-10:45 Break
8. 10:45-12:00 Futility

Course Outline (2)

All times are Pacific Daylight Savings Time (PDT)

Day 3

1. 8:30-9:30 Handling missing data
9:30-9:45 Break
2. 9:45-10:45 Multiple Comparisons
10:45-11:00 Break
3. 11:00-11:55 Adaptive design
4. 11:55-12:00 Wrap Up

Overall aim

That you will gain a set of simple tools and principles that go a long way towards robust clinical trial design and analysis.

- ▶ Emphasis will be on practical application
- ▶ Examples will be used throughout
- ▶ Key references provided

Lecture 1: Choice of primary outcome and analysis

- ▶ A **Randomized Controlled Trial (RCT)** is a study of a novel intervention in human subjects where the intervention assignment is randomized
- ▶ An RCT is the gold standard for clinical evidence for establishing efficacy
- ▶ International Council for Harmonisation (ICH)/FDA Guidance provide universally adopted guidelines to maintain rigorous standards for the ethical and scientific integrity of the trial
 - ▶ ICH E9 Statistical Principles
 - ▶ <https://www.ich.org/page/efficacy-guidelines>
- ▶ A central pillar to the scientific rigor of the RCT is the choice of a relevant clinical endpoint that will reliably and efficiently capture the treatment effect of interest

A few definitions....

Types of randomized studies

Parallel group - subjects randomized to one of k treatments

Cross-over - each subject used as their own control. Patients receive each treatment sequentially, and the order is randomized

Factorial design - Multiple treatments under study, where each has a control. So for k treatments, subjects randomized to one of 2^k possible arms in a full factorial design

Cluster randomized - groups instead of individuals are randomized (eg. schools, building, clinic)

Phases of clinical trials

<https://www.fda.gov/patients/drug-development-process/step-3-clinical-research>

- ▶ Phase 1: Evaluates safety and dosage of drug. Generally 20-100 subjects, either healthy or with condition
- ▶ Phase 2: Evaluating efficacy and side effects. Up to 300 subjects with condition
- ▶ Phase 3: Evaluating efficacy and monitoring adverse events. 300-3000 with disease condition

Note, these sample size ranges can vary based on disease setting, e.g. cancer treatment trials tend to be smaller, cancer prevention tend to be large

The RCT Gold Standard

Key features that contribute to the strength of evidence of the RCT:

- ▶ Randomization of the treatment allocation allows for causality to be established
- ▶ A single primary outcome to evaluate efficacy is chosen
- ▶ Outcomes and analyses are pre-specified
- ▶ Analyses are done as intent-to-treat

Intent-to-treat analyses

An **intent-to-treat (ITT)** analysis is one where randomized individuals are analyzed in the group they were randomized to, regardless of what happens during the trial. Analyze as you randomize!

- ▶ Randomization ensures that there are no systematic differences between the treatment groups
- ▶ The exclusion of patients from the analysis on a systematic basis (e.g., lack of compliance with assigned treatment) can introduce systematic differences between treatment groups, thereby biasing the comparison
- ▶ Sometimes a modified ITT analysis (mITT) is considered, which would consider very limited exceptions to ITT
 - ▶ Shortly after randomization and before any intervention was given (control or otherwise), trial participant drops out
 - ▶ Other exceptions to ITT not widely accepted (more details in Lecture 2)

Implication of ITT for primary endpoint

- ▶ ITT means representing patients in the analysis even if they have missing data
 - ▶ Missing data must be imputed for an ITT or IPW approach to be considered (Day 3 topic)
- ▶ Too much missing data will degrade the integrity/acceptability of the trial results
- ▶ A fundamental consideration of primary endpoint is that it be something that can be reliably obtained on all subjects

Considerations for the primary outcome

- ▶ Should be measured similarly in both treatment arms
- ▶ Less is more (benefits of choosing 1 primary)
- ▶ Reliability/feasibility
- ▶ Clinical relevance (surrogate endpoints, composite endpoints)
- ▶ Primary analysis: Efficiency versus robustness. (phase 1 vs phase 3)
 - ▶ In phase 1 avoid type II error, in phase 3 avoid type I error. Phase 2 you are somewhere in between
- ▶ Composite outcomes
- ▶ Interpretability (Clearly stated estimand)

The Measurement Principle

The process of measurement of the primary outcome should not be influenced by treatment

- ▶ The primary outcome should be measurable in all subjects
- ▶ There should be similar monitoring of events in both treatment arms
- ▶ Sometimes violations of the measurement principle can be subtle
 - ▶ Example: Suppose a viral vaccine causes mild disease. Then comparison of viral load between arms may suggest a treatment benefit, but point is moot since vaccine caused the disease

So why choose only 1 endpoint?

Probability of at least one false positive test assuming multiple independent tests **under the null**

Number of tests	Prob of ≥ 1 significant test
1	0.0500
2	0.0975
3	0.1426
4	0.1855
5	0.2262
6	0.2649
7	0.3017
8	0.3366
9	0.3698
10	0.4013

Maintaining type I error without loss of power

The dominant paradigm:

1. Pick one efficacy outcome as the primary outcome
 - ▶ Formal hypothesis test maintains α level
2. Consider other endpoints as secondary or exploratory
3. If rigorous standards are sought for more than primary outcome, consider adjustment for multiple comparisons (Day 3 topic)

Reliability

- ▶ Clinical outcomes that are more variable between patients, such as those affected by more factors than just the treatment, will have less power
- ▶ Difficult to measure quantities will have more missing data (e.g. more assay failure)
 - ▶ This can introduce bias, particularly if say lower levels more likely to be missing
 - ▶ Worry if too many values below limit of detection, it will be difficult to detect arm differences
 - ▶ If your trial involves a novel assay, having some pilot data will be important before launching the trial

Efficacy and Safety of Metronidazole for Pulmonary Multidrug Resistant Tuberculosis (MDR-TB)

Study NCT00425113

Background

- ▶ MDR TB is a difficult to treat disease. Individuals have been observed to fail first line therapies (Isoniazid and Rifampicin)
- ▶ Standard MDR-TB treatment is 18-24 months of 2nd-line antibiotics
- ▶ In vitro data showed that metronidazole is active against *Mycobacterium tuberculosis* (MTB) maintained under anaerobic conditions
- ▶ Pre-clinical studies (non-human primates, rabbits) also showed metronidazole may have unique activity against an anaerobic sub-population of bacilli in human disease

Design of the Metronidazole for MDR-TB Trial

- ▶ A double-blinded RCT with a planned 60 patients with MDR-TB randomized to one of placebo or 500 mg MTZ for first 8 weeks of 2nd-line TB therapy. 2nd line therapy to continue for 18-24 months.
- ▶ Primary outcome was "Changes in TB lesion sizes" at 6 months
- ▶ In humans, TB disease characterized by aerobic (cavities) and anaerobic (caseous necrotic nodules) areas
- ▶ Hypothesis was that MTZ would reduce the volume of nodules in the lung, which would be quantified at baseline and follow-up, using FDG-PET HRCT
- ▶ FDG-PET HRCT was a relatively novel tool for assessing extent of TB disease

Problematic Primary outcome

- ▶ As scans were evaluated on the patients in the trial it became clear that the primary outcome was not a good measure of change
- ▶ For some patients, volume of lesions decreased because lesions were reducing in size as patients improved
- ▶ For some patients, volume of lesions decreased because lesions collapsed into cavities as patients got worse
- ▶ Number of lesions was discussed as a secondary endpoint, but the number of lesions could increase or decrease as patients got better
- ▶ Investigators had no choice but to alter the primary and other outcome measures of the trial
- ▶ Changing primary endpoint mid-trial is problematic

Transparency on Clinical Trials.gov

Study NCT00425113

5 years after trial opened and after study had closed early, primary outcome was changed

Changes in TB Lesion Sizes Using High Resolution Computed Tomography (HRCT). [Time Frame: 6 months.] Lesions were defined as nodules (<2 mm, 2-<4 mm, and 4–10 mm), consolidations, collapse, cavities, fibrosis, bronchial thickening, tree-in-bud opacities, and ground glass opacities. Each CT was divided into six zones (upper, middle, and lower zones of the right and left lungs) and independently scored for the above lesions by three separate radiologists blinded to treatment arm. A fourth radiologist adjudicated any scores that were widely discrepant among the initial three radiologists. The HRCT score was determined by visually estimating the extent of the above lesions in each lung zone as follows: 0=0% involvement; 1= 1-25% involvement; 2=26-50% involvement; 3=51-75% involvement; and 4=76-100% involvement. A composite score for each lesion was calculated by adding the score for each specific abnormality in the 6 lung zones and dividing by 6, with the change in composite score measured at 2 and 6 months compared to baseline. Composite sums of all 10 composite scores are reported.

RCT Example: Effect of Ranitidine on Hyper-IgE Recurrent Infection (Job's) Syndrome

NCT00527878

Background

- ▶ Hyper-IgE syndrome (HIES) is an immunological disorder caused by a genetic mutation (STAT3) characterized by recurrent infections of the ears, sinuses, lungs and skin, and abnormal levels of the antibody immunoglobulin E (IgE).
- ▶ Patients with hyper-IgE syndrome also tend to have skeletal abnormalities: characteristic face, retained teeth, and recurrent fractures from minimal trauma
- ▶ An early phase RCT was launched in 2007 at NIAID to study whether ranitidine would reduce infections

Considerations for an endpoint for this diverse disease

One possibility: A patient-reported score of severity of symptoms.

Problem: Patients with more severe disease less bothered by mild to moderate symptoms, and high functioning patients bothered by relatively minor symptoms

Alternative: A numeric score was considered that would capture the number of new infections

- ▶ The number of infections that required new antibiotics was reported on a quarterly basis, to balance burden and accuracy (require recall over shorter period)
- ▶ Total number in a year is prone to missingness
 - ▶ Rate of infections per month is a more flexible endpoint
- ▶ Disease had many other chronic morbidities (e.g. recurrent fracture), but ranitidine only expected to affect infections

Final: Primary endpoint chosen was the rate of infections (i.e. avg # per month during first year). Primary endpoint to require at least 2 of the 4 quarters to give a robust estimate of the yearly rate.

Clinical relevance

- ▶ When weighing possible outcomes/endpoints want to consider the relative seriousness of the different conditions/symptoms the drug could be affecting
- ▶ Also need to consider the mechanisms of action for the intervention under study and the outcomes expected to have the biggest change
- ▶ Often a trade-off between clinical relevance and power: frequent less serious events and infrequent serious events
- ▶ In some trials it is more practical to observe a surrogate outcome
 - ▶ In TB trials the short-term endpoint of sputum conversion or change in first 6 months used in place of the gold standard “cure” outcome: 6 months after end of therapy need to be disease free

Surrogate Outcome

- ▶ Various definitions exist for a surrogate endpoint. Ellenberg and Hamilton (1989) lay out a general definition: A “Surrogate endpoint captures an intermediate endpoint on the disease pathway, which is informative of the true outcome”
- ▶ Generally, the point of a surrogate endpoint is to have an expected reduction in sample size or trial duration, such as when a rare or distal endpoint is replaced by a more frequent or proximate endpoint
- ▶ In 1989, Prentice laid out conditions for a surrogate outcome (known as the Prentice criterion), as well as a working definition, that assumes a treatment Z effect on the true endpoint Y is completely captured by the surrogate endpoint X
 - ▶ $E(Y|Z, X) = E(Y|X)$
- ▶ Prentice criterion has come under criticism as not practical and various other discussions have ensued regarding the definition of and how to validate a surrogate

Examples of surrogate endpoints

- ▶ Cholesterol, when ultimate goal to reduce cardiovascular events (e.g., heart attacks and strokes)
- ▶ Blood pressure, when ultimate goal to reduce stroke risk
- ▶ CD4 or HIV viral load, when ultimate goal to reduce serious infections AIDS infections or death
- ▶ Hemoglobin A1c, when ultimate goal to reduce serious complications of diabetes

CAST Example: Caution is needed when working with surrogate endpoints

- ▶ Arrhythmias can lead to cardiac arrest, which is fatal a high percentage of time
- ▶ Given that arrhythmia is on the causal pathway to cardiac arrest and sudden death, arrhythmia could be considered a surrogate endpoint for cardiac arrest/sudden death
- ▶ In mid-80s to 1990, encainide, flecainide and moricizine were approved by FDA on basis of their effect on arrhythmias
- ▶ Anti-arrhythmia drugs were in broad use at the time of Cardiac Arrhythmia Suppression Trial (CAST)
 - ▶ Led to difficulties in recruitment

Cardia Arrhythmia Suppression Trial (CAST)

CAST Investigators, 1989

CAST would test hypothesis that *suppression* of ventricular premature complexes after a myocardial infarction would improve survival

- ▶ Patients at high risk for death from cardiac arrest were eligible (recent MI, low ejection fraction)
- ▶ Three suppression drugs considered, with matching placebos: encainide, flecainide and moricizine
- ▶ The primary endpoint of the trial was death or cardiac arrest with resuscitation, either of which was due to arrhythmia
- ▶ During titration phase analysis of Holter recordings required to show that a drug had indeed suppressed arrhythmias adequately before a patient could be randomized
- ▶ Randomization was to the agent that achieved successful suppression or matching placebo
- ▶ Trial launched in June 1987 with 3 year planned recruitment

CAST Results

Ruskin (1989)

- ▶ In April 1989 after 1498 patients randomized, the ecainide and flecainide arms were stopped due to higher overall cardiac mortality and higher mortality due to arrhythmia
- ▶ In April 1989 CAST II - placebo-controlled trial was launched with moricizine as only active drug
 - ▶ Only 277 patients to date had been randomized
- ▶ Titration phase now had a blinded placebo
- ▶ Early exposure to moricizine was shown to have higher death rates than the placebo arm
- ▶ Anti-arrhythmia drugs were no longer routinely recommended (Greene et al., 1992)
- ▶ *Deadly Medicine: Why tens of thousands of heart patients died in America's worst drug disaster Moore (1995a)*

Surrogate Endpoints: Controversy continues

- ▶ In early phase trials, need biologically motivated intermediate endpoints
- ▶ Many would argue in large phase III trials, need to move to the target clinical (non-surrogate) endpoint
- ▶ Those in pharmaceutical industry would argue that validated surrogates would mean smaller, faster cheaper trials
 - ▶ Fewer patients are exposed during testing, and beneficial new medications reach the market faster
- ▶ Problem: No universal way to validate a surrogate. Some advocate meta-analyses (Molenberghs et al., 2002)
- ▶ Buyse and Molenberghs (1998); Buyse et al. (2000) reviews different methods to validate a surrogate, with extension to meta-analyses

Many examples of misleading surrogates

- ▶ Cyclic adenosine monophosphate– enhancing agents, such as milrinone, were considered a “particularly rational approach to the treatment of chronic heart failure.” Milrinone was later found to increase mortality by 28% over placebo. (Svensson et al. (2013))
- ▶ Estrogen in pre-menopausal women thought to be protective against heart disease. Hormone replacement therapy used for decades in post-menopausal women before found to be harmful in the WHI
- ▶ High blood sugar in diabetics can lead to bad outcome. Hemoglobin A1c used to monitor diabetes mellitus therapy (short-term effects of treatment). In ACCORD, over suppression of hbA1c led to increased mortality (Action to Control Cardiovascular Risk in Diabetes Study Group, 2008)
- ▶ Svensson et al. (2013) give multiple examples of treatment approved based on surrogate, later found harmful on true outcome
 - ▶ Even when new drug under consideration is a member of an already established class, adequate safety cannot be assumed (cerivastatin)
 - ▶ Demonstrated value for one indication does not necessarily extend to a related indication (Dronedarone hydrochloride)

Composite outcomes are another way to improve practicality

Composite outcomes are an outcome that combine multiple clinical endpoints. General idea behind composite endpoints is to increase power through an increased event-rate

Examples

- ▶ Time-to-first of disease progression or death (Progression-free survival)
- ▶ Relapse-free survival
- ▶ Major adverse cardiovascular events (MACE)
- ▶ Time to first serious AIDS or serious non-AIDS event in the Strategic Timing of AntiRetroviral Treatment (START) trial
- ▶ Time to first of cardiac arrest or arrhythmic death (CAST)

Note: Some composite endpoints are surrogate endpoints

Considerations for composite endpoints

Neaton et al. (2005)

- ▶ The definition of the endpoint should be clearly established a priori
- ▶ Endpoints should have similar seriousness
- ▶ In order to interpret the results of the trial, should look at treatment effect on the individual components of the composite (secondary endpoints)
- ▶ All endpoints should be affected by the drug
 - ▶ You could decrease power if you expand composite to include things not affected by treatment just for sake of higher event rate

Example: SOLVD Trial

NEJM 1991, 325: 293-302.

Background

- ▶ SOLVD a RCT examining novel treatment for prevention of mortality/hospitalization in patients with congestive heart failure (CHF) and weak left ventricle ejection fraction (EF)
- ▶ In 1986-89, 2569 patients randomized to enalapril or placebo
- ▶ Enalapril found beneficial for mortality ($p = 0.0036$) and time to first hospitalization/death ($p < 0.0001$)

Subgroup Analysis (Shaw and Fay (2016))

- ▶ Seek to evaluate treatment effect on subset of 662 diabetic subjects
- ▶ Considered alternative to time to first that considers overall severity

SOLVD: Results

Endpoint	Enalapril (N=319)		Placebo (N=343)		Cox PH	Score Test
	Yes	No	Yes	No	HR	(P-value)
Death	137	182	145	198	0.99	(0.91)
Hospitalization	94	225	148	195	0.60	(< 0.0001)
TTF	174	145	229	114	0.71	(0.0007)

- ▶ Treatment arm: 57/94 (61%) hospitalization followed by death
- ▶ Placebo arm: 64/148 (43%) hospitalization followed by death

Shaw and Fay severity score test $p = 0.07$ (Shaw and Fay, 2016)

SOLVD: Results

Endpoint	Enalapril (N=319)		Placebo (N=343)		Cox PH	Score Test
	Yes	No	Yes	No	HR	(P-value)
Death	137	182	145	198	0.99	(0.91)
Hospitalization	94	225	148	195	0.60	(< 0.0001)
TTF	174	145	229	114	0.71	(0.0007)

- ▶ Treatment arm: 57/94 (61%) hospitalization followed by death
 - ▶ Placebo arm: 64/148 (43%) hospitalization followed by death
- Shaw and Fay severity score test $p = 0.07$ (Shaw and Fay, 2016)

Alternative to Time-to-First: Prioritized severity score (Shaw and Fay, 2016; Shaw, 2018)

- ▶ General idea: rank individuals according to *clinical severity*
- ▶ Depending on setting, clinical severity could consider two or more outcomes or event times
- ▶ Shaw and Fay (2016) Proposed ranking considered surrogate and "true" event of interest
 - ▶ Rank the time to event of interest (death) if it is observed
 - ▶ Rank time to surrogate event (MI hospitalization) for the survivors
 - ▶ Surrogate time does not affect clinical severity when event of interest is observed
 - ▶ Perform two sample test on clinical severity which incorporates bivariate survival information
 - ▶ Resulting test is average of two log-rank tests (aids interpretation)
- ▶ Prioritization endpoints have grown in popularity in recent years. Examples: win ratio (Pocock et al., 2012), Desirability of outcome ranking (DOOR) (Evans et al., 2015). See review Shaw (2018).

Choice of primary analysis

- ▶ Want to choose an efficient analysis
- ▶ Need to consider the interpretation of the parameter for your test statistic
- ▶ If no one understands the method or parameter interpretation, then unlikely to affect clinical practice
- ▶ There may be some trade-off between efficiency and interpretability.

Common test statistics for a parallel two-arm trial

- ▶ Continuous outcome: t-test (assuming unequal variance) is a common choice
 - ▶ Note non-parametric tests like Wilcoxon Rank sum test will be more robust, particularly for modest sample sizes.
- ▶ Binary outcome: difference of proportions often of interest - exact test will be more robust and often preferred particularly for small samples sizes
- ▶ Survival outcome: simple log-rank test
- ▶ If anticipate missing data, good to consider how your primary test statistic will be calculated

Interpretability: Wilcoxon

- ▶ There are different ways to interpret a test, and some may be more relevant than others.
- ▶ Example: Wilcoxon rank sum and Mann-Whitney tests are equivalent.
 - ▶ Wilcoxon, assumes one distribution is shifted relative to other, and estimates size of shift.
 - ▶ Mann Whitney compares (treatment, control) pairs and estimates the following probability for outcome of randomly picked treatment/control patients ($>$ means better):

$$P(\text{treatment} > \text{control}) + (1/2)P(\text{treatment} = \text{control})$$

- ▶ Latter helpful in COVID-19 trial with ordinal score like WHO-8. Even if proportional odds assumption is violated, still asymptotically equivalent to Mann-Whitney (Wang and Tian (2017)), so still estimates above probability parameter.
- ▶ Another example: Hazard ratio versus restricted mean survival time (RMST). Many believe RMST is easier to interpret.

Change from baseline

- ▶ Frison and Pocock (1992) generalize the following.
- ▶ With continuous outcome Y , at least 3 ways to analyze:
 - ▶ T-test on end of study value, treatment effect estimate $\hat{\delta} = \bar{Y}_T - \bar{Y}_C$.
 - ▶ T-test on change from baseline, treatment effect estimator $\hat{\delta} = (\bar{Y}_T - \bar{X}_T) - (\bar{Y}_C - \bar{X}_C)$.
 - ▶ Analysis of covariance (ANCOVA) regression using baseline value as covariate:

$$Y = \beta_0 + \beta_1 X + \beta_2 Z + \varepsilon,$$

where z is treatment indicator and ε is a random error independent of X . Treatment effect estimator $\hat{\delta} = \bar{Y}_T - \bar{Y}_C - \hat{\beta}_1(\bar{X}_T - \bar{X}_C)$.

- ▶ Which one is best?
- ▶ Assume ANCOVA model is correct.
- ▶ Unconditionally (averaged over distribution of X), all 3 estimate the same parameter, $E(Y_T) - E(Y_C)$ because X is baseline variable, so $E(X_T) = E(X_C)$.

Change from baseline

- ▶ Asymptotically,
 - ▶ T-test on Y , $\text{var}(Y) = \beta_1^2 \sigma_X^2 + \sigma_\varepsilon^2$.
 - ▶ T-test on $Y - X$,
 $\text{var}(Y - X) = \text{var}\{\beta_0 + (\beta_1 - 1)X + \varepsilon\} = (\beta_1 - 1)^2 \sigma_X^2 + \sigma_\varepsilon^2$.
 - ▶ ANCOVA is essentially t-test on $Y - \beta_1 X$, and
 $\text{var}(Y - \beta_1 X) = \text{var}(\varepsilon) = \sigma_\varepsilon^2$. **Smallest variance, so best.**

Asymptotic power when $\sigma_X = \sigma_Y = 1$, $\rho = \text{cor}(X, Y)$.

ρ	Post	Change	ANCOVA
0.00	0.50	0.28	0.50
0.20	0.50	0.34	0.52
0.40	0.50	0.43	0.57
0.60	0.50	0.59	0.69
0.80	0.50	0.87	0.90
0.90	0.50	0.99	0.99
0.95	0.50	1.00	1.00

Note: Post is better than change if $\rho < 0.50$.

Special considerations for cluster RCT

Public Access Defibrillation (PAD) Trial (Hallstrom et al. (2004))

- ▶ Cardiac arrest has very low survival probability (10%). Can we improve survival by putting defibrillators in communities and letting lay people use them?
- ▶ Note: No guarantees because lay people might make mistakes with defibrillator and fail to call 911.
- ▶ Communities (shopping malls, apartment buildings, etc.) randomized to CPR training of lay people (like managers) or CPR training of lay people plus defibrillators.
- ▶ Primary outcome: number of people saved after cardiac arrest.
- ▶ In community-randomized trial, think of community like we think of individuals in individual-randomized trial. Primary endpoint is measured in each community (number of saves).
- ▶ 993 communities! Most community-randomized trials have only about 20 communities.

Conclusions

- ▶ The choice of a good primary outcome is paramount. Ultimately, an RCT will be judged a success or failure based on the primary outcome results
- ▶ For RCT's there are rigorous standards for the primary outcome: single endpoint with a pre-specified ITT, analysis
- ▶ Reliability/feasibility of measurement need to be considered
- ▶ Surrogate endpoints are often used in early phase trials, but a definitive trial on the clinical endpoint is required to truly understand treatment effect (remember CAST!)
- ▶ When choosing a primary analysis, robustness, power and interpretability all come into play

Lecture 2: Randomization

Outline

- ▶ Basic principles
- ▶ Randomization Methods
 - simple, permuted block, stratified
- ▶ Cluster vs individual designs
- ▶ Platform trials
- ▶ Adaptive randomization
- ▶ Threats to integrity of randomization

Basic definitions (1)

What do we mean by **random**?

- ▶ **In everyday speech**, we describe a process as random if there was no discernible pattern
- ▶ **In statistics**, random characterizes a process of selection that is governed by a known, probability rule
 - i.e., 2 treatments have equal chance of being assigned
 - Random in statistics does not mean haphazard

Basic Definitions (2)

What is **random treatment allocation**?

By random allocation, we mean that each patient has a known chance, usually equal chance, of being given each treatment, but the treatment to be given cannot be predicted. (Altman 1991)

- ▶ **Randomization** is the act of allocating a random treatment assignment.
- ▶ A patient is said to be **randomized** to a treatment arm or group when they are assigned to the treatment group using random allocation
- ▶ Clinical trials that use random treatment allocation are referred to as **randomized clinical trials**

Random Examples

Flip a coin

- “Heads” and “Tails” have equal chance for a fair coin

Rolling a die

- The numbers 1 through 6 have equal chance of coming up

Draw one ball out of an urn filled with 10 red balls and 10 blue balls

- The chance of drawing a red or blue ball are equal

Random Examples

Everybody pick a random number from this list

0 1 2 3 4 5 6 7 8 9 10

Examples of Randomized Designs

- ▶ Parallel 1-1 randomized 2-arm trial
- ▶ Parallel k-1 randomized 2-arm trial
- ▶ Factorial designs: Two or more treatments given in combination:
AB, aB, Ab, ab
- ▶ Crossover trials: every patient gets all treatments under study
- ▶ Cluster randomized trials: entire communities are randomized to receive a treatment (example: anti-smoking campaign for high schools)

Motivation Behind Randomization

- ▶ Randomization tries to ensure that only one factor is different between two or more study groups.
- ▶ Provides basis for valid statistical tests between treatment groups
- ▶ Randomization means we can attribute causality, i.e. any between group difference in outcomes can be attributed to the treatment
- ▶ In truth, randomization does not guarantee causality, but it increases the likelihood that causality is the main driving factor

Ethics of Randomization

Equipoise – uncertainty about which intervention under study in a clinical trial would have a better outcome for the participant

- The fundamental principle underlying the ethics of random treatment allocation

Masking/Blinding: Key Components of Randomization

- ▶ **Double-blinded trial**: the treatment assignment is masked so neither the investigator nor participants know the treatment assignment
 - Treatment assignments are masked, individuals are blinded
- ▶ **Single-blinded trial** : only one of investigator/participant (usually investigator) knows the treatment assignment
- ▶ Unpredictability of treatment allocation prevents selection bias
 - Even when treatment can't be blinded, it is helpful to have a blinded randomization process
- ▶ Maintaining blind throughout the trial prevents **evaluation/response bias**

Ways to Randomize

- ▶ Standard ways:
 - **Computer programs** (R, stata, sas, REDCap...)
 - Random number tables
 - Online tools (e.g., randomization.com)
- ▶ NOT legitimate
 - Odd vs even birth dates
 - Last digit of the medical record number
 - Alternate as patients enroll
- ▶ Theoretically legitimate, but not so in practice
 - Flipping a coin
 - Rolling dice
 - Drawing balls (m&ms) out of an urn (bag)

Summary of Important Features of Randomization

- ▶ Random Allocation
 - Known chance receiving a treatment
 - Cannot predict the treatment to be given
 - Scheme is reproducible
- ▶ Minimizes the risk of selection bias
- ▶ In double-blinded trials, no response/evaluation bias
- ▶ Similar treatment groups
 - Patient characteristics will tend to be balanced across study arms
 - Chance baseline imbalances between groups may still occur

Types of Randomization

- ▶ Simple
- ▶ Blocked Randomization
- ▶ Stratified Randomization
- ▶ Cluster Randomization
- ▶ Baseline Covariate Adaptive Allocation
- ▶ Response Adaptive Allocation (using interim data)

Simple Randomization

- ▶ Randomize each patient to a treatment with a known probability
 - For example, to assign one of (T,C) with equal chance then:
Use a random number generator to generate a number in (0,1);
If $u < 0.5$ assign C; If $u \geq 0.5$ assign T
- ▶ Advantage: Simple to conduct
- ▶ Advantage: Simple to analyze. The usual two-group tests (t-test, Wilcoxon, Fisher's exact, etc) are appropriate
- ▶ Disadvantage: Could have imbalance in # per arm or trends in group assignment
 - No guarantee equal number of heads and tails
 - Could have runs of heads or tails
 - Could have different distributions of a trait like gender in the different arms
- ▶ Particularly good for large trials

Chance of Imbalance Decreases with Sample Size

E.g., suppose 1000 women; expected & “worse case” allocation across T and C:

	% assigned to control	% assigned to treatment
Expected	50%	50%
95% extremes:	47%	53% or
	53%	47%

Block Randomization

- ▶ Each block would contain the desired treatment ratio. For example: equal numbers of patients assigned to each treatment within a block
 - Sample size 24, Block size = 6, 2 study interventions- A & B
BAABAB AAABBB ABABAB BBABAA
- ▶ Exactly balanced after each completed block
- ▶ Ensures treatment number on each arm at any given time is not that not far out of balance
 - Maintaining balance over time protects against unintended patterns created by changes in patient population over time
- ▶ Good for small and modest sample sizes

Block Randomization (2)

- ▶ Block size can be fixed or random
- ▶ Variable block size (permuted) adds an additional layer of blindness, especially if not masked
- ▶ Does not protect against possibility of an imbalance of a trait like gender in the two arms possible
- ▶ Any complication means more ways to make a mistake: Test algorithm!!!
 - Archive code and results (preserve reproducibility)

Issues for Block Randomization

- ▶ If blocking is not masked, the sequence can get predictable

Example: Block size 4

A B A B B A B ? Must be A.

 A A ? ? Must be B B.

- ▶ If block too small, unblinding one subject can reveal rest of block
 - i.e. if block size is 2, knowing one reveals a second
 - Solution: use random block sizes, don't use block size of 2
- ▶ Predictability can lead to selection bias
- ▶ Simple solution to selection bias
 - Do not reveal blocking mechanism
 - Use random block sizes
- ▶ Proper analysis would incorporate the blocking used in randomization, such as a test stratified on the randomization blocks (Matts and Lachin, 1988)
 - ▶ This is rarely done
 - ▶ Why some have advocated for simple randomization for larger trials, allows for simpler analysis (Lachin et al., 1988)

Sample Code in R

```
> library(blockrand)
> set.seed(31415)
> list<-blockrand(24,num.levels=2,
levels=c("T","C"),id.prefix="CCP2-",block.sizes=2:4)
> list
id block.id block.size treatment
1 CCP2-01 1 6 T
2 CCP2-02 1 6 T
3 CCP2-03 1 6 C
...
28 CCP2-28 5 8 T
29 CCP2-29 5 8 T
30 CCP2-30 5 8 C
> table(list$treatment)
C T
15 15
```


Stratified Randomization

- ▶ A priori certain factors known to be important predictors of outcome (e.g. age, gender, diabetes)
- ▶ AABB BABA BABA BAAB, balanced trial of 16 but what if women are patients 1,2,6,8 and 16?
- ▶ **Stratified randomization**: Randomize within strata so different levels of the factor are balanced between treatment groups
- ▶ **Stratified blocked randomization** is a useful way to achieve balance
 - For each subgroup or strata, perform a separate block randomization

```
## stratified by sex, 100 in stratum, 2 treatments  
male <- blockrand(n=100, id.prefix='M',  
block.prefix='M',stratum='Male')  
female <- blockrand(n=100, id.prefix='F',  
block.prefix='F',stratum='Female')
```

Considerations for Stratified/Blocked Randomization

- ▶ Common choices for strata
 - Strong prognostic variables: age, gender, diabetes
 - Logistics and politics can motivate stratification by center
- ▶ Balance will be defeated if you choose too many strata and wind up with many incomplete blocks
 - Strata add up quickly: 5 age groups, 2 genders, 3 centers = 30 strata
- ▶ Stratification should be taken into account in the data analysis
 - Blocks commonly ignored due to preference for a simple (easy to understand) analysis
 - Adjusting for strong prognostic variables can help with precision (Pocock et al., 2002; Tsiatis et al., 2008)
 - Not adjusting for stratification variables can result in inflated standard errors and incorrect nominal confidence interval coverage (Kahan and Morris, 2012)
 - Adjusting for too many factors could be a concern for small trials, another reason to keep strata # small (Kahan and Morris, 2012)

Stratified Block Randomization Example

Preexposure Prophylaxis Initiative (iPrEx) Trial

REF: NEJM 2010 v363 (27): 2587-2599

- ▶ Double-blinded placebo-controlled randomized trial examining safety and efficacy of a chemoprophylaxis regimen (once-daily oral FTC–TDF) for HIV prevention
- ▶ International multi-center study
 - 9 sites: US: Boston, San Francisco; Peru: Iquitos, Lima; Brazil: São Paulo, Rio de Janeiro (2 sites); Ecuador; Guayaquil; Thailand: Chiang Mai
 - Multiple advantages to achieving balanced allocation by site
- ▶ 2499 HIV- men or transgender women were randomly assigned in blocks of 10, stratified according to site
 - Main analysis an unadjusted logrank test for HIV seroconversion

Design consideration: Who/What to Randomize

- ▶ Person
 - Most common unit of randomization in RCTs
- ▶ Provider
 - Doctor
 - Nursing station
- ▶ Locality
 - School
 - Community
- ▶ The sample size is predominantly determined by the number of randomized units
 - This is due to correlation of repeated samples within a person/doctor/community

Cluster Randomization

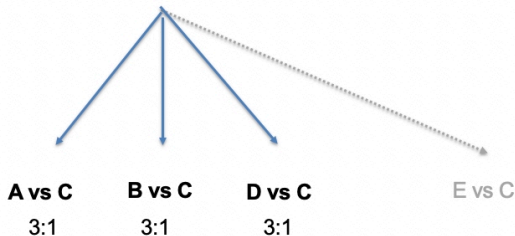
- ▶ Same ideas as before
- ▶ Unit of randomization
 - School/Clinic/Hospital/Providers/Community
- ▶ Outcome measurement
 - Students/Patients
- ▶ Need to use special models for analysis when those reporting outcomes are nested within a cluster, to account for within cluster correlation
- ▶ Best for interventions meant to be implemented at community level (smoking cessation program) and relatively quick and easy to assess outcome
 - Cost can often be an issue
 - In today's world, isolated communities harder to find

Randomization in Platform Trials

- ▶ Platform trials compare multiple intervention arms to the same control to treat a single disease
 - Different from umbrella trials that might be studying multiple indications for a single drug
- ▶ A common strategy is to randomize first to a component [(A,C) (B,C) (D,C)] and then randomize to arm in that component (Drug vs Control)
- ▶ Randomization probabilities are generally set so that you have approximately equal sized groups for each drug and control
 - From power standpoint and fixed total sample size, $n_C/n_A = \sqrt{\text{\#active arms}}$ is optimal
 - But a preference often to have equal sample size

Platform Example

Randomization to Eligible Components



Note: When 4 patients have been randomized to each of these 3 arms, would have 3 on A, B, D and C

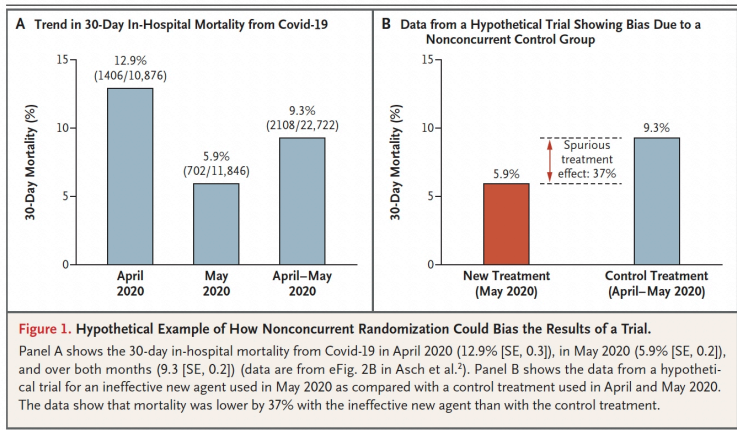
Considerations for Platform Trials

Gold et al. (2022); Berry et al. (2015)

- ▶ Attractive in settings where there may be multiple novel candidates, potentially evolving over time
 - COVID 19: Solidarity, Recovery, ACTIV-k
 - Cancer
- ▶ Analytical Downsides:
 - Comparisons are little tricky between active drugs: Unless all patients eligible for all components, not a randomized comparison if lumping all exposed patients
 - If adding interventions over time, need to worry about issue of non-concurrent controls
- ▶ Upsides
 - Can take more patients into trial.
 - Efficient infrastructure

The Danger of Non-concurrent Controls

Dodd et al. (2021)



What is Adaptive Randomization?

