Design and Analysis of Clinical Trials

Pamela Shaw & Michael Proschan

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

Virtual Course, July 10-12, 2023

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)
- Biostatistics Research Branch
- proscham@niaid.nih.gov

Course Outline

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)

Day 3

1. 8:30-9:40 Handling missing data

   9:40-9:55 Break

2. 9:55-10:45 Multiple Comparisons

   10:45-11:00 Break

3. 11:00-11:55 Adaptive design

4. 11:55-12:00 Wrap Up

Course overview

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

Key Features of Randomized Controlled Trial (RCT)

- A **Randomized Controlled Trial** is a study of a novel intervention in human subjects where the intervention assignment is randomized
- A randomized controlled trial (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)

Group sequential designs - Studies with prespecified methods to analyze data partway through trial for potential early stopping (Day 2)

Adaptive designs - Studies with pre-specified methods to use within trial data to change aspects of design (Day 3)

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

Handling missing data in an RCT

Little et al. (2012)

The best way to handle missing data is to prevent it

Preventing missing data

- A well-written, complete and understandable informed consent is vital: not just to protect ethics but to make sure participants know what they are getting into
- Poor understanding of trial procedures and expectations for follow-up will lead to missing data
  - Off-treatment does not mean off-study
- Highly burdensome procedures will have drop-out regardless of how good the informed consent is
- When constructing endpoint, it is worth considering what is minimally necessary to obtain the clinically necessary information to evaluate the treatment
- Any procedures above the minimum, will need to think carefully about tradeoff of participant burden

Three-prong Approach to Minimizing Impact of Missing Data

Design

- Avoid endpoints that are more likely to be missing
- Choose the smallest time frame for primary analysis that still yields clinically relevant information on treatment effects
- Consider a run-in period to ensure commitment
  - Particularly important for long/complicated studies

Conduct

- Make extensive efforts to retain subjects
- Continue follow-up for outcomes even if subject stops treatment

Analysis

- Choose analyses that require minimally problematic assumptions

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 $\geq$ 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)
4. Rigorous control of type 1 error generally does not apply when evaluating safety outcomes in an early phase trial. Generally do not want to miss a safety signal (may evaluate several AEs)
5. Type 1 error control is needed when you win if at least one test is statistically significant.
   - When all tests must be significant (e.g., showing that a combination drug beats each of its constituents), no multiple comparison adjustment is needed.

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 Resistent 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
- At time trial done only about 76 known cases in US.

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.

HIES Ranitidine Trial Study: Double trouble

- A cross-over design was chosen given the rarity of disease. 20 patients were to be followed on ranitidine and control arm (usual care) for one year each, in random order
- In a cross-over design, when someone drops out, it's like losing two subjects - particularly if drop out during the first period
- For this trial, complete follow-up over two years was essential
- This trial closed early: higher than expected drop-out and a diminishing interest in Ranitidine

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 (1995)*

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.
- 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 ($p < 0.0001$)

Analysis

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

SOLVD: Results

	Enalapril (N=319)		Placebo (N=343)		Cox PH	Score Test
Endpoint	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

	Enalapril (N=319)		Placebo (N=343)		Cox PH	Score Test
Endpoint	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-Whiney (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, $\mathrm{E}(Y_T) - \mathrm{E}(Y_C)$ because X is baseline variable, so $\mathrm{E}(X_T) = \mathrm{E}(X_C)$.

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

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.

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.

Ellenberg, S. S. and Hamilton, J. M. (1989). Surrogate endpoints in clinical trials: cancer. Statistics in Medicine **8,** 405–413.

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.

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.

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.

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.

References II

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.

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.

Moore, T. (1995). Deadly medicine: Why tens of thousands of heart patients died in the America's worst drug disaster. Simon and Schuster.

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.

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.

Ruskin, J. N. (1989). The cardiac arrhythmia suppression trial (CAST).

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.