- ▶ All previously discussed methods of randomization were examples of **fixed allocation** schemes
 - Order of treatment assignments can be completely determined in advance of the trial
- ▶ **Adaptive randomization** schemes “adapt” or change according to characteristics of subjects enrolled in trial
 - Sequence of treatment assignments cannot be determined in advance
 - Probability of assigning a new participant a particular treatment can change over time
 - Two major classes: adaptive with respect to baseline characteristics or with respect to patient outcomes
- ▶ Note not all adaptive trials involve adaptive randomization, namely group sequential trials

Baseline Adaptive Schemes (1)

- ▶ **Biased coin randomization**: allocates treatment for the next participant with a probability that depends on current balance between arms
 - Introduced by Brad Efron, suggested $p=2/3$ for the arm with fewer participants
 - **Benefits**: low probability of long runs, maintaining simple coin-flip type randomization, avoids the potential unmasking problems of permuted blocks
 - **Con**: statistical analysis less straight forward. Familiar tests lose their asymptotic normality and exact inference is recommended (Markaryan and Rosenberger, 2010)

Baseline Adaptive Schemes (2)

- ▶ **Dynamic allocation** algorithms based on maintaining balance across multiple important prognostic variables
 - Develop an index of imbalance across multiple baseline covariates
 - **Minimization**: next treatment assignment minimizes current imbalance
 - Other dynamic allocation schemes give the treatment which minimizes the imbalance a higher probability of assignment
 - **Benefits**: can maintain balance across several prognostic variables, without worrying about lots of incomplete blocks. Maintains balance better than stratified permuted block, particularly in small trials and/or many covariates
 - **Cons**: statistical analysis less straight forward, easy to screw up, hard to document. Classic problem: what happens if find an error in allocation or participant's data

Eye-Opening Experience for Minimization

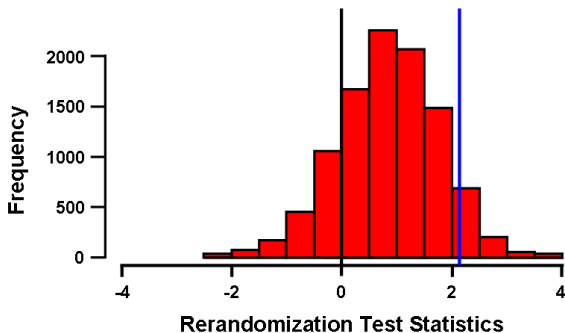
- ▶ Genzyme conducted Late Onset Treatment Study (LOTS)
 - 90 patients with late onset Pompe' disease
 - Primary outcome: 6 minute walk test
 - 2:1 allocation to drug/placebo using minimization
 - ▶ Site
 - ▶ BL 6 minute walk ($\leq 300\text{m}$, $>300\text{m}$)
 - ▶ Forced vital capacity ($\leq 55\%$ pred., $>55\%$ pred.)
 - One analysis requested by FDA: re-randomization test

Eye-Opening Experience for Minimization

- ▶ At the time, FDA was skeptical about minimization, so they require companies to use a re-randomization test
- ▶ Proponents of minimization argue that you can do a re-randomization test, but it is unnecessary because you get about same answer as t-test
- ▶ Wrong!

Problem with Application of Rerandomization Test in Analysis of 6MWT

- Distribution of 6MWT ANCOVA test statistics



ANCOVA $p = 0.035$

Rerandomization $p = 0.06$

Eye-Opening Experience for Minimization

- ▶ The problem is that minimization severely limits amount of randomization
- ▶ The particular randomization scheme for unequal allocation was flawed (see Kuznetsova and Tymofyeyev (2012) for how to fix it)

For the statistical geeks:

- ▶ Big problem: mean of re-randomization distribution is NOT 0
 - It is 0 for standard randomization methods
- ▶ Nonzero mean causes loss of efficiency of re-randomization test: no longer close to t-test even for very large sample sizes

Eye-Opening Experience for Minimization

- ▶ For more details on LOTS trial, see Van der Ploeg et al (2010) NEJM 362, 1396-1406
- ▶ For more details about statistical problems minimization caused see Proschan et al. (2011), and for how to fix them, see Kuznetsova and Tymofyeyev (2012)
- ▶ For more details about mathematics of randomization see: Rosenberger, W. F., and Lachin, J. M. (2015). Randomization in clinical trials: theory and practice. John Wiley & Sons.

Response Adaptive Schemes

- ▶ **Response adaptive allocation**: responses of participants enrolled to date are taken into account when randomizing next participant
 - Relies on assumption that response to treatment can be assessed fairly quickly and cohort is not changing over time
- ▶ **Zelen's Play the Winner**: assigns same treatment if previous patient a success and the other treatment if otherwise
- ▶ **Randomized Play the Winner**: gives more successful treatment a higher chance of allocation (but $p < 1$)
- ▶ Many other methods, including Bayesian approaches

Goals of Response Adaptive Schemes can vary

- ▶ Some may seek to relate probability of the treatment arm with probability of a positive response
 - Lots of algorithms. Methods vary whether this may be deterministic, probabilistic
- ▶ Some response adaptive schemes may target increasing power

ECMO Trial: A Cautionary Tale for Response Adaptive Allocation

- ▶ **ECMO Trial**: study of extracorporeal membrane oxygenator in newborns suffering from respiratory failure
- ▶ Play the winner type algorithm used for treatment allocation
- ▶ First baby randomly assigned to active arm and was a success; 2nd baby randomized to control and died; next 9? babies assigned to experimental ECMO arm and survived
- ▶ Trial stopped after 2 more babies non-randomly assigned ECMO
- ▶ By chance, control baby was the sickest
- ▶ After much controversy, a second trial was launched
- ▶ More controversy and debates over methodology and ethics

Challenges of Response Adaptive Schemes:

Analytical properties are hard to decipher

- ▶ **Some argue these trials are more ethical**, because they aim to maximize number of people on the better treatment
 - There have been statistical efficiency claims, but actually now shown to be false
- ▶ **Adaptive allocation designs are difficult** to implement without mistakes or problems with blinding
- ▶ Inference for response-adaptive randomization is very complicated because both the treatment assignment and responses are correlated (Rosenberger and Lachin, 2015)
- ▶ Analytical properties are not well-established, especially of new designs
- ▶ **Advice**: These methods are controversial and prone to problems, avoid unless you are an expert and willing to repeat your trial

Lessons from ECMO If you must use RAR

Proschan and Evans (2020); Chandereng and Chappell (2020)

- ▶ Randomization should have fairly long run in of standard randomization before you start the adaptive allocations
- ▶ In multi-arm trial, you can change assigned probability of being assigned to a component but you keep the probability of being assigned to the control the same
- ▶ In trial analysis, need to have methods to adjust for time trends
 - Essentially doing a stratified analysis within time buckets

What Randomization scheme is best?

- ▶ Depends on the study and resources available
 - Currently likely never to recommend response-adaptive
 - Best scheme likely dictated by what is practical given resources, including programming resources and other infrastructure
- ▶ Keep it simple
 - Simple randomization: hard to mess up, large trials will be balanced
 - Permuted block randomization: simple, widely used, widely understood
 - Stratified by site: common choice for multi-site trials
 - Choose block size(s) appropriate to sample size
 - Randomize at last possible second

Maintaining Randomization Integrity

- ▶ Fundamental motivation of randomization: create comparable treatment groups
 - Allows causality inference
- ▶ To maintain comparability, primary analysis is an intent-to-treat (ITT) analysis
 - All subjects are analyzed according to randomization assignment, regardless of what treatment they actually get

Flavors of ITT

- ▶ **ITT** analysis
 - Analyze according to the study regimen assigned
 - Requires models to weight observed or impute missing outcomes, requires sensitivity analysis
 - **Only analysis which preserves randomization**
- ▶ **Modified ITT (MITT)** analysis
 - ITT, but only include people who take the first dosage
 - In well-implemented trials few people drop out before first dose
 - Potentially minor departure from ITT if blinded

Analysis Choices (2)

▶ **Per Protocol** Analysis:

- Analysis includes data only from completers/adherers
- Subject to bias, analyzes only the well behaved and potentially only the healthiest participants (no adverse events)
- Especially problematic when drop out rates different by treatment arm

Threats to Randomization Integrity

- ▶ Improper masking or blinding
 - Bias will creep into data
- ▶ Excluding subjects who withdraw from treatment: can lead to bias
- ▶ Drop-out/missing data: breaks randomization without ITT; weakens treatment result with ITT
 - Long trials may need to have a screening period to assess commitment of subjects before randomization

“Analyze as you randomize”

- ▶ Analysis of study results generally should take into account the method of randomization
 - Adjusting for stratification is recommended (to avoid overly wide confidence intervals)
 - Adaptive procedures need to be accounted for
 - Ignoring “blocks” is standard and generally considered okay

Summary

- ▶ Permuted block randomization often the best
 - Stratify on only a few factors, usually one or two
 - Choose block size(s) appropriate to sample size
- ▶ Randomize smallest independent element at last possible second
- ▶ Masking/blinding is key for preventing bias
- ▶ ITT (intent to treat) analysis necessary to preserve randomization and infer causality, or lack thereof
- ▶ Proper documentation as important as proper implementation

Conclusion

- ▶ **Randomized Studies** are the Gold Standard of Clinical Research
- ▶ Randomization to treatments separates clinical trials from all other studies; don't muck it up!
- ▶ Randomization
 - Eliminates selection bias
 - **Forms basis for statistical tests**
 - Balances arms with respect to prognostic variables (known and unknown)

Lecture 3: Sample Size/Power

Outline

- ▶ Introduction to Power/Sample Size
- ▶ Introduction to EZ Principle
- ▶ Where Does the Key Formula Come From?
- ▶ General EZ Principle and Applications
 - ▶ T-test
 - ▶ Test of Proportions
 - ▶ Survival
 - ▶ Noninferiority
 - ▶ Lack of Reproducibility
- ▶ Sample Size: Practical Aspects
 - ▶ Treatment Effect
 - ▶ Nuisance Parameters
- ▶ Sample Size: Estimation
- ▶ Sample Size: Safety

Introduction to Power/Sample Size

Introduction to Power/Sample Size

- ▶ Clinical trials are the gold standard of evidence.
- ▶ Clinical trials use hypothesis testing to choose between null and alternative hypotheses:

H_0 : treatment has no effect

H_1 : treatment has an effect.

- ▶ We hope to reject H_0 and conclude that treatment works.
- ▶ α , the probability of falsely rejecting H_0 (making a **type 1 error**) is set low to avoid approving an ineffective treatment.
- ▶ If H_0 is rejected, there is strong evidence against the null hypothesis and in favor of treatment benefit.

Introduction to Power/Sample Size

- ▶ But abandoning an effective treatment by failing to reject H_0 when H_1 is true (making a **type 2 error**) is also a serious error.
- ▶ To be confident we are not making a type 2 error, we should make β , the probability of a type 2 error, low.
- ▶ Equivalently, we should make **power**, namely $1 - \beta = P(\text{rejecting } H_0 \text{ when } H_1 \text{ is true})$, high.
- ▶ If power is high and we still do not reject H_0 , treatment probably did not have its intended effect.

Introduction to Power/Sample Size

- ▶ See chapter 8 of Proschan (2022).
- ▶ Standardized test statistics (z-scores) in clinical trials are often:
 - ▶ Of form $Z = \hat{\delta}/\text{se}(\hat{\delta})$, where $\hat{\delta}$ is a treatment effect estimator.
 - ▶ Approximately $N(\theta, 1)$ for large sample sizes, where $\theta = 0$ under H_0 .
- ▶ Examples:
 - ▶ T-statistic: $\hat{\delta} = \bar{Y}_T - \bar{Y}_C$; $\text{se}(\hat{\delta}) = \sqrt{2\sigma^2/n}$.
 - ▶ Z-score for proportions: $\hat{\delta} = \hat{p}_T - \hat{p}_C$; $\text{se}(\hat{\delta}) \approx \sqrt{2p(1-p)/n}$.
 - ▶ Z-score for logrank statistic: $\hat{\delta} = \sum(O_i - E_i)/\sum V_i$ estimates log hazard ratio; $\text{se}(\hat{\delta}) = 1/\sum V_i$, where O_i, E_i, V_i are observed, expected, and variance from 2×2 table at each event.
 - ▶ Z-scores for maximum likelihood estimators (MLEs), minimum variance unbiased estimators, Cox models, etc.

Introduction to EZ Principle

Introduction to EZ Principle

- ▶ There is really only one power/sample size formula.
- ▶ EZ principle (its easy!): Power depends on $E(Z)$, the expected z-score.
- ▶ Parameterize so that large z-scores mean treatment is beneficial. E.g., may need to change $\delta = \mu_T - \mu_C$ to $\delta = \mu_C - \mu_T$.
- ▶ For a 2-sided test at $\alpha = 0.05$ (or a 1-sided test at $\alpha = 0.025$), $E(Z)$ must be:
 - ▶ 3.24 for 90% power.
 - ▶ 3.00 for 85% power.
 - ▶ 2.80 for 80% power.
- ▶ We will see justification after some examples.

Introduction to EZ Principle

- ▶ Makes checking sample size calculations quick and easy.
- ▶ Continuous outcome example: You compare a new treatment to standard treatment for hepatitis C virus.
 - ▶ Primary outcome: log viral load at end of study minus log viral load at baseline. Use t-test.
 - ▶ Want 80% power for difference $\delta = 0.5$ and you expect $\sigma = 1.25$.
 - ▶ Investigator says you need 50/arm. Is that correct?
- ▶ $Z = \hat{\delta} / \sqrt{2\sigma^2/n}$, $\hat{\delta} = \bar{Y}_C - \bar{Y}_T$ (treatment beneficial if > 0).
- ▶ Expected z-score is

$$E(Z) = \frac{\mu_C - \mu_T}{\sqrt{2\sigma^2/n}} = \frac{0.5}{\sqrt{2(1.25)^2/50}} = 2.$$

- ▶ $E(Z)$ much less than 2.80. Power much less than 0.80.

Introduction to EZ Principle

- ▶ Binary outcome example: New treatment for hospitalized COVID-19 patients on mechanical ventilation/ECMO.
 - ▶ Primary endpoint: 60-day mortality.
 - ▶ Want 85% power to detect improvement in 60-day mortality from 0.20 to 0.12.
 - ▶ Statistician reports you need $n = 1,000$ per arm. Is it correct?

Parameterize so positive values are good: $\delta = p_C - p_T$.

$$Z = \frac{\hat{\delta}}{\text{se}(\hat{\delta})} = \frac{\hat{p}_C - \hat{p}_T}{\sqrt{2p(1-p)/n}}, \quad p = \frac{0.20 + 0.12}{2} = 0.16.$$

- ▶ Expected z-score is:

$$E(Z) = \frac{p_C - p_T}{\sqrt{2p(1-p)/n}} = \frac{0.20 - 0.12}{\sqrt{2(0.16)(1-0.16)/1000}} = 4.880.$$

- ▶ $E(Z)$ much larger than 3. Power much larger than 85%.

Introduction to EZ Principle

- ▶ Before looking at more examples, let's look at the basis for the EZ principle.
- ▶ This will show us a general formula that allows us to compute, for any alpha :
 - ▶ Sample size for a given treatment effect and power.
 - ▶ Power for a given treatment effect and sample size.
 - ▶ Treatment effect for a given sample size and power.

Where Does The Key Formula Come From?

Where Does The Key Formula Come from?

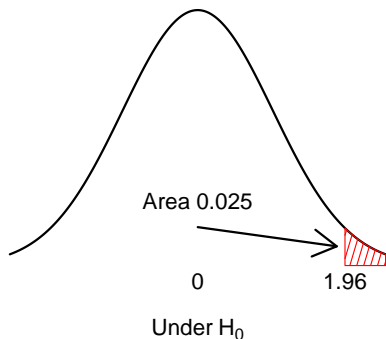


Figure: The standard normal null density for the z-statistic. For a 1-tailed test at $\alpha = 0.025$, we reject H_0 if $Z > 1.96$.

Where Does The Key Formula Come from?

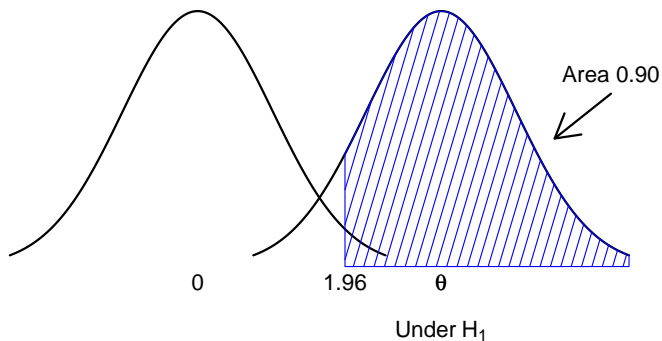


Figure: The alternative $N(\theta, 1)$ density for Z . For power 0.90, we want the blue shaded area to be 0.90.

Where Does The Key Formula Come from?

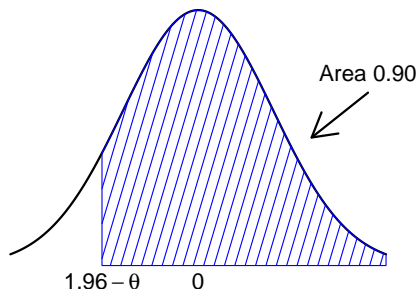


Figure: The blue shaded area in Figure 2 equals the blue shaded area to the right of $1.96 - \theta$ under the standard normal curve. For power 0.90, $1.96 - \theta = -1.28$, so $\theta = 1.96 + 1.28 = 3.24$.

General EZ Principle and Applications

General EZ Principle and Applications

- ▶ Same reasoning applies for different levels of α and β .
- ▶ For 2-tailed test at level α and power $1 - \beta$, set

$$E(Z) = z_{\alpha/2} + z_{\beta}, \text{ (EZ Principle)} \quad (1)$$

where, for $0 < a < 1$, z_a denotes the $(1 - a)$ th quantile of a standard normal distribution.

- ▶ $z_{\alpha/2} = 1.96$ for $\alpha = 0.05$, 2-sided test.
- ▶ $z_{\beta} = 0.84, 1.04$, or 1.28 for $\beta = 0.20, 0.15$, or 0.10 .
- ▶ $z_{\alpha/2} + z_{\beta} = 2.80, 3.00$, or 3.24 for 80%, 85%, or 90% power.
- ▶ 1 formula, for sample size, power, or detectable effect.

General EZ Principle and Applications

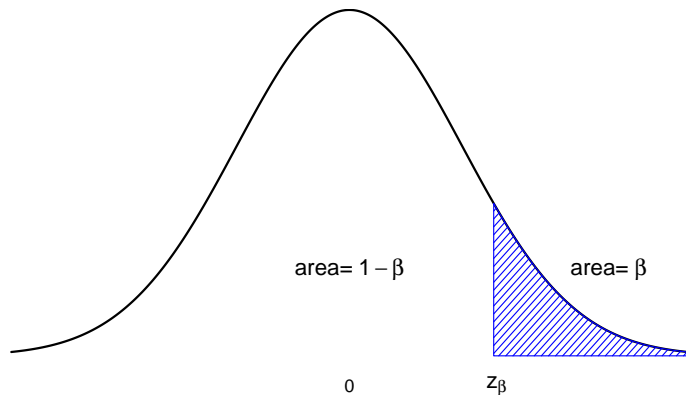


Figure: The area to the right of z_β is β , so the area to the left of z_β is $1 - \beta = \text{power}$.

General EZ Principle and Applications

- ▶ Continuous outcome example: return to hepatitis C (HCV) trial.
 - ▶ Primary outcome: Change in log viral load from baseline. T-test.
 - ▶ Want sample size for 80% power for 2-sided test at $\alpha = 0.05$.
 - ▶ $\delta = 0.5$ and $\sigma = 1.25$.

$$Z = \frac{\hat{\delta}}{se(\hat{\delta})} = \frac{\bar{Y}_C - \bar{Y}_T}{\sqrt{2\sigma^2/n}}; E(Z) = \frac{\delta}{\sqrt{2\sigma^2/n}} = \frac{0.5}{\sqrt{\frac{2(1.25)^2}{n}}}$$

$$E(Z) = z_{\alpha/2} + z_{\beta} \text{ (EZ Principle).}$$

$$\frac{0.5}{\sqrt{\frac{2(1.25)^2}{n}}} = 1.96 + 0.84 = 2.80. \text{ Solve for } n:$$

$$n = \frac{2(1.25)^2(2.80)^2}{0.5^2} = 98.$$

Need 98/arm.

General EZ Principle and Applications

- ▶ Suppose you can only recruit 75/arm. What is the detectable effect (the δ) with 80% power?

$$Z = \frac{\hat{\delta}}{se(\hat{\delta})} = \frac{\bar{Y}_C - \bar{Y}_T}{\sqrt{2\sigma^2/n}}$$

$$E(Z) = \frac{\delta}{\sqrt{\frac{2(1.25)^2}{75}}} = (z_{\alpha/2} + z_{\beta}) \text{ (EZ Principle)}$$

$$\frac{\delta}{\sqrt{\frac{2(1.25)^2}{75}}} = 1.96 + 0.84 = 2.80. \text{ Solve for } \delta :$$

$$\delta = 2.80 \sqrt{\frac{2(1.25)^2}{75}} = 0.57.$$

Detectable effect is 0.57 logs.

General EZ Principle and Applications

- ▶ What is power for detecting 0.5 log if you only recruit 75/arm?

$$Z = \frac{\hat{\delta}}{se(\hat{\delta})} = \frac{\bar{Y}_C - \bar{Y}_T}{\sqrt{2\sigma^2/n}}; \quad E(Z) = \frac{\delta}{\sqrt{2\sigma^2/n}}$$

$$E(Z) = \frac{0.5}{\sqrt{\frac{2(1.25)^2}{75}}} = z_{\alpha/2} + z_{\beta} \quad (\text{EZ Principle})$$

$$2.449 = 1.96 + z_{\beta}$$

$$0.489 = z_{\beta}. \quad \text{Let } \Phi(x) \text{ denote } N(0,1) \text{ distribution function at } x.$$

$$\Phi(0.489) = \Phi(z_{\beta}) = 1 - \beta = \text{power},$$

Power is $\Phi(0.489) \approx 0.69$.

General EZ Principle and Applications

- ▶ Binary outcome example: Return to COVID-19 example:
 - ▶ Primary endpoint: 60-day mortality.
 - ▶ Want 85% power to detect improvement in 60-day mortality from 0.20 to 0.12.

$$Z = \frac{\hat{p}_C - \hat{p}_T}{\sqrt{\frac{2p(1-p)}{n}}}$$

$$E(Z) = \frac{p_C - p_T}{\sqrt{\frac{2p(1-p)}{n}}} = \frac{0.20 - 0.12}{\sqrt{\frac{2(0.16)(1-0.16)}{n}}}$$

$$\frac{0.20 - 0.12}{\sqrt{\frac{2(0.16)(1-0.16)}{n}}} = z_{\alpha/2} + z_{\beta} = 3 \text{ for 85\% power (EZ Principle).}$$

$$n \approx \frac{2(0.16)(0.84)(3)^2}{(0.08)^2} = 378. \quad (2)$$

Need 378/arm.

General EZ Principle and Applications

- ▶ Exercise: Show that if you can only get 300/arm, power is approximately 0.76.

General EZ Principle and Applications

- ▶ Schoenfeld (1981) derives sample size for survival tests.

Table: Table at i th death.

	Dead	Alive	
Control	O_i		n_{Ci}
Treatment			n_{Ti}
	1	$n_i - 1$	n_i

n_{Ci}, n_{Ti} = numbers at risk in treatment, control just prior to i th death.

O_i = indicator that i th death is in control arm.

Under H_0 , no difference in survival, the i th death is equally likely to be from any of the $n_i = n_{Ci} + n_{Ti}$ people at risk.

O_i is Bernoulli with parameter $p_i = n_{Ci}/n_i$ under H_0 .

General EZ Principle and Applications

Table: Table at i th death.

	Dead	Alive	
Control	O_i		n_{Ci}
Treatment			n_{Ti}
	1	$n_i - 1$	n_i

n_{Ci}, n_{Ti} = numbers at risk in treatment, control just prior to i th death.

O_i = indicator that i th death is in control arm.

E_i = null expected value of O_i , given marginal totals: $E_i = p_i = n_{Ci}/n_i$.

V_i = null variance of O_i , given marginal totals: $V_i = p_i(1 - p_i) = \frac{n_{Ci}n_{Ti}}{n_i^2}$.

General EZ Principle and Applications

- ▶ FUN FACT: “O minus E except after V principle” (justified in Section 7.7 of Proschan (2022)): $\hat{\delta}_i = (O_i - E_i)/V_i$ is crude estimate of log hazard ratio with variance (conditioned on marginal totals) $1/V_i$. Combine these to get a better estimate of log hazard ratio.
- ▶ Optimal weighted average of the $\hat{\delta}_i$ weights inversely proportional to variances, $w_i = 1/\text{var}(\hat{\delta}_i) = V_i$.

$$\begin{aligned}\hat{\delta} &= \frac{\sum_{i=1}^d w_i \hat{\delta}_i}{\sum_{i=1}^d w_i} = \frac{\sum_{i=1}^d V_i \{(O_i - E_i)/V_i\}}{\sum_{i=1}^d V_i} \\ &= \frac{\sum_{i=1}^d (O_i - E_i)}{\sum_{i=1}^d V_i}.\end{aligned}\tag{3}$$

- ▶ $\hat{\delta}$ estimates log hazard ratio and $\text{var}(\hat{\delta} \mid \text{marginals}) = 1/\sum_{i=1}^d V_i$.

General EZ Principle and Applications

Logrank z-statistic is

$$\begin{aligned} Z &= \frac{\hat{\delta}}{\sqrt{\text{var}(\hat{\delta})}} = \frac{\sum_{i=1}^d (O_i - E_i)}{\sqrt{\sum_{i=1}^d V_i}} = \left(\frac{\sum_{i=1}^d (O_i - E_i)}{\sum_{i=1}^d V_i} \right) \sqrt{\sum_{i=1}^d V_i} \\ &= \hat{\delta} \sqrt{\sum_{i=1}^d V_i}, \end{aligned} \quad (4)$$

With 1-1 randomization, $V_i \approx (1/2)(1 - 1/2) = 1/4$, so

$\sum_{i=1}^d V_i \approx \sum_{i=1}^d (1/4) = d/4$ and

$$\begin{aligned} Z &\approx \hat{\delta} \sqrt{d/4} \\ E(Z) &\approx \delta \sqrt{d/4}, \end{aligned} \quad (5)$$

d = number of deaths, $\hat{\delta}$, δ are estimated and true log hazard ratios.

General EZ Principle and Applications

- ▶ For power $1 - \beta$, equate $E(Z)$ to $z_{\alpha/2} + z_{\beta}$ and solve for number of deaths (events) d :

$$E(Z) = \delta \sqrt{d/4} = (z_{\alpha/2} + z_{\beta}) \text{ (EZ Principle)}$$

$$d = \frac{4(z_{\alpha/2} + z_{\beta})^2}{\delta^2}, \quad (6)$$

δ = log hazard ratio (parameterized so that large hazard ratios show that treatment works).

- ▶ Continue the trial until this number of deaths.

General EZ Principle and Applications

- ▶ Example: return to COVID-19 trial, but suppose you use logrank test instead of test of proportions.
- ▶ Want 85% power to detect a control-to-treatment hazard ratio of 1.333. Set

$$Z = \frac{\sum_{i=1}^d (O_i - E_i)}{\sqrt{\sum_{i=1}^d V_i}} = \left(\frac{\sum_{i=1}^d (O_i - E_i)}{\sum_{i=1}^d V_i} \right) \sqrt{\sum_{i=1}^d V_i} \approx \hat{\delta} \sqrt{d/4}.$$

$$E(Z) = \delta \sqrt{d/4} = (z_{\alpha/2} + z_{\beta}) = (1.96 + 1.04) = 3 \text{ (EZ Principle)}$$

$$d = \frac{4(3)^2}{\delta^2 \{\ln(1.333)\}^2} \approx 436 \text{ events.} \quad (7)$$

General EZ Principle and Applications

- ▶ Suppose you get only 350 events (deaths).
- ▶ For power to detect hazard ratio of 1.333,

$$E(Z) = \ln(1.333)\sqrt{350/4} = 2.689$$

$$2.689 = (z_{\alpha/2} + z_{\beta}) = 1.96 + z_{\beta} \text{ (EZ Principle)}$$

$$2.689 - 1.96 = z_{\beta}$$

$$\Phi(2.689 - 1.96) = \Phi(z_{\beta}) = 1 - \beta = \text{power}. \quad (8)$$

Power is approximately $\Phi(0.729) = \text{pnorm}(0.729) = 0.77$.

General EZ Principle and Applications

- ▶ Exercise: Suppose you get only 350 events (deaths). What hazard ratio can be detected with 85% power?,
- ▶ Answer: Control to treatment hazard ratio is approximately 1.378.

General EZ Principle and Applications

- ▶ In noninferiority trials, not trying to prove new treatment (N) is better than standard treatment (S), but that it is almost as good.
- ▶ One application of NI testing: Standard treatment might be onerous (3 injections/day) and new treatment is easier to take (1 pill/day).
- ▶ Prefer new treatment provided it is not worse than standard by more than some small amount known as the ***noninferiority (NI) margin***.
- ▶ Let p_N and p_S be probability of event on new and standard treatment.

General EZ Principle and Applications

- ▶ NI trials often use **1-sided** $\alpha = 0.05$ and test nonzero null.
- ▶ E.g., if willing to tolerate new treatment being worse than standard by 0.10 (NI margin=0.10), test:

$$H_0 : p_S - p_N < -0.10 \text{ versus } H_1 : p_S - p_N \geq -0.10.$$

- ▶ Convert to zero-null by:

$$H_0 : p_S - p_N + 0.10 < 0 \text{ versus } H_1 : p_S - p_N + 0.10 \geq 0.$$

- ▶ Suppose we want 90% probability of showing noninferiority if truth is that $p_N = p_S$.
- ▶ Again use EZ principle:

General EZ Principle and Applications

$$Z = \frac{\hat{p}_S - \hat{p}_N + 0.10}{\sqrt{\frac{\hat{p}_S(1-\hat{p}_S) + \hat{p}_N(1-\hat{p}_N)}{n}}} \quad (9)$$

If $p_S = p_N$,

$$E(Z) \approx \frac{0.10}{\sqrt{\frac{2p(1-p)}{n}}} = (z_{\alpha/2} + z_{\beta}) \text{ (EZ Principle)}$$

$$\frac{0.10}{\sqrt{\frac{2p(1-p)}{n}}} = 1.645 + 1.282 = 2.927.$$

$$n = 2p(1-p)(2.927)^2 / (0.1)^2.$$

If p is expected to be 0.5, $n = 429$ per arm.

General EZ Principle and Applications

- ▶ One important implication of EZ principle: Inability to replicate results (see, e.g, Goodman (1992); Halsey et al. (2015)).
- ▶ Suppose $Z = 1.96$, and you believe this reflects the truth; i.e., you believe that $E(Z) = 1.96$.
- ▶ If we repeat trial, can compute power using EZ principle:

$$\begin{aligned}E(Z) &= 1.96 + z_\beta \\1.96 &= 1.96 + z_\beta \\0 &= z_\beta \\ \Phi(0) &= \Phi(z_\beta) = 1 - \beta = \text{power} \\0.5 &= \text{power}\end{aligned}\tag{10}$$

Only a 50% chance of replicating the result!

General EZ Principle and Applications

- ▶ Why?
- ▶ $Z = 1.96$ is much smaller than its expected value of 3.24 for 90% power.
- ▶ **Lesson: Can reach statistical significance even though observed treatment effect is much smaller than expected.**
 - ▶ **Stresses need to supplement hypothesis test with treatment effect estimate and confidence interval**
- ▶ Not surprisingly, if true treatment effect is much smaller than expected, may not reproduce statistically significant result.

Sample Size: Practical Aspects

Sample Size: Practical Aspects

- ▶ Sample size depends on treatment effect and nuisance parameters.
 - ▶ Nuisance for t-test: σ .
 - ▶ Nuisance for test of proportions: p_C (or overall p).
- ▶ The treatment effect and nuisance parameter are very different.
 - ▶ We can **specify** treatment effect either as minimal relevant effect or the anticipated effect based on other studies.
 - ▶ We must **estimate** the nuisance parameter accurately for power calculations.
- ▶ **Underestimating** σ in a t-test or **overestimating** p_C in a test of proportions to detect a given **relative effect** (e.g., 25%) will result in an underpowered trial.

Sample Size: Practical Aspects: Treatment Effect

- ▶ If treatment has many side effects or is difficult (e.g., several injections a day), then treatment effect should be large to justify its use.
- ▶ If treatment has few side effects (e.g., a diet), even a small effect is worthwhile.
- ▶ Dietary Approaches to Stop Hypertension (DASH) trial (Appel et al. (1997))
 - ▶ Compared 3 diets: (1) control, (2) fruits & vegetables, (3) combination fruits and vegetables and lowfat dairy.
 - ▶ Primary endpoint: change in diastolic blood pressure from baseline.
 - ▶ Powered for 2mmHg difference because even a small effect has public health benefit with few expected side effects.

Sample Size: Practical Aspects: Treatment Effect

- ▶ In early phase trials, type 2 may be more serious than type 1 error.
 - ▶ Type 2 error ends further testing– may be abandoning a good drug.
 - ▶ Type 1 error is not tragic because definitive test is in phase 3.
- ▶ Therefore, want to ensure high power in early phase trials.
- ▶ Stack deck in your favor by:
 - ▶ Picking population especially expected to benefit (might use a *run-in* phase before randomization to weed out people who cannot tolerate drug)
 - ▶ Using an intermediate outcome that treatment should affect.

Sample Size: Practical Aspects: Nuisance Parameters

- ▶ With t-test, power depends on standard deviation, σ .
- ▶ Err on side of overestimating σ to avoid underpowered trial.
- ▶ Use estimates based on similar trials, if possible.
- ▶ If σ estimated from observational study, **increase it**.
- ▶ When standard deviation is of a change from baseline, use a trial with a similar or longer duration (standard deviation of a change usually increases with duration)

Sample Size: Practical Aspects: Nuisance Parameters

- ▶ Useful formula for variance of change from baseline (BL) to end of study (EOS):

$$\text{var}(Y_{\text{EOS}} - Y_{\text{BL}}) = 2\sigma^2(1 - \rho), \text{ where}$$

σ^2 is variance at fixed time (BL or EOS) and $\rho = \text{cor}(Y_{\text{BL}}, Y_{\text{EOS}})$.

- ▶ E.g., if variance at single time is 65, and $\rho = 0.80$, use

$$\text{var}(Y_{\text{EOS}} - Y_{\text{BL}}) = 2(65)(1 - 0.80) = 26.$$

- ▶ NOTE: If $\rho < 0.5$, then you should use Y_{EOS} , **NOT** $Y_{\text{EOS}} - Y_{\text{BL}}$. Even better, use baseline value as covariate (also called analysis of covariance—ANCOVA).
- ▶ Err on side of **underestimating** ρ to avoid underpowered trial.

Sample Size: Practical Aspects: Nuisance Parameters

- ▶ With binary endpoint, nuisance parameter is control event probability, p_C .
- ▶ If treatment effect is a **relative reduction** (e.g., 25%, so $RR=0.75$), err on side of **underestimating** p_C to avoid an underpowered trial.
- ▶ If estimate of p_C comes from observational trial, **decrease** it! Clinical trial participants tend to be more health conscious & have lower event rates (healthy volunteer effect).

Sample Size: Practical Aspects: Nuisance Parameters

- ▶ Sample size is often a negotiation between principal investigator and statistician.
- ▶ Statistician: “You will need 10,000 people”.
- ▶ Options:
 - ▶ Increase treatment effect. PI: “A larger effect is unrealistic.”
 - ▶ Use a different primary endpoint. E.g., add stroke to composite of coronary heart disease/death. Statistician: “That will work as long as treatment has a similar effect on added component. Otherwise, you could **decrease** power.”

Sample Size: Estimation

Sample Size: Estimation

- ▶ In early phase, may do 1-arm trial to get a reasonable estimate of effect. Set sample size to achieve given accuracy.
- ▶ Example: How large does n need to be to estimate the proportion of successes on new treatment to within 0.15?
- ▶ 95% confidence interval: $\hat{p} \pm 1.96\sqrt{p(1-p)/n}$
- ▶ Set

$$1.96\sqrt{p(1-p)/n} = 0.15$$

and solve for n :

Sample Size: Estimation

$$n = \frac{(1.96)^2 p(1-p)}{(0.15)^2} = 341.4756p(1-p). \quad (11)$$

- ▶ If we expect $p = 0.3$, substitute 0.3 into Equation (11) to get $n = 72$.
- ▶ Could also use $p = 0.5$ as a worst case scenario: Substituting 0.5 into (11) gives $n = 86$.

Sample Size: Safety

Sample Size: Safety

- ▶ In safety studies, want to calculate probability of seeing at least one adverse event (AE) of a given probability.

$P(\text{see at least one AE of probability } p)$

$$\begin{aligned} &= 1 - P(0 \text{ AEs of probability } p) \\ &= 1 - (1 - p)^n. \end{aligned} \tag{12}$$

- ▶ Example: in a study of 20 people, the probability of at least one AE of probability 0.10 is $1 - (1 - 0.10)^{20} = 0.88$.
- ▶ Confident we will see at least one if true probability is 0.10.

Summary

- ▶ High power is essential for avoiding type 2 errors.
- ▶ EZ principle: You need only 1 formula for sample size/power/detectable effect for 2-sided test at level α (or 1-sided at $\alpha/2$). For power $1 - \beta$, set

$$E(Z) = z_{\alpha/2} + z_{\beta}.$$

- ▶ For 2-sided $\alpha = 0.05$, $E(Z)$ must be
 - ▶ 3.24 for 90% power.
 - ▶ 3.00 for 85% power.
 - ▶ 2.80 for 80% power.
- ▶ Can apply to any statistic that is asymptotically $N(\theta, 1)$.

Summary

- ▶ Sample size depends on treatment effect and nuisance parameters.
- ▶ Nuisance parameters are:
 - ▶ σ for continuous outcome using t-test.
 - ▶ p_C or overall p for binary outcome.
- ▶ For t-test, err on side of overestimating σ .
- ▶ For binary outcome when trying to detect a given **relative** treatment effect (i.e., 25% reduction in event rate), err on side of underestimating p_C .

Lecture 4: Interim Monitoring: Efficacy

Outline

- ▶ Introduction to Efficacy Monitoring
- ▶ Historical Monitoring Boundaries
 - ▶ Haybittle-Peto
 - ▶ Pocock
 - ▶ O'Brien-Fleming
- ▶ Unified Approach: Information, Z-Scores, B-Values
- ▶ Alpha Spending Functions
 - ▶ Definition
 - ▶ O'Brien-Fleming-Like Spending Function
 - ▶ Pocock-Like Spending Function
 - ▶ Families of Spending Functions
- ▶ Effect of Efficacy Monitoring on Power
- ▶ Case Study: CAST

Introduction to Efficacy Monitoring

Introduction to Efficacy Monitoring

- ▶ The ACTG 076 trial in France and the U.S. (Connor et al. (1994)) compared AZT to placebo to prevent mother to infant transmission of HIV.
- ▶ Primary endpoint: HIV in infant. Planned 636 mother/infant pairs.
- ▶ After 363 live births with known HIV status:
 1. 13 AZT infants infected.
 2. 40 placebo infants infected.
- ▶ $Z = 4.03$. Enough evidence, or could this be the play of chance?
- ▶ Who decides and how?
- ▶ Must consider welfare of trial participants and whether results will change clinical practice.

Introduction to Efficacy Monitoring

- ▶ Clinical trials are monitored by a **Data and Safety Monitoring Board (DSMB)** (also called a DSMC or DMC).
- ▶ Committee of 3-9 EXTERNAL experts (MDs, 1-2 statisticians, an ethicist). Keeps study team blinded to results.
- ▶ Typically meet 1-2 times a year.
- ▶ Review general trial conduct (accrual, data quality, missing data, etc.), safety (serious adverse events, unexpected events, etc.), futility, and efficacy.
- ▶ Make recommendations to trial sponsor, sponsor makes final decision (but almost always accepts recommendations).

Introduction to Efficacy Monitoring

- ▶ Futility: Are data so unpromising or is trial conduct so poor that a null result is almost assured?
- ▶ Efficacy: If one arm is clearly superior, may stop trial or recommend change (e.g., announce result, make the superior treatment the new control, etc.).
- ▶ Problem: If we reject H_0 whenever **nominal p-value** (not adjusted for monitoring) is $\leq \alpha$, the familywise 1 error rate (probability of rejecting H_0 at some point) is inflated.
- ▶ Even with only 1 interim and 1 final analysis,
 $P(\text{rejecting } H_0 | H_0) = 0.083$ for 2-sided test at $\alpha = 0.05$ if looks are equally-spaced.
- ▶ Armitage et al. (1969) showed inflation of type 1 error rate.

Introduction to Efficacy Monitoring

- ▶ Situation can be worse if looks are not equally-spaced.
- ▶ Example: Suppose looks are after 10 and 10,000 observations.
 - ▶ The 2 p-values are nearly independent because the overlap is only 10 out of 10,000 people.
 - ▶ Independence is the worst case; the type 1 error rate is

$$\begin{aligned}P(\text{Reject } H_0 | H_0) &= P(P_1 \leq 0.05 \cup P_2 \leq 0.05) \\&= 1 - P(P_1 > 0.05 \cap P_2 > 0.05) \\&\approx 1 - P(P_1 > 0.05)P(P_2 > 0.05) \\&= 1 - (1 - 0.05)^2 = 0.0975. \quad (13)\end{aligned}$$

Introduction to Efficacy Monitoring

- ▶ Situation can be worse if looks are not equally-spaced.
- ▶ Example: Suppose looks are after 10 and 10,000 observations.
 - ▶ The 2 p-values are nearly independent because the overlap is only 10 out of 10,000 people.
 - ▶ Independence is the worst case; the type 1 error rate is

$$\begin{aligned}P(\text{Reject } H_0 | H_0) &= P(P_1 \leq 0.05 \cup P_2 \leq 0.05) \\&= 1 - P(P_1 > 0.05 \cap P_2 > 0.05) \\&\approx 1 - P(P_1 > 0.05)P(P_2 > 0.05) \\&= 1 - (1 - 0.05)^2 = 0.0975. \quad (13)\end{aligned}$$

Introduction to Efficacy Monitoring

- ▶ Situation can be worse if looks are not equally-spaced.
- ▶ Example: Suppose looks are after 10 and 10,000 observations.
 - ▶ The 2 p-values are nearly independent because the overlap is only 10 out of 10,000 people.
 - ▶ Independence is the worst case; the type 1 error rate is

$$\begin{aligned}P(\text{Reject } H_0 | H_0) &= P(P_1 \leq 0.05 \cup P_2 \leq 0.05) \\&= 1 - P(P_1 > 0.05 \cap P_2 > 0.05) \\&\approx 1 - P(P_1 > 0.05)P(P_2 > 0.05) \\&= 1 - (1 - 0.05)^2 = 0.0975. \quad (13)\end{aligned}$$

Introduction to Efficacy Monitoring

Table: Type 1 error rate for unadjusted monitoring for 2-sided test at $\alpha = 0.01$ or $\alpha = 0.05$. Note: $2p \leq 0.05$ means 2-sided p-value ≤ 0.05 .

# Looks (k)	Reject H_0 if $2p \leq 0.01$		Reject H_0 if $2p \leq 0.05$	
	Equally Spaced	Worst Case	Equally Spaced	Worst Case
2	0.018	0.020	0.083	0.098
3	0.024	0.030	0.107	0.143
4	0.029	0.039	0.126	0.185
5	0.033	0.049	0.142	0.226
10	0.047	0.096	0.193	0.401
20	0.064	0.182	0.248	0.642
∞	1	1	1	1

- ▶ This table applies to many different tests: t-test, test of proportions, logrank test, Cox model, etc.

Introduction to Efficacy Monitoring

- ▶ Because efficacy (upper) boundary could differ from “harm” (lower) boundary, we focus **for the rest of this lecture on 1-sided efficacy (upper) boundaries.**
- ▶ For symmetric, 2-sided z-score boundaries at level α , use $\pm c_i$, where c_i is 1-sided boundary at level $\alpha/2$.
 - ▶ This is slightly conservative. Actual 2-sided error rate is infinitesimally less than α .

Historical Monitoring Boundaries

Historical Efficacy Boundaries: Haybittle-Peto

- ▶ The earliest boundary was the Haybittle-Peto boundary (Haybittle (1971)).
- ▶ Original suggestion used a very large z-statistic boundary (3) for interim looks, and 1.96 for final look.
- ▶ Haybittle-Peto was modified using the Bonferroni inequality:
 - ▶ Use p-value threshold 0.001 at interim looks.
 - ▶ Use p-value threshold $\alpha - (k - 1)(0.001)$ at final look.
 - ▶ E.g., with 3 interim and 1 final analysis, reject at interim if $p \leq 0.001$, and at end if $p \leq 0.025 - 3(0.001) = 0.022$.
- ▶ By Bonferroni, the type 1 error rate, $P(\text{reject } H_0 \text{ sometime})$, is
$$\leq 0.001 + 0.001 + \dots + 0.001 + \alpha - (k - 1)(0.001) = \alpha.$$

Historical Efficacy Boundaries: Haybittle-Peto

- ▶ Desirable properties of Haybittle-Peto.
 - ▶ Simple to implement.
 - ▶ Can use regardless of timing of analyses.
 - ▶ Valid for **any** test statistic (don't need to know joint distribution of test statistic over time because Bonferroni inequality is used).
 - ▶ Final z-statistic boundary is close to what it would be with no monitoring (for a reasonable number of analyses).
- ▶ Undesirable property of Haybittle-Peto: Reversal of fortune
 - ▶ Z-statistic boundary drops drastically at the end, causing a logical inconsistency: Could be under boundary at penultimate look, see a partial reversal, and be over boundary at end. How could you be convinced now that you've seen a partial reversal?

Historical Efficacy Boundaries: Haybittle-Peto

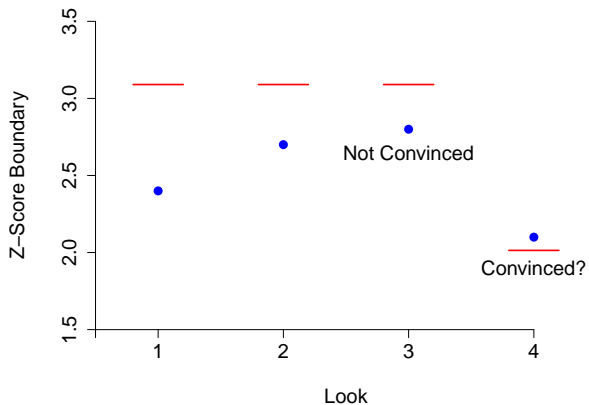


Figure: Reversal of fortune problem with Haybittle-Peto.

Historical Efficacy Boundaries: Pocock

- ▶ Pocock (1977) raised the z-statistic boundary by the same amount for each look.
- ▶ Pre-specify number of looks, k , and **assume they are equally spaced.**
- ▶ Use z-score boundary $c = c(k)$ such that

$$P\left(\bigcup_{i=1}^k Z_i \geq c\right) = \alpha.$$

Historical Efficacy Boundaries: Pocock

Table: 1-sided z-score/p-value boundaries for Pocock procedure.

# Looks (k)	$\alpha = 0.005$	$\alpha = 0.025$	$\alpha = 0.05$
1	2.576 0.0050	1.960 0.0250	1.645 0.0500
2	2.772 0.0028	2.178 0.147	1.875 0.0304
3	2.873 0.0020	2.289 0.0110	1.992 0.0232
4	2.939 0.0016	2.361 0.0091	2.067 0.0194
5	2.986 0.0014	2.413 0.0079	2.122 0.0169
6	3.023 0.0013	2.453 0.0071	2.164 0.0152
7	3.053 0.0011	2.485 0.0065	2.197 0.0140
8	3.078 0.0010	2.512 0.0060	2.225 0.0130
9	3.099 0.0010	2.535 0.0056	2.249 0.0123
10	3.117 0.0009	2.555 0.0053	2.270 0.0116

Historical Efficacy Boundaries: Pocock

- ▶ Problem with Pocock: Z-statistic boundary at end is too high (equivalently, p-value boundary is too low).
- ▶ Example: for $k = 5$ looks, 1-sided 0.025 final boundary for z-statistic (p-value) is 2.413 (0.0079).
- ▶ **Causes loss of power, requiring larger sample sizes.**
- ▶ Also practical reasons for wanting high early boundaries and lower late boundaries. Early in trial, staff may not understand protocol.
- ▶ Pocock recommends against his own procedure.

Historical Efficacy Boundaries: O'Brien-Fleming

- ▶ Haybittle-Peto had a desirable final z-statistic boundary, but dropped so abruptly that it allowed a logical inconsistency.
- ▶ **Assume looks are equally-spaced.**
- ▶ What is the steepest descending boundary that avoids the logical inconsistency? Answer: O'Brien and Fleming (1979) boundary.
- ▶ Z-statistic boundary at look i is proportional to $1/\sqrt{i}$.
- ▶ Z-statistic boundaries at looks $1, 2, \dots, k$ are

$$\frac{a}{\sqrt{1}} = a, \frac{a}{\sqrt{2}}, \dots, \frac{a}{\sqrt{k}},$$

where $a = a(k)$ is a constant making the type 1 error rate α .

Historical Efficacy Boundaries: O'Brien-Fleming

Table: 1-sided O'Brien-Fleming z-score/p-value boundaries for $\alpha = 0.025$.

k	1	2	3	4	5	6	7	8	9	10
1	1.960 0.0250									
2	2.796 0.0026	1.977 0.0240								
3	3.471 0.0003	2.454 0.0071	2.004 0.0225							
4	4.048 2.6×10^{-5}	2.862 0.0021	2.337 0.0097	2.024 0.0215						
5	4.562 2.5×10^{-6}	3.226 0.0006	2.634 0.0042	2.281 0.0113	2.040 0.0207					
6	5.029 2.5×10^{-7}	3.556 0.0002	2.903 0.0018	2.514 0.0060	2.249 0.0123	2.053 0.0200				
7	5.458 2.4×10^{-8}	3.860 0.0001	3.151 0.0008	2.729 0.0032	2.441 0.0073	2.228 0.0129	2.063 0.0196			
8	5.861 2.3×10^{-9}	4.144 1.7×10^{-5}	3.384 0.0004	2.930 0.0017	2.621 0.0044	2.393 0.0084	2.215 0.0134	2.072 0.0191		
9	6.240 2.2×10^{-10}	4.412 5.1×10^{-6}	3.603 0.0002	3.120 0.0009	2.791 0.0026	2.547 0.0054	2.358 0.0092	2.206 0.0137	2.080 0.0188	
10	6.600 2.1×10^{-11}	4.667 1.5×10^{-6}	3.810 0.0001	3.300 0.0005	2.951 0.0016	2.694 0.0035	2.494 0.0063	2.333 0.0098	2.200 0.0139	2.087 0.0184

Historical Efficacy Boundaries

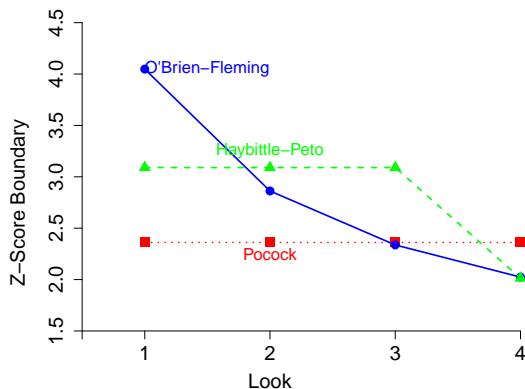


Figure: Haybittle-Peto, Pocock, and O'Brien-Fleming boundaries for 4 equally-spaced looks.

Unified Approach: Information, Z-Scores, B-Values

Unified Approach: Information, Z-Scores, B-Values

- ▶ We want to unify monitoring so same boundaries apply to many different testing settings.
- ▶ We will see that for large sample sizes, joint distribution of test statistics over time is same for different tests.
 - ▶ Lan and Zucker (1993).
 - ▶ Proschan et al. (2006).
 - ▶ Jennison and Turnbull (2000)
- ▶ **Warning: This section is technical.**
 - ▶ **May be difficult to absorb the first time, and you may need to return to this material.**
 - ▶ **We summarize important points in blue.**

Unified Approach: Information, Z-Scores, B-Values

- ▶ First step: Think about simple setting of iid $N(\delta, 1)$ data Y_1, Y_2, \dots, Y_N and we monitor after n , $n < N$:
- ▶ Estimator is $\hat{\delta}_n = \bar{Y}_n = S_n/n$, $S_n = \sum_{i=1}^n Y_i$.
- ▶ Sample size n measures amount of information contained in $\hat{\delta}_n$.
 - ▶ Note: $\text{var}(\hat{\delta}_n) = 1/n$, so $n = 1/\text{var}(\hat{\delta}_n)$ is information in $\hat{\delta}_n$.
- ▶ Fraction of information at interim analysis, $t = n/N$, is called **information time** or **information fraction**.

Unified Approach: Information, Z-Scores, B-Values

- ▶ For $t = n/N$, $Z(t) = \frac{S_n}{\sqrt{n}}$. Note that

$$\begin{aligned} E\{Z(t)\} &= E\left(\frac{S_n}{\sqrt{n}}\right) = \frac{n\delta}{\sqrt{n}} = \sqrt{n}\delta \\ &= (\sqrt{N}\delta) \sqrt{\frac{n}{N}} = \theta\sqrt{t}, \end{aligned}$$

$$\text{where } \theta = \sqrt{N}\delta = E\left(\frac{S_N}{\sqrt{N}}\right) = E\{Z(1)\}. \quad (14)$$

- ▶ Can also find variances and covariances of $Z(t)$ process.

Unified Approach: Information, Z-Scores, B-Values

- ▶ **Summary of z-score process: Joint distribution of $Z(t_1), \dots, Z(t_k)$ is multivariate normal with:**
 - ▶ $E\{Z(t)\} = \theta\sqrt{t}$, where $\theta = E\{Z(1)\}$.
 - ▶ $\text{var}\{Z(t)\} = 1$.
 - ▶ $\text{cov}\{Z(s), Z(t)\} = \sqrt{s/t}$, $s \leq t$.
- ▶ Note that z-scores become more correlated the closer their information times are to each other.
- ▶ Can use joint distribution of z-scores to find boundaries that control the type 1 error rate.

Unified Approach: Information, Z-Scores, B-Values

- ▶ We can instead monitor using 'B-values'.
- ▶ **Let $B(t) = \sqrt{t}Z(t) = \sqrt{\frac{n}{N}} \left(\frac{S_n}{\sqrt{n}} \right) = \frac{S_n}{\sqrt{N}}$.**
- ▶ $B(t)$ is proportional to a sum of iid $N(\delta, 1)$ observations, where the proportionality constant makes $B(1) = Z(1)$, (z-score at end).
- ▶ **Think of $B(t)$ like a sum.**

$$\frac{S_n}{X_1 + \dots + X_n} \quad \frac{S_N - S_n}{X_{n+1} + \dots + X_N}$$

- ▶ S_n and $S_N - S_n$ independent because sums don't overlap, and similarly for $S_{n_1}, S_{n_2} - S_{n_1}, S_{n_3} - S_{n_2}, \dots$
- ▶ Likewise, $B(t_1), B(t_2) - B(t_1), B(t_3) - B(t_2) \dots$ are independent.

Unified Approach: Information, Z-Scores, B-Values

- ▶ Independent increments make calculations of conditional power much easier. E.g.,

$$\frac{S_n}{X_1 + \dots + X_n} \quad \frac{S_N - S_n}{X_{n+1} + \dots + X_N}$$

$$\begin{aligned} P(S_N \geq c | S_n = s) &= P(s + S_N - S_n \geq c) \\ &= P(S_N - S_n \geq c - s). \end{aligned} \quad (15)$$

Same approach works when using B-values.

- ▶ Also, $E\{B(t)\} = \sqrt{t}E\{Z(t)\} = \sqrt{t}\theta\sqrt{t} = \theta t$, where $\theta = E\{Z(1)\}$.
- ▶ Easy to see if treatment is better ($B(t) > \theta t$) or worse ($B(t) < \theta t$) than expected.

Unified Approach: Information, Z-Scores, B-Values

- ▶ **Summary of B-values:** $B(t) = \sqrt{t}Z(t)$ has the following properties:
 - ▶ The joint distribution of $B(t_1), \dots, B(t_k)$ is multivariate normal.
 - ▶ $E\{B(t)\} = \theta t, \theta = E\{Z(1)\}$.
 - ▶ $\text{var}\{B(t)\} = t$ and, for $s \leq t$, $\text{cov}\{B(s), B(t)\} = s$.
 - ▶ $B(t)$ has independent increments.
- ▶ $B(t)$ is called a *Brownian motion with drift* θ .

Unified Approach: Information, Z-Scores, B-Values

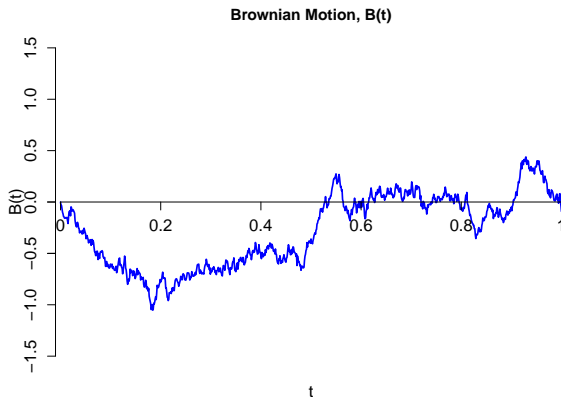


Figure: Brownian motion $B(t)$ with drift 0. t is information time. Paths are continuous everywhere but differentiable nowhere!

Another Interpretation of O'Brien-Fleming Procedure

- ▶ B-values also help us understand another motivation for O'Brien-Fleming.
- ▶ Pocock is a constant boundary for $Z(t)$, whereas O'Brien-Fleming is a constant boundary for $B(t)$.
- ▶ To see this, note that if we use constant B-value boundary a at times $t_i = i/k$, $i = 1, \dots, k$, then

$$B(i/k) \geq a \Leftrightarrow Z(i/k) \geq \frac{a}{\sqrt{i/k}},$$

so z-score boundary is proportional to $1/\sqrt{i}$, which is how we defined the O'Brien-Fleming boundary.

Unified Approach: Information, Z-Scores, B-Values

Table: **Properties of $Z(t)$ and $B(t)$ processes, where $\theta = E\{Z(1)\} = E\{B(1)\}$. See Appendix 1 at end of this lecture for additional details.**

	$Z(t)$	$B(t)$
$E\{Z(t)\}$ or $E\{B(t)\}$	$\theta \sqrt{t}$	θt
$\text{Var}\{Z(t)\}$ or $\text{var}\{B(t)\}$	1	t
$\text{Cov}\{Z(s), Z(t)\}$ or $\text{Cov}\{B(s), B(t)\}$, $s \leq t$	$\sqrt{\frac{s}{t}}$	s
Independent increments?	No	Yes

- ▶ We can monitor with $B(t)$ or $Z(t)$, but calculations of probabilities are easier with $B(t)$ because of independent increments and $B(t)$ tells whether treatment trend is continuing or reversing.

Unified Approach: Information, Z-Scores, B-Values

Table: **Properties of $Z(t)$ and $B(t)$ processes, where $\theta = E\{Z(1)\} = E\{B(1)\}$. See Appendix 1 at end of this lecture for additional details.**

	$Z(t)$	$B(t)$
$E\{Z(t)\}$ or $E\{B(t)\}$	$\theta \sqrt{t}$	θt
$\text{Var}\{Z(t)\}$ or $\text{var}\{B(t)\}$	1	t
$\text{Cov}\{Z(s), Z(t)\}$ or $\text{Cov}\{B(s), B(t)\}$, $s \leq t$	$\sqrt{\frac{s}{t}}$	s
Independent increments?	No	Yes

- ▶ We can monitor with $B(t)$ or $Z(t)$, but calculations of probabilities are easier with $B(t)$ because of independent increments and $B(t)$ tells whether treatment trend is continuing or reversing.

Unified Approach: Information, Z-Scores, B-Values

- ▶ **Key Fact: The same joint distribution of tests statistics holds for many different test statistics, provided we define information time correctly**
 - ▶ One- and two-sample t-tests.
 - ▶ One and two-sample z-tests of proportions.
 - ▶ The logrank test and Cox model.
 - ▶ Large sample tests using an MLE.
 - ▶ Tests based on a complete, sufficient statistic.
 - ▶ Many more.
- ▶ **Information time is n/N , except for survival, when it is d/D , where n and N are interim and final sample sizes and d and D are interim and final numbers of people with events.**
- ▶ See Lan and Zucker (1993); chapter 2 of Proschan et al. (2006); chapters 3 and 11 of Jennison and Turnbull (2000).

Unified Approach: Information, Z-Scores, B-Values

- ▶ Key idea is that many estimators $\hat{\delta}$ behave just like a mean of *some number, I* , of iid $N(\delta, \mathbf{1})$ observations.
- ▶ Just as $\hat{\delta}$ behaves like a mean, $I\hat{\delta}$ behaves like a sum.
- ▶ **We just have to define I (information) appropriately; $I = 1/\text{var}(\hat{\delta})$.**
- ▶ What really matters is information fraction, $t = I/I_{\text{end}}$, where I and I_{end} are the information at interim analysis and end of trial.
- ▶ **For most estimators, information is proportional to sample size, so $t = n/N$, where n and N are interim and final sample sizes.**
- ▶ **In survival, information is proportional to # events, so $t = d/D$, where d and D are interim and final # people with events.**

Alpha Spending Functions

Alpha Spending Functions

- ▶ For any z-score boundary c_1, \dots, c_k , we can compute probabilities of crossing boundaries by different times.
- ▶ Likewise, if we know probabilities of crossing by different times, we can re-construct boundaries.
- ▶ Lan and DeMets (1983): Instead of specifying boundaries, specify an **alpha spending function** $\alpha^*(t)$ giving cumulative alpha spent by information time t .
 - ▶ $\alpha^*(t)$ increases as t increases.
 - ▶ $\alpha^*(0) = 0$, $\alpha^*(1) = \alpha$ (spend no alpha at beginning and all alpha by the end).
- ▶ Then use information times to construct boundaries.

Alpha Spending Functions

- ▶ Properties depend on which spending function we choose.

- ▶ **Most popular spending function for a 1-sided test at level α :**

$$\alpha^*(t) = 2 \left\{ 1 - \Phi \left(\frac{z_{\alpha/2}}{\sqrt{t}} \right) \right\}, \quad (16)$$

where $z_{\alpha/2} = \Phi^{-1}(1 - \alpha/2)$.

- ▶ For 1-sided, $\alpha = 0.025$,

$$\alpha^*(t) = 2 \left\{ 1 - \Phi \left(\frac{2.2414}{\sqrt{t}} \right) \right\}. \quad (17)$$

- ▶ Generates boundaries similar to O'Brien-Fleming's for equally-spaced t , but **spending function approach does not require equal spacing**

Alpha Spending Functions

- ▶ We illustrate boundary construction using this spending function and 3 looks for a survival trial planned for 200 deaths by end.
- ▶ Suppose first look occurs at 58th death, $t = 58/200 = 0.29$.
- ▶ Cumulative alpha to spend by $t = 0.29$ is

$$\alpha^*(0.29) = 2 \left\{ 1 - \Phi \left(\frac{2.2414}{\sqrt{0.29}} \right) \right\} = 3.15 \times 10^{-5}. \quad (18)$$

- ▶ Need to find c_1 such that $P\{Z(0.29) \geq c_1\} = 3.15 \times 10^{-5}$.
- ▶ In R, can use `ldBounds` command in `ldbounds` function.

Alpha Spending Functions

- ▶ We illustrate boundary construction using this spending function and 3 looks for a survival trial planned for 200 deaths by end.
- ▶ Suppose first look occurs at 58th death, $t = 58/200 = 0.29$.
- ▶ Cumulative alpha to spend by $t = 0.29$ is

$$\alpha^*(0.29) = 2 \left\{ 1 - \Phi \left(\frac{2.2414}{\sqrt{0.29}} \right) \right\} = 3.15 \times 10^{-5}. \quad (18)$$

- ▶ Need to find c_1 such that $P\{Z(0.29) \geq c_1\} = 3.15 \times 10^{-5}$.
- ▶ In R, can use `ldBounds` command in `ldbounds` function.

Alpha Spending Functions

- ▶ We illustrate boundary construction using this spending function and 3 looks for a survival trial planned for 200 deaths by end.
- ▶ Suppose first look occurs at 58th death, $t = 58/200 = 0.29$.
- ▶ Cumulative alpha to spend by $t = 0.29$ is

$$\alpha^*(0.29) = 2 \left\{ 1 - \Phi \left(\frac{2.2414}{\sqrt{0.29}} \right) \right\} = 3.15 \times 10^{-5}. \quad (18)$$

- ▶ Need to find c_1 such that $P\{Z(0.29) \geq c_1\} = 3.15 \times 10^{-5}$.
- ▶ In R, can use `ldBounds` command in `ldbounds` function.

Alpha Spending Functions

```
library(ldbounds);t<-c(.29);ldBounds(t, iuse=1, alpha=0.025, sides=1)
```

- ▶ `iuse=1` is O'Brien-Fleming spending function. R responds with:

```
Lan-DeMets bounds for a given spending function
```

```
n = 1
```

```
Overall alpha: 0.025
```

```
Type: One-Sided Bounds
```

```
alpha: 0.025
```

```
Spending function: O'Brien-Fleming
```

```
Boundaries:
```

```
Time Upper  
0.2900 4.0011
```

- ▶ First z-score boundary is $c_1 = 4.0011$.

Alpha Spending Functions

- ▶ Suppose $Z(0.29) < 4.0011$, so go to second look.
- ▶ Second look occurs after 110th death, so $t = 110/200 = 0.55$.
- ▶ Cumulative alpha to spend by $t = 0.55$ is

$$\alpha^*(0.55) = 2 \left\{ 1 - \Phi \left(\frac{2.2414}{\sqrt{0.55}} \right) \right\} = 0.0025. \quad (19)$$

- ▶ Need to find c_2 such that

$$P\{Z(0.29) \geq 4.0011 \cup Z(0.55) \geq c_2\} = 0.0025$$

(cumulative error rate 0.0025).

Alpha Spending Functions

- ▶ Suppose $Z(0.29) < 4.0011$, so go to second look.
- ▶ Second look occurs after 110th death, so $t = 110/200 = 0.55$.
- ▶ Cumulative alpha to spend by $t = 0.55$ is

$$\alpha^*(0.55) = 2 \left\{ 1 - \Phi \left(\frac{2.2414}{\sqrt{0.55}} \right) \right\} = 0.0025. \quad (19)$$

- ▶ Need to find c_2 such that

$$P\{Z(0.29) \geq 4.0011 \cup Z(0.55) \geq c_2\} = 0.0025$$

(cumulative error rate 0.0025).

Alpha Spending Functions

```
t<-c(.29,0.55); ldBounds(t, iuse=1, alpha=0.025, sides=1)
```

Lan-DeMets bounds for a given spending function

```
n = 2
```

```
Overall alpha: 0.025
```

```
Type: One-Sided Bounds
```

```
alpha: 0.025
```

```
Spending function: O'Brien-Fleming
```

```
Boundaries:
```

	Time	Upper
1	0.29	4.0011
2	0.55	2.8074

▶ $c_2 = 2.8074$.

Alpha Spending Functions

- ▶ Suppose $Z(t_2) < 2.8074$, so go last look at $t = 1$:
- ▶ Cumulative alpha to spend by $t = 1$ is

$$\alpha^*(1) = 2 \left\{ 1 - \Phi \left(\frac{2.2414}{\sqrt{1}} \right) \right\} = 0.025. \quad (20)$$

- ▶ Need to find c_3 such that

$$P\{Z(0.29) \geq 4.0011 \cup Z(0.55) \geq 2.8074 \cup Z(1) \geq c_3\} = 0.025$$

(cumulative error rate 0.025).

Alpha Spending Functions

```
t<-c(.29,0.55,1); ldBounds(t, iuse=1, alpha=0.025, sides=1)
```

Lan-DeMets bounds for a given spending function

```
n = 3
```

```
Overall alpha: 0.025
```

```
Type: One-Sided Bounds
```

```
alpha: 0.025
```

```
Spending function: O'Brien-Fleming
```

```
Boundaries:
```

	Time	Upper
1	0.29	4.0011
2	0.55	2.8074
3	1.00	1.9740

Alpha Spending Functions

- ▶ Last boundary is $c_3 = 1.9740$.
- ▶ Boundaries at the 3 looks are:
 - ▶ $t = 0.29$: $c_1 = 4.0011$.
 - ▶ $t_2 = 0.55$: $c_2 = 2.8074$.
 - ▶ $t = 1$: $c_3 = 1.9740$.
- ▶ Note: Could also use Free software at U. Wisconsin (see Appendix 2 at end of this lecture).

Alpha Spending Functions

- ▶ For 2-sided test at $\alpha = 0.05$, change R command to:

```
t<-c(.29,0.55,1); ldBounds(t, iuse=1, alpha=0.05, sides=2)
```

Lan-DeMets bounds for a given spending function

```
n = 3
```

```
Overall alpha: 0.05
```

```
Type: Two-Sided Symmetric Bounds
```

```
Lower alpha: 0.025
```

```
Upper alpha: 0.025
```

```
Spending function: O'Brien-Fleming
```

Boundaries:

	Time	Lower	Upper
1	0.29	-4.001115	4.001115
2	0.55	-2.807364	2.807364
3	1.00	-1.973987	1.973987

Alpha Spending Functions

- ▶ We have been using the cumulative alpha formulation. An equivalent way to compute c_i uses the first crossing formulation:

$$P(Z(t_1) < c_1 \cap \dots \cap Z(t_{i-1}) < c_{i-1} \cap Z(t_i) \geq c_i) = \alpha^*(t_i) - \alpha^*(t_{i-1})$$

(for 2-sided symmetric test, replace $Z(t_i)$ with $|Z(t_i)|$).

- ▶ That is, probability of **first** crossing boundary at time t_i is $\alpha^*(t_i) - \alpha^*(t_{i-1})$.
- ▶ Then cumulative probability of crossing by t_i is

$$\alpha^*(t_1) + \{\alpha^*(t_2) - \alpha^*(t_1)\} + \dots + \{\alpha^*(t_i) - \alpha^*(t_{i-1})\} = \alpha^*(t_i).$$

Alpha Spending Functions

- ▶ **Big advantages of spending function approach:**
 - ▶ Looks need not be equally-spaced.
 - ▶ Don't even have to pre-specify number of looks (but number and timing of looks assumed independent of data).
- ▶ Nonetheless, pre-specification of number and timing of looks is advisable.

Alpha Spending Functions

- ▶ Pocock-like spending function: Lan and DeMets noticed that amount spent by Pocock boundaries looked like a log function for a large number of looks.
- ▶ To get similar spending function:

$$\alpha^*(t) = \alpha \ln(a + bt),$$

$$\alpha^*(0) = 0 \Rightarrow a = 1$$

$$\alpha^*(1) = \alpha \Rightarrow b = e - 1$$

$$\alpha^*(t) = \alpha \ln\{1 + (e - 1)t\}. \quad (21)$$

(21) is Pocock-like spending function. Generates boundaries similar to those of Pocock's when looks are equally spaced.

Alpha Spending Functions

- ▶ Pocock-like spending function: Lan and DeMets noticed that amount spent by Pocock boundaries looked like a log function for a large number of looks.
- ▶ To get similar spending function:

$$\alpha^*(t) = \alpha \ln(a + bt),$$

$$\alpha^*(0) = 0 \Rightarrow a = 1$$

$$\alpha^*(1) = \alpha \Rightarrow b = e - 1$$

$$\alpha^*(t) = \alpha \ln\{1 + (e - 1)t\}. \quad (21)$$

(21) is Pocock-like spending function. Generates boundaries similar to those of Pocock's when looks are equally spaced.

Alpha Spending Functions

- ▶ Pocock-like spending function: Lan and DeMets noticed that amount spent by Pocock boundaries looked like a log function for a large number of looks.
- ▶ To get similar spending function:

$$\alpha^*(t) = \alpha \ln(a + bt),$$

$$\alpha^*(0) = 0 \Rightarrow a = 1$$

$$\alpha^*(1) = \alpha \Rightarrow b = e - 1$$

$$\alpha^*(t) = \alpha \ln\{1 + (e - 1)t\}. \quad (21)$$

(21) is Pocock-like spending function. Generates boundaries similar to those of Pocock's when looks are equally spaced.

Alpha Spending Functions

- ▶ Pocock-like spending function: Lan and DeMets noticed that amount spent by Pocock boundaries looked like a log function for a large number of looks.
- ▶ To get similar spending function:

$$\alpha^*(t) = \alpha \ln(a + bt),$$

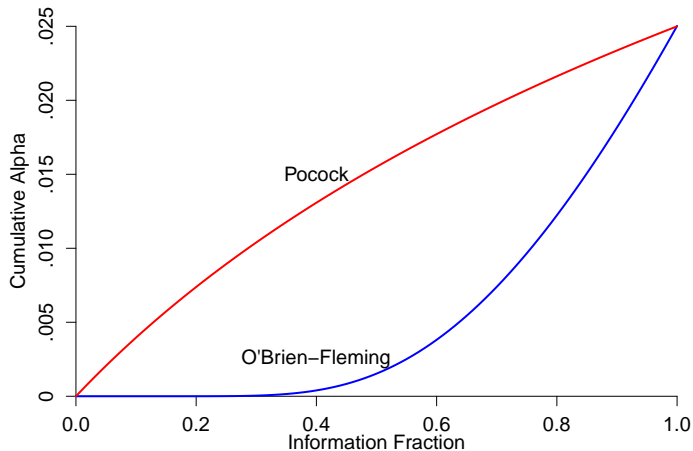
$$\alpha^*(0) = 0 \Rightarrow a = 1$$

$$\alpha^*(1) = \alpha \Rightarrow b = e - 1$$

$$\alpha^*(t) = \alpha \ln\{1 + (e - 1)t\}. \quad (21)$$

(21) is Pocock-like spending function. Generates boundaries similar to those of Pocock's when looks are equally spaced.