Svensson, S., Menkes, D., and Lexchin, J. (2013). Surrogate outcomes in clinical trials: A cautionary tail. JAMA Internal Medicine **173,** 611–612.

References III

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.

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 understudy 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 $u >= 0.5$ 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
```

Blocked Randomization Example: Flu Vaccine Dose Escalation Study

- An early phase I dose-finding study for a flu vaccine candidate sought to investigate 6 dose levels in a blinded, placebo controlled study
  - Considered single (one nare) and double dose (both nares) of 0.25mg, 0.5 mg, and 1mg administered intranasally
  - Primary outcome was safety and tolerability
- Dose cohorts had 5 active and 2 placebo subjects
  - Randomized block design, with a block of size 7
  - The order of the 5 active and 2 placebo assignments are randomly permuted for each dose group
  - Note: if want 5:2 ratio, then block sizes have to be multiples of 7

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

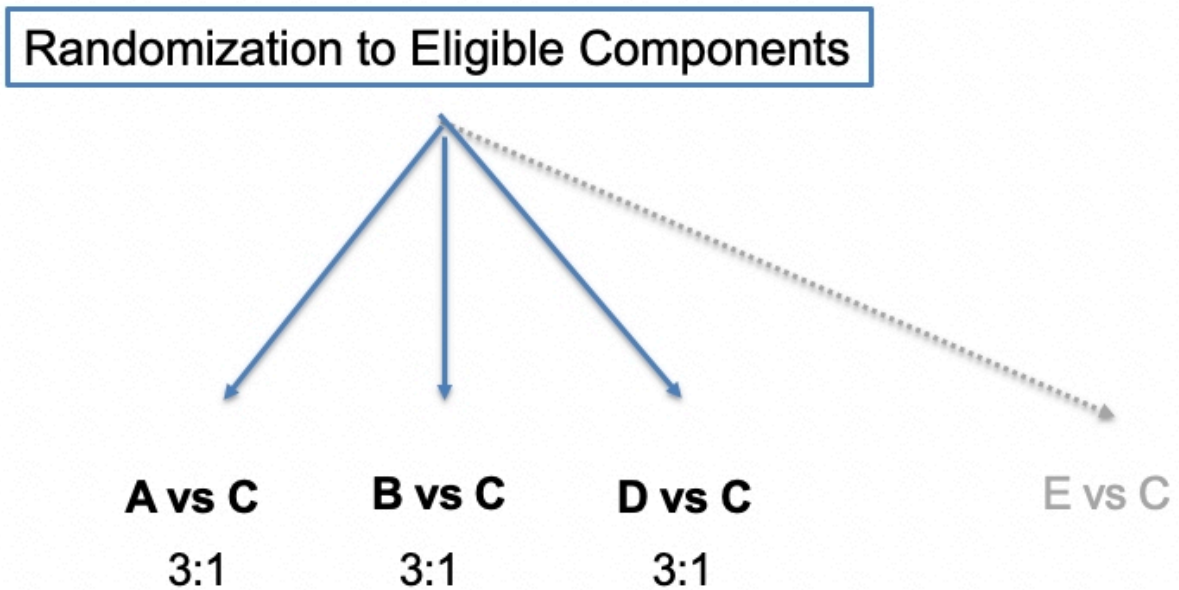
Cluster Randomization

- 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 todays world, isolated communities harder to find

Randomization in Platform Trials

- Platforms 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, nC/nA = sqrt(#active arms) Is optimal