Alpha Spending Functions



Alpha Spending Functions

- ▶ O'Brien-Fleming-like spending function is convex and spends almost no α early in trial, then rises quickly toward end (desirable).
- ▶ In contrast, Pocock-like spending function is concave and spends more aggressively early (undesirable).
- ▶ **Consequence: O'Brien-Fleming-like spending function (good) creates high early boundaries and boundaries close to 1.96 at end, whereas Pocock (bad) has lower early boundaries but higher late boundaries.**

Alpha Spending Functions

- ▶ There are also families of spending functions like the Kim DeMets power family (Kim and DeMets (1987)):

$$\alpha^*(t) = \alpha t^\phi,$$

where small values of power parameter ϕ spend α aggressively early, whereas larger values spend alpha conservatively until close to $t = 1$.

- ▶ Another family is the Hwang, Shih-DeCani family, also called the gamma family (Hwang et al. (1990)):

$$\alpha^*(0) = 0, \quad \alpha^*(t) = \alpha \times \left\{ \frac{1 - \exp(-\gamma t)}{1 - \exp(-\gamma)} \right\}, \quad t > 0.$$

- ▶ γ can be positive (aggressive early spending) or negative (conservative early spending).
- ▶ $\gamma = -4$ is like O'Brien-Fleming.

Alpha Spending Functions

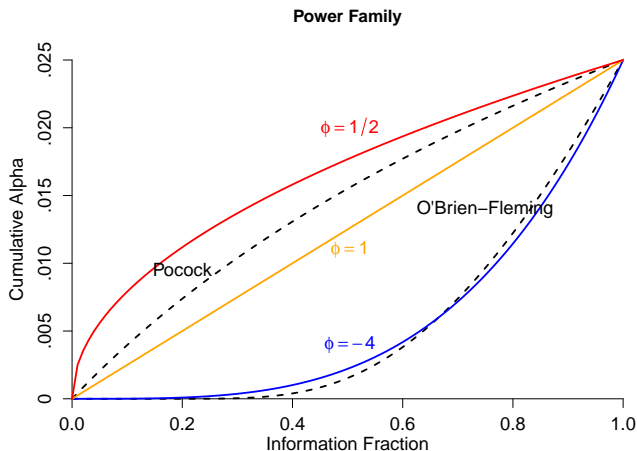


Figure: The power family $\alpha^*(t) = 0.025t^\phi$ for different values of ϕ .

Alpha Spending Functions

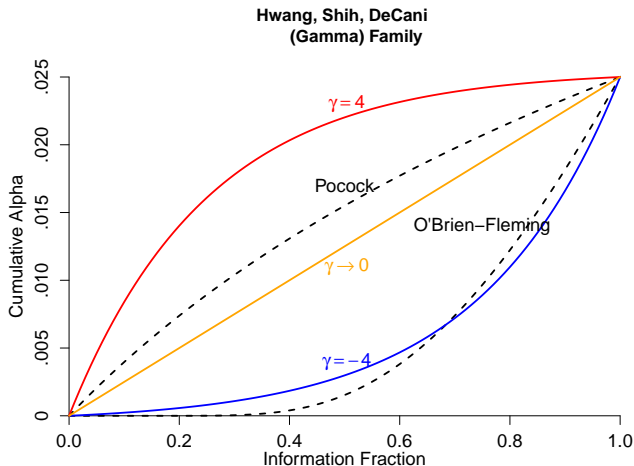


Figure: Wang, Shih, DeCani family of spending functions, $\{1 - \exp(-\gamma t)\} / \{1 - \exp(-\gamma)\}$ for different values of γ .

Effect of Efficacy Monitoring on Power

Effect of Efficacy Monitoring on Power

- ▶ Just like in non-monitoring setting, power in monitoring setting depends only on $E\{Z(1)\} = \theta$, the expected z-score at end (drift parameter of Brownian motion).
- ▶ Monitoring must incur a sample size penalty; always more powerful to spend all alpha at end (Neyman-Pearson lemma).
- ▶ The size of the penalty depends on the spending function
 - ▶ Pocock–big penalty
 - ▶ O'Brien-Fleming–small penalty.
- ▶ Why?

Effect of Efficacy Monitoring on Power

- ▶ Consider O-F-like spending function with 4 equally-spaced looks. Z-score boundaries are:

Table: Boundaries for O-F-like spending function, 4 equally spaced looks.

	$t = 1/4$	$t = 2/4$	$t = 3/4$	$t = 1$
Z boundary	4.3326	2.9631	2.3590	2.0141

$$P_{\theta}(Z(1) \geq 2.0141) \leq \text{Power, monitoring} \leq P_{\theta}(Z(1) \geq 1.96).$$

- ▶ Left and right sides are almost equal, so power with O-F almost same as power with no monitoring! **Not true for Pocock.**
- ▶ **O-F spending function has minimal effect on power.**

Effect of Efficacy Monitoring on Power

- ▶ We can compute sample size/power for monitoring using the EZ principle.
- ▶ Just like in non-monitoring setting, power depends on EZ, namely $E\{Z(1)\}$ (the drift parameter, θ).
- ▶ Can use R to compute drift parameter θ for given power.
- ▶ Then equate $E\{Z(1)\}$ to given value of drift and solve for N .

Effect of Efficacy Monitoring on Power

- ▶ For example, suppose we want 4 equally-spaced looks.
- ▶ $t = (1/4, 2/4, 3/4, 1)$.
- ▶ To use R, first compute boundaries:

```
t<-c(1/4,2/4,3/4,1)
bdry<-ldBounds(t, iuse=1, alpha=0.05, sides=2)
lwr<-bdry$lower.bounds
upr<-bdry$upper.bounds
```

- ▶ Now lwr and upr contain lower and upper boundaries

Effect of Efficacy Monitoring on Power

- ▶ Now compute drift parameter using:

```
ldPower(t, za=lwr, zb=upr, pow=0.90, drift=NULL)
```

- ▶ Note: Can use `ldpower` to compute either the drift parameter for given power or power for given drift parameter.
- ▶ Whichever you give `R` (drift parameter or power level), it will supply the other.
- ▶ `R` responds with the following output:

Effect of Efficacy Monitoring on Power

Lan-DeMets method for group sequential boundaries

n = 4

Boundaries:

	Time	Lower	Upper	Lower probs	Upper probs
1	0.25	-4.332634	4.332634	0.000000e+00	0.003497291
2	0.50	-2.963112	2.963112	6.586691e-08	0.254380134
3	0.75	-2.359023	2.359023	9.702757e-08	0.427384452
4	1.00	-2.014059	2.014059	5.061840e-08	0.214737908

Power : 0.9

Drift: 3.271063

- ▶ Tells us we need a drift parameter of 3.2711.

Effect of Efficacy Monitoring on Power

- ▶ Tells us that in sample size calculations, make expected z-score 3.2711 instead of $1.96 + 1.28 = 3.24$.
- ▶ Example: For a t-test, expected z-score at end is

$$\frac{\delta}{\sqrt{\frac{2\sigma^2}{N}}}$$

- ▶ Equate to 3.2711 and solve for N . Per-arm sample size is

$$N = \frac{2\sigma^2(3.2711)^2}{\delta^2} \text{ per arm.}$$

Effect of Efficacy Monitoring on Power

- ▶ If $\delta = 5$ and $\sigma = 14$, $N \approx 168$ per arm.
- ▶ Compare to 165 per arm with no monitoring.
- ▶ **For O'Brien-Fleming, only require 3 more people per arm than if we did not monitor the trial!**
- ▶ For " Pocock" spending function, only difference is replace

```
bdry<-ldBounds(t, iuse=1, alpha=0.05, sides=2)
```

with

```
bdry<-ldBounds(t, iuse=2, alpha=0.05, sides=2)
```

Effect of Efficacy Monitoring on Power

- ▶ For Pocock, drift parameter needed for 90% power is 3.5177.
- ▶ New per-arm sample size is

$$N = \frac{2(14)^2(3.5177)^2}{5^2} \approx 195.$$

- ▶ **Bottom line: Much more substantial sample size penalty for Pocock spending function than O'Brien-Fleming spending function.**

CASE Study: CAST

Case Study: CAST

- ▶ Recall the Cardiac Arrhythmia Suppression Trial (Case Study: CAST).
- ▶ Patients with prior heart attack and arrhythmias.
- ▶ Several papers showed arrhythmias are associated with increased risk of death after heart attack (see, e.g., CDP (1973), Ruberman et al. (1977)).
- ▶ Class I antiarrhythmic drugs encainide, flecainide, and moricizine were approved based on surrogate arrhythmia endpoint.
- ▶ Goal of CAST: Determine whether suppressing arrhythmias leads to fewer sudden deaths/cardiac arrests.
- ▶ CAST titrated encainide, flecainide, and moricizine until they found one that worked (did not care which drug used).
- ▶ Wanted to ensure arrhythmias could be suppressed, so patients whose arrhythmias not suppressed were not randomized

Case Study: CAST

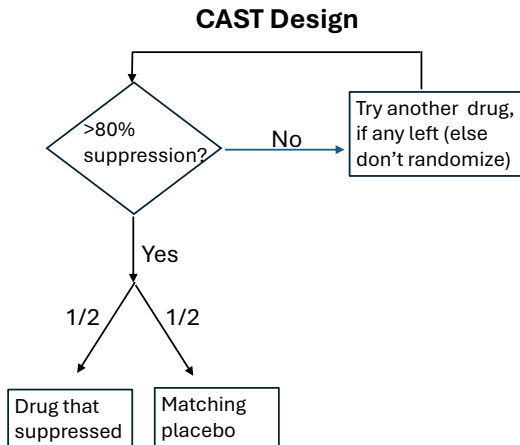


Figure: CAST design.

Case Study: CAST

- ▶ Some died during titration, but that's expected in sick population.
- ▶ Doctors already believed suppression hypothesis.
- ▶ In fact, some felt it was unethical to withhold treatment after finding a drug that suppressed arrhythmias.
 - ▶ Led to recruitment problems in CAST.
- ▶ So convinced were doctors that results could only be beneficial that CAST investigators proposed a 1-sided test at $\alpha = 0.05$.
- ▶ At first DSMB meeting 3/14/87, DSMB reviewed protocol and recommended using 1-sided $\alpha = 0.025$ instead of 0.05 for efficacy. Asked for monitoring guideline in future.
- ▶ Next meeting January 1988: Board chose to be blinded.

Case Study: CAST

- ▶ Next DSMB meeting: 9/16/88. DSMB discussed and approved monitoring plan.
- ▶ Plan used $\alpha = 0.025$ for efficacy.
- ▶ DSMB decided to use symmetric lower 0.025 boundary for harm.
- ▶ Spending function spent $\alpha = 0.0125$ linearly until just before end, then spent remaining 0.0125 at end:

$$\alpha^*(t) = \begin{cases} 0.0125t & \text{if } t < 1, \\ 0.025 & \text{if } t = 1. \end{cases} \quad (22)$$

Case Study: CAST

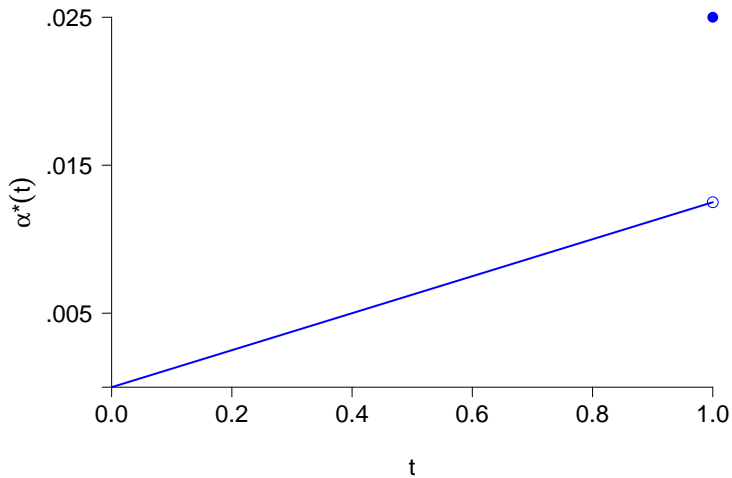


Figure: Spending function for efficacy used in CAST. Used symmetric lower boundary.

Case Study: CAST

- ▶ After approving monitoring plan, the DSMB examined data.
- ▶ Proportion with sudden death/cardiac arrest:
 - ▶ Arm X: 3/576.
 - ▶ Arm Y: 19/571.
- ▶ Information fraction after 22 of 425 expected events:
 $t = 22/425 = 0.05$.
- ▶ No boundary was yet in place, but suppose it were.
 - ▶ Can spend $\alpha^*(0.05) = 0.0125(0.05) = 0.000625$. Boundary would have been -3.22 and logrank z-score was -3.43 .
 - ▶ Would have crossed harm boundary!
- ▶ Board decided no matter which arm was placebo, they wouldn't stop. Remained blinded.

Case Study: CAST

- ▶ Next DSMB meeting: 4/16-4/17, 1989.
- ▶ There were now 48 sudden deaths/cardiac arrests, 35 of which were on arm Y!
- ▶ Now clear they overestimated final number events. Now expected to be 300, so $t = 48/300 = 0.16$.
- ▶ Boundary ± 2.97 .
- ▶ DSMB unblinded and discovered arm Y was active.
- ▶ Logrank z-score: $Z(0.16) = -3.22$. Harm boundary crossed!
- ▶ DSMB was shocked.

Case Study: CAST

- ▶ Maybe pills were mis-labeled? No. In preparation for meeting, they analyzed pills and found no mis-labeling.
- ▶ Maybe randomization failed? No. Baseline characteristics were similar across arms.
- ▶ Maybe harm was confined to a certain subgroup? No. Results were consistent across subgroups and secondary endpoint of mortality.
- ▶ Maybe harm was confined to 1 or 2 drugs? It appeared that encainide and flecainide were the “bad actors” and moricizine was good.
- ▶ Decided to discontinue encainide and flecainide and continue CAST II with moricizine.

Case Study: CAST

- ▶ DSMB chose to be blinded again. Saw results by Arm P versus Arm Q.
- ▶ Changed entry criteria to get sicker patients: Thought drugs should work in sicker patients.
- ▶ Decided to spend 0.05 for harm and 0.025 for benefit.
- ▶ Included 2-week placebo titration phase to see if too many patients were dying while titrating drugs.
- ▶ On July 31, 1991, there were 3 events on arm P and 15 on arm Q in 2-week titration phase.
- ▶ Arm Q was moricizine!
- ▶ CAST II ended.

Case Study: CAST

- ▶ **Lessons from CAST:**
 - ▶ **Can have benefit on surrogate (arrhythmias) and harm on endpoint of real interest (sudden death/cardiac arrest).**
 - ▶ **Harm can never be ruled out. Include upper and lower boundaries in clinical trials.**
 - ▶ **May make sense to use asymmetric boundaries for harm versus benefit.**
 - ▶ **Blinding DSMBs is ill-advised. Members think this makes them more objective, but the opposite is true.**
 - ▶ **Pre-conceived ideas enter maximally if blinded.**
 - ▶ **Why waste time pondering what you would do if results were in one direction or the other, when they are only in one direction?**
- ▶ **CAST references: CAST (1989) and CAST (1992).**
- ▶ **Moore (1995b) gives a history of the suppression hypothesis and development and testing of antiarrhythmic drugs.**

Summary

- ▶ Unified monitoring:
 - ▶ Information I in treatment effect estimator $\hat{\delta}$ is $1/\text{var}(\hat{\delta})$, and information time is $t = I_{\text{current}}/I_{\text{final}}$.
 - ▶ $B(t)$ and $\hat{\delta}(t)$ behave like sum and sample mean of I iid $N(\delta, 1)$ observations.
 - ▶ t often reduces to ratio of current to final sample size (non-survival) or current to final number of people with events (survival).
 - ▶ Same boundaries apply to different test statistics.
- ▶ Can monitor using $Z(t)$ or Brownian motion $B(t)$. Joint distribution over time is multivariate normal with:
 - ▶ $E\{Z(t)\} = \theta\sqrt{t}$, $\text{cov}\{Z(s), Z(t)\} = \sqrt{s/t}$, $s \leq t$.
 - ▶ $E\{B(t)\} = \theta t$, $\text{cov}\{B(s), B(t)\} = s$, $s \leq t$. $\theta = E\{Z(1)\} = E\{B(1)\}$ is drift parameter.

Summary

- ▶ classic boundaries (Haybittle-Peto, Pocock, O'Brien-Fleming) require equal spacing and pre-specification of number of looks.
- ▶ Desirable z-score boundaries (e.g., O'Brien-Fleming) are high early and close to z_α at end. Effect on power is minimal (unlike Pocock).
- ▶ Haybittle-Peto simple and valid regardless of joint distribution of test statistic, but has reversal of fortune problem.
- ▶ Alpha spending functions are flexible and preferable to classic boundaries. Neither number nor timing of looks must be pre-specified. Can pick shape to ensure desired properties.

Appendix 1: Unified Approach

- ▶ Consider trial with continuous outcome and paired differences D_1, D_2, \dots, D_N , one member on treatment, other on control.

$$\text{Interim: } Z_n = \frac{S_n}{\sqrt{n\sigma^2}}, \quad \text{Final: } Z_N = \frac{S_N}{\sqrt{N\sigma^2}}.$$

$$\begin{aligned} \text{cov}(Z_n, Z_N) &= \text{cov}\left\{\frac{S_n}{\sqrt{n\sigma^2}}, \frac{S_N}{\sqrt{N\sigma^2}}\right\} \\ &= \frac{\text{cov}\{S_n, S_N\}}{\sqrt{n\sigma^2}\sqrt{N\sigma^2}} \\ &= \frac{\text{cov}\{S_n, S_n + (S_N - S_n)\}}{\sqrt{n\sigma^2}\sqrt{N\sigma^2}} \end{aligned}$$

Appendix 1: Unified Approach

- ▶ Consider trial with continuous outcome and paired differences D_1, D_2, \dots, D_N , one member on treatment, other on control.

$$\text{Interim: } Z_n = \frac{S_n}{\sqrt{n\sigma^2}}, \quad \text{Final: } Z_N = \frac{S_N}{\sqrt{N\sigma^2}}.$$

$$\begin{aligned} \text{cov}(Z_n, Z_N) &= \text{cov}\left\{\frac{S_n}{\sqrt{n\sigma^2}}, \frac{S_N}{\sqrt{N\sigma^2}}\right\} \\ &= \frac{\text{cov}\{S_n, S_N\}}{\sqrt{n\sigma^2}\sqrt{N\sigma^2}} \\ &= \frac{\text{cov}\{S_n, S_n + (S_N - S_n)\}}{\sqrt{n\sigma^2}\sqrt{N\sigma^2}} \end{aligned}$$

Appendix 1: Unified Approach

- ▶ Consider trial with continuous outcome and paired differences D_1, D_2, \dots, D_N , one member on treatment, other on control.

$$\text{Interim: } Z_n = \frac{S_n}{\sqrt{n\sigma^2}}, \quad \text{Final: } Z_N = \frac{S_N}{\sqrt{N\sigma^2}}.$$

$$\begin{aligned} \text{cov}(Z_n, Z_N) &= \text{cov}\left\{\frac{S_n}{\sqrt{n\sigma^2}}, \frac{S_N}{\sqrt{N\sigma^2}}\right\} \\ &= \frac{\text{cov}\{S_n, S_N\}}{\sqrt{n\sigma^2}\sqrt{N\sigma^2}} \\ &= \frac{\text{cov}\{S_n, S_n + (S_N - S_n)\}}{\sqrt{n\sigma^2}\sqrt{N\sigma^2}} \end{aligned}$$

Appendix 1: Unified Approach

$$\begin{aligned} &= \frac{\text{cov}\{S_n, S_n\} + \text{cov}\{S_n, (S_N - s_n)\}}{\sqrt{n\sigma^2}\sqrt{N\sigma^2}} \\ &= \frac{\text{var}(S_n) + 0}{\sqrt{n\sigma^2}\sqrt{N\sigma^2}} = \frac{n\sigma^2}{\sqrt{n\sigma^2}\sqrt{N\sigma^2}} \\ &= \sqrt{n/N} = \sqrt{t}, \quad t = n/N. \end{aligned}$$

- ▶ t is called the **information fraction** or **information time** of the interim analysis.
- ▶ Note that $t = 0$ and 1 at the beginning and end of the trial.

Appendix 1: Unified Approach

$$\begin{aligned} &= \frac{\text{cov}\{S_n, S_n\} + \text{cov}\{S_n, (S_N - s_n)\}}{\sqrt{n\sigma^2}\sqrt{N\sigma^2}} \\ &= \frac{\text{var}(S_n) + 0}{\sqrt{n\sigma^2}\sqrt{N\sigma^2}} = \frac{n\sigma^2}{\sqrt{n\sigma^2}\sqrt{N\sigma^2}} \\ &= \sqrt{n/N} = \sqrt{t}, \quad t = n/N. \end{aligned}$$

- ▶ t is called the *information fraction* or *information time* of the interim analysis.
- ▶ Note that $t = 0$ and 1 at the beginning and end of the trial.

Appendix 1: Unified Approach

$$\begin{aligned} &= \frac{\text{cov}\{S_n, S_n\} + \text{cov}\{S_n, (S_N - s_n)\}}{\sqrt{n\sigma^2}\sqrt{N\sigma^2}} \\ &= \frac{\text{var}(S_n) + 0}{\sqrt{n\sigma^2}\sqrt{N\sigma^2}} = \frac{n\sigma^2}{\sqrt{n\sigma^2}\sqrt{N\sigma^2}} \\ &= \sqrt{n/N} = \sqrt{t}, \quad t = n/N. \end{aligned}$$

- ▶ t is called the **information fraction** or **information time** of the interim analysis.
- ▶ Note that $t = 0$ and 1 at the beginning and end of the trial.

Appendix 1: Unified Approach

- ▶ Similarly, at interim analyses with n_1 and n_2 observations, $n_1 < n_2$,

$$\text{cov}(Z_{n_1}, Z_{n_2}) = \sqrt{\frac{n_1}{n_2}} = \sqrt{\frac{n_1/N}{n_2/N}} = \sqrt{t_1/t_2}. \quad (23)$$

- ▶ The closer the two interim analyses are, the higher the correlation of z-statistics. The mean of the z-statistic is:

$$\begin{aligned} E\{Z(t)\} &= E\left\{\frac{S_n}{\sqrt{n\sigma^2}}\right\} \\ &= \frac{n\mu}{\sqrt{n\sigma^2}} = \frac{\sqrt{n}\mu}{\sigma} \end{aligned}$$

$$\theta \equiv E\{Z(1)\} = \frac{\sqrt{N}\mu}{\sigma}, \text{ so}$$

$$E\{Z(t)\} = \theta\sqrt{t}. \quad (24)$$

Appendix 1: Unified Approach

- ▶ Similarly, at interim analyses with n_1 and n_2 observations, $n_1 < n_2$,

$$\text{cov}(Z_{n_1}, Z_{n_2}) = \sqrt{\frac{n_1}{n_2}} = \sqrt{\frac{n_1/N}{n_2/N}} = \sqrt{t_1/t_2}. \quad (23)$$

- ▶ The closer the two interim analyses are, the higher the correlation of z-statistics. The mean of the z-statistic is:

$$\begin{aligned} E\{Z(t)\} &= E\left\{\frac{S_n}{\sqrt{n\sigma^2}}\right\} \\ &= \frac{n\mu}{\sqrt{n\sigma^2}} = \frac{\sqrt{n}\mu}{\sigma} \end{aligned}$$

$$\theta \equiv E\{Z(1)\} = \frac{\sqrt{N}\mu}{\sigma}, \text{ so}$$

$$E\{Z(t)\} = \theta\sqrt{t}. \quad (24)$$

Appendix 1: Unified Approach

- ▶ Similarly, at interim analyses with n_1 and n_2 observations, $n_1 < n_2$,

$$\text{cov}(Z_{n_1}, Z_{n_2}) = \sqrt{\frac{n_1}{n_2}} = \sqrt{\frac{n_1/N}{n_2/N}} = \sqrt{t_1/t_2}. \quad (23)$$

- ▶ The closer the two interim analyses are, the higher the correlation of z-statistics. The mean of the z-statistic is:

$$\begin{aligned} E\{Z(t)\} &= E\left\{\frac{S_n}{\sqrt{n\sigma^2}}\right\} \\ &= \frac{n\mu}{\sqrt{n\sigma^2}} = \frac{\sqrt{n}\mu}{\sigma} \end{aligned}$$

$$\theta \equiv E\{Z(1)\} = \frac{\sqrt{N}\mu}{\sigma}, \text{ so}$$

$$E\{Z(t)\} = \theta\sqrt{t}. \quad (24)$$

Appendix 1: Unified Approach

- ▶ By the central limit theorem, $Z(t_i)$ is approximately normal for large sample size.
- ▶ Moreover, the joint distribution of $\{Z(t_1), \dots, Z(t_k)\}$ is also approximately multivariate normal.
- ▶ Sometimes it is more convenient to look at the ***B-value*** rather than the z-statistic.

$$B(t) = \sqrt{t}Z(t). \quad (25)$$

Appendix 1: Unified Approach

- ▶ The B-value has mean

$$\begin{aligned} E\{B(t)\} &= \sqrt{t}E\{Z(t)\} \\ &= \sqrt{t}(\theta\sqrt{t}) = \theta t. \end{aligned} \tag{26}$$

Also, for $s \leq t$,

$$\begin{aligned} \text{cov}\{B(s), B(t)\} &= \text{cov}\{\sqrt{s}Z(s), \sqrt{t}Z(t)\} \\ &= \sqrt{s}\sqrt{t}\text{cov}\{Z(s), Z(t)\} \\ &= \sqrt{s}\sqrt{t}\sqrt{\frac{s}{t}} = s. \end{aligned} \tag{27}$$

Appendix 1: Unified Approach

- ▶ The B-value has mean

$$\begin{aligned} E\{B(t)\} &= \sqrt{t}E\{Z(t)\} \\ &= \sqrt{t}(\theta\sqrt{t}) = \theta t. \end{aligned} \tag{26}$$

Also, for $s \leq t$,

$$\begin{aligned} \text{cov}\{B(s), B(t)\} &= \text{cov}\{\sqrt{s}Z(s), \sqrt{t}Z(t)\} \\ &= \sqrt{s}\sqrt{t} \text{cov}\{Z(s), Z(t)\} \\ &= \sqrt{s}\sqrt{t} \sqrt{\frac{s}{t}} = s. \end{aligned} \tag{27}$$

Appendix 1: Unified Approach

- ▶ The B-value has mean

$$\begin{aligned} E\{B(t)\} &= \sqrt{t}E\{Z(t)\} \\ &= \sqrt{t}(\theta\sqrt{t}) = \theta t. \end{aligned} \tag{26}$$

Also, for $s \leq t$,

$$\begin{aligned} \text{cov}\{B(s), B(t)\} &= \text{cov}\{\sqrt{s}Z(s), \sqrt{t}Z(t)\} \\ &= \sqrt{s}\sqrt{t} \text{cov}\{Z(s), Z(t)\} \\ &= \sqrt{s}\sqrt{t} \sqrt{\frac{s}{t}} = s. \end{aligned} \tag{27}$$

Appendix 1: Unified Approach

- ▶ Notice that, for $s < t$,

$$\begin{aligned}\operatorname{cov}\{B(s), B(t) - B(s)\} &= \operatorname{cov}\{B(s), B(t)\} - \operatorname{cov}\{B(s), B(s)\} \\ &= s - s = 0.\end{aligned}\tag{28}$$

- ▶ Because $B(s)$ and $B(t) - B(s)$ are bivariate normal with 0 correlation, $B(s)$ and $B(t) - B(s)$ are independent.
- ▶ (**Independent increments property**) More generally, for $t_1 < t_2 < \dots < t_k$, the increments $B(t_1)$, $B(t_2) - B(t_1)$, \dots , $B(t_k) - B(t_{k-1})$ are independent.
- ▶ $B(t)$ is continuous everywhere, differentiable nowhere.

Appendix 1: Unified Approach

- ▶ Notice that, for $s < t$,

$$\begin{aligned}\operatorname{cov}\{B(s), B(t) - B(s)\} &= \operatorname{cov}\{B(s), B(t)\} - \operatorname{cov}\{B(s), B(s)\} \\ &= s - s = 0.\end{aligned}\tag{28}$$

- ▶ Because $B(s)$ and $B(t) - B(s)$ are bivariate normal with 0 correlation, $B(s)$ and $B(t) - B(s)$ are independent.
- ▶ (***Independent increments property***) More generally, for $t_1 < t_2 < \dots < t_k$, the increments $B(t_1)$, $B(t_2) - B(t_1)$, \dots , $B(t_k) - B(t_{k-1})$ are independent.
- ▶ $B(t)$ is continuous everywhere, differentiable nowhere.

Appendix 2: U. Wisconsin Software

- ▶ To use U. Wisconsin software instead of R to compute boundaries for O'Brien-Fleming-like spending function at information times $t = (0.29, 0.55, 1)$, use the following steps.

Appendix 2: U. Wisconsin Software

Lan: DeMets Group Sequential Calculations

File **Compute** Help

No Computation Selected

Analysis Parameters

Interim Analyses (k): (1 < k ≤ 25)

Information times(t): (0 < t ≤ 1)

Test Boundaries:

Calculate

Time	Lower Bound	Upper Bound		
1	0.20			
2	0.40			
3	0.60			
4	0.80			
5	1.00			

This is the statusbar

Type here to search

Appendix 2: U. Wisconsin Software

Lan-DeMets Group Sequential Calculations

File Compute Help

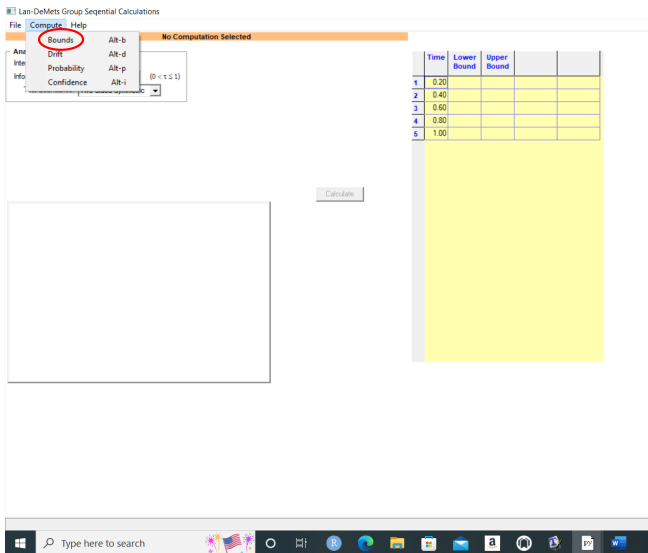
No Computation Selected

Bounds Alt-b
Drift Alt-d
Probability Alt-p
Confidence Alt-i

(0 <= t <= 1)

Time	Lower Bound	Upper Bound		
1	0.20			
2	0.40			
3	0.60			
4	0.80			
5	1.00			

Calculate



Appendix 2: U. Wisconsin Software

Lan-DeMets Group Sequential Calculations

File Compute Help

Compute Bounds

Analysis Parameters

Interim Analyses (k): (k ≤ 25)

Information times (t): (0 < t ≤ 1)

Test Boundaries:

Z-Score

Observed Z?

Spending Function

Overall Alpha: (0 < α ≤ 1.0)

Function:

Truncate bounds?

	Time	Lower Bound	Upper Bound	Nominal Upr Alpha	Cum Alpha
1	0.20				
2	0.40				
3	0.60				
4	0.80				
5	1.00				

Calculate

Appendix 2: U. Wisconsin Software

Lan-DeMets Group Sequential Calculations

File Compute Help

Compute Bounds

Analysis Parameters

Interim Analyses (k): 3 (0 < k ≤ 25)

Information times (t): Equally Spaced (0 < t ≤ 1)

Test Boundaries: Two-Sided Symmetric

Z-Score

Observed Z? No

Spending Function

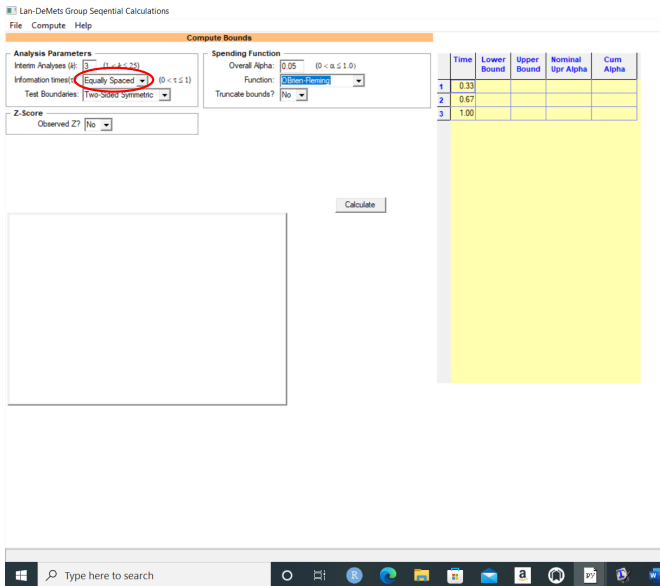
Overall Alpha: 0.05 (0 < α ≤ 1.0)

Function: Blin-Rimm

Truncate bounds? No

Calculate

Time	Lower Bound	Upper Bound	Nominal Upr Alpha	Cum Alpha
1	0.33			
2	0.67			
3	1.00			



Appendix 2: U. Wisconsin Software

Lan-DeMets Group Sequential Calculations

File Compute Help

Compute Bounds

Analysis Parameters

Interim Analyses (k): ($1 < k \leq 25$)
Information times (t): ($0 < t \leq 1$)
Test Boundaries: ($Use Input$)

Z-Score

Observed Z?

Spending Function

Overall Alpha: ($0 < \alpha \leq 1.0$)
Function: ($Use Input$)
Truncate bounds?

Calculate

Time	Lower Bound	Upper Bound	Nominal Upr Alpha	Cum Alpha
1				
2				
3				

Appendix 2: U. Wisconsin Software

Lan-DeMets Group Sequential Calculations

File Compute Help

Compute Bounds

Analysis Parameters

Interim Analyses (k): 3 (1 < k ≤ 25)

Information times (t): User Input (0 < t ≤ 1)

Test Boundaries: Two-Sided Symmetric

Z-Score

Observed Z? No

Spending Function

Overall Alpha: 0.05 (0 < α ≤ 1.0)

Function: O'Brien-Fleming

Truncate bounds? No

Time	Lower Bound	Upper Bound	Nominal Upr Alpha	Cum Alpha
1	0.29			
2	0.55			
3	1.00			

Calculate

Appendix 2: U. Wisconsin Software

Lan-DeMets Group Sequential Calculations

File Compute Help

Compute Bounds

Analysis Parameters

Interim Analysis (k): ($1 < k \leq 25$)

Information Times (t): ($0 < t \leq 1$)

Test Boundaries:

Z-Score

Observed Z?

Spending Function

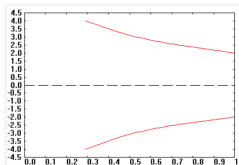
Overall Alpha: ($0 < \alpha \leq 1.0$)

Function:

Truncate bounds?

Time	Lower Bound	Upper Bound	Nominal Upr Alpha	Cam Alpha	
1	0.29	-4.0011	-4.0011	0.00003	0.00006
2	0.55	-2.8074	2.8074	0.00250	0.00502
3	1.00	-1.9740	1.9740	0.02419	0.05000

Calculate



Appendix 2: U. Wisconsin Software

- ▶ Free software says $c_2 = 2.8074$.
- ▶ Last look is at 200th death, $t = 200/200 = 1$.
- ▶ Cumulative alpha to spend by $t = 1$ is

$$\alpha^*(1) = 2 \left\{ 1 - \Phi \left(\frac{2.2414}{\sqrt{1}} \right) \right\} = 0.025. \quad (29)$$

- ▶ Need to find c_3 such that

$$P[\{Z(0.29) \geq 4.0011\} \cup \{Z(0.55) \geq 2.8074\} \cup \{Z(1) \geq c_3\}] = 0.025.$$

- ▶ Free software says $c = 1.9740$.

Appendix 2: U. Wisconsin Software

- ▶ Free software says $c_2 = 2.8074$.
- ▶ Last look is at 200th death, $t = 200/200 = 1$.
- ▶ Cumulative alpha to spend by $t = 1$ is

$$\alpha^*(1) = 2 \left\{ 1 - \Phi \left(\frac{2.2414}{\sqrt{1}} \right) \right\} = 0.025. \quad (29)$$

- ▶ Need to find c_3 such that

$$P[\{Z(0.29) \geq 4.0011\} \cup \{Z(0.55) \geq 2.8074\} \cup \{Z(1) \geq c_3\}] = 0.025.$$

- ▶ Free software says $c = 1.9740$.

Appendix 2: U. Wisconsin Software

- ▶ Free software says $c_2 = 2.8074$.
- ▶ Last look is at 200th death, $t = 200/200 = 1$.
- ▶ Cumulative alpha to spend by $t = 1$ is

$$\alpha^*(1) = 2 \left\{ 1 - \Phi \left(\frac{2.2414}{\sqrt{1}} \right) \right\} = 0.025. \quad (29)$$

- ▶ Need to find c_3 such that

$$P[\{Z(0.29) \geq 4.0011\} \cup \{Z(0.55) \geq 2.8074\} \cup \{Z(1) \geq c_3\}] = 0.025.$$

- ▶ Free software says $c = 1.9740$.

Appendix 2: U. Wisconsin Software

Lan-DeMets Group Sequential Boundaries Calculations

Compute Spending

Analysis Parameters

Interim Analyses (k): 3

Information times(t): User Input

Test Boundaries: Two-Sided Symmetric

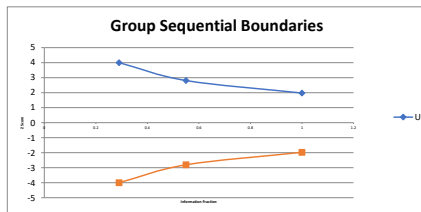
Spending Function

Overall Significance Level: 0.05

Spending Function: O'Brien-Fleming

Truncate Bounds? No

Time	Lower	Upper	Nominal Upr Alpha	Cum Alpha
0.29	-4.0011	4.0011	0.00003	0.00006
0.55	-2.8074	2.8074	0.00250	0.00502
1.00	-1.9740	1.9740	0.02419	0.05000



Appendix 2: U. Wisconsin Software

- ▶ Now use University of Wisconsin software to compute sample size/power for 4 equally-spaced looks using either the O'Brien-Fleming-like spending function or Pocock-like spending function.

Appendix 2: U. Wisconsin Software

- ▶ Use WinLD software at U. of Wisconsin.
- ▶ Click on “Compute” menu and click on “Drift”. Choose “Interim Analyses (k)”, select 4 and hit enter (must hit enter).
- ▶ Table at upper right shows 4 equally-spaced looks (to change to unequal spacings, use “Use Input” under “Information Times”).
- ▶ Choose power level (default is 0.90) under “Power and Bounds Parameters”.
- ▶ Choose spending function under “Spending Function” (default is O’Brien-Fleming).

Appendix 2: U. Wisconsin Software

- ▶ Click on the “Calculate” button.
- ▶ In lower left, under “Drift” is number 3.2711.
- ▶ Tells us that in sample size calculations, make expected z-score 3.2711 instead of $1.96 + 1.28 = 3.24$.
- ▶ Example: For a t-test, expected z-score at end is

$$\frac{\delta}{\sqrt{\frac{2\sigma^2}{N}}}$$

- ▶ Equate to 3.2711 and solve for N . Per-arm sample size is

$$N = \frac{2\sigma^2(3.2711)^2}{\delta^2} \text{ per arm.}$$

Appendix 2: U. of Wisconsin Software

- ▶ If $\delta = 5$ and $\sigma = 14$, $N \approx 168$ per arm.
- ▶ Compare to 165 per arm with no monitoring.
- ▶ Now choose “Pocock” spending function and hit “Calculate”.
 - ▶ New drift is 3.5177.
 - ▶ New per-arm sample size is

$$N = \frac{2(14)^2(3.5177)^2}{5^2} \approx 195.$$

- ▶ More substantial sample size penalty.

Appendix 3: Post-Monitoring Inference: Panoply of Problems

- ▶ Assume nuisance parameters are known.
- ▶ With no monitoring, z-statistic is sufficient statistic.
 - ▶ Inferences should be based solely on Z .
 - ▶ Likelihood ratio for testing $H_0 : \theta = 0$ versus $H_1 : \theta > 0$ is monotone.
 - ▶ Most powerful test against each simple alternative $H_1 : \theta = \theta_1$ is same for all $\theta_1 > 0$.
 - ▶ $Z > z_\alpha$ is UMP level α test for $H_1 : \theta > 0$.
 - ▶ MLE of θ is Z and is unbiased.

Appendix 3: Post-Monitoring Inference: Panoply of Problems

- ▶ These are all problematic with monitoring.

- ▶ If τ is info time when trial stopped, likelihood ratio is

$$L(\theta)/L(0) = \exp \left\{ \theta \sqrt{\tau} Z(\tau) - (\theta^2/2)\tau \right\}.$$

- ▶ Consequently, sufficient statistic is **pair** $\{\tau, Z(\tau)\}$, so inferences should be based solely on $\{\tau, Z(\tau)\}$ or, equivalently, $\{\tau, B(\tau)\}$.
 - ▶ No monotone likelihood ratio, so most powerful test against $H_1 : \theta = 1$ could be different from most powerful test against $H_1 : \theta = 2$ (no UMP test against $H_1 : \theta > 0$).
 - ▶ Must specify how to order sample space to compute p-value to test $H_1 : \theta > 0$.
 - ▶ MLE of θ , $B(\tau)/\tau$, is biased high.

Appendix 3: Post-Monitoring Inference: Panoply of Problems

- ▶ Regarding calculating 1-sided p-value adjusted for monitoring, consider all outcomes consistent with boundaries, and order them in some way:
 - ▶ **MLE ordering** orders by $B(\tau)/\tau$. $\{\tau_2, Z(\tau_2)\}$ more extreme than $\{\tau_1, Z(\tau_1)\}$ if $B(\tau_2)/\tau_2 \geq B(\tau_1)/\tau_1$.
 - ▶ **Z-score ordering** orders by $Z(\tau)$. $\{\tau_2, Z(\tau_2)\}$ more extreme than $\{\tau_1, Z(\tau_1)\}$ if $Z(\tau_2) \geq Z(\tau_1)$.
 - ▶ **B-value ordering** orders by $B(\tau)$. $\{\tau_2, Z(\tau_2)\}$ more extreme than $\{\tau_1, Z(\tau_1)\}$ if $B(\tau_2) \geq B(\tau_1)$.
 - ▶ **Stagewise ordering** orders first by τ , then by $Z(\tau)$. $\{\tau_2, Z(\tau_2)\}$ more extreme than $\{\tau_1, Z(\tau_1)\}$ if $\tau_2 < \tau_1$ or if $\tau_2 = \tau_1$ and $Z(\tau_2) \geq Z(\tau_1)$.
 - ▶ My favorite because does not force us to consider future events to compute p-value.

Appendix 3: Post-Monitoring Inference: Panoply of Problems

- ▶ Regarding estimation, suppose I have 2 independent unbiased estimators, $\hat{\delta}_1$ and $\hat{\delta}_2$, of treatment effect δ .
- ▶ I peek at $\hat{\delta}_1$.
 - ▶ If $\hat{\delta}_1$ is very large, I report only $\hat{\delta}_1$.
 - ▶ If $\hat{\delta}_1$ is not large, I average it with $\hat{\delta}_2$ and report the average.
- ▶ Intuitively clear that this overestimates δ .
- ▶ That is monitoring! Suppose 1 interim analysis at halfway point.
 - ▶ If interim estimate is very large, stop trial and report $\hat{\delta}_1$.
 - ▶ If interim estimate is not large, continue to end, so final estimate is average, $(\hat{\delta}_1 + \hat{\delta}_2)/2$, of 1st and 2nd half estimates.

Appendix 4: Small Sample Sizes

- ▶ Methods so far have considered large samples sizes.
- ▶ What can we do if sample sizes are small?
- ▶ Consider continuous outcomes.
- ▶ Applying Z-score boundaries to the t-statistic is bad because:
 - ▶ Marginal distribution of T-score differs from that of Z-score.
 - ▶ Joint distribution of T-scores differs from that of Z-scores.
- ▶ Better approach: Apply p-value boundaries to p-value from t-distribution. Marginal distribution of p-value is correct (uniform).

Appendix 4: Small Sample Sizes

- ▶ Example: Suppose you monitor continuous outcome using t-statistic with 4 equally-spaced looks and Pocock boundary for 1-sided $\alpha = 0.025$.
- ▶ At first look after 5 people per-arm, compute t-statistic

$$T = \frac{\bar{Y}_T - \bar{Y}_C}{\sqrt{2s^2/5}}.$$

and compute p-value using the t-distribution with $2(5 - 1) = 8$ degrees of freedom.

- ▶ Convert the z-score boundary of 2.361 to p-value boundary 0.0091.
- ▶ Declare benefit if the p-value from t-test is ≤ 0.0091 .

Appendix 4: Small Sample Sizes

- ▶ Another approach with small sample sizes: Use permutation test. Pawitan and Hallstrom (1990)
- ▶ E.g, with spending function $\alpha^*(t)$, determine boundary c_i such that.

$$P_{\text{perm}}(Z(t_1) \geq c_1 \cup \dots \cup Z(t_i) \geq c_i) = \alpha^*(t_i),$$

where $P_{\text{perm}}(A)$ denotes permutation probability of A :

$$P_{\text{perm}}(A) = \frac{\# \text{ permutations such that event } A \text{ occurs}}{\text{total } \# \text{ permutations}}.$$

Lecture 5: Monitoring for Futility

What is futility monitoring?

Interim look at the analysis of the primary endpoint for the purposes of examining whether the trial has a reasonable chance of providing useful scientific evidence

- ▶ Futility analyses often consider the current trend in the data and whether the trial has a reasonable chance of producing a statistically significant result at end of study
- ▶ There could be many factors contributing to possible futility
 - ▶ Lack of treatment effect
 - ▶ Lower than expected recruitment rate
 - ▶ Lower than expected event rate
 - ▶ Emerging evidence from other trials
- ▶ Futility caused by poor recruitment, too much loss to follow-up or non-adherence, is called *operational futility*

When is futility monitoring worth considering?

- ▶ Futility monitoring gives a trial a chance to stop early if there appears to be little to no chance the trial will provide useful evidence
- ▶ Trials for which a failure to show an advantage for a new treatment would not lead to changes in medical practice would be candidates for interim futility assessments
- ▶ Stopping trials on a path to failure prevents patients from taking unnecessary risks
- ▶ Stopping trials on a path to failure saves resources that can be redirected to therapies with better potential
- ▶ Some have argued large publicly funded trials should always consider futility monitoring, saving costs and recruiting fewer patients to failed trials (Sully et al., 2014)

Table 1. Trials That May Incorporate Interim Futility Analysis in Their Monitoring Plans.

Trial Type

Placebo-controlled trials of an investigational treatment for a serious medical condition

Trials comparing an investigational treatment with a standard treatment

Trials comparing a drug combination with one or more of its components

Trials comparing a higher dose with the standard dose of an available treatment

Trials involving considerable participant burden or cost

The Cardiovascular Inflammation Reduction Trial (CIRT)

Background

- ▶ Inflammation plays a key role in atherothrombosis
- ▶ The Canakinumab Anti-inflammatory Thrombosis Outcomes Study (CANTOS) found use of a monoclonal antibody reduced cardiovascular events over placebo, without lowering blood pressure or lipids
 - ▶ Largest reduction in events were in subjects with largest reductions in interleukin-6 and high-sensitivity C-reactive protein
- ▶ There was interest to study another anti-inflammatory drug
 - ▶ Low-dose methotrexate is an inexpensive, effective, and widely used treatment for inflammatory conditions, including rheumatoid arthritis, psoriatic and juvenile arthritis
 - ▶ In observational studies, patients with rheumatoid/psoriatic arthritis who received low-dose methotrexate had fewer cardiovascular events than patients receiving other therapies.

The Cardiovascular Inflammation Reduction Trial (CIRT)

NCT01594333

- ▶ Randomized placebo-controlled trial examining whether low-dose methotrexate reduces heart attacks, strokes, or death
- ▶ NIH-sponsored trial launched in 2013 with goal of randomizing 7000 men and women
- ▶ Trial included medically stable participants with type 2 diabetes or metabolic syndrome and history of a heart attack or multiple coronary blockages
- ▶ Study protocol included a statistical plan for early termination for “futility”
 - ▶ If emerging data indicate trial unlikely to demonstrate benefit, this would not lead to changes in clinical practice

CIRT Futility Analysis

NCT01594333

- ▶ There were two planned looks, when 50% and 75% of the planned primary events had accrued
- ▶ In March of 2018, the DSMB recommended trial termination.
 - ▶ Median follow-up of 2.3 years in 4786 participants
- ▶ The Methotrexate HR crossed a prespecified inefficacy boundary (Freidlin et al. (2010))
- ▶ Methotrexate did not reduce high-sensitivity C-reactive protein levels during the run-in phase
- ▶ Methotrexate elevated liver enzymes

ACTIV-4B Example

A trial that ended for logistical reasons

- ▶ Accelerating Covid-19 Therapeutic Interventions and Vaccines (ACTIV-4B), a placebo-controlled trial testing antithrombotic agents given prophylactically to people with Covid-19 who had not yet been hospitalized
- ▶ Basis for trial was that thrombosis was a known risk for subjects with Covid-19
- ▶ DSMB recommended trial end for futility when it was observed event rate far too low to demonstrate benefit for treatment
 - ▶ 3/558 participants had a thrombotic event (Connors et al., 2021)
- ▶ Decision to conclude for futility also that such a low event rate does not justify the use of anticoagulant or antithrombotic drugs (Ellenberg and Shaw, 2022)

When is futility monitoring **NOT** worth considering?

- ▶ Even if a treatment is showing little benefit partway through the trial, additional safety evidence may be desired
- ▶ Trials comparing two or more widely used therapies to see whether one has advantages over the other
 - ▶ Non-inferiority trials typically would not need futility monitoring, since more about exploring safety profiles and not establishing superiority
- ▶ Some settings would require full evidence to accrue in order to have a convincing null result (as long as ethical)
 - ▶ Trials of therapies already in use may need a convincing result to change practice
- ▶ Although futility monitoring may not have been pre-specified, unexpected events or trends in trial may lead a DSMB to recommend consideration of futility

Testosterone Trials

Snyder et al. (2016)

- ▶ Testosterone Trials studied testosterone therapy in older men with documented subnormal testosterone levels
- ▶ Trial evaluated a widely used product not well studied for many of the functional outcomes it was advertised for
- ▶ It was deemed important to collect as much data as possible
- ▶ Stopping early for futility would be less likely to persuade providers and consumers than a larger database, and full safety information was particularly important given concerns about testosterone's effects on prostate and cardiovascular health
- ▶ No futility plan was provided by the study investigators

Women's Health Initiative (WHI)

Women's Health Initiative Study Group and others (1998)

The WHI Hormone Replacement Therapy (HRT) Trials studied the risks and benefits of estrogen therapy in post-menopausal women.

- ▶ HRT came into wide use in 1960s, based largely on presumptions and observational data
 - ▶ Observational studies suggested HRT reduced a women's risk of coronary heart disease (CHD) by 40-50%
 - ▶ Evidence from trials suggested estrogen prevented hip fracture
 - ▶ Some concern regarding increased risk of breast cancer
- ▶ Two trials launched in 1993 that would enroll a 27,347 women.
 - ▶ 16,608 in the progestin+ estrogen trial (EP) and 10,739 in the estrogen alone (E alone) trial.
- ▶ Primary outcome was CHD; secondary outcome hip fracture; safety outcome breast cancer
- ▶ The WHI HRT trials monitored for efficacy and harm
- ▶ The results of these trials changed clinical practice (Rossouw et al., 2002; Anderson et al., 2004)

Monitoring the Women's Health Initiative (WHI)

Wittes et al. (2007)

- ▶ There were a number of twists and turns in the monitoring of the WHI HRT, including early indications of increased risk of venous thrombosis and stroke
- ▶ In May 2001, the DSMB was convinced neither trial would show a benefit of HRT on CHD
 - ▶ Study continued because there would need to be unequivocal results in order to change practice
 - ▶ More data would also help elucidate some contrary trends: EP seemed to increase or breast cancer, E alone decreased risk
- ▶ In July 2002, 3 years before expected end, its EP arm was halted because compared to placebo, experimental developed more heart disease, invasive breast cancer, and other harmful outcomes such that risks outweighed benefits
- ▶ In Feb 2004, the E alone arm was halted due to concerns of stroke, with no apparent benefit on CHD

Methods for monitoring futility

Conditional Power

Conditional Power (CP) is the probability that a trial would successfully reject the null hypothesis if the trial continued until planned completion.

- ▶ CP useful in settings in which a convincing positive result at the end of the trial is needed to change clinical practice
- ▶ If the conditional power falls below a pre-specified threshold (commonly 10 to 20%), termination for futility may be considered
- ▶ CP must be computed by an unblinded statistician
 - ▶ Calculated using the observed trend in the data so far and a hypothesized trend for the data not yet collected
 - ▶ A DSMB may wish to see several estimates of conditional power based on a range of assumptions about the true treatment effect.
 - ▶ CP using the originally hypothesized trend generally the main estimate of CP
 - ▶ A binding rule w.r.t. CP (i.e., stop if $CP \leq \gamma$ is called **stochastic curtailment**)

Conditional Power (CP)

To calculate the CP at an interim analysis, we again can make use of the **B-value** $B(t) = \sqrt{t}Z(t)$ and **information fraction** t (see Lecture 4)

- ▶ Where $Z(t)$ is the Z-score at the interim analysis and
- ▶ n observations out of the planned N are available at the interim analysis, yielding $t = n/N$

Conditional Power Calculation (part 1)

Proschan (2021)

Suppose c is the critical value needed for significance at end of trial

$$\begin{aligned} CP &= P\{Z(1) > c \mid Z(t) = z\} = P\{B(1) > c \mid B(t) = z\sqrt{t}\} \\ &= P\{B(1) - B(t) > c - z\sqrt{t} \mid B(t) = z\sqrt{t}\} \\ &= \\ &= P\{B(1) - B(t) > c - z\sqrt{t}\} \text{ (independent increments).} \end{aligned}$$

Also, $B(1) - B(t)$ is normal with mean $\theta \cdot 1 - \theta \cdot t = \theta(1 - t)$, where $\theta = E\{Z(1)\}$, and variance

$$\begin{aligned} \text{var}\{B(1) - B(t)\} &= \text{var}\{B(1)\} + \text{var}\{B(t)\} - 2\text{cov}\{B(1), B(t)\} \\ &= 1 + t - 2t \\ &= 1 - t. \end{aligned}$$

Conditional Power Calculation (part 2)

- ▶ Therefore, conditional power is

$$\begin{aligned} CP &= P \left\{ \frac{B(1) - B(t) - \theta(1-t)}{\sqrt{1-t}} > \frac{c - z\sqrt{t} - \theta(1-t)}{\sqrt{1-t}} \right\} \\ &= 1 - \Phi \left\{ \frac{c - z\sqrt{t} - \theta(1-t)}{\sqrt{1-t}} \right\} \\ &= \Phi \left\{ \frac{z\sqrt{t} + \theta(1-t) - c}{\sqrt{1-t}} \right\}, \end{aligned}$$

where z is value of the z -statistic at information time t .

$\theta = E(Z(1))$ and c is critical value at end of study

CP Example: t-test

Proschan (2021)

Consider trial randomizing patients with hepatitis B to new drug (N) and standard (S). Primary outcome: change in \log_{10} viral load between baseline and 1 week expected SD = 2.8, N=250 per arm gives 85% power to detect 0.8 log difference.

At interim analysis: $n_S=108$ and $n_N = 111$, mean change:

$\bar{Y}_S = -0.30$ and $\bar{Y}_N = -0.18$, sd of change: $s_S = 2.35$ and $s_N = 2.60$

so pooled variance = $\frac{(108-1)s_S^2 + (111-1)s_N^2}{108+111-2} = 6.15$

information fraction $t = \frac{1/\text{var}(\hat{\delta})}{1/\text{var}(\hat{\delta}_{\text{end}})} = \frac{1/\{\sigma^2(1/108+1/111)\}}{1/(2\sigma^2/250)}$

or $t \approx (108 + 111)/2/250 = 0.438$

$Z(0.438) = \frac{\bar{Y}_S - \bar{Y}_N}{\sqrt{s^2(1/n_S + 1/n_N)}} = \frac{-0.30 - (-0.18)}{\sqrt{6.15(1/108 + 1/111)}} = -0.358$

$B(0.438) = \sqrt{t}Z(t) = \sqrt{0.438}(-0.358) = -0.237$

CP Example: t-test (part 2)

Conditional power under original alternative hypothesis:

Because original power was 85%, expected Z score at end of trial is

$$\theta = 1.96 + 1.04 = 3$$

$$CP_3 = \Phi\left(\frac{-0.237 + 3(1 - 0.438) - 1.96}{\sqrt{1 - 0.438}}\right) = \Phi(-0.682) = 0.25$$

Conditional power under the null hypothesis:

$$CP_0 = \Phi\left(\frac{-0.237 + 0(1 - 0.438) - 1.96}{\sqrt{1 - 0.438}}\right) = \Phi(-2.931) = 0.002$$

PREVAIL II: Binomial CP Example

Proschan (2021)

PREVAIL II was a trial of Ebola virus disease that compared triple monoclonal antibody product ZMapp + standard of care (SOC) with SOC alone

Primary endpoint was 28-day mortality, target ss was 100/arm that gave $\approx 87\%$ power to detect a 50% reduction in mortality: 40% to 20%. Trial ended with 71 observations because epidemic ended, but an interesting question is would the results have been significant at end of trial?

There were: 13/35 deaths on SOC and 8/36 deaths on SOC+ZMapp.

$$t = \frac{1/\{p(1-p)(1/35+1/36)\}}{100/\{2p(1-p)\}} = 0.355$$

$$Z(.355) = \frac{13/35 - 8/36}{\sqrt{(21/71)(1-21/71)(1/35+1/36)}} = 1.377$$

$$B(.355) = \sqrt{(.355)}Z(.355) = .820$$

$$\theta = E(Z(1)) = \frac{0.40 - 0.20}{\sqrt{(2(.30)(1-.30))/100}} = 3.086$$

PREVAIL II: Binomial CP Example (part 2)

Proschan (2021)

Conditional power under original alternative hypothesis:

$$CP_{orig} = \Phi\left(\frac{0.820 + 3.086(1 - 0.355) - 1.96}{\sqrt{1 - 0.355}}\right) = \Phi(1.059) = 0.86$$

Conditional power under current trend

$$CP_{trend} = \Phi\left(\frac{0.820 + 2.31(1 - 0.355) - 1.96}{\sqrt{1 - 0.355}}\right) = \Phi(0.436) = 0.67$$

PREVAIL II: Survival CP Example

Suppose PREVAIL II had planned to use the logrank test instead of test of proportions and that the trial had 80% power to detect a control to treatment hazard ratio of 1.6, with 142 events.

Suppose at interim analysis, there were 21 deaths observed and that the log-rank zscore was $Z=0.83$

The expected Z score at end $\theta = 1.96 + .84 = 2.80$

Information fraction is the ratio of current number of deaths to the number expected at end of trial $d/D = 21/142 = 0.148$

Then, $B(0.148) = \sqrt{0.148}(0.830) = 0.319$

$$CP_{2.8} \approx \Phi\left(\frac{0.319+2.8(1-0.148)-1.96}{\sqrt{1-0.148}}\right) = \Phi(0.807) = 0.79$$

When amount of information is so low (15% in this case), nearly impossible to stop for futility using conditional power. Note, conditional power is close to the original power in this case.

Recall CAST Example

Cardiac Arrhythmia Suppression Trial (CAST)

- ▶ Tested whether suppressing arrhythmias in patients with prior heart attack reduces composite of sudden deaths/cardiac arrests.
- ▶ Planned for 425 sudden deaths/cardiac arrests by end of trial.
- ▶ At DSMB meeting 4/17/1989, 35 sudden deaths/cardiac arrests in active arm, 13 in placebo arm.

CP can help decision making

CAST Example (Proschan, 2021)

- ▶ Expected z-score at end if treatment reduces sudden deaths/cardiac arrests by 25% $\theta = \ln(4/3) * \sqrt{425/4} = 2.965$.
Note, 25% reduction means treatment/control hazard ratio is 3/4, so control/treatment hazard ratio is $1/(3/4)=4/3$.
- ▶ Information time at interim analysis was $(35+13)/425=0.113$.
- ▶ Z-score was $Z(0.113) = -3.22 = z$ (suggesting harm).
- ▶ Conditional power assuming 25% reduction is

$$\begin{aligned}\Phi\left\{\frac{z\sqrt{t} + \theta(1-t) - c}{\sqrt{1-t}}\right\} &\approx \Phi\left\{\frac{-3.22\sqrt{.113} + 2.965(1-.113) - 1.96}{\sqrt{1-.113}}\right\} \\ &= \Phi(-.438) = 0.33.\end{aligned}$$

- ▶ Only 33% chance of proving benefit at end, given current results and optimistic 25% reduction. CP under null hypothesis = 0.0006.
- ▶ Final # events now predicted to be 300, not 425.
- ▶ Using 300 final events and recomputing θ and t , $CP \approx 0.10$ under 25% reduction and $CP = 0.0002$ under null hypothesis.

Planning for Futility Monitoring

- ▶ Though futility monitoring tends to be less formal, generally good idea to specify out a plan in the protocol
- ▶ How often to monitor will likely depend on setting
 - ▶ Timing often tied to achieving some threshold, such as when 50% and 75% of event rates have accrued
 - ▶ Hard to have low conditional power when less than 25% of data accrued
- ▶ DSMB may ask for an evaluation of futility partway thru the trial, even if not pre-specified

Other factors at play

When evaluating futility, DSMB may consider emerging results from other trials.

- ▶ In the CIRT trial described earlier, the conditional power to detect the originally targeted effect size was 28%.
- ▶ The observed effect sizes in recently completed trials of anti-inflammatory agents in similar populations were much smaller than the targeted effect size in CIRT, increasing the DSMB's concern that CIRT was very unlikely to show a benefit.
- ▶ This emerging outside evidence together with the lack of effect on inflammatory markers, in addition to low CP, contributed to the DSMB's recommendation to terminate the trial for futility.

Predicted Intervals

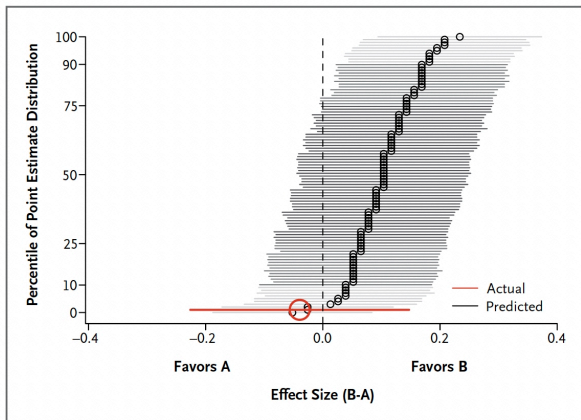
Predicted intervals predict the confidence interval that might be observed at trial's end under a given assumption about the future data. (Evans et al., 2007; Li et al., 2009)

- ▶ Can enhance the interpretation of the conditional power at an interim analysis
- ▶ To calculate predicted intervals you simulate the remaining unobserved data for the trial, under a hypothesized trend, and combine with interim data
- ▶ Generate a large number of such confidence intervals (CI) and plot, ordering by effect size
- ▶ The comparisons of the width of the CI based on observed interim data alone with the width of the predicted interval sheds light on precision that could be gained with trial continuation, a potentially valuable tool for a DSMB
- ▶ Easily done in the R Software with package PIPS

Predicted Intervals: Hypothetical Example

Ellenberg and Shaw (2022)

Predicted interval plot: 100 simulations of a hypothetical trial. At interim analysis: Deaths were 10/40 in group A vs 8/38 in group B halfway through trial. Hypothesized trend: 25% mortality for group A vs 50% mortality for group B for future data. Can see $CP = 20\%$



Are there downsides to monitoring for fertility?

Effect of monitoring for futility on Type I and Type II errors

- ▶ Monitoring for futility increases the probability that there will be a null result at the end of the trial (i.e. increased probability of failing to reject null hypothesis for primary endpoint)
- ▶ Type I error can not be inflated by monitoring for futility
- ▶ Type II error can be inflated

The effect of stochastic curtailment on Power

- ▶ A **stochastic curtailment** rule like stop if CP under originally hypothesized treatment effect < 0.20 lowers power because if you had continued, you might have gotten significant result at end
- ▶ Lan et al. (1982) showed that even if you monitored continuously, power is reduced only by small amount.
 - ▶ If type 2 error rate with no monitoring is β and stochastic curtailment rule stops if CP (**computed under original hypothesis**) $\leq \gamma$, then actual type 2 error rate is no greater than $\beta/(1 - \gamma)$.
- ▶ E.g., suppose stop if $CP \leq 0.20$ and type 2 error rate was $\beta = 0.10$ (power = 0.90) with no monitoring. Actual type 2 error rate is at most $0.10/(1 - 0.20) = 0.125$, so actual power is at least $1 - 0.125 = 0.875$ even if you monitor continuously!

Beta spending

- ▶ Another approach to futility monitoring is to consider a **beta-spending** approach. Analogous to alpha-spending, one could create a lower futility boundary which would guide stopping for futility if this boundary is crossed.
- ▶ The idea is that you can allow for repeated monitoring for futility while controlling the trial's false negative rate, thereby retaining trial power despite the multiple testing
- ▶ Advantage: Can choose the beta spending function of your liking (making it easier or harder to stop early)
- ▶ Tying upper boundary to lower futility boundary allows for slight increase in α , to offset fact that futility monitoring lowers probability of falsely concluding significant result
 - ▶ This must be followed as a binding rule and so generally is NOT recommended.
- ▶ Better to not have upper tied to the lower boundary, since may ignore the lower boundary.

Another downside to futility monitoring: Data in pipeline could change results

Whether the study team preparing report or DSMB member, need to think about a few things regarding stopping for futility

- ▶ Were the data used for the futility analysis the same for the endpoint used for final analysis: i.e. adjudicated endpoints?
- ▶ How much outstanding data was in pipeline at time of interim analysis?
- ▶ How much follow-up time/events will accrue between time of stopping for futility and final analysis?

Example: LUME-2 Lung Trial

Hanna et al. (2016)

- ▶ LUME-Lung 2, a phase III trial of treatment of non–small cell lung cancer
- ▶ This trial was stopped early for futility on the basis of low conditional power (approximately 10%)
- ▶ Final trial results showed a significant benefit for the novel treatment on the primary end point of progression-free survival
- ▶ How did this happen?
 - ▶ Interim analysis based on the investigators' evaluation of disease progression, while final analysis based on centrally adjudicated determination of progression
 - ▶ There was also additional follow-up between stopping for futility and final analysis
 - ▶ Nice discussion by Lesaffre et al. (2017)
 - ▶ Conditional power might have been a useful tool to consider chances for result to turn around.

Other Tools

Revised Power

A **revised power** calculation can be done part-way through a trial using updated information on important design considerations, such as the variance of a continuous outcome, event rate, or the expected recruitment rate.

- ▶ The purpose of revised power is to determine whether a null result at the end of the trial would be informative under the updated assumptions.
- ▶ If the revised power is low, a null result would not rule out the original hypothesized treatment effect, suggesting that continuing the trial may be futile.
- ▶ **Revised power is done blinded**, unlike conditional power. It does not rely on interim trends in the data with respect to treatment effect.

Revised Power: Don't always have to stop...

Love et al. (2015); Hade et al. (2019)

- ▶ An international breast cancer trial was launched in 2013 to study the effects of surgical timing during the menstrual phase on disease-free survival
 - ▶ Prior studies suggested adjuvant oophorectomy surgery during the luteal phase of the menstrual cycle may improve disease-free survival and overall survival compared to the follicular phase
- ▶ Concern arose from emergent data, including another trial, that the event rate assumed for the placebo arm during the planning stage may have been too high, potentially resulting in an underpowered trial
- ▶ In cooperation with the DSMB, investigators agreed to implement a blinded sample size re-estimation that relied on blinded trial data
- ▶ These calculations resulted in an increase in trial size from 340 to 510 due to the lower expected event rate.

Bayesian Approaches

Snapinn et al. (2006); Dmitrienko and Wang (2006)

There are a number of Bayesian approaches to futility monitoring

- ▶ One can compute **predictive power** (Dmitrienko and Wang, 2006), which is the conditional power averaged over a range of assumptions about the treatment difference that will be observed in the future data.
- ▶ The **predictive probability** framework is a fully Bayesian approach that specifies a prior probability for the treatment effect and, using the observed interim data, determines the posterior probability of a clinically important treatment difference
- ▶ For Bayesian approaches, careful thought must be given to specifying the prior probability.
 - ▶ Weak priors may give too much weight to early data
 - ▶ Dmitrienko and Wang (2006) argue that weak priors are advisable in large mortality trials to lessen the exposure of critically ill patients to ineffective interventions

Summary

- ▶ Futility monitoring (FM) allows DSMB to evaluate mid-trial of whether scientifically useful information will be gained if trial continues until the planned end
- ▶ FM makes sense for some settings, not others
- ▶ Conditional Power one of the most common tools used for futility monitoring
- ▶ FM boundaries are generally considered advisory and not binding
- ▶ DSMB will always consider totality of evidence before recommending stopping for futility

Lecture 6: Handling Missing Data

Outline for lecture

- ▶ Review of concepts regarding impact of missing data
 - ▶ Impact on bias and precision
 - ▶ Taxonomy of missing data: MCAR, MAR, MNAR
- ▶ Approaches to analysis in presence of missing data
 - ▶ Approaches to avoid
 - ▶ Recommended approaches
- ▶ Sample R code
- ▶ Misconceptions about missing data (If time allows)

Overview: Impact of Missing data

Conducting an analysis that simply excludes the observations with missing data suffers from:

- ▶ Introduction of bias
- ▶ Loss of precision
- ▶ Lack of internal and face validity, particularly if a large amount of data are missing

Key Questions

- ▶ What are some ways that missing data can cause bias in a study analysis?
- ▶ How do we preserve intent-to-treat principle with missing data?
- ▶ What are the differences between these types of missing data: MCAR, MAR, MNAR?
- ▶ What are the methods of analysis to handle missing data and the assumptions they rely on to be appropriate?

Common Causes of Missing Data

- ▶ Dropout
- ▶ Loss to follow-up
- ▶ Noncompliance with measurement procedure
 - Missed visit
 - Refused procedure/skipped questions or surveys
- ▶ Error
 - Test not done
 - Test done improperly
 - Results lost
- ▶ **Deliberate exclusion from analysis**

THE BIG WORRY ABOUT MISSING DATA

- ▶ Missingness may be associated with outcome and confound analysis
- ▶ We don't know the form of this association
 - Any fix we apply relies on untestable assumptions
- ▶ Nevertheless, if we fail to account for the (true) association, we may bias our results

ANTURANE REINFARCTION TRIAL (ART)

Temple and Pledger (1980)

- ▶ Randomized, double-blind, placebo-controlled trial in post-MI subjects
- ▶ Primary endpoint: mortality

	<u>Anturane</u>	<u>Placebo</u>	<u>Total</u>
Randomized	813	816	1629
Ineligible	38	33	71
Analyzed	775	783	1558

REASONS FOR INELIGIBILITY

- ▶ 1/3 - time since MI: < 25 days or > 35 days
- ▶ 1/3 - enzymes not elevated
- ▶ 1/3 - other: age, enlarged heart, prolonged hospitalization,
- ▶ Number ineligible about the same in each treatment group (38 vs 33)

But....

ANTURANE MORTALITY RESULTS

Temple & Pledger (1980) NEJM, p. 1488

	<u>Anturane</u>	<u>Placebo</u>	<u>P-Value</u>
Randomized	74/813 (9.1%)	89/816 (10.9%)	0.20
“Eligible”	64/775 (8.3%)	85/783(10.9%)	0.07
“Ineligible”	10/38 (26.3%)	4/33 (12.1%)	0.12
P-Values for eligible vs. ineligible	0.0001	0.92	

FDA CONCERNS

- ▶ Apparently haphazard application of “eligibility” criteria by adjudication group
- ▶ Some deaths excluded on drug arm were very similar to deaths not excluded on placebo arm
- ▶ Overall, cause of death classification deemed very unreliable
- ▶ No pre-specified plan to exclude any randomized subjects from analysis
- ▶ No issue of deliberate bias; adjudicators were blinded to treatment arm

HISTORICAL IMPORTANCE OF ART

- ▶ Anturane example established the ITT principle
 - Account for all participants randomized
 - Account for all events during follow up
- ▶ Extremely influential in FDA approach to review

INTENT-TO-TREAT (ITT) PRINCIPLE

All randomized patients should be included in the (primary) analysis, in their assigned treatment groups

A PERENNIAL PROBLEM: NONCOMPLIANCE

- ▶ There will always be noncompliant subjects in clinical trials
- ▶ There will always be people who don't use medical treatment as prescribed
- ▶ Conflict between the question we may WANT to ask, and the question we CAN ask

PRAGMATIC VS EXPLANATORY

- ▶ “Explanatory” question: is the product safe and effective when used appropriately in carefully defined population
 - Scientific question, what pharmaceutical companies and FDA are trying to establish
 - Evaluation of treatment product (treatment efficacy)
- ▶ “Pragmatic” question: Is this product safe and effective when prescribed for its generally intended purpose by practicing physicians?
 - Public health perspective: if we put this product out there, what are the benefits and what are the risks?
 - Evaluation of treatment policy (treatment effectiveness)

Estimands, Estimates, and Estimators

Little and Lewis (2021)

- ▶ **Estimand** - a target quantity; ie, what the study aspires to measure
- ▶ **Estimator** - a formula or algorithm used to estimate the target quantity from the clinical trial data
- ▶ **Estimate** - the numeric value obtained when the estimator is applied to the actual data from the trial.