  - But a preference often to have equal sample size

Platform Example



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)

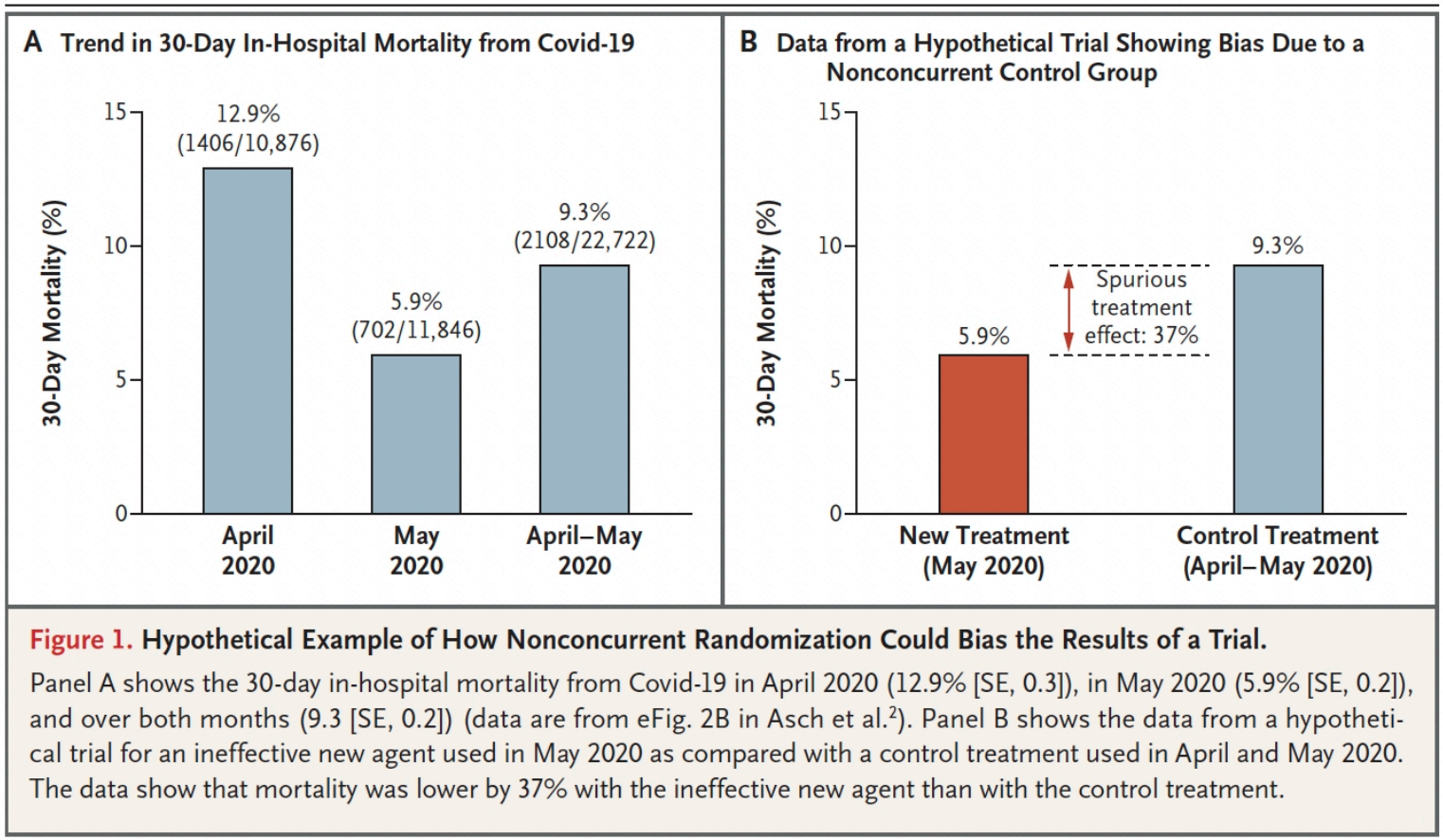


Figure 1. Hypothetical Example of How Nonconcurrent Randomization Could Bias the Results of a Trial.

Panel A shows the 30-day in-hospital mortality from Covid-19 in April 2020 (12.9% [SE, 0.3]), in May 2020 (5.9% [SE, 0.2]), and over both months (9.3 [SE, 0.2]) (data are from eFig. 2B in Asch et al.[2]). Panel B shows the data from a hypothetical trial for an ineffective new agent used in May 2020 as compared with a control treatment used in April and May 2020. The data show that mortality was lower by 37% with the ineffective new agent than with the control treatment.

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

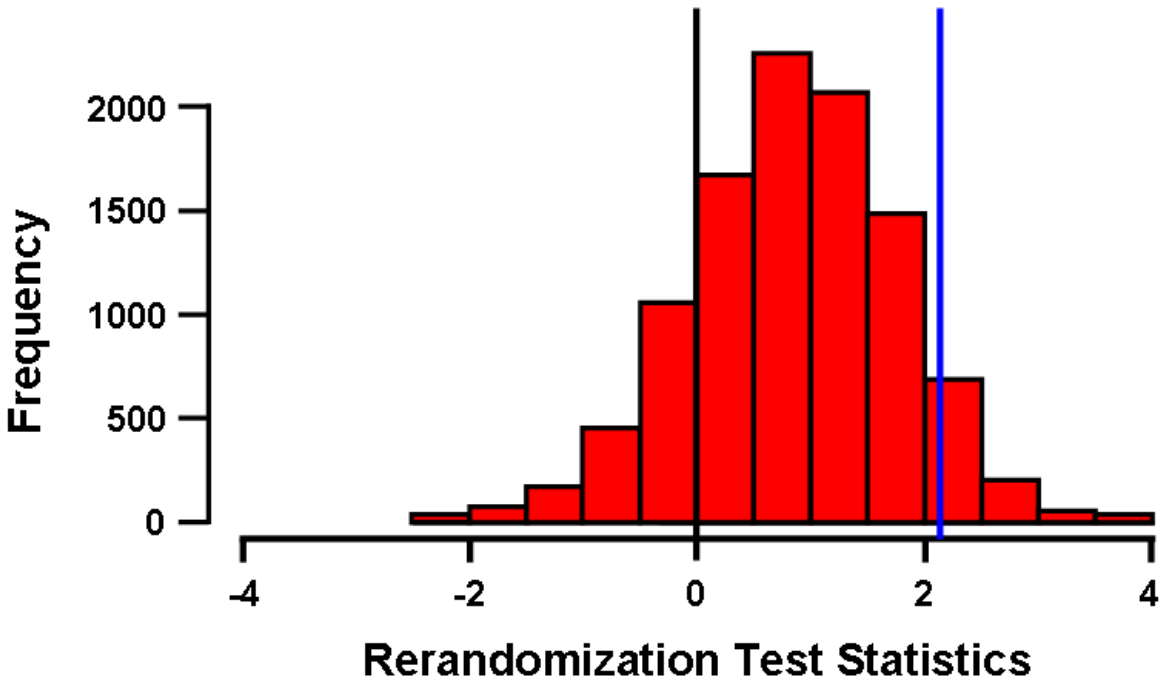
- Genzyme conducted Late Onset Treatment Study (LOTS)
  - 90 patients with late onset Pompe' disease
  - Primary outcome: 6 minute walk test
  - 2:1 allocation to drug/placebo using minimization
    - Site
    - BL 6 minute walk ($\leq$300m, $>$300m)
    - Forced vital capacity ($\leq$55% pred., $>$55%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

References I

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.

Chandereng, T. and Chappell, R. (2020). How to do response-adaptive randomization (rar) if you really must. Clinical Infectious Diseases **73,** 560.