Another good reference: <https://www.fda.gov/regulatory-information/search-fda-guidance-documents/e9r1-statistical-principles-clinical-trials-addendum-estimands-and-sensitivity-analysis-clinical>

ITT and the Causal Estimand

- ▶ The ITT analysis addresses the treatment policy estimand—difference between randomized groups regardless of compliance
 - ▶ Most common choice in RCTs
- ▶ Some argue ITT analysis is not estimating the ideal causal estimand
 - When analyze outcome regardless of compliance, not getting at “pure” treatment effect
- ▶ An estimand evaluating the causal treatment effect can be much more difficult to estimate
 - Have to address missing data
 - Will involve an unobserved counterfactual outcome
 - May involve a “hypothetical” outcome
- ▶ See Olarte Parra et al. (2023); Lipkovich et al. (2020) for nice discussion of alternative estimands in clinical trials.
 - The *average treatment effect* (ATE) and *average treatment effect among the treated* (ATT) are common alternatives

So why cant we just exclude those who did not take the treatment?

- ▶ Better outcomes in those who follow prescribed regimen could be due to the fact that they actually took the treatment
- ▶ Could be the reverse: those who are better prognostically may be more likely to be good adherers

CLASSIC EXAMPLE

- ▶ Coronary Drug Project (CDP)
 - Large RCT conducted by NIH in 1970's
 - Compared several treatments to placebo
 - Goal: improve survival in patients at high risk of death from heart disease
- ▶ Results disappointing
- ▶ Investigators recognized that many subjects did not fully comply with treatment protocol

CDP: Five-Year Mortality by Treatment Arm

Coronary Drug Project Research Group (1975)

Treatment	N	mortality
clofibrate	1065	18.2
placebo	2695	19.4

CDP: Five-Year Mortality by Adherence to Clofibrate

Coronary Drug Project Research Group, JAMA, 1975

Adherence	N	% mortality
< 80%	357	24.6
≥80%	708	15.0

CDP: Five-Year Mortality by Adherence to Clofibrate and Placebo

Coronary Drug Project Research Group, JAMA, 1975

Clofibrate

Placebo

Adherence	N	%mortality	N	%mortality
<80%	357	24.6	882	28.2
>80%	708	15.0	1813	15.1

Implication of ITT for Study Design and Analysis

- ▶ Implication for design: collect all required data on all patients, regardless of compliance
 - Keep on study even if off treatment
- ▶ Implication for analysis: Need to model the missingness
 - Common approaches: Have to impute either the probability that each participant would have complete data or impute the missing outcome or do an analysis that adjusts for the confounding introduced by missingness
 - To be successful, need to observe the variables associated with missingness
 - This is hard. Thus, avoid missing data

Best method for handling missing data

Little et al. (2012)

PREVENTION

- ▶ The less missing data, the less chance that missing data can impact your analysis
- ▶ There are many practical approaches through study design and implementation to prevent missing data
 - ▶ Have a detailed informed consent process
 - ▶ Minimize participant burden
 - ▶ Think about practicality of study length and endpoints
 - ▶ For long trials or complicated/burdensome procedures, consider a run-in period
 - ▶ Maintain routine contact with participants to maintain interest and up-to-date contact information

HANDLING MISSING VALUES IN ANALYSIS

Many different approaches for handling missing data

- ▶ Exclude subjects with missing values (“Completers” analysis or “complete case” analysis)
- ▶ Last Observation Carried Forward (LOCF)
- ▶ Baseline Observation Carried Forward (BOCF)
- ▶ Group means
- ▶ Inverse probability weighting (IPW)
- ▶ Multiple imputation (MI)
- ▶ Confounder adjustment
- ▶ Causal modeling for desired estimand

Taxonomy for Missing Data

Little et al. (2012); Rubin (1976)

- ▶ Missing completely at random (MCAR)
 - Missingness mechanism depends neither on the observed or unobserved data
- ▶ Missing at random (MAR)
 - Missingness mechanism depends on data that are observed, but not data that are not observed
- ▶ Missing not at random (MNAR), or nonignorable missing

MISSING COMPLETELY AT RANDOM (MCAR)

- ▶ Missing data are independent of any covariates and independent of the true outcome
- ▶ Fact that the data are missing provides no information about outcome, nor could it be predicted from other measured variables
- ▶ Under MCAR, completers analysis will give unbiased estimate of treatment effect
- ▶ Rare that we can be confident that missing data in clinical trials are MCAR

MISSING AT RANDOM (MAR)

- ▶ Missingness is associated with other variables we measure, but once we account for those variables missingness is not associated with the missing value
- ▶ Implication: we can predict the missing outcome in an unbiased way from characteristics of subject, of other subjects, and on observed outcomes of other subjects
 - Two common approaches: multiple imputation (MI), inverse probability weighting (IPW)

MISSING NOT AT RANDOM (MNAR)

- ▶ Missingness depends on unobserved events/characteristics
- ▶ The MNAR setting of concern is when missingness is dependent on the outcome (either directly or through unobserved prognostic variables)
 - ▶ When outcome has MNAR missingness, one cannot make unbiased estimates of the outcome using other measured variables
 - ▶ There can be MNAR missingness in covariates, but if missingness is conditionally independent of outcome given observed data, then wont bias treatment association of interest (Bartlett et al. (2014); Daniel et al. (2012))
- ▶ Incorporation of subjects with missing data in the analysis requires modeling the missing data mechanism
 - Generally takes form of sensitivity analysis of results under different scenarios or a “pattern mixture” modeling

IMPLICATIONS FOR ANALYSIS

▶ MCAR

- Inference for subjects with complete data will be same as for subjects with missing data; can just exclude dropouts

▶ MAR

- Inference for subjects with missing data will be the same after accounting for baseline characteristics and observed data for other participants

▶ MNAR

- Inference requires additional assumptions about missingness mechanism

Simple Example

Clinical Trial outcomes

Arm 1: 155 135 120 175 160 144 190 185 210 180

Arm 2: 125 145 120 185 170 134 200 175 205 185

mean Arm 1: 165; mean Arm 2: 164

Observed Clinical Trials Outcomes

Arm 1: 175 160 144 185 210 180

Arm 2: 125 145 120 185 170 134 175 205 185

mean Arm 1: 176; mean Arm 2: 160

Simple Example

Clinical Trial outcomes

Arm 1: 155 135 120 175 160 144 190 185 210 180

Arm 2: 125 145 120 185 170 134 200 175 205 185

mean Arm 1: 165; mean Arm 2: 164

Observed Clinical Trials Outcomes

Arm 1: 175 160 144 185 210 180

Arm 2: 125 145 120 185 170 134 175 205 185

mean Arm 1: 176; mean Arm 2: 160

EXCLUSIONS OF SUBJECTS

- ▶ Excluding subjects with missing values “completers analysis” is simplest approach
 - Requires assumption that excluded subjects are a random subset of all randomized subjects: MCAR or a special case of MNAR
- ▶ Approach generally not recommended unless amount of missing data is minimal (<5%)
 - Assumptions too difficult to justify
 - Complete case data does not take advantage of the partial data available on excluded participants
- ▶ Unfortunately very common approach; often, no acknowledgement of basic assumptions

IMPUTATION

- ▶ Determine a value that is “best guess” of true value of missing data point
- ▶ Several approaches proposed and/or in use
- ▶ Simplest approaches are generally statistically most problematic
- ▶ Any approach involving “made-up” data is problematic to some degree

SINGLE IMPUTATION APPROACHES

- ▶ Last Observation Carried Forward (LOCF)
 - Use last measurement available in patients with missing data after a certain point
- ▶ Baseline Observation Carried Forward (BOCF)
 - Use the measurement at baseline as final data point
- ▶ Group means
 - Assign average value of outcome variable among those in that treatment group (or in total study population) with complete data

Historically popular, Statistically not recommended

LAST OBSERVATION CARRIED FORWARD (LOCF)

- ▶ Used in trials with repeated measures but where primary comparison is final value to baseline value
- ▶ Concept: whatever the last measurement was prior to dropout, use that as the final value
- ▶ Commonly used in trials evaluating symptom relief for new products

EXAMPLE

- ▶ Trials of new antidepressants generally use the Hamilton Depression (Ham-D) scale as the primary outcome
- ▶ Trials typically last 4-8 weeks; subjects are evaluated weekly
- ▶ Primary comparison is value at final week to baseline value
- ▶ If subject drops out after week 3 evaluation, week 3 score is used to assess effect in that subject

APPEAL OF LOCF

- ▶ It's simple
- ▶ Allows all randomized subjects receiving at least one evaluation to be included in final analysis: looks sort of like ITT
- ▶ Assumption is that dropout is related to lack of effect; scores after dropout may improve if subject initiates active medication
- ▶ In certain settings, some (including regulators) have argued this approach is conservative (unlikely to inflate Type 1 error) but. . . .

PROBLEMS WITH LOCF

- ▶ The last observation before dropout is not a reliable estimate of the effect of the treatment at the desired time point
- ▶ Variance underestimated: get same credit for full sample size as if all data were available
- ▶ Will not always be conservative!
- ▶ Generally regarded as a suboptimal method

WHY IS LOCF NOT ALWAYS CONSERVATIVE?

- ▶ If active treatment has side effects that cause subjects to discontinue, efficacy data may look OK at time of treatment stop but would be expected to worsen once subject is off treatment
- ▶ If tolerability is a major cause of treatment stop, LOCF analysis could overstate treatment benefit
- ▶ Allowing use of entire sample size increases power for detecting differences (that may not be real)

BASELINE OBSERVATION CARRIED FORWARD

- ▶ Same idea as LOCF, but use baseline as final value
- ▶ Often used in pain studies; assumption is that subjects abandoning treatment are getting no relief
- ▶ Idea again is to be conservative; assume no improvement from baseline
- ▶ Big problem: people might have worsened on assigned treatment, so could be anticonservative
- ▶ As with LOCF, get credit for more data than actually have
- ▶ **Generally regarded as a suboptimal method**

GROUP MEANS

- ▶ Assumption: best estimate of effect for individual is the average effect of that individual's treatment group
- ▶ This is equivalent to simply excluding dropouts from analysis, except worse—we give ourselves credit for a larger sample size and so reduce variance
- ▶ **Generally regarded as a suboptimal method**
- ▶ Perhaps more conservative: use average effect of both groups combined or average effect from the control group –but still not accounting for added uncertainty from missing data

MORE SOPHISTICATED: MODEL-BASED Imputation APPROACH (MAR)

- ▶ Model-based imputation: Predict missing outcome on basis of outcomes for other patients with similar characteristics
- ▶ These approaches will yield comparisons with little to no bias if the data are “missing at random” and the models correct

Multiple Imputation vs Single Imputation

- ▶ LOCF, BOCF, group means are all examples of single imputation methods
 - Choose an estimate for the missing data point, and insert that, so that the subject can be included in the analysis
 - This essentially gives you credit for an observation you don't have; results in an underestimate of variability, artificially increases precision of estimate
- ▶ Multiple imputation (MI)
 - A way to account for the extra variability that is inherent in estimating the missing data point

MI: Basic Idea

- ▶ Create a model to predict missing value on basis of covariates and outcomes
 - ▶ Develop this model on those with complete data
- ▶ Apply this model to obtain multiple versions of the “completed data” – impute multiple estimates of possible data points for each missing observation
- ▶ Data analysis proceeds by using each of these completed datasets to get the target estimate, then aggregates results into an overall results
- ▶ Variability among outcomes for subjects with missing data incorporated into overall variability

Simple Example: Regression based imputation

Molenberghs et al. (2014)

Suppose only a continuous outcome is missing

Step 1: Use complete cases to build model and predict outcome Y based on covariates observed on everyone

- ▶ That prediction model will have regression coefficients (α) that have uncertainty and residual unexplained variance (σ^2)

Step 2: Draw $\hat{\alpha}^{(m)}$ and $\hat{\sigma}^{(m)}$ from “posterior” distributions

- Note, a “proper” imputation would take into account the uncertainty of the model parameters and prediction error

Step 3: For each missing y_i , draw $\varepsilon_i^{(m)} \sim N(0, \hat{\sigma}^{(m)})$ and impute (predict) the missing outcome: $y_i^{(m)} = X_i^t \hat{\alpha}^{(m)} + \varepsilon_i^{(m)}$

Step 4: Do outcome regression on completed data to obtain parameter estimate $\hat{\beta}^{(m)}$ of interest

Step 5: Repeat Step 2-4 m times (10,25,50...)

Step 6 : Average β and get Rubin's variance/df

Rubin's Rules

- ▶ Each imputation yields $\hat{\beta}^m$ and $\text{var}(\hat{\beta}^m)$
- ▶ Pool the estimates across the M imputations:
$$\hat{\beta}_{MI} = \text{Avg}(\hat{\beta}^m)$$
$$\text{var}(\hat{\beta}_{MI}) = \text{Avg}(\text{var}(\hat{\beta}^m)) + \text{var}(\hat{\beta}_{MI}) \text{ (Rubin's rules)}$$
- ▶ Note the typical formula for Rubin's rules was for a small number of imputations ($M \approx 5$) and had an adjustment to the 2nd variance term that isn't needed for large M
- ▶ For reproducibility of results, today $M \approx 25$ or larger is much more typical and better for reproducibility
- ▶ A few references for practical considerations are White et al. (2011) and Von Hippel (2020)

More advanced multiple imputation techniques

- ▶ Multivariate Imputation by Chained Equations (MICE) is a popular, flexible and powerful approach (mice package in R (Van Buuren and Groothuis-Oudshoorn (2011)), proc MI in SAS) - this involves specifying a set of equations for how each variable can be predicted.
 - ▶ Don't assume the package defaults are okay for your problem
 - ▶ Generally a Gibbs sampler type algorithm and needs a "burn-in"
 - ▶ There are pitfalls to be aware of with common algorithms: in that you could specify conditional models that are not compatible with a joint distribution, which can lead to inflated variance estimates.
 - ▶ Imputation with survival endpoints needs special consideration
- ▶ *Multiple Imputation of Covariates by Substantive Model Compatible Fully Conditional Specification* is one alternative in R and Stata (smcfcs package) that addresses common pitfalls of MICE (Bartlett et al., 2015)
- ▶ This is a deep topic that needs its own short course

Adjustment Approach for Handling MAR in RCTs

- ▶ In the simple case of missing outcome only, adjustment of the outcome model for covariates will address bias from missingness under MAR and be fully efficient (Groenwold et al. (2012))
 - ▶ Some argue (Sullivan et al., 2018; Little et al., 2022) that non-MI approaches should be considered more often under MAR (e.g. mixed-effects modeling or adjustment method), with caveat MI still offers chance to be more precise
 - ▶ Common approach in longitudinal data (e.g. mixed models)
- ▶ MI approaches allow you to handle MAR and still do the original statistical test (e.g. two group comparison).
- ▶ MI approaches directly satisfy ITT
- ▶ Groenwold et al. (2012) work suggested when MNAR MI could address bias better than adjustment approach, but both biased
- ▶ Generally good to do multiple approaches as sensitivity analyses

INVERSE PROBABILITY WEIGHTING

- ▶ If the probability of having an observed value (i.e., not missing) is closely related to the values of other measured variables, this method can be useful
- ▶ Weights observed values inversely by the probability of being observed
- ▶ Can reduce bias if the assumption that the probability of being observed is a function of other measured variables is reasonable
 - In other words, if we are in a “missing at random” situation
- ▶ Easy to do in software, many standard estimators accept weighted observations in R, Stata, and ...
 - ▶ the R package `survey` has additional tools to handle “design weights”
 - ▶ the IPW package in R has tools for developing IPW weights (van der Wal and Geskus (2011))

Nitty Gritty: IPW

Step 1: Create a binary variable indicating completeness outcome

- 1=complete case
- 0=non-complete case (missing at least one of: outcome or key covariate)

Step 2: Model this completeness outcome (e.g. with logistic regression)

- Include characteristics thought to be related to missingness (need to be observed on most people)

Step 3: Create IPW weights for each person= $1/\text{predicted probability}$ from logistic model

Step 4: Perform a weighted regression using IPW

Step 5: Final analysis should consider variability in the weights: either using a bootstrap or sandwich approach ((Carpenter and Kenward, 2007))

You can examine the success

- ▶ Should evaluate the predictive accuracy of the IPW model.
 - Could consider AUC for the ROC for that to predict the binary outcome complete (yes/no)
 - AUC close to 0.50 means your model could not predict which data were complete
 - Poor AUC means won't have a good adjustment for missingness
- ▶ Analytical concern: sometimes large weights can create instability
 - Look at distribution of weights, check for extreme values
 - Look for influential points in regression
 - More sophisticated algorithms are available that stabilize weights (van der Wal and Geskus (2011))
 - Sometimes large weights will be truncated (e.g. at 10 or 50), but difficult to choose cut off
- ▶ IPW approach can be more variable than MI
 - Classic Bias/variance tradeoff

Approaches for MAR: Take aways

- ▶ Validity of analytical approaches under MAR depend on assumption that we can model the systematic mechanisms driving the missing data
- ▶ Build models to predict missing outcome or to predict probability of being observed
- ▶ IPW seen as more “robust” (relying on fewer assumptions) than MI
- ▶ MI seen as more “efficient” (less variable) than IPW
- ▶ MI approach more common, more intuitive way to preserve ITT and efficiency is attractive
 - ▶ Very flexible MI algorithms now in standard software (e.g. MICE)
- ▶ If equally confident about imputation and missingness models, then choose MI for efficiency

Sample R code: MICE

See Day 3 handout

See Day 3 handout

What to do if data are MNAR?

SENSITIVITY ANALYSES

- ▶ Analyze data under variety of different assumptions regarding missing data—see how much the inference changes
- ▶ Sensitivity analyses could involve any of approaches already described, and others
- ▶ When multiple sensitivity analyses are done without changing conclusions, we can be more comfortable that missing data are not obscuring important information about treatment effect

SIMPLE SENSITIVITY ANALYSIS: Impute the most extreme scenarios

- ▶ Provide bounds for the “true” results if all planned data points had been observed
- ▶ Provides a sense of how far off any of the other analyses could be
- ▶ Does not provide a single “answer” but aids in interpretation of other “answers”

EXAMPLE

- ▶ Suppose outcome for each subject is “response” or “nonresponse”
- ▶ In worst case analysis, assume each missing subject in treatment group has “nonresponse” and each missing subject in control group has “response”
- ▶ In best case analysis, reverse
- ▶ Less straightforward but still feasible when outcome is a continuous rather than a binary variable
 - In weight loss trial, there are biological limits to how much weight could change in 6 months.

More Complex Sensitivity Analyses

- ▶ “Pattern mixture” methods consider how patterns in the missing data may be different than the observed patterns
- ▶ Can show how departures from the MAR assumption can influence conclusions
- ▶ Tipping point analysis: e.g. conjecture differential outcome models until results are reversed. Judge to what extent that scenario would be possible/believable.
- ▶ Using different models and showing minimal impact on conclusions will support primary findings

Great reference:

Little RJ, Carpenter JR, Lee KJ. A comparison of three popular methods for handling missing data: complete-case analysis, inverse probability weighting, and multiple imputation. *Sociological Methods & Research*, 2022.

MISCONCEPTIONS ABOUT HANDLING MISSING DATA

MISCONCEPTIONS ABOUT HANDLING OF MISSING DATA

1. If a dropout rate of $x\%$ is expected, the sample size should be adjusted upward by $x\%$ to compensate.

MISCONCEPTIONS ABOUT HANDLING OF MISSING DATA

1. If a dropout rate of $x\%$ is expected, the sample size should be adjusted upward by $x\%$ to compensate.

Problem: This provides the desired number of data points for precision, but the results may be biased

Example: Suppose dropouts are those who are not benefiting from treatment. Then we are hiding treatment failures by replacing such patients.

Not wrong to enlarge sample size; it just doesn't make the problem go away

Key consideration: to safeguard power, likely need to assume a certain % on tx arm wont respond /has a worse outcome. . . reduces power by bringing arms closer

MISCONCEPTIONS ABOUT HANDLING OF MISSING DATA

2. If the amount of missing data is the same in each study arm, everything is OK

MISCONCEPTIONS ABOUT HANDLING OF MISSING DATA

2. If the amount of missing data is the same in each study arm, everything is OK

Problem: The reasons for data being missing may not be the same for each arm. On one arm, drop outs may be those doing well; on the other, those doing poorly

MISCONCEPTIONS ABOUT HANDLING OF MISSING DATA

3. If the baseline characteristics of subjects with missing data are similar to those of subjects with complete data, it is probably safe to limit analysis to “completers”

MISCONCEPTIONS ABOUT HANDLING OF MISSING DATA

3. If the baseline characteristics of subjects with missing data are similar to those of subjects with complete data, it is probably safe to limit analysis to “completers”

Problem: Many aspects of prognosis are not understood and therefore not measured. The balance on unknown prognostic factors provided by randomization is no longer assured when randomized subjects are excluded.

Note: this is the big concern about methods assuming MAR.

Final note on Reporting

- ▶ CONSORT Guidelines for Clinical Trial Reporting (Altman et al., 2001) have detailed recommendations on how to report missing data
- ▶ Must account for all randomized subjects/all individuals in original cohort
- ▶ Describe pre-specified approach to handling missing data in analysis
- ▶ Address extent to which results may be biased due to missing data
- ▶ Report results of sensitivity analyses

SUMMARY

- ▶ The less missing data, and the lower the rate of noncompliance, the less concern about
 - bias
 - unreliable conclusions
 - inappropriate methods of analysis
- ▶ Methods to replace/account for missing data are all problematic in important ways – but put your best foot forward! (In analysis, try to address bias, incorporate uncertainty)
- ▶ Sensitivity analyses are essential to evaluating reliability of conclusions

KEY REFERENCES

- ▶ Little et al. (2022)
- ▶ Missing data in randomised controlled trials— a practical guide by John Carpenter (https://researchonline.lshtm.ac.uk/id/eprint/4018500/1/rm04_jh17_mk.pdf)
- ▶ National Research Council report: “The Prevention and Treatment of Missing Data in Clinical Trials” (National Academy Press, 2010)
 - Detailed discussion of different methods with examples
- ▶ Little RJ, D’Agostino R, Cohen ML, Dickersin K, Emerson SS, Farrar JT, Frangakis C, Hogan JW, Molenberghs G, Murphy SA, Neaton JD. The prevention and treatment of missing data in clinical trials. *NEJM* 2012;367(14):1355-60.
 - Overview discussion of different methods for handling missing data, with examples
- ▶ Schulz KF, Altman DG, Moher D. CONSORT 2010 statement: updated guidelines for reporting parallel group randomised trials. *Trials*. 2010 Dec;11(1):1-8.

Lecture 7: Multiple Comparisons

Outline

- ▶ Introduction
- ▶ Strong/Weak Control of FWER
- ▶ Methods for Strong Control of FWER
 - ▶ Bonferroni
 - ▶ Holm's Sequentially Rejective Bonferroni
 - ▶ Graphical Alpha Transfer
 - ▶ The Closure Principle
- ▶ Multiple Arms
 - ▶ Dunnett and Variants
- ▶ Multiple Subgroups
 - ▶ Forest Plots
 - ▶ Permutation Methods
- ▶ Multiple Endpoints
 - ▶ Hochberg
- ▶ False Discover Rate (FDR)

Introduction

Introduction

- ▶ See Hochberg and Tamhane (2009) for thorough treatment and Proschan and Brittain (2020) for a tutorial on multiple comparisons.
- ▶ Multiple comparisons arise in many ways in clinical trials: Arms, endpoints, subgroups, analyses, covariates, time points (monitoring).
- ▶ Problem: With enough comparisons, some will be significant even if nothing is going on.
“If you torture the data long enough, it will confess to anything.”
Ronald H. Coase, British economist.
- ▶ Sometimes the size of the multiplicity problem is not clear.

Introduction: Magnitude of Problem

- ▶ In 2003, VaxGen announced their AIDSVAX HIV vaccine may be effective in Blacks and Asian-Americans.
 - ▶ 5,108 men who have sex with men (MSM) and 309 high risk women in North America and the Netherlands.
 - ▶ Overall, only a 3.8% reduction in HIV incidence in vaccine arm compared to placebo ($p=0.76$).
 - ▶ Blacks: 4/203 vaccine versus 9/111 placebo participants got HIV.
 - ▶ Asians: 2/53 vaccine versus 2/20 placebo participants got HIV.
“If we announced to the world that we were abandoning the project because the study failed in whites, we’d be crucified.”
VaxGen CEO Lance Gordon.
- ▶ Very controversial.
- ▶ How many subgroups did they examine?

Introduction: Magnitude of Problem

- ▶ ISIS-2 trial evaluated effect of aspirin on mortality in heart patients (ISIS-2 (1988)).
- ▶ Journal asked for subgroup results, but authors thought it would mislead. Journal insisted, so authors included signs of zodiac to prove a point:

“...for patients born under Gemini or Libra there was a slightly adverse effect of aspirin on mortality (9% SD 13 increase; NS), while for patients born under all other astrological signs there was a strikingly beneficial effect (28% SD 5 reduction; $2p < 0.00001$).”

- ▶ They combined 2 astrological signs and compared to others; if that had not worked, they probably would have combined 3 signs. Number of potential comparisons huge!

Strong/Weak Control of FWER

Strong/Weak Control of FWER

- ▶ Consider testing hypotheses H_1, \dots, H_k .
- ▶ **Global null hypothesis** is $\cap_i H_i$, meaning set of parameter vectors such that all null hypothesis are true. E.g., for comparing all pairs among 3 means, global null is $\{(\mu_1, \mu_2, \mu_3) : \mu_1 = \mu_2, \mu_1 = \mu_3, \mu_2 = \mu_3\} = \{(\mu_1, \mu_2, \mu_3) : \mu_1 = \mu_2 = \mu_3\}$.
- ▶ **Familywise error rate (FWER)** is $P(\text{at least 1 false rejection of a null hypothesis})$.
- ▶ Can calculate FWER under global null or under other nulls.
 - ▶ **Weak control of FWER** means FWER is controlled if global null is true.
 - ▶ **Strong control of FWER** means FWER is controlled no matter which nulls are true.

Strong control \Rightarrow weak control.

Strong/Weak Control of FWER: Fisher's LSD

- ▶ Example: Comparing all pairs of means of arms 1,2,3,4.
- ▶ ***Fisher's least significant difference (LSD) procedure***: If F-test comparing all arms is significant at level α , compare each pair with level α t-test.
- ▶ **For 4 (or more) arms, Fisher's LSD controls FWER weakly, but not strongly.**
- ▶ Weak control: To declare any differences, F must be significant, and $P(F \text{ significant} | \text{global null}) = \alpha$.
- ▶ Therefore, Fisher's LSD controls FWER weakly.

Strong/Weak Control of FWER: Fisher's LSD



Figure: Configuration of 4 population means that inflates the FWER for Fisher's LSD: $\mu_1 = \mu_2 = \mu_3 = L$ and $\mu_4 = U$, where $U - L$ is huge.

- ▶ Lack of strong control of Fisher's LSD with 4 arms: Suppose $\mu_1 = \mu_2 = \mu_3 = L$, $\mu_4 = U$, where $U - L$ is huge.
 - ▶ F is almost guaranteed significant.
 - ▶ In above scenario, Fisher's LSD almost same as doing unadjusted t-tests, which inflates FWER among μ_1, μ_2, μ_3 .
 - ▶ So Fisher's LSD does not strongly control FWER for ≥ 4 arms.

Strong/Weak Control of FWER: Fisher's LSD

- ▶ Sometimes easy to show a procedure strongly controls FWER by enumerating all possibilities.
- ▶ Can use enumeration method to show **Fisher's LSD with 3 arms does strongly control FWER.**
- ▶ Why? Consider all possible configurations of parameters.
 - ▶ If global null is true, $P(F \text{ significant}) \leq \alpha$, so FWER protected.
 - ▶ If global null is false, then at most one pairwise null is true and its t-test protects type 1 error rate if that pairwise null is true.
- ▶ Argument shows FWER protected under all configurations of parameters, so Fisher's LSD strongly controls FWER when $k = 3$ arms.

Methods for Strong Control of FWER

Strong Control Methods: Bonferroni

- ▶ **Bonferroni method always strongly controls FWER.**
- ▶ Bonferroni method requires $P_i \leq \alpha/k$ to reject H_i , where P_i is p-value for i th comparison and k = number comparisons.
- ▶ Why does Bonferroni strongly control FWER?
 - ▶ Suppose t is number of true nulls, & without loss of generality, assume they are the first t .
 - ▶ By Bonferroni inequality,

$$\begin{aligned} P(\cup_{i=1}^t \text{reject } H_i) &\leq \sum_{i=1}^t P(\text{reject } H_i) \\ &\leq t(\alpha/k) \leq k(\alpha/k) \\ &= \alpha. \end{aligned} \tag{30}$$

Strong Control Methods: Bonferroni

- ▶ **Bonferroni too conservative if statistics highly correlated.**
- ▶ For example, if statistics had correlation 1, then could use level α for each.
- ▶ **If statistics are independent, Bonferroni is only slightly conservative:**

$$\alpha - \alpha^2/2 \leq \text{FWER} \leq \alpha. \text{ (proof in Appendix 1)} \quad (31)$$

- ▶ If $\alpha = 0.05$, then $0.04875 \leq \text{FWER} \leq 0.05$ for independent comparisons.

Strong Control Methods: Holm's Sequentially Rejective Bonferroni

▶ *Holm's sequentially rejective Bonferroni method (Holm (1979)):*

- ▶ Order p-values $P_{(1)} < P_{(2)} < \dots < P_{(k)}$.
- ▶ Compare $P_{(1)}$ to α/k .
 - ▶ If $P_{(1)} > \alpha/k$, stop and declare nothing significant.
 - ▶ if $P_{(1)} \leq \alpha/k$, reject null associated with $P_{(1)}$ and proceed to next step.
- ▶ Compare $P_{(2)}$ to $\alpha/(k-1)$.
 - ▶ If $P_{(2)} > \alpha/(k-1)$, stop and declare nothing else significant.
 - ▶ If $P_{(2)} \leq \alpha/(k-1)$, reject null associated with $P_{(2)}$ and proceed to next step.
- ▶ Compare $P_{(3)}$ to $\alpha/(k-2)$.
 - ▶ If $P_{(3)} > \alpha/(k-2)$, stop and declare nothing else significant.
 - ▶ If $P_{(3)} \leq \alpha/(k-2)$, reject null associated with $P_{(3)}$ and proceed to next step.

⋮

Strong Control Methods: Holm's Sequentially Rejective Bonferroni

▶ *Holm's sequentially rejective Bonferroni method (Holm (1979)):*

- ▶ Order p-values $P_{(1)} < P_{(2)} < \dots < P_{(k)}$.
- ▶ Compare $P_{(1)}$ to α/k .
 - ▶ If $P_{(1)} > \alpha/k$, stop and declare nothing significant.
 - ▶ if $P_{(1)} \leq \alpha/k$, reject null associated with $P_{(1)}$ and proceed to next step.
- ▶ Compare $P_{(2)}$ to $\alpha/(k-1)$.
 - ▶ If $P_{(2)} > \alpha/(k-1)$, stop and declare nothing else significant.
 - ▶ If $P_{(2)} \leq \alpha/(k-1)$, reject null associated with $P_{(2)}$ and proceed to next step.
- ▶ Compare $P_{(3)}$ to $\alpha/(k-2)$.
 - ▶ If $P_{(3)} > \alpha/(k-2)$, stop and declare nothing else significant.
 - ▶ If $P_{(3)} \leq \alpha/(k-2)$, reject null associated with $P_{(3)}$ and proceed to next step.

⋮

Strong Control Methods: Holm's Sequentially Rejective Bonferroni

▶ *Holm's sequentially rejective Bonferroni method (Holm (1979)):*

- ▶ Order p-values $P_{(1)} < P_{(2)} < \dots < P_{(k)}$.
- ▶ Compare $P_{(1)}$ to α/k .
 - ▶ If $P_{(1)} > \alpha/k$, stop and declare nothing significant.
 - ▶ if $P_{(1)} \leq \alpha/k$, reject null associated with $P_{(1)}$ and proceed to next step.
- ▶ Compare $P_{(2)}$ to $\alpha/(k-1)$.
 - ▶ If $P_{(2)} > \alpha/(k-1)$, stop and declare nothing else significant.
 - ▶ If $P_{(2)} \leq \alpha/(k-1)$, reject null associated with $P_{(2)}$ and proceed to next step.
- ▶ Compare $P_{(3)}$ to $\alpha/(k-2)$.
 - ▶ If $P_{(3)} > \alpha/(k-2)$, stop and declare nothing else significant.
 - ▶ If $P_{(3)} \leq \alpha/(k-2)$, reject null associated with $P_{(3)}$ and proceed to next step.

⋮

Strong Control Methods: Holm's Sequentially Rejective Bonferroni

▶ *Holm's sequentially rejective Bonferroni method (Holm (1979)):*

- ▶ Order p-values $P_{(1)} < P_{(2)} < \dots < P_{(k)}$.
- ▶ Compare $P_{(1)}$ to α/k .
 - ▶ If $P_{(1)} > \alpha/k$, stop and declare nothing significant.
 - ▶ if $P_{(1)} \leq \alpha/k$, reject null associated with $P_{(1)}$ and proceed to next step.
- ▶ Compare $P_{(2)}$ to $\alpha/(k-1)$.
 - ▶ If $P_{(2)} > \alpha/(k-1)$, stop and declare nothing else significant.
 - ▶ If $P_{(2)} \leq \alpha/(k-1)$, reject null associated with $P_{(2)}$ and proceed to next step.
- ▶ Compare $P_{(3)}$ to $\alpha/(k-2)$.
 - ▶ If $P_{(3)} > \alpha/(k-2)$, stop and declare nothing else significant.
 - ▶ If $P_{(3)} \leq \alpha/(k-2)$, reject null associated with $P_{(3)}$ and proceed to next step.

⋮

Strong Control Methods: Holm's Sequentially Rejective Bonferroni

Death	Mech vent/death	WHO 8-pt ord score
0.042	0.020	0.013

- ▶ Suppose p-values for 3 endpoints in COVID-19 trial are as shown in above table.
- ▶ For Holm:
 - ▶ $p_{(1)} = 0.013 \leq 0.05/3$, so WHO 8-point scale is significant.
 - ▶ $p_{(2)} = 0.020 \leq 0.05/2$, so mechanical ventilation/death is significant.
 - ▶ $p_{(3)} = 0.042 \leq 0.05/1$, so mortality is significant.
- ▶ For ordinary Bonferroni, only WHO 8-point scale is significant.

Strong Control Methods: Holm's Sequentially Rejective Bonferroni

Death	Mech vent/death	WHO 8-pt ord score
0.042	0.020	0.013

- ▶ Equivalent formulation: Adjust p-values and compare to 0.05.

- ▶ In R, first store p-values in a variable:

```
pvals<-c(0.042,0.020,0.013)
```

- ▶ Now use function p.adjust. E.g., for Holm:

```
p.adjust(pvals, method="holm")
```

Adjusted p-values (0.042,0.040,0.039) \leq 0.05; all significant.

- ▶ For ordinary Bonferroni:

```
p.adjust(pvals, method="bonferroni")
```

Gives 0.126, 0.060, 0.039; Only 0.039 is \leq 0.05, so only WHO 8=point ordinal score is significant by ordinary Bonferroni.

Strong Control Methods: Holm's Sequentially Rejective Bonferroni

- ▶ **Holm strongly controls FWER and is more powerful than Bonferroni.**
- ▶ Informal rationale for Holm: Once we reject null associated with $P_{(1)}$, we have either made a type 1 error or not.
 - ▶ If so, it doesn't matter what happens next because we already made ≥ 1 type 1 error.
 - ▶ If not, then null associated with $P_{(1)}$ was false, so there were at most $k - 1$ true nulls. Can use Bonferroni with $\alpha/(k - 1)$.
- ▶ Once we reject null associated with $P_{(2)}$ we have either made ≥ 1 type 1 error or not.
 - ▶ If so, it doesn't matter what happens next because we already made ≥ 1 type 1 error.
 - ▶ If not, then nulls associated with $P_{(1)}$ and $P_{(2)}$ were false, so there were at most $k - 2$ true nulls. Can use Bonferroni with $\alpha/(k - 2)$.

Strong Control Methods: Holm's Sequentially Rejective Bonferroni

- ▶ **Holm strongly controls FWER and is more powerful than Bonferroni.**
- ▶ Informal rationale for Holm: Once we reject null associated with $P_{(1)}$, we have either made a type 1 error or not.
 - ▶ If so, it doesn't matter what happens next because we already made ≥ 1 type 1 error.
 - ▶ If not, then null associated with $P_{(1)}$ was false, so there were at most $k - 1$ true nulls. Can use Bonferroni with $\alpha/(k - 1)$.
- ▶ Once we reject null associated with $P_{(2)}$ we have either made ≥ 1 type 1 error or not.
 - ▶ If so, it doesn't matter what happens next because we already made ≥ 1 type 1 error.
 - ▶ If not, then nulls associated with $P_{(1)}$ and $P_{(2)}$ were false, so there were at most $k - 2$ true nulls. Can use Bonferroni with $\alpha/(k - 2)$.

Strong Control Methods: Holm's Sequentially Rejective Bonferroni

- ▶ **Holm strongly controls FWER and is more powerful than Bonferroni.**
- ▶ Informal rationale for Holm: Once we reject null associated with $P_{(1)}$, we have either made a type 1 error or not.
 - ▶ If so, it doesn't matter what happens next because we already made ≥ 1 type 1 error.
 - ▶ If not, then null associated with $P_{(1)}$ was false, so there were at most $k - 1$ true nulls. Can use Bonferroni with $\alpha/(k - 1)$.
- ▶ Once we reject null associated with $P_{(2)}$ we have either made ≥ 1 type 1 error or not.
 - ▶ If so, it doesn't matter what happens next because we already made ≥ 1 type 1 error.
 - ▶ If not, then nulls associated with $P_{(1)}$ and $P_{(2)}$ were false, so there were at most $k - 2$ true nulls. Can use Bonferroni with $\alpha/(k - 2)$.

Strong Control Methods: Holm's Sequentially Rejective Bonferroni

- ▶ **Holm strongly controls FWER and is more powerful than Bonferroni.**
- ▶ Informal rationale for Holm: Once we reject null associated with $P_{(1)}$, we have either made a type 1 error or not.
 - ▶ If so, it doesn't matter what happens next because we already made ≥ 1 type 1 error.
 - ▶ If not, then null associated with $P_{(1)}$ was false, so there were at most $k - 1$ true nulls. Can use Bonferroni with $\alpha/(k - 1)$.
- ▶ Once we reject null associated with $P_{(2)}$ we have either made ≥ 1 type 1 error or not.
 - ▶ If so, it doesn't matter what happens next because we already made ≥ 1 type 1 error.
 - ▶ If not, then nulls associated with $P_{(1)}$ and $P_{(2)}$ were false, so there were at most $k - 2$ true nulls. Can use Bonferroni with $\alpha/(k - 2)$.

Strong Control Methods: Holm's Sequentially Rejective Bonferroni

- ▶ **Holm strongly controls FWER and is more powerful than Bonferroni.**
- ▶ Informal rationale for Holm: Once we reject null associated with $P_{(1)}$, we have either made a type 1 error or not.
 - ▶ If so, it doesn't matter what happens next because we already made ≥ 1 type 1 error.
 - ▶ If not, then null associated with $P_{(1)}$ was false, so there were at most $k - 1$ true nulls. Can use Bonferroni with $\alpha/(k - 1)$.
- ▶ Once we reject null associated with $P_{(2)}$ we have either made ≥ 1 type 1 error or not.
 - ▶ If so, it doesn't matter what happens next because we already made ≥ 1 type 1 error.
 - ▶ If not, then nulls associated with $P_{(1)}$ and $P_{(2)}$ were false, so there were at most $k - 2$ true nulls. Can use Bonferroni with $\alpha/(k - 2)$.

Strong Control Methods: Graphical Alpha Transfer

- ▶ Same reasoning to transfer alpha from significant comparisons to other comparisons (Bretz et al. (2011)).
 - ▶ Pre-specify initial alpha levels $\alpha_1, \dots, \alpha_k$ with $\sum_i \alpha_i \leq \alpha$.
 - ▶ Pre-specify plan for transferring alpha from significant comparisons to other comparisons.
- ▶ Example: 3 endpoints: (1) Cardiovascular disease (CVD), (2) Coronary heart disease (CHD), (3) stroke.
 - ▶ CVD primary, so initial allocation of alpha: $(\alpha, 0, 0)$ to (CVD, CHD, stroke).
 - ▶ If CVD is significant, transfer all of its alpha to secondary endpoint CHD.
 - ▶ If reach CHD & it is significant, transfer all of its alpha to secondary endpoint stroke.

Strong Control Methods: Graphical Alpha Transfer

- ▶ Called a **gatekeeping procedure**. Must pass through each gate to test subsequent hypothesis.
- ▶ Great if you are confident of passing through gates. E.g., comparing doses of drug to placebo.
 - ▶ Compare high dose to placebo at level α . Stop if not significant.
 - ▶ If significant, compare middle dose to placebo at same level α . Stop if not significant.
 - ▶ If significant, compare low dose to placebo at same level α .
- ▶ Downside: Must stop testing if fail to pass through gate.
- ▶ Another option in 3-endpoint example: If CVD significant, transfer half of alpha to each of CHD & stroke.
- ▶ If reach CHD/stroke & one is significant, transfer all of its alpha to other one.

Strong Control Methods: Graphical Alpha Transfer

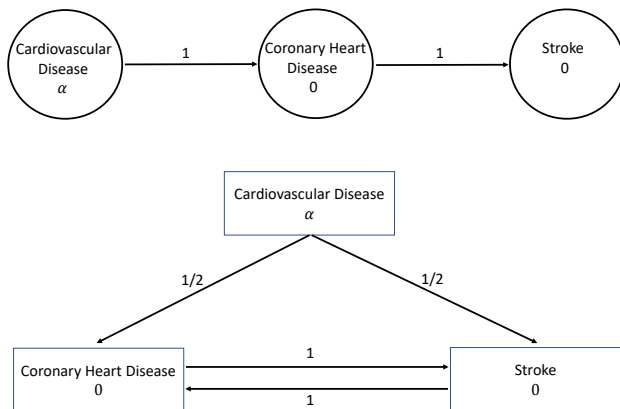


Figure: Two alpha transfer options for primary endpoint (cardiovascular disease) and two secondary endpoints (coronary heart disease, stroke).

Strong Control Methods: Graphical Alpha Transfer

- ▶ Can represent with graph: Initial alpha allocation inside nodes, proportion of alpha transferred shown on arrows.
- ▶ Top option: Gatekeeping procedure.
- ▶ Bottom option with $\alpha = 0.05$: If CVD not significant at $\alpha = 0.05$, stop. If CVD significant at $\alpha = 0.05$, test both secondary endpoints at $\alpha = 0.05/2 = 0.025$.
- ▶ If either secondary is significant at 0.025, transfer all of 0.025 to other secondary.
- ▶ I.e., if primary and one secondary significant, last endpoint is tested at $0.025 + 0.025 = 0.05$.

Strong Control Methods: Graphical Alpha Transfer

- ▶ Bretz et al. (2011) show that **graphical alpha transfer method strongly controls FWER.**
- ▶ Same reason that Holm method strongly controls FWER:
- ▶ **Key Principle: Once you reject a null hypothesis, you either erred or you didn't.**
 - ▶ If you erred, it doesn't matter what happens next because you've already made at least 1 type 1 error.
 - ▶ If you didn't err, then that null hypothesis was false and you never needed to spend any alpha on it. Therefore, you can transfer its alpha! (but only if you follow pre-specified method for transferring).

Strong Control Methods: Graphical Alpha Transfer

- ▶ Another advantage of graphical transfer method: Graph succinctly summarizes what would take pages of text to explain, and the text would be harder to understand!
- ▶ Order of testing doesn't matter provided you continue until no longer able to transfer any alpha, but some orders are more efficient than others.
- ▶ Bretz et al. (2011) give efficient way to carry out procedure.

Strong Control Methods: The Closure Principle

- ▶ Marcus et al. (1976).
- ▶ Testing finite family \mathcal{F} of all intersections of $\{H_{0i}, i = 1, \dots, n\}$.
- ▶ Suppose that for each subset $P \subset \{1, \dots, n\}$, there is an α level test of $H_P = \bigcap_{i \in P} H_i$.
- ▶ **Closure principle:** Reject H_P if and only if the test of H_R is significant at $\alpha = 0.05$ for each H_R that implies H_P (i.e., for each R such that $R \supset P$).

Theorem

(Marcus et al. (1976)) **Following the closure principle strongly controls the FWER among the family \mathcal{F} .**

Strong Control Methods: The Closure Principle

- ▶ Very powerful tool for proofs (used to prove strong FWER control for Holm, gatekeeping, etc.), but sometimes misunderstood.
- ▶ E.g., suppose compare 4 means, $H_{ij} : \mu_i = \mu_j$.
- ▶ **Newman-Keuls procedure:** for comparing pairs of means, $H_{ij} : \mu_i = \mu_j$. Reject H_{12} if and only if T_{12} , F_{123} , F_{124} , and F_{1234} are all significant at level α , where T and F denote t and F statistics. Similarly for other pairwise comparisons.
- ▶ Does Newman-Keuls follow closure principle?
- ▶ No! Closure principle also requires an α level test of $H_{12} \cap H_{34}$.
- ▶ In fact, Newman-Keuls does not strongly control FWER for $k = 4$.

Strong Control Methods: The Closure Principle

● μ_1

● μ_3

● μ_2

● μ_4

L

U

Figure: Configuration of 4 population means that inflates the FWER for Newman-Keuls: $\mu_1 = \mu_2 = L$ and $\mu_3 = \mu_4 = U$, where $U - L$ is huge.

- ▶ F_{123} , F_{124} , F_{134} , F_{234} , and F_{1234} nearly guaranteed to be statistically significant, so FWER of Newman Keuls \approx FWER of rejecting if either T_{12} or T_{34} is significant at level α .

$$\begin{aligned} \text{FWER} &\approx P(T_{12} \text{ signif. or } T_{34} \text{ signif.}) \\ &= 1 - P(\text{neither signif.}) = 1 - (1 - 0.05)^2 \\ &= 0.0975. \end{aligned} \tag{32}$$

- ▶ Doesn't violate closure principle: To truly follow closure, must have α -level test of $H_{12} \cap H_{34}$. Newman-Keuls doesn't do this.

Multiple Arms

Multiple Arms: Dunnett and Variants

- ▶ Consider 1-sided problem of comparing k arms to same control (platform trial) on continuous endpoint Y .
- ▶ $H_i : \mu_i = \mu_0$, μ_0 and μ_i are means in control and arm i .
- ▶ Global null is $H = \cap H_i : \mu_0 = \mu_1 = \dots = \mu_k$.
- ▶ Poor option: Bonferroni.
 - ▶ Reject hypothesis associated with $p_{(i)}$ if $p_{(i)} \leq \alpha/k$.
- ▶ Better option: Holm sequentially rejective Bonferroni using p-value thresholds of $\alpha/k, \alpha/(k-1), \dots, \alpha/1$ for $p_{(1)}, \dots, p_{(k)}$.
- ▶ Better still: **Dunnett procedure** Dunnett (1955): Find actual joint distribution of test statistics using clever trick.

Multiple Arms: Dunnett and Variants

- ▶ Assume Y normally distributed with same variance σ^2 in different arms, and assume n large enough to treat σ^2 as known.
- ▶ $Z_i = \frac{\bar{Y}_i - \bar{Y}_0}{\sqrt{2\sigma^2/n}}$, $i = 1, \dots, k$.
- ▶ Z_i are dependent only because of shared control. Can condition on control sample mean to remove dependence, then uncondition at end.
- ▶ See Appendix 2 at end of this lecture for details of derivation of Dunnett critical value.

Multiple Arms: Dunnett and Variants

Table: Dunnett critical values c_k for 1-sided test at $\alpha = 0.025$ for large n .

k	Z-score boundary
1	1.960
2	2.212
3	2.349
4	2.442
5	2.511
6	2.567

- ▶ Substitute pooled variance $\hat{\sigma}^2$ for σ^2 in Z_j .

Multiple Arms: Dunnett and Variants

- ▶ **Even better option: Use sequential Dunnett based on closure principle. Let $Z_{(1)} < Z_{(2)} < \dots < Z_{(k)}$ be order statistics:**
 - ▶ Compare $Z_{(k)}$ to c_k .
 - ▶ If $Z_{(k)} < c_k$, stop.
 - ▶ If $Z_{(k)} \geq c_k$, reject hypothesis associated with $Z_{(k)}$ and proceed to next step.
 - ▶ Compare $Z_{(k-1)}$ to c_{k-1} .
 - ▶ If $Z_{(k-1)} < c_{k-1}$, stop.
 - ▶ If $Z_{(k-1)} \geq c_{k-1}$, reject hypothesis associated with $Z_{(k-1)}$ and proceed to next step.
 - ▶ Compare $Z_{(k-2)}$ to c_{k-2} .
 - ▶ If $Z_{(k-2)} < c_{k-2}$, stop.
 - ▶ If $Z_{(k-2)} \geq c_{k-2}$, proceed to next step.
 - ▶ \vdots

Multiple Arms: Dunnett and Variants

Table: Dunnett critical values c_k for 1-sided test at $\alpha = 0.025$ for large n .

k	Z-score boundary
1	1.960
2	2.212
3	2.349
4	2.442
5	2.511
6	2.567

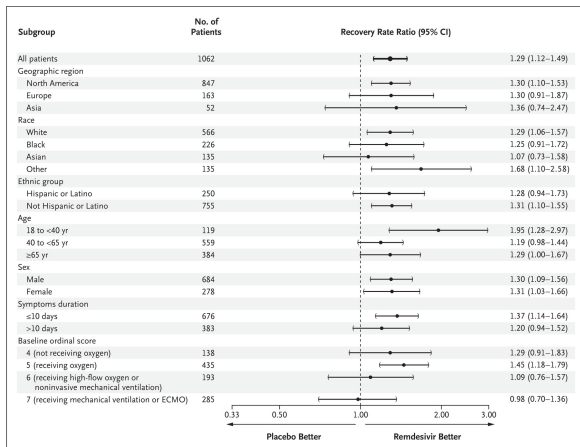
- ▶ Example of sequential Dunnett comparing 3 arms to control: Start by comparing $\max_{i=1,2,3} Z_i$ to 2.349.
- ▶ Suppose max is Z_2 and $Z_2 \geq 2.349$. Declare Arm 2 better than control and compare $\max_{i=1,3} Z_i$ to 2.212.
- ▶ Suppose $\max_{i=1,3} Z_i$ is Z_1 and $Z_1 \geq 2.212$. Declare Arm 1 better than control and compare Z_3 to 1.96.
- ▶ Declare Arm 3 better than control if $Z_3 \geq 1.96$.

Multiple Subgroups

Multiple Subgroups

- ▶ Subgroups difficult to deal with because of their sheer number:
 - ▶ Demographics: race, sex, age, etc.
 - ▶ Prognostic characteristics: prior heart attack or family history in heart disease trial, risky behaviors in HIV vaccine trial, etc.
- ▶ Two opposing problems: (1) Multiplicity leads to false positives and (2) Low power to detect interactions, exacerbated by adjusting for multiple comparisons!
- ▶ Typical approaches either don't control, or weakly control, FWER.
- ▶ E.g., for each factor, first test (factor) \times (treatment) interaction. Only if interaction is significant, test treatment effect separately within subgroups. Doesn't control FWER, even weakly.
- ▶ Accompanying plot: ***Forest plot***.

Multiple Subgroups: Forest Plots



NEJM 2020;
383:1813-26

Figure: Forest plot from the ACTT-1 clinical trial of Remdesivir for treatment of hospitalized COVID-19.

Multiple Subgroups: Forest Plots

- ▶ Forest plots often show good agreement of treatment effect across subgroups, but there is no FWER control.
- ▶ Could control FWER using Bonferroni on number of subgroup comparisons.
- ▶ Problem: Too conservative because subgroups based on prognostic factors can have big overlap.
 - ▶ E.g., in HIV, big overlap of patients in worst viral load and worst CD4 count subgroups.
- ▶ How do we control FWER weakly, accounting for overlap?
- ▶ There are bootstrap (Rosenkranz (2014)) and permutation (Dane et al. (2019); Lipkovich et al. (2011)) methods. We focus on one permutation method.

Multiple Subgroups: SEAMOS

- ▶ **Standardized Effects Adjusted for Multiple Subgroups (SEAMOS)** (Dane et al. (2019)).

- ▶ Compute standardized difference between effect $\hat{\delta}_{ij}$ in category j of factor i and overall effect $\hat{\delta}$ that ignores subgroups, then take max:

$$Z_{ij} = \frac{\hat{\delta}_{ij} - \hat{\delta}}{\text{se}(\hat{\delta}_{ij})}, \quad M = \max_{ij}(Z_{ij}). \quad (33)$$

- ▶ Fix outcome & treatment vectors, randomly interchange covariate vectors across people, compute M for covariate-interchanged data.
- ▶ Repeat many times to get null reference distribution of M . Declare effect in a subgroup better than average if original (unpermuted) M exceeds 97.5th percentile of permutation distribution.
- ▶ SEAMOS preserves overall treatment effect and dependence of covariates.

Multiple Subgroups: SEAMOS

SEAMOS Method

Patient	Treatment Assignment	Outcome (28-day mortality)	Ordinal Score	Other Covariates
1	T	0	4	x_1
2	P	1	6	x_2
3	P	0	5	x_3
4	T	0	5	x_4
5	T	0	4	x_5
6	P	1	7	x_6
\vdots	\vdots	\vdots	\vdots	

Fix

Figure: SEAMOS method illustrated for mortality comparison in ACCT-1 trial: Fix outcomes and treatment indicators, randomly permute patient covariate vectors, compare original $M = \max_{ij}(Z_{ij})$ from Equation (33) to its permutation distribution.

Multiple Subgroups: SEAMOS

SEAMOS applied to Adaptive COVID-19 Treatment Trial (ACTT 1)

Table: Mortality results from ACTT-1.

Subgroup	Remdesivir	Placebo	HR (95% CI)
OS-4 (no ox.)	3/75	3/63	0.82 (0.17, 4.07)
OS-5 (suppl. ox.)	9/232	25/203	0.30 (0.14, 0.64)
OS-6 (non-inv. vent.	19/95	20/98	1.02 (0.54, 1.91)
OS-7 (inv. mech. vent.)	28/131	29/154	1.13 (0.67, 1.89)

Multiple Subgroups: SEAMOS

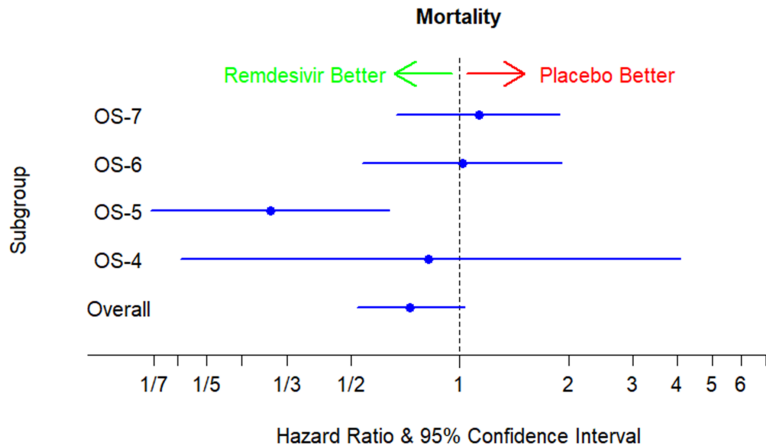


Figure: Forest plot of hazard ratios for mortality in ACTT-1 by baseline ordinal score..

Multiple Subgroups: SEAMOS

SEAMOS applied to Adaptive COVID-19 Treatment Trial (ACTT 1)

Table: Mortality results from ACTT-1.

Subgroup	Remdesivir	Placebo	HR (95% CI)
OS-4 (no ox.)	3/75	3/63	0.82 (0.17, 4.07)
OS-5 (suppl. ox.)	9/232	25/203	0.30 (0.14, 0.64)
OS-6 (non-inv. vent.	19/95	20/98	1.02 (0.54, 1.91)
OS-7 (inv. mech. vent.)	28/131	29/154	1.13 (0.67, 1.89)

Adjusted p-value comparing OS-5 to overall effect using SEAMOS:
0.005: Effect in OS-5 is different from overall effect

Multiple Endpoints

Multiple Endpoints: Hochberg*

- ▶ Hochberg (1988) is used to test whether effect in any subgroup.
- ▶ Holm's procedure started with most significant (smallest p-value) result and proceeded to less significant results.
- ▶ Can also go in reverse direction.
- ▶ Start with largest p-value, $p_{(k)}$ and compare to α .
 - ▶ If $p_{(k)} \leq \alpha$, reject ALL hypotheses and stop.
 - ▶ If $p_{(k)} > \alpha$, do not reject hypothesis associated with $p_{(k)}$, but move to next step.
- ▶ Compare $p_{(k-1)}$ to $\alpha/2$.
 - ▶ If $p_{(k-1)} \leq \alpha/2$, reject hypotheses associated with $p_{(k-1)}, \dots, p_{(1)}$.
 - ▶ If $p_{(k-1)} > \alpha/2$, do not reject hypothesis associated with $p_{(k-1)}$, but move to next step.
- ▶ Compare $p_{(k-2)}$ to $\alpha/3 \dots$

*Not guaranteed to strongly control FWER without further conditions

Multiple Endpoints: Hochberg*

- ▶ Called **Hochberg** procedure.
- ▶ **P-value thresholds for Holm and Hochberg are same, but Holm starts with most significant, while Hochberg starts with least significant.**



Figure: Holm and Hochberg procedures.

*Not guaranteed to strongly control FWER without further conditions

Multiple Endpoints: Hochberg*

- ▶ Hochberg always more powerful than Holm, which is more powerful than Bonferroni.
- ▶ Problem: Hochberg does not always strongly control FWER.
- ▶ However, Hochberg DOES strongly control FWER when test statistics independent or have certain types of positive dependence.

*Not guaranteed to strongly control FWER without further conditions

Multiple Endpoints: Hochberg*

- ▶ **Hochberg is attractive for multiple positively correlated endpoints.**
- ▶ Example: COVID-19 trial with primary endpoints 1. mechanical ventilation or death, 2. death.
- ▶ Declare benefit on at least one endpoint if
 - ▶ smaller p-value ≤ 0.025 for Bonferroni/Holm procedure,
 - ▶ smaller p-value ≤ 0.025 **or** both p-values ≤ 0.05 for Hochberg procedure.
- ▶ If the two endpoints are positively correlated, then Hochberg procedure strongly controls FWER.

*Not guaranteed to strongly control FWER without further conditions

False Discovery Rate (FDR)

False Discovery Rate (FDR)

- ▶ Another error rate is **false discovery rate**.

Table:

	Declared not significant	Declared significant	
H_0 True	T	V	m_0
H_0 Untrue	U	W	$m - m_0$
	$m - X$	X	m

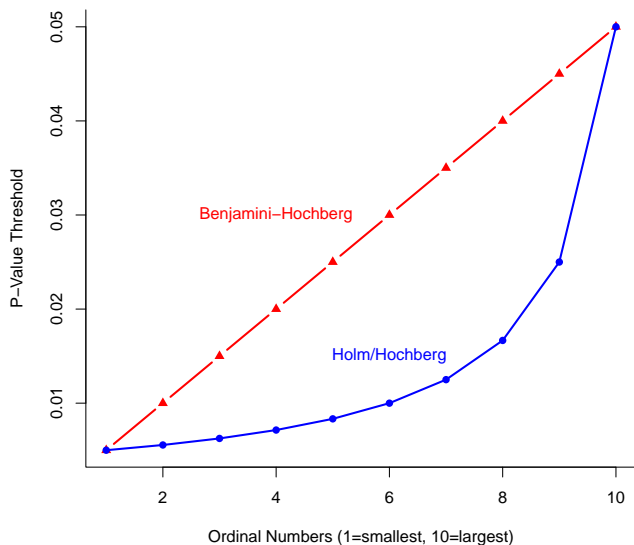
- ▶ $Q = V/X$, proportion of nulls that are true among rejections (defined as 0 if $X = 0$).
- ▶ $FDR = E(Q)$, expected proportion of nulls that are true among rejections.
- ▶ $FDR \leq FWER$ (FDR less stringent). See Appendix 3

False Discovery Rate (FDR)

- ▶ Can have $FDR \leq \alpha$ and $FWER > \alpha$ but not vice-versa.
- ▶ FDR control is reasonable when # comparisons very large.
- ▶ **Benjamini-Hochberg procedure** (Benjamini and Hochberg (1995)).
- ▶ Reject hypotheses associated with $p_{(1)}, \dots, p_{(j)}$ at level q^* if $p_{(j)} \leq (j/k)q^*$.
- ▶ If test statistics are independent, Benjamini-Hochberg controls FDR at level q^* .

False Discovery Rate (FDR)

Benjamini–Hochberg Versus
Holm/Hochberg, k=10 Comparisons



False Discovery Rate (FDR)

- ▶ Note that under global null, if test statistics independent and have continuous distribution function, $E(P_{(j)}) = E(U_{(j)}) = j/(k+1)$, where U_j are iid uniforms,
- ▶ In that case, Benjamini-Hochberg rejects hypotheses associated with $p_{(1)}, \dots, p_{(j)}$ at level q^* if

$$\frac{p_{(j)}}{E(P_{(j)})} \leq \left(\frac{k+1}{k}\right) q^* \approx q^* \text{ if } k \text{ large.}$$

- ▶ In terms of stringency,

Benjamin-Hochberg \leq Hochberg \leq Holm \leq Bonferroni.

- ▶ FDR usually reserved for large # comparisons (e.g., gene association studies).

Summary

- ▶ Multiple comparisons in clinical trials caused by multiple arms, endpoints, analyses, subgroups, time points (monitoring), etc.
- ▶ $P(\text{at least one type 1 error}) = \text{FWER}$. Can control it:
 - ▶ Weakly: Under global null.
 - ▶ Strongly: Regardless of which nulls are true.
- ▶ Bretz et al. (2011) graphical method of transferring alpha from rejected hypotheses to other hypotheses is very powerful. Several techniques are special cases, including
 - ▶ Gatekeeping procedures.
 - ▶ Sequentially rejective Bonferroni.
- ▶ Closure principle also useful for proving strong control of FWER.
- ▶ With huge number of comparisons, may want to control FDR (less conservative) instead of FWER.

Appendix 1: Bonferroni with Independent Test Statistics

Use Bonferroni's other inequality (p. 100 of Feller (1957)):

$$\begin{aligned}FWER &= P(\cup_{i=1}^k \text{reject } H_i) \\ &\geq \sum_{i=1}^k P(\text{reject } H_i) - \sum_{1 \leq i < j \leq k} P(\text{reject } H_i \cap \text{reject } H_j) \\ &= k(\alpha/k) - \binom{k}{2}(\alpha/k)^2 \\ &\geq \alpha - \alpha^2/2. \text{ Thus,} \\ &\quad \alpha - \alpha^2/2 \leq FWER \leq \alpha.\end{aligned}\tag{34}$$

Appendix 2: Deriving Distribution of Dunnett

- ▶ Assume Y normally distributed with same variance σ^2 in different arms, and for simplicity, assume n large enough to treat σ^2 as known.
- ▶ $Z_i = \frac{\bar{Y}_i - \bar{Y}_0}{\sqrt{2\sigma^2/n}}, i = 1, \dots, k.$
- ▶ To compute FWER, assume without loss of generality that $Y_i \sim N(0, 1), i = 1, \dots, k.$
- ▶ Z_i are dependent only because of shared control. Condition on control sample mean to remove dependence.

Appendix 2: Deriving Distribution of Dunnett

Given $\bar{Y}_0 = y_0$,

$$Z_1 = \frac{\bar{Y}_1 - y_0}{\sqrt{2/n}}, Z_2 = \frac{\bar{Y}_2 - y_0}{\sqrt{2/n}}, \dots, Z_k = \frac{\bar{Y}_k - y_0}{\sqrt{2/n}} \text{ are iid.}$$

Let $M = \max(Z_1, \dots, Z_k)$. Then

$$\begin{aligned} P(M \leq c_k | \bar{Y}_0 = y_0) &= P\left\{ \max(\sqrt{n}\bar{Y}_1, \dots, \sqrt{n}\bar{Y}_k) \leq \sqrt{2}c_k + \sqrt{n}y_0 \right\} \\ &= \left\{ \Phi\left(\sqrt{2}c_k + \sqrt{n}y_0\right) \right\}^k. \end{aligned}$$

Appendix 2: Deriving Distribution of Dunnett

- ▶ Now integrate over density, $f_n(y_0)$, of \bar{Y}_0 , which is normal with mean 0 and variance $1/n$.

$$P(M \leq c_k) = \int_{-\infty}^{\infty} \left\{ \Phi \left(\sqrt{2} c_k + \sqrt{n} y_0 \right) \right\}^k f_n(y_0) dy_0. \quad (35)$$

- ▶ Make substitution $z_0 = \sqrt{n} y_0$. Because $Z_0 = \sqrt{n} \bar{Y}_0 \sim N(0, 1)$, get

$$P(M \leq c_k) = \int_{-\infty}^{\infty} \left\{ \Phi \left(\sqrt{2} c_k + z_0 \right) \right\}^k \phi(z_0) dz_0,$$

where $\phi(z_0)$ is standard normal density.

- ▶ Equate to $1 - \alpha$ and solve for c_k .

Appendix 3: Proof that if $\text{FWER} \leq \alpha$, then $\text{FDR} \leq \alpha$

Table:

	Declared not significant	Declared significant	
H_0 True	T	V	m_0
H_0 Untrue	U	W	$m - m_0$
	$m - X$	X	m

- ▶ Claim: $I(V \geq 1) \geq Q = V/X$ because:
 - ▶ If $V = 0$, then $I(V \geq 1) = 0$ and $Q = V/X = 0$ (even if $X = 0$).
 - ▶ If $V \geq 1$, then $I(V \geq 1) = 1$ and $Q = V/X \leq 1$.
- ▶ Because $I(V \geq 1) \geq Q$, $\text{FWER} = E\{I(V \geq 1)\} \geq E(Q) = \text{FDR}$.

Lecture 8: Adaptive Procedures

- ▶ Introduction
- ▶ Blinded Sample Size Methods
 - ▶ Binary Outcomes
 - ▶ Continuous Outcomes
- ▶ **Unblinded Sample Size Methods**
- ▶ **Unplanned Changes**

Introduction

Introduction

- ▶ **WARNING:** This is a brief overview of a vast topic! Only the main ideas are provided. See chapter 11 of Proschan (2022).
- ▶ **Assume 1-sided, level α testing rejecting H_0 for large Z .**
- ▶ Clinical trial planning requires information such as:
 - ▶ The control event rate (for binary outcome) or variance (for continuous outcome).
 - ▶ The treatment effect.
 - ▶ Whether data will be skewed, etc.
- ▶ This information is lacking in new disease.
- ▶ What do we do?
- ▶ Adaptive methods: **Pre-specified** procedures to use within-trial data to change design.

Introduction

- ▶ Examples:
 - ▶ Sample size re-estimation based on lumped event rate (binary outcome trial) or lumped variance (continuous endpoint trial).
 - ▶ Examine lumped data to detect outliers. If none, use t-test; if outliers, use Wilcoxon rank sum test.
 - ▶ **Sample size re-estimation based on treatment effect.**
 - ▶ **Change of primary endpoint after examining treatment effect.**
- ▶ First two are very different from last two: They don't require unblinding.
- ▶ Unblinding is undesirable because could lead to:
 - ▶ Background treatment bias, selection bias, etc.
 - ▶ Problems with regulators.
 - ▶ Inflation of type 1 error rate if not careful.

Blinded Sample Size Methods

Blinded Sample Size Methods

- ▶ We begin with binary outcomes and safest adaptations—blinded ones (Gould (1992)).
- ▶ In trials with binary endpoint like 28-day mortality, we often over-estimate event rate. Why?
- ▶ Event rate estimates may be based on observational data. People volunteering for clinical trials:
 - ▶ May be more health-conscious.
 - ▶ May get better care in clinical trial.
 - ▶ Must satisfy entry criteria that could exclude sickest patients.
 - ▶ May undergo run-in to weed out non-adherers.
- ▶ **If power is based on a given relative effect (e.g., 25% reduction in 28-day mortality), overestimating event rate means underpowering trial.**

Blinded Sample Size Methods: Binary Outcomes

- ▶ Want specified power to detect **fixed** relative risk $R < 1$. E.g., for 25% reduction, $R = 0.75$.
- ▶ **Procedure:**
 - ▶ Calculate initial sample size n_0 /arm based on fixed relative risk $R < 1$ and pre-trial estimate of control probability p_C .
 - ▶ At planned halfway point, $n_1 = n_0/2$ per arm (or another fraction), calculate overall event rate (blinded)
 $\hat{p}_1 = (\# \text{ with event}) / (\# \text{ evaluated})$.
 - ▶ Equate

$$\begin{aligned}(p_T + p_C)/2 &= \hat{p}_1 \\ p_T/p_C &= R,\end{aligned}\tag{36}$$

and solve for p_T and p_C .

$$p_C = \frac{2\hat{p}_1}{R+1}, \quad p_T = \frac{2R\hat{p}_1}{R+1}.\tag{37}$$

Blinded Sample Size Methods: Binary Outcomes

► Review of Notation:

- n_0 : Initial (pre-trial) guess of per-arm sample size at end (fixed).
- n_1 : Per-arm sample size at interim analysis (stage 1 sample size),
 $n_1 = n_0/2$ (fixed).
- n_2 : Per-arm sample size in 2nd stage (random).
- $n = n_1 + n_2$: Actual per-arm sample size at end (random).
- p_T and p_C : True treatment and control probabilities (fixed),
- $\hat{p}_1 = (\hat{p}_{C1} + \hat{p}_{C2})/2$: Lumped empirical event probability in stage 1 (random).
- R : Assumed relative risk (fixed).

Blinded Sample Size Methods: Binary Outcomes

- ▶ Example: COVID-19 trial of hospitalized patients, usual care+new drug versus usual care+placebo.
- ▶ Primary endpoint: Mechanical ventilation/death by day 60.
- ▶ Initial estimate of control probability: $p_C = 0.20$.
- ▶ Want 85% power to detect 25% reduction ($R = 1 - 0.25 = 0.75$) using 2-tailed test at $\alpha = 0.05$.
- ▶ Initial sample size: $n_0 = 1,036/\text{arm}$.
- ▶ After $n_1 = 500/\text{arm}$, 150 have events (combined across arms), so proportion with events is $\hat{p}_1 = 150/1,000 = 0.15$.

Blinded Sample Size Methods: Binary Outcomes

- ▶ Compute

$$p_C = \frac{2\hat{p}_1}{R+1} = \frac{2(0.15)}{0.75+1} = 0.1714$$

$$p_T = \frac{2R\hat{p}_1}{R+1} = \frac{2(0.75)(0.15)}{0.75+1} = 0.1286.$$

- ▶ New sample size based on $p_C = 0.1714$ and $p_T = 0.1286$:
 $n = 1249/\text{arm}$.
- ▶ **Increase sample size to 1249/arm and use z-test of proportions at end as if sample size were fixed in advance.**
- ▶ Could also stratify analysis by before/after sample size change (always good to compare results pre- and post-adaptation).

Blinded Sample Size Methods: Binary Outcomes

- ▶ One problem with blinded sample size method: Overall event rate could be low for at least 2 reasons:
 1. Control event rate is lower than expected.
 2. Treatment effect is much higher than expected.
- ▶ If second reason is true, then don't need to increase sample size!
- ▶ But very large treatment effects are unusual in clinical trials.
- ▶ Why not avoid problem by looking at control event rate instead of overall rate?
- ▶ **Problem 1 with peeking at control event rate: Even blinded investigators may know overall event rate.**
 - ▶ **knowledge of control event rate plus overall rate reveals the observed treatment effect!**

Blinded Sample Size Methods: Binary Outcomes

- ▶ **Problem 2 with peeking at control event rate: Even if investigators don't know overall event rate, control event rate gives some information about treatment effect because control event rate is correlated with observed treatment effect:**

$$\begin{aligned}\text{cov}(\hat{p}_{C1}, \hat{p}_{T1} - \hat{p}_{C1}) &= \text{cov}(\hat{p}_{C1}, \hat{p}_{T1}) - \text{cov}(\hat{p}_{C1}, \hat{p}_{C1}) \\ &= 0 - \text{var}(\hat{p}_{C1}) \\ &= -p_C(1 - p_C)/n_1.\end{aligned}\tag{38}$$

- ▶ Dividing by product of standard deviations gives correlation $\rho \approx -0.71$ under the null hypothesis.
- ▶ Nonzero correlation implies that the type 1 error rate could be inflated by procedure that allows peeking at control event rate.

Blinded Sample Size Methods: Binary Outcomes

- ▶ **Problem 2 with peeking at control event rate: Even if investigators don't know overall event rate, control event rate gives some information about treatment effect because control event rate is correlated with observed treatment effect:**

$$\begin{aligned}\text{cov}(\hat{p}_{C1}, \hat{p}_{T1} - \hat{p}_{C1}) &= \text{cov}(\hat{p}_{C1}, \hat{p}_{T1}) - \text{cov}(\hat{p}_{C1}, \hat{p}_{C1}) \\ &= 0 - \text{var}(\hat{p}_{C1}) \\ &= -p_C(1 - p_C)/n_1.\end{aligned}\tag{38}$$

- ▶ Dividing by product of standard deviations gives correlation $\rho \approx -0.71$ under the null hypothesis.
- ▶ Nonzero correlation implies that the type 1 error rate could be inflated by procedure that allows peeking at control event rate.