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.

Gold, S. M., Bofill Roig, M., Miranda, J. J., Pariante, C., Posch, M., and Otte, C. (2022). Platform trials and the future of evaluating therapeutic behavioural interventions. Nature Reviews Psychology **1,** 7–8.

Kahan, B. C. and Morris, T. P. (2012). Improper analysis of trials randomised using stratified blocks or minimisation. Statistics in medicine **31,** 328–340.

Kuznetsova, O. M. and Tymofyeyev, 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.

Markaryan, T. and Rosenberger, W. F. (2010). Exact properties of efron's biased coin randomization procedure. The Annals of Statistics **38,** 1546–1567.

Matts, J. P. and Lachin, J. M. (1988). Properties of permuted-block randomization in clinical trials. Controlled clinical trials **9,** 327–344.

References II

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 practiceand problems. Statistics in medicine **21,** 2917–2930.

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.

Rosenberger, W. F. and Lachin, J. M. (2015). Randomization in clinical trials: theory and practice. John Wiley & Sons.

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.

Lecture 3: Sample Size/Power

Outline

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:

$$H_0 : \text{treatment has no effect}$$
$$H_1 : \text{treatment has an effect.}$$

- 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 $1-\beta = P($ rejecting H_0 when H_1 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 $Z = \hat{\delta}/\mathrm{se}(\hat{\delta})$, 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: $\hat{\delta} = \bar{Y}_T - \bar{Y}_C$; $\mathrm{se}(\hat{\delta}) = \sqrt{2\sigma^2/n}$.
  - Z-score for proportions: $\hat{\delta} = \hat{p}_T - \hat{p}_C$; $\mathrm{se}(\hat{\delta}) \approx \sqrt{2p(1-p)/n}$.
  - Z-score for logrank statistic: $\hat{\delta} = \sum(O_i - E_i)/\sum V_i$ estimates log hazard ratio; $\mathrm{se}(\hat{\delta}) = 1/\sum V_i$.
  - Z-scores for maximum likelihood estimators (MLEs), minimum variance unbiased estimators, Cox models, etc.

Outline

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 $\mu_T - \mu_C$ to $\mu_C - \mu_T$.
- 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.
- Example: You compare a new treatment to standard treatment for hepatitis C virus.
  - Primary outcome: change in log viral load from 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?
- $Z = \hat{\delta}/\sqrt{2\sigma^2/n}$, $\hat{\delta} = \bar{Y}_C - \bar{Y}_T$.
- Expected z-score is

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

Introduction to EZ Principle

- Expected z-score is lower than 2.80.
- The trial is underpowered.
- Investigator: “Sorry, I meant 100 per arm.”
- Check:

$$E(Z) = \frac{\delta}{\sqrt{2\sigma^2/n}} = \frac{0.5}{\sqrt{2(1.25)^2/100}} = 2.828.$$

- Close to 2.80. Sample size is accurate.

Introduction to EZ Principle

- 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? Paramerize so large values are good.

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

- Expected z-score is:

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

Introduction to EZ Principle

- Expected z-score is much greater than 3.00.
- Trial is overpowered.
- Statistician: “My bad. I meant 400 per arm.”
- Check:

$$E(Z) = \frac{p_C - p_T}{\sqrt{2p(1-p)/n}} = \frac{0.20 - 0.12}{\sqrt{2(0.16)(1-0.16)/400}} = 3.086.$$

- Still slightly overpowered, but not much because $E(Z)$ is not too far from 3.00.

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.

Outline

Where Does The Key Formula Come from?

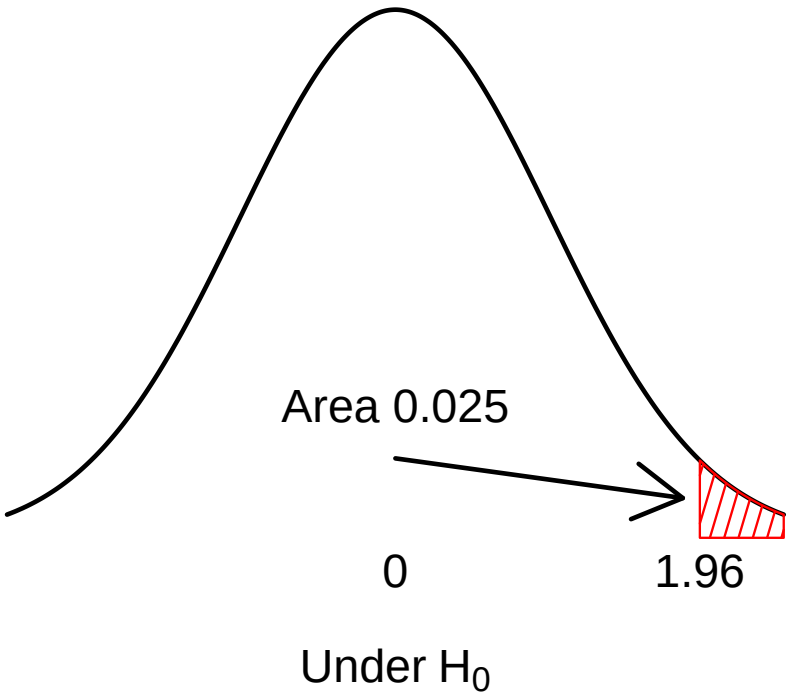


Figure: The standard normal null density for the z-statistic. For a 1-tailed test at $\alpha = 0.025$, we reject H_0 if $Z > 1.96$.

Where Does The Key Formula Come from?

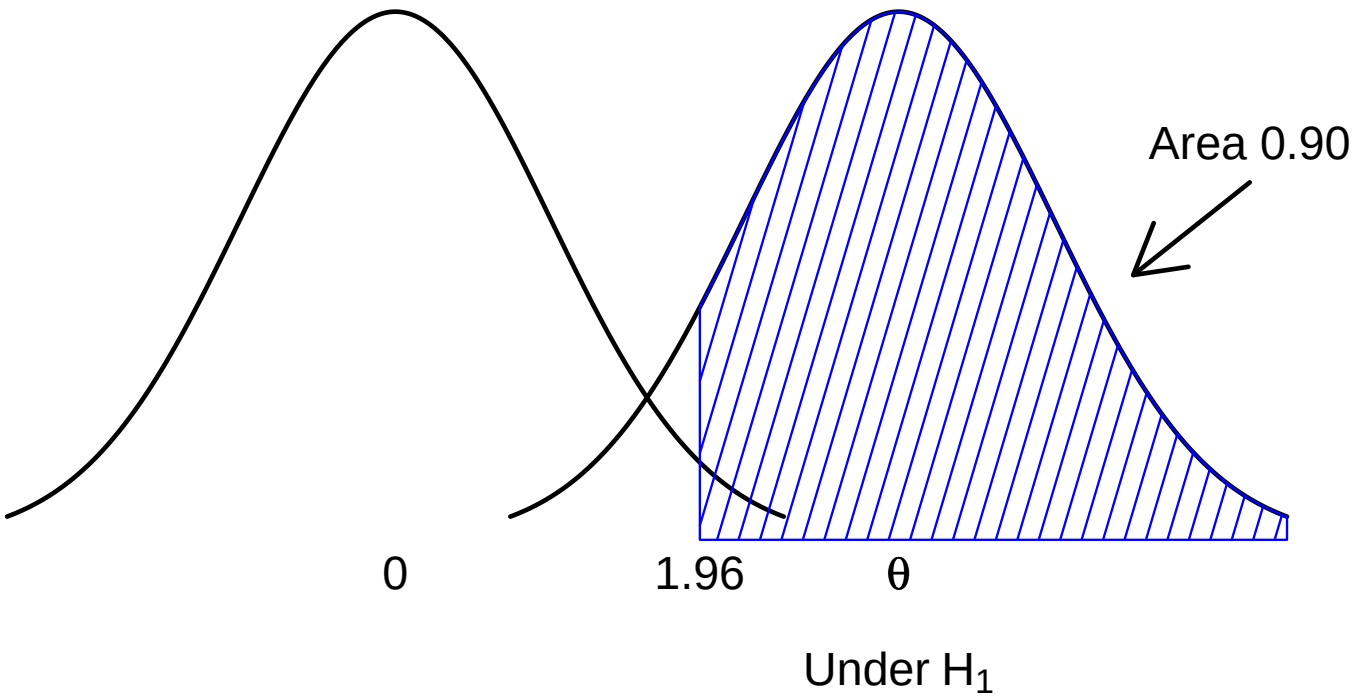


Figure: The alternative $N(\theta, 1)$ density for Z. For power 0.90, we want the blue shaded area to be 0.90.

Where Does The Key Formula Come from?

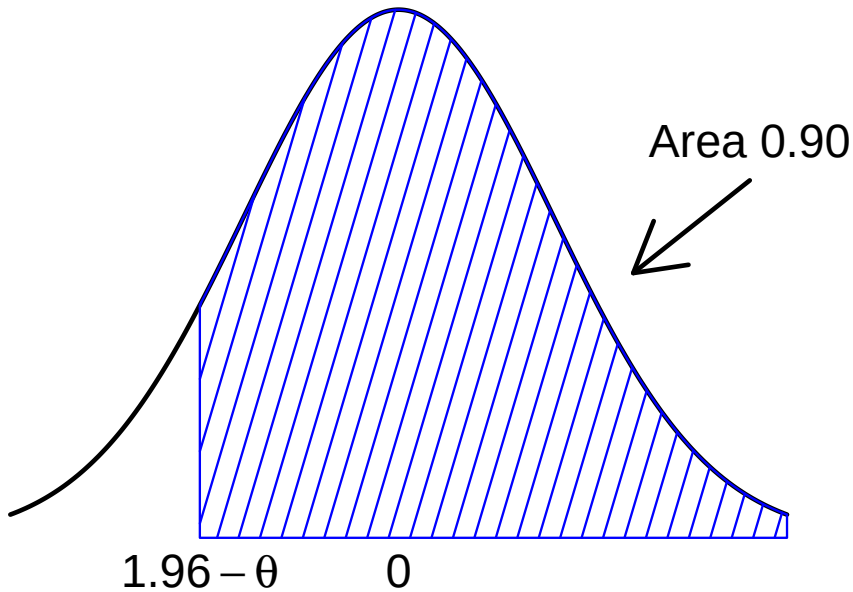


Figure: The blue shaded area in Figure 2 equals the blue shaded area to the right of $1.96-\theta$ under the standard normal curve. For power 0.90, $1.96-\theta=-1.28$, so $\theta=1.96+1.28=3.24$.

Outline

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