Blinded Sample Size Methods: Binary Outcomes

- ▶ **Problem 2 with peeking at control event rate: Even if investigators don't know overall event rate, control event rate gives some information about treatment effect because control event rate is correlated with observed treatment effect:**

$$\begin{aligned}\text{cov}(\hat{p}_{C1}, \hat{p}_{T1} - \hat{p}_{C1}) &= \text{cov}(\hat{p}_{C1}, \hat{p}_{T1}) - \text{cov}(\hat{p}_{C1}, \hat{p}_{C1}) \\ &= 0 - \text{var}(\hat{p}_{C1}) \\ &= -p_C(1 - p_C)/n_1.\end{aligned}\tag{38}$$

- ▶ Dividing by product of standard deviations gives correlation $\rho \approx -0.71$ under the null hypothesis.
- ▶ Nonzero correlation implies that the type 1 error rate could be inflated by procedure that allows peeking at control event rate.

Blinded Sample Size Methods: Binary Outcomes

- ▶ Not a problem with blinded method because:

$$\begin{aligned} & \text{cov} \left\{ (\hat{p}_{T1} + \hat{p}_{C1})/2, \hat{p}_{T1} - \hat{p}_{C1} \right\} \\ &= (1/2) \left\{ \text{cov}(\hat{p}_{T1}, \hat{p}_{T1}) - \text{cov}(\hat{p}_{T1}, \hat{p}_{C1}) + \text{cov}(\hat{p}_{C1}, \hat{p}_{T1}) - \text{cov}(\hat{p}_{C1}, \hat{p}_{C1}) \right\} \\ &= (1/2) \left\{ \text{var}(\hat{p}_{T1}) - 0 + 0 - \text{var}(\hat{p}_{C1}) \right\} \\ &= (1/2) \left\{ p_T(1 - p_T)/n_1 - p_C(1 - p_C)/n_1 \right\} \\ &= (1/2) \left\{ p(1 - p)/n_1 - p(1 - p)/n_1 \right\} \quad (\text{under } H_0) \\ &= 0. \end{aligned} \tag{39}$$

Blinded Sample Size Methods: Continuous Outcomes

- ▶ Same ideas apply to trials with continuous outcomes analyzed using unpaired t-test. Gould and Shih (1992).
- ▶ Sample size depends on treatment effect and variance σ^2 .
- ▶ **Blinded method: Keep treatment effect fixed.**
 - ▶ Use pre-trial estimate of σ^2 for initial sample size n_0 /per arm.
 - ▶ At planned halfway point, $n_1 = n_0/2$ /arm, re-estimate σ^2 using variance of combined data across arms.
 - ▶ Modify sample size and treat as fixed in final analysis (use ordinary t-test at end).
 - ▶ Alternatively, could use t-test stratified by stage.
- ▶ Similar issue as with binary outcomes: Blinded variance could be large because treatment effect is huge.

Blinded Sample Size Methods: Continuous Outcomes



Figure: Lumped variance is inflated when there is large treatment effect.

- ▶ **Still, unless treatment effect is huge, the variance inflation problem is minor.**
- ▶ E.g., if σ^2 is within-arm variance and $E = \delta/\sigma$ is treatment effect relative to standard deviation (some authors call this the **effect size**, then $\text{var}(Y) = \sigma^2(1 + E^2/4)$.
- ▶ Even large treatment effect of $E = 1/2$ results in only 6% inflation of variance, 3% inflation of standard deviation.

Unblinded Sample Size Methods

Unblinded Sample Size Methods

- ▶ **Now consider unblinded sample size methods: Much riskier and error rate can be inflated if not careful.**
- ▶ Example: Suppose original sample size is 1,000/arm and we peek at z-score after 10/arm.
 - ▶ If $Z_{10} \geq 1.96$, change final sample size to 10/arm & reject H_0 .
 - ▶ If $Z_{10} < 1.96$, continue to 1,000/arm and reject if $Z_{1,000} \geq 1.96$.
- ▶ Z_{10} and $Z_{1,000}$ are nearly independent because they share only 1% of data; type 1 error rate, $P\{Z_{10} \geq 1.96 \cup Z_{1,000} \geq 1.96\}$, is

$$\begin{aligned} &= 1 - P\{Z_{10} < 1.96 \cap Z_{1,000} < 1.96\} \\ &\approx 1 - (1 - 0.025)^2 = 0.0494. \end{aligned} \tag{40}$$

- ▶ Proschan and Hunsberger (1995) showed that the most nefarious method more than doubles the type 1 error rate.

Unblinded Sample Size Methods

- ▶ How can we protect the type 1 error rate?
- ▶ Assume n is large enough to treat nuisance parameters (like σ or ρ) as known.
- ▶ Again specify initial sample size n_0/arm and assume sample size reassessment is after $n_1 = n_0/2/\text{arm}$ (fixed number).
- ▶ Let Z_1 be z-score at stage 1 with n_1/arm .
- ▶ Choose $n_2 = n_2(Z_1)$ (random, as is $n = n_1 + n_2$) based on Z_1 . We give more details on choice of n_2 later.

Unblinded Sample Size Methods

- ▶ If make no change in sample size ($n_2 = n_0/2$), then usual, fixed sample size z-score using all data equally weights z-scores:

$$Z = \frac{Z_1 + Z_2}{\sqrt{2}}. \quad (41)$$

- ▶ **Key to adaptive method: Even if we change n_2 , still equally weight Z_1 and Z_2 , as in (41), and reject if Z in (41) is $\geq z_\alpha$.**
- ▶ Justification:
 - ▶ Under H_0 , $Z_1 \sim N(0, 1)$ and $Z_2|Z_1 \sim N(0, 1)$, regardless of n_2 .
 - ▶ Because distribution of $Z_2|Z_1$ does not depend on Z_1 , (Z_1, Z_2) are independent.
 - ▶ Conclusion: (Z_1, Z_2) are iid $N(0, 1)$ under H_0 in this adaptive setting, just as in fixed sample setting.

Unblinded Sample Size Methods

- ▶ **More generally, any level α rejection region $(Z_1, Z_2) \in R$ when n_2 is fixed remains level α in adaptive n_2 setting.**
- ▶ Weakness of adaptive method is clear: Equally weighting z-scores from stage 1 and 2 even though n_2 is much larger (or smaller) than n_1 is inefficient.
- ▶ Reasonable if n_2 doesn't change much, but not if it does!
- ▶ Extreme case: Can even reject H_0 for right-tailed alternative hypothesis when conventionally computed z-score is negative!
 - ▶ Proschan and Hunsberger (1995) showed this.
 - ▶ Burman and Sonesson (2006) "rediscovered" this fact.
- ▶ Other adaptive procedures have similar drawbacks.

Unblinded Sample Size Methods

- ▶ **Note: procedure would NOT work if weighted Z_1 and Z_2 by sample sizes instead of equally weighting them.**
- ▶ Why not? Because then we would be taking weighted combination $w_1 Z_1 + w_2 Z_2$, where
 - ▶ if Z_1 is small, n_2 (and therefore w_2) would be large.
 - ▶ If Z_1 is large, n_2 (and therefore w_2) would be small.
 - ▶ Intuitively clear that giving more weight to a very large observed z-score would create bias and inflate type 1 error rate.

Unblinded Sample Size Methods

- ▶ Could combine 1-sided p-values, $P_i = 1 - \Phi(Z_i)$, $i = 1, 2$ instead of z-scores.
- ▶ Called a ***p-value combination function***.
- ▶ **Any level α rejection region $(P_1, P_2) \in R$ when n_2 is fixed remains level α in adaptive n_2 setting.**
- ▶ **Z-score and p-value combination procedures are equivalent:**
 - ▶ **Any z-score combination procedure corresponds to some p-value combination function procedure and vice versa.**
- ▶ One such p-value combination function: Fisher's method of combining p-values:

$$-2\ln(P_1 P_2) \sim \chi_4^2 \text{ under } H_0. \quad (42)$$

Unblinded Sample Size Methods

- ▶ **Bauer and Köhne (1994) method:**
 - ▶ **After observing first stage p-value, change sample size but continue to use (42) at end.**
 - ▶ **Reject H_0 if $-2\ln(P_1 P_2) \geq \chi_4^2(\alpha)$, the $(1 - \alpha)$ th quantile of a chi-squared distribution with 4 degrees of freedom.**
 - ▶ **They also modified procedure to allow stopping at first stage for futility/benefit.**
- ▶ Note: If $-2\ln(P_1) \geq \chi_4^2(\alpha)$, the $(1 - \alpha)$ th quantile of a χ_4^2 distribution, no need for second stage.
- ▶ Procedure has same drawback as equally weighting z-scores: Giving equal weight to p-values is inefficient if one p-value is based on much more information than the other.

Unblinded Sample Size Methods

- ▶ Another equivalent way to formulate adaptive methods, the **conditional error function** approach of Proschan and Hunsberger (1995), works as follows.

- ▶ Pre-specify a function $A(z_1)$, $0 \leq A(z_1) \leq 1$, telling how much conditional error rate you can spend after seeing $Z_1 = z_1$, where

$$\int_{-\infty}^{\infty} A(z_1)\phi(z_1)dz_1 = \alpha \text{ and } \phi(z_1) \text{ denotes standard normal density.}$$

- ▶ After seeing $Z_1 = z_1$, do a test using second stage z-score Z_2 only, but use alpha level $A(z_1)$.
- ▶ $A(z_1)$ is called a **conditional error function** because it is the conditional type 1 error rate given $Z_1 = z_1$ (conditional power under H_0).

Unblinded Sample Size Methods

- ▶ **Z-score combinations, p-value combinations, and conditional error functions are mathematically equivalent. They are different formulations of the same idea.**
- ▶ Conditional error formulation is useful because we can choose n_2 by considering conditional power under an alternative hypothesis.
- ▶ At end of first stage, alpha level $A = A(z_1)$ to use in second stage is fixed.
- ▶ To achieve (conditional) power $1 - \beta$, use EZ principle: Equate

$$E(Z_2) = z_A + z_\beta,$$

and solve for n_2 .

Unblinded Sample Size Methods

- ▶ For example, in t-test setting, if $A(z_1) = 0.15$,

$$E(Z_2) = \frac{\delta}{\sqrt{2\sigma^2/n_2}}.$$

- ▶ Estimate σ^2 and δ . Suppose estimates are $\hat{\sigma}^2 = 25$ and $\hat{\delta} = 1$. If we want 90% conditional power, set

$$\begin{aligned} \frac{\sqrt{n_2} \delta}{\sqrt{2\sigma^2}} &= z_{0.15} + z_{0.10} = 1.036 + 1.282 = 2.318 \\ n_2 &= \frac{2\sigma^2(2.318)^2}{\delta^2} = \frac{2(25)(2.318)^2}{1^2} \approx 269. \end{aligned} \quad (43)$$

- ▶ Choose second stage sample size 269/arm.

Unblinded Sample Size Methods

- ▶ Adaptive methods based on treatment effect have been criticized on grounds of inefficiency and potentially strange behavior (Burman and Sonesson (2006); Jennison and Turnbull (2003); Jennison and Turnbull (2006)).
- ▶ Additionally, Tsiatis and Mehta (2003) showed that it is more powerful to set large initial sample size and use group-sequential monitoring to stop early.
- ▶ Therefore, in general, adaptive methods based on treatment effect should be avoided.

Unplanned Changes

Unplanned Changes

- ▶ In rare cases, unplanned changes are needed (Posch and Proschan (2012)).
- ▶ Recall Pam's example of Metronidazole for TB (NCT00425113): Investigators reviewed blinded lung scans, discovered primary endpoint was not meaningful, and changed it.
- ▶ Can anything be done to protect the type 1 error rate?

Unplanned Changes

- ▶ Can anything be done to protect type 1 error rate in Metronidazole example? Assume endpoint change was made at end of trial.
 - ▶ **A re-randomization test would still protect type 1 error rate under strong null hypothesis that treatment has no effect on any outcome examined.**
 - ▶ Get null distribution by treating data as fixed constants, re-randomizing many times, and computing new value of test statistic for each re-randomization.
 - ▶ Compute p-value as proportion of re-randomized trials with result at least as extreme as observed result.
 - ▶ Valid because re-randomization tests already condition on outcomes. **Tests strong null hypothesis: H_0 : Treatment has no effect on **any** endpoint you examined.**

Unplanned Changes

- ▶ What about unplanned sample size increase after examining data by arm? Can anything be done to protect type 1 error rate?
- ▶ **One potentially useful method is due to Chen et al. (2004).**
 - ▶ **Compute conditional power (CP) under the current trend estimate, $B(t)/t = Z(t)/\sqrt{t}$ of drift parameter, θ (remember, $\theta = E\{Z(1)\}$).**
 - ▶ **If CP is greater than 0.50, you can increase sample size with no penalty.**
- ▶ Justification: If $CP \geq 0.50$, then increasing the sample size **decreases** the null conditional power.
- ▶ Proven to protect type 1 error rate with at most 1 interim analysis, and simulations suggest control of error rate with more interim analyses.

Unplanned Changes

- ▶ Müller and Schäfer (2004) proposed a method to make any design change (number or timing of interim analyses, sample size, primary endpoint, analysis method, etc.).
- ▶ Let CRP_{orig} be conditional rejection probability at interim analysis under original design:

$$CRP_{orig} = P_0(\text{Cross future boundary with original design} \mid Z = z),$$

where z is interim z-score.

- ▶ Make design change but make sure $CRP_{new} \leq CRP_{orig}$.

Unplanned Changes

- ▶ Controls type 1 error rate because

$$\begin{aligned} & P_0(\text{reject } H_0 \text{ with new design}) \\ &= \int P_0(\text{reject } H_0 \text{ with new design} \mid Z = z)\phi(z)dz \\ &= \int \text{CRP}_{\text{new}}(z)\phi(z)dz \leq \int \text{CRP}_{\text{orig}}(z)\phi(z)dz \\ &= \int P(\text{reject } H_0 \text{ with original design} \mid Z = z)\phi(z)dz \\ &= P(\text{reject } H_0 \text{ with original design}) = \alpha. \end{aligned} \tag{44}$$

- ▶ **Still, use only in emergencies and for minor changes!**

Summary

- ▶ Adaptive methods use within-trial data to make design changes.
- ▶ Adaptive methods should be **preplanned**, although emergencies sometimes happen.
- ▶ Blinded methods are safer & less controversial than unblinded methods.
- ▶ Blinded sample size re-estimation is common.
 - ▶ Binary outcomes: Gould (1992): Use overall event rate and hypothesized treatment effect to compute p_T and p_C and re-compute sample size.
 - ▶ Continuous outcomes Gould and Shih (1992): Treat overall variance as within-arm variance and re-compute sample size.

Summary

- ▶ Unblinded sample size increase based on promising trend also safe if use Chen et al. (2004).
 - ▶ Can increase sample size if conditional power under current treatment effect estimate is ≥ 0.50 .
 - ▶ Boundary at end is unchanged.
 - ▶ Logistical issues: Who makes sample size decision? Investigators should remain blinded.
- ▶ Müller and Schäfer (2004) can be used **in emergencies**.
 - ▶ Allows **any** design change after looking at data by arm.
 - ▶ Type 1 error rate is protected if conditional rejection probability (CRP) is \leq CRP of original design.
 - ▶ Will generate controversy unless design change is minor.

Thank you!

References I

- Action to Control Cardiovascular Risk in Diabetes Study Group (2008). Effects of intensive glucose lowering in type 2 diabetes. New England Journal of Medicine **358**, 2545–2559.
- Altman, D. G., Schulz, K. F., Moher, D., Egger, M., Davidoff, F., Elbourne, D., Gøtzsche, P. C., Lang, T., and Group, C. (2001). The revised consort statement for reporting randomized trials: explanation and elaboration. Annals of Internal Medicine **134**, 663–694.
- Anderson, G. L., Limacher, M., Assaf, A. R., Bassford, T., Beresford, S. A., Black, H., Bonds, D., Brunner, R., Brzyski, R., Caan, B., et al. (2004). Effects of conjugated equine estrogen in postmenopausal women with hysterectomy: the women's health initiative randomized controlled trial. JAMA **291**, 1701–1712.
- Appel, L., Moore, T., Obarzanek, E., et al. (1997). A clinical trial of the effects of dietary patterns on blood pressure. New England Journal of Medicine **336**, 1117–1124.
- Armitage, P., McPherson, C., and Rowe, B. (1969). Repeated significance tests on accumulating data. Journal of the Royal Statistical Society: Series A (General) **132**, 235–244.
- Bartlett, J. W., Carpenter, J. R., Tilling, K., and Vansteelandt, S. (2014). Improving upon the efficiency of complete case analysis when covariates are mnr. Biostatistics **15**, 719–730.
- Bartlett, J. W., Seaman, S. R., White, I. R., Carpenter, J. R., and Initiative*, A. D. N. (2015). Multiple imputation of covariates by fully conditional specification: accommodating the substantive model. Statistical Methods in Medical Research **24**, 462–487.
- Bauer, P. and Köhne, K. (1994). Evaluation of experiments with adaptive interim analyses. Biometrics **50**, 1029–1041.

References II

- Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal statistical society: series B (Methodological) **57**, 289–300.
- Berry, S. M., Connor, J. T., and Lewis, R. J. (2015). The platform trial: an efficient strategy for evaluating multiple treatments. JAMA **313**, 1619–1620.
- Bretz, F., Posch, M., Glimm, E., Klinglmueller, F., Maurer, W., and Rohmeyer, K. (2011). Graphical approaches for multiple comparison procedures using weighted bonferroni, simes, or parametric tests. Biometrical Journal. Biometrische Zeitschrift **53**, 894 – 913.
- Burman, C.-F. and Sonesson, C. (2006). Are flexible designs sound? Biometrics **62**, 664–669.
- Buyse, M. and Molenberghs, G. (1998). Criteria for the validation of surrogate endpoints in randomized experiments. Biometrics pages 1014–1029.
- Buyse, M., Molenberghs, G., Burzykowski, T., Renard, D., and Geys, H. (2000). The validation of surrogate endpoints in meta-analyses of randomized experiments. Biostatistics **1**, 49–67.
- Carpenter, J. R. and Kenward, M. G. (2007). Missing data in randomised controlled trials: a practical guide.
- CAST (1989). Preliminary report: effect of encainide and flecainide on mortality in a randomized trial of arrhythmia suppression after myocardial infarction. New England Journal of Medicine **321**, 406–412.

References III

- CAST (1992). Effect of the antiarrhythmic agent moricizine on survival after myocardial infarction. New England Journal of Medicine **327**, 227–233.
- CDP (1973). Prognostic importance of premature beats following myocardial infarction. JAMA **223**, 1116–1124.
- Chandereng, T. and Chappell, R. (2020). How to do response-adaptive randomization (rar) if you really must. Clinical Infectious Diseases **73**, 560.
- Chen, Y. H., DeMets, D. L., and Lan, K. K. (2004). Increasing the sample size when the unblinded interim result is promising. Statistics in Medicine **23**, 1023–1038.
- Connor, E. M., Sperling, R. S., Gelber, R., Kiselev, P., Scott, G., O'Sullivan, M. J., VanDyke, R., Bey, M., Shearer, W., Jacobson, R. L., Jimenez, E., O'Neill, E., Bazin, B., Delfraissy, J.-F., Culnane, M., Coombs, R., Elkins, M., Moye, J., Stratton, P., and Balsley, J. (1994). Reduction of maternal-infant transmission of human immunodeficiency virus type 1 with zidovudine treatment. New England Journal of Medicine **331**, 1173–1180.
- Connors, J. M., Brooks, M. M., Sciruba, F. C., Krishnan, J. A., Bledsoe, J. R., Kindzelski, A., Baucom, A. L., Kirwan, B.-A., Eng, H., Martin, D., et al. (2021). Effect of antithrombotic therapy on clinical outcomes in outpatients with clinically stable symptomatic covid-19: the activ-4b randomized clinical trial. JAMA **326**, 1703–1712.
- Coronary Drug Project Research Group (1975). Clofibrate and niacin in coronary heart disease. J Am Med Assoc **231**, 360–381.

References IV

- Dane, A., Spencer, A., Rosenkranz, G., Lipkovich, I., and Parke, T. (2019). Subgroup analysis and interpretation for phase 3 confirmatory trials: White paper of the efspi/psi working group on subgroup analysis. Pharmaceutical Statistics **18**, 126–139.
- Daniel, R. M., Kenward, M. G., Cousens, S. N., and De Stavola, B. L. (2012). Using causal diagrams to guide analysis in missing data problems. Statistical methods in medical research **21**, 243–256.
- Dmitrienko, A. and Wang, M.-D. (2006). Bayesian predictive approach to interim monitoring in clinical trials. Statistics in Medicine **25**, 2178–2195.
- Dodd, L. E., Freidlin, B., and Korn, E. L. (2021). Platform trials—beware the noncomparable control group. New England Journal of Medicine **384**, 1572–1573.
- Dunnett, C. W. (1955). A multiple comparison procedure for comparing several treatments with a control. Journal of the American Statistical Association **50**, 1096–1121.
- Ellenberg, S. S. and Hamilton, J. M. (1989). Surrogate endpoints in clinical trials: cancer. Statistics in Medicine **8**, 405–413.
- Ellenberg, S. S. and Shaw, P. A. (2022). Early termination of clinical trials for futility—considerations for a data and safety monitoring board. NEJM Evidence **1**, EVIDctw2100020.
- Evans, S. R., Li, L., and Wei, L. (2007). Data monitoring in clinical trials using prediction. Drug information journal: DIJ/Drug Information Association **41**, 733–742.

References V

- Evans, S. R., Rubin, D., Follmann, D., Pennello, G., Huskins, W. C., Powers, J. H., Schoenfeld, D., Chuang-Stein, C., Cosgrove, S. E., Fowler Jr, V. G., et al. (2015). Desirability of outcome ranking (door) and response adjusted for duration of antibiotic risk (radar). Clinical Infectious Diseases **61**, 800–806.
- Feller, W. (1957). An Introduction to Probability Theory and Its Applications. John Wiley & Sons.
- Freidlin, B., Korn, E. L., and Gray, R. (2010). A general inefficacy interim monitoring rule for randomized clinical trials. Clinical Trials **7**, 197–208.
- Frison, L. and Pocock, S. J. (1992). Repeated measures in clinical trials: analysis using mean summary statistics and its implications for design. Statistics in Medicine **11**, 1685–1704.
- Gold, S. M., Bofill Roig, M., Miranda, J. J., Pariente, C., Posch, M., and Otte, C. (2022). Platform trials and the future of evaluating therapeutic behavioural interventions. Nature Reviews Psychology **1**, 7–8.
- Goodman, S. (1992). A comment on replication, p-values and evidence. Statistics in Medicine **11**, 875–879.
- Gould, A. (1992). Interim analyses for monitoring clinical trials that do not materially affect the type I error rate. Statistics in Medicine **11**, 55–66.
- Gould, A. L. and Shih, W. J. (1992). Sample size re-estimation without unblinding for normally distributed outcomes with unknown variance. Communications in Statistics-Theory and Methods **21**, 2833–2853.

References VI

- Greene, H. L., Roden, D. M., Katz, R. J., Woosley, R. L., Salerno, D. M., and Henthorn, R. W. (1992). The cardiac arrhythmia suppression trial: First cast. . . then cast-ii. Journal of the American College of Cardiology **19**, 894–898.
- Groenwold, R. H., Donders, A. R. T., Roes, K. C., Harrell Jr, F. E., and Moons, K. G. (2012). Dealing with missing outcome data in randomized trials and observational studies. American Journal of Epidemiology **175**, 210–217.
- Hade, E. M., Young, G. S., and Love, R. R. (2019). Follow up after sample size re-estimation in a breast cancer randomized trial for disease-free survival. Trials **20**, 1–9.
- Hallstrom, A., Ornato, J., Weisfeldt, M., Travers, A., Christenson, J., McBurnie, M., Zalenski, R., Becker, L., and Proschan, M. (2004). Public-access defibrillation and survival after out-of-hospital cardiac arrest. New England Journal of Medicine **351**, 637–646.
- Halsey, L. G., Curran-Everett, D., Vowler, S. L., and Drummond, G. B. (2015). The fickle p value generates irreproducible results. Nature Methods **12**, 179–185.
- Hanna, N. H., Kaiser, R., Sullivan, R. N., Aren, O. R., Ahn, M.-J., Tiangco, B., Voccia, I., von Pawel, J., Kovcin, V., Agulnik, J., et al. (2016). Nintedanib plus pemetrexed versus placebo plus pemetrexed in patients with relapsed or refractory, advanced non-small cell lung cancer (lume-lung 2): a randomized, double-blind, phase iii trial. Lung Cancer **102**, 65–73.
- Haybittle, J. (1971). Repeated assessment of results in clinical trials of cancer treatment. The British journal of radiology **44**, 793–797.

References VII

- Hochberg, Y. (1988). A sharper bonferroni procedure for multiple tests of significance. Biometrika **75**, 800–802.
- Hochberg, Y. and Tamhane, A. C. (2009). Multiple Comparison Procedures. John Wiley & Sons, Inc.
- Holm, S. (1979). A simple sequentially rejective multiple test procedure. Scandinavian journal of statistics pages 65–70.
- Hwang, I., Shih, W., and De Cani, J. (1990). Group sequential designs using a family of type i error probability spending functions. Statistics in Medicine **9**, 1439–1445.
- ISIS-2 (1988). Randomised trial of intravenous streptokinase, oral aspirin, both, or neither among 17,187 cases of suspected acute myocardial infarction: Isis-2. The Lancet **332**, 349–360.
- Jennison, C. and Turnbull, B. (2003). Mid-course sample size modification in clinical trials based on the observed treatment effect. Statistics in Medicine **22**, 971–993.
- Jennison, C. and Turnbull, B. (2006). Adaptive and nonadaptive group sequential tests. Biometrika **93**, 1–21.
- Jennison, C. and Turnbull, B. W. (2000). Group Sequential Methods with Applications to Clinical Trials. CRC Press.
- Kahan, B. C. and Morris, T. P. (2012). Improper analysis of trials randomised using stratified blocks or minimisation. Statistics in Medicine **31**, 328–340.

References VIII

- Kim, K. and DeMets, D. (1987). Design and analysis of group sequential tests based on the type 1 error spending function. Biometrika **74**, 149–154.
- Kuznetsova, O. M. and Tymofeyev, Y. (2012). Preserving the allocation ratio at every allocation with biased coin randomization and minimization in studies with unequal allocation. Statistics in Medicine **31**, 701–723.
- Lachin, J. M., Matts, J. P., and Wei, L. (1988). Randomization in clinical trials: conclusions and recommendations. Controlled Clinical Trials **9**, 365–374.
- Lan, K., Simon, R., and Halperin, M. (1982). Stochastically curtailed tests in long-term clinical trials. Communications in Statistics—Sequential Analysis **1**, 207–219.
- Lan, K. G. and Zucker, D. M. (1993). Sequential monitoring of clinical trials: the role of information and brownian motion. Statistics in Medicine **12**, 753–765.
- Lan, K. K. G. and DeMets, D. L. (1983). Discrete sequential boundaries for clinical trials. Biometrika **70**, 659–663.
- Lesaffre, E., Edelman, M., Hanna, N., Park, K., Thatcher, N., Willemsen, S., Gaschler-Markefski, B., Kaiser, R., and Manegold, C. (2017). Statistical controversies in clinical research: fertility analyses in oncology—lessons on potential pitfalls from a randomized controlled trial. Annals of Oncology **28**, 1419–1426.
- Li, L., Evans, S. R., Uno, H., and Wei, L. (2009). Predicted interval plots (pips): a graphical tool for data monitoring of clinical trials. Statistics in Biopharmaceutical Research **1**, 348–355.

References IX

- Lipkovich, I., Dmitrienko, A., Denne, J., and Enas, G. (2011). Subgroup identification based on differential effect search—a recursive partitioning method for establishing response to treatment in patient subpopulations. Statistics in medicine **30**, 2601–2621.
- Lipkovich, I., Ratitch, B., and Mallinckrodt, C. H. (2020). Causal inference and estimands in clinical trials. Statistics in Biopharmaceutical Research **12**, 54–67.
- Little, R., D’Agostino, R., Cohen, M., Dickersin, K., Emerson, S., Farrar, J., Frangakis, C., Hogan, J., Molenberghs, G., Murphy, S., and Neaton, J. (2012). The prevention and treatment of missing data in clinical trials. NEJM **367**, 1355–60.
- Little, R. J., Carpenter, J. R., and Lee, K. J. (2022). A comparison of three popular methods for handling missing data: complete-case analysis, inverse probability weighting, and multiple imputation. Sociological Methods & Research page 00491241221113873.
- Little, R. J. and Lewis, R. J. (2021). Estimands, estimators, and estimates. JAMA **326**, 967–968.
- Love, R. R., Laudico, A. V., Van Dinh, N., Allred, D. C., Uy, G. B., Quang, L. H., Salvador, J. D. S., Siguan, S. S. S., Mirasol-Lumague, M. R., Tung, N. D., et al. (2015). Timing of adjuvant surgical oophorectomy in the menstrual cycle and disease-free and overall survival in premenopausal women with operable breast cancer. JNCI: Journal of the National Cancer Institute **107**,.
- Marcus, R., Peritz, E., and Gabriel, K. R. (1976). On closed testing procedures with special reference to ordered analysis of variance. Biometrika **63**, 655–660.
- Markaryan, T. and Rosenberger, W. F. (2010). Exact properties of efron’s biased coin randomization procedure. The Annals of Statistics **38**, 1546–1567.

References X

- Matts, J. P. and Lachin, J. M. (1988). Properties of permuted-block randomization in clinical trials. Controlled Clinical Trials **9**, 327–344.
- Molenberghs, G., Buyse, M., Geys, H., Renard, D., Burzykowski, T., and Alonso, A. (2002). Statistical challenges in the evaluation of surrogate endpoints in randomized trials. Controlled Clinical Trials **23**, 607–625.
- Molenberghs, G., Fitzmaurice, G., Kenward, M. G., Tsiatis, A., and Verbeke, G. (2014). Handbook of missing data methodology. CRC Press.
- Moore, T. (1995a). Deadly medicine: Why tens of thousands of heart patients died in the America's worst drug disaster. Simon and Schuster.
- Moore, T. J. (1995b). Deadly Medicine: Why Tens of Thousands of Heart Patients Died in America's Worst Drug Disaster. Simon & Schuster.
- Müller, H. H. and Schäfer, H. (2004). A general statistical principle for changing a design any time during the course of a trial. Statistics in Medicine **23**, 2497–2508.
- Neaton, J. D., Gray, G., Zuckerman, B. D., and Konstam, M. A. (2005). Key issues in end point selection for heart failure trials: composite end points. Journal of Cardiac Failure **11**, 567–575.
- O'Brien, P. C. and Fleming, T. R. (1979). A multiple testing procedure for clinical trials. Biometrics pages 549–556.
- Olarte Parra, C., Daniel, R. M., and Bartlett, J. W. (2023). Hypothetical estimands in clinical trials: a unification of causal inference and missing data methods. Statistics in Biopharmaceutical Research **15**, 421–432.

References XI

- Pawitan, Y. and Hallstrom, A. (1990). Statistical interim monitoring of the cardiac arrhythmia suppression trial. Statistics in Medicine **9**, 1081–1090.
- Pocock, S. J. (1977). Group sequential methods in the design and analysis of clinical trials. Biometrika **64**, 191–199.
- Pocock, S. J., Ariti, C. A., Collier, T. J., and Wang, D. (2012). The win ratio: a new approach to the analysis of composite endpoints in clinical trials based on clinical priorities. European Heart Journal **33**, 176–182.
- Pocock, S. J., Assmann, S. E., Enos, L. E., and Kasten, L. E. (2002). Subgroup analysis, covariate adjustment and baseline comparisons in clinical trial reporting: current practice and problems. Statistics in Medicine **21**, 2917–2930.
- Posch, M. and Proschan, M. A. (2012). Unplanned adaptations before breaking the blind. Statistics in Medicine **31**, 4146–4153.
- Proschan, M., Brittain, E., and Kammerman, L. (2011). Minimize the use of minimization with unequal allocation. Biometrics **67**, 1135–1141.
- Proschan, M. and Evans, S. (2020). Resist the temptation of response-adaptive randomization. Clinical Infectious Diseases **71**, 3002–3004.
- Proschan, M. A. (2021). Statistical Thinking in Clinical Trials. Chapman and Hall/CRC.
- Proschan, M. A. (2022). Statistical Thinking in Clinical Trials. Chapman and Hall/CRC.

References XII

- Proschan, M. A. and Brittain, E. H. (2020). A primer on strong vs weak control of familywise error rate. Statistics in Medicine **39**, 1407–1413.
- Proschan, M. A. and Hunsberger, S. A. (1995). Designed extension of studies based on conditional power. Biometrics pages 1315–1324.
- Proschan, M. A., Lan, K. G., and Wittes, J. T. (2006). Statistical Monitoring of Clinical Trials: A Unified Approach. Springer Science & Business Media.
- Rosenberger, W. F. and Lachin, J. M. (2015). Randomization in clinical trials: theory and practice. John Wiley & Sons.
- Rosenkranz, G. K. (2014). Bootstrap corrections of treatment effect estimates following selection. Computational Statistics & Data Analysis **69**, 220–227.
- Rossouw, J., Anderson, G., Prentice, R., et al. (2002). Risks and benefits of estrogen plus progestin in healthy postmenopausal women: principal results from the women's health initiative randomized controlled trial. JAMA **288**, 321–333.
- Ruberman, W., Weinblatt, E., Goldberg, J. D., Frank, C. W., and Shapiro, S. (1977). Ventricular premature beats and mortality after myocardial infarction. New England Journal of Medicine **297**, 750–757.
- Rubin, D. B. (1976). Inference and missing data. Biometrika **63**, 581–592.
- Ruskin, J. N. (1989). The cardiac arrhythmia suppression trial (CAST).

References XIII

- Schoenfeld, D. (1981). The asymptotic properties of nonparametric tests for comparing survival distributions. Biometrika **68**, 316–319.
- Shaw, P. A. (2018). Use of composite outcomes to assess risk–benefit in clinical trials. Clinical Trials **15**, 352–358.
- Shaw, P. A. and Fay, M. P. (2016). A rank test for bivariate time-to-event outcomes when one event is a surrogate. Statistics in Medicine **35**, 3413–3423.
- Snapinn, S., Chen, M.-G., Jiang, Q., and Koutsoukos, T. (2006). Assessment of futility in clinical trials. Pharmaceutical Statistics: The Journal of Applied Statistics in the Pharmaceutical Industry **5**, 273–281.
- Snyder, P. J., Bhasin, S., Cunningham, G. R., Matsumoto, A. M., Stephens-Shields, A. J., Cauley, J. A., Gill, T. M., Barrett-Connor, E., Swerdloff, R. S., Wang, C., et al. (2016). Effects of testosterone treatment in older men. New England Journal of Medicine **374**, 611–624.
- Sullivan, T. R., White, I. R., Salter, A. B., Ryan, P., and Lee, K. J. (2018). Should multiple imputation be the method of choice for handling missing data in randomized trials? Statistical Methods in Medical Research **27**, 2610–2626.
- Sully, B. G., Julious, S. A., and Nicholl, J. (2014). An investigation of the impact of futility analysis in publicly funded trials. Trials **15**, 1–9.
- Svensson, S., Menkes, D., and Lexchin, J. (2013). Surrogate outcomes in clinical trials: A cautionary tail. JAMA Internal Medicine **173**, 611–612.

References XIV

- Temple, R. and Pledger, G. W. (1980). The fda's critique of the anturane reinfarction trial. New England Journal of Medicine **303**, 1488–1492.
- Tsiatis, A. and Mehta, C. (2003). On the inefficiency of the adaptive design for monitoring clinical trials.
- Tsiatis, A. A., Davidian, M., Zhang, M., and Lu, X. (2008). Covariate adjustment for two-sample treatment comparisons in randomized clinical trials: a principled yet flexible approach. Statistics in Medicine **27**, 4658–4677.
- Van Buuren, S. and Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in r. Journal of statistical software **45**, 1–67.
- van der Wal, W. M. and Geskus, R. B. (2011). ipw: an r package for inverse probability weighting. Journal of Statistical Software **43**, 1–23.
- Von Hippel, P. T. (2020). How many imputations do you need? a two-stage calculation using a quadratic rule. Sociological Methods & Research **49**, 699–718.
- Wang, Y. and Tian, L. (2017). The equivalence between mann-whitney wilcoxon test and score test based on the proportional odds model for ordinal responses. In 2017 4th International Conference on Industrial Economics System and Industrial Security Engineering (IEIS), pages 1–5. IEEE.
- White, I. R., Royston, P., and Wood, A. M. (2011). Multiple imputation using chained equations: issues and guidance for practice. Statistics in Medicine **30**, 377–399.

References XV

- Wittes, J., Barrett-Connor, E., Braunwald, E., Chesney, M., Cohen, H. J., DeMets, D., Dunn, L., Dwyer, J., Heaney, R. P., Vogel, V., et al. (2007). Monitoring the randomized trials of the women's health initiative: the experience of the data and safety monitoring board. Clinical Trials **4**, 218–234.
- Women's Health Initiative Study Group and others (1998). Design of the women's health initiative clinical trial and observational study. Control Clinical Trials **19**, 61–109.