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

  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

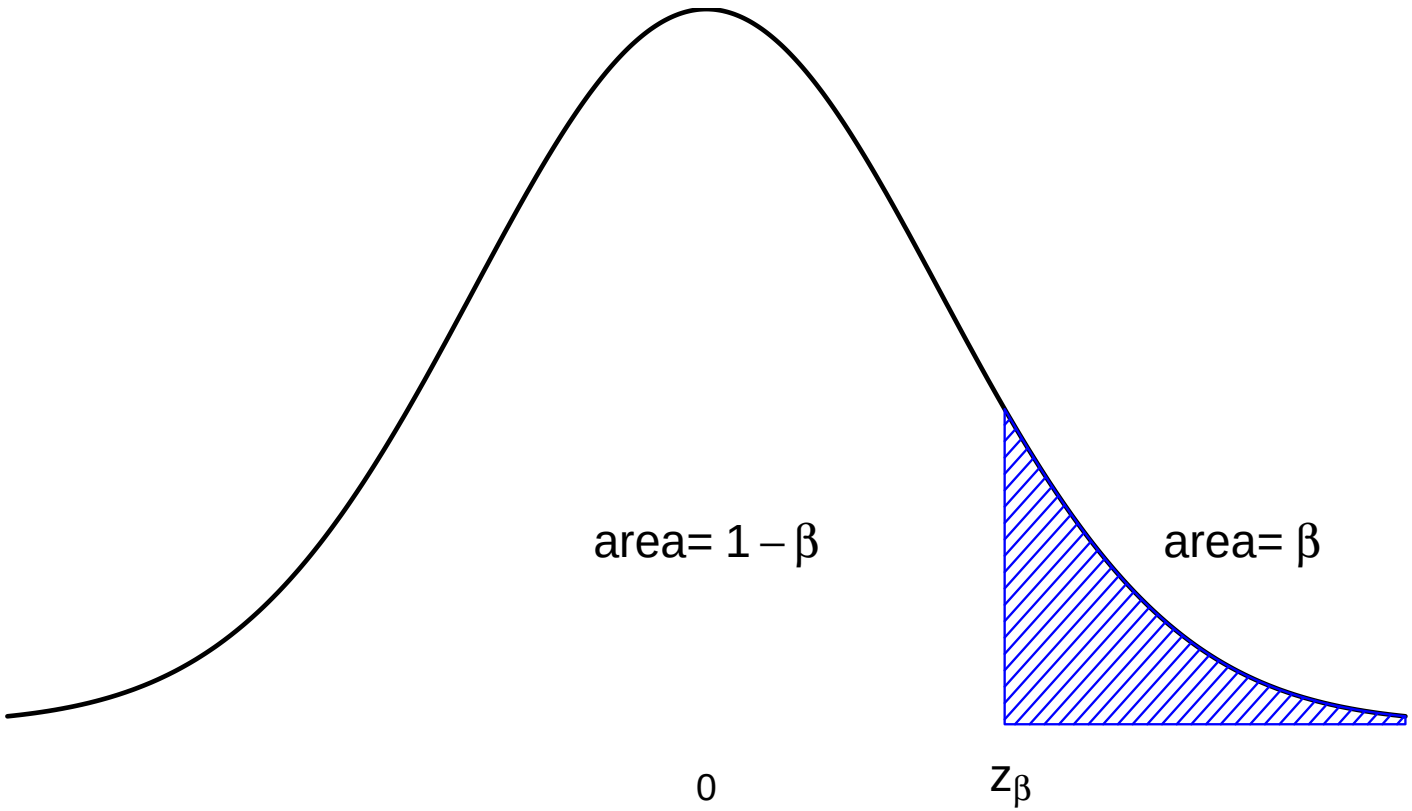


Figure: The area to the right of z_β is β, so the area to the left of z_β is $1-\beta$ =power.

General EZ Principle and Applications

- 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 $\sigma = 1.25$.

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

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

$$\frac{0.5}{\sqrt{\frac{2(1.25)^2}{n}}} = 1.96 + 0.84 = 2.80$$

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

Need 98/arm.

General EZ Principle and Applications

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

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

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

$$\frac{\delta}{\sqrt{\frac{2(1.25)^2}{75}}} = 1.96 + 0.84 = 2.80$$

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

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?

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

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

$$2.449 = 1.96 + z_\beta$$

$$0.489 = z_\beta$$

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

$\Phi(0.489)$ is N(0,1) distribution at 0.489 (in R, pnorm(0.489)).

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

General EZ Principle and Applications

- 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.

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

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

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

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

Need 378/arm.

General EZ Principle and Applications

- If you can only get 300/arm, what is power?

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

$$E(Z) = \frac{p_C - p_T}{\sqrt{\frac{2p(1-p)}{n}}} = \frac{0.20 - 0.12}{\sqrt{\frac{2(0.16)(1-0.16)}{300}}} = 2.673$$

$$\mathrm{E}(Z) = 2.673 = (z_{\alpha/2} + z_\beta) = 1.96 + z_\beta \quad \text{(EZ Principle)}$$

$$\Phi(2.673 - 1.96) = \Phi(z_\beta) = 1 - \beta = \text{power}. \tag{3}$$

Power is approximately $\Phi(0.713) = \mathrm{pnorm}(0.713) = 0.76$.

General EZ Principle and Applications

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

Table: Table at ith death.

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

$n_{C_i}, n_{T_i} =$ numbers at risk in treatment, control just prior to ith death.

$O_i =$indicator that ith death is in control arm.

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

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

General EZ Principle and Applications

Table: Table at ith death.

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

n_{C_i}, n_{T_i} = numbers at risk in treatment, control just prior to ith death.

O_i =indicator that ith death is in control arm.

$E_i = p_i = n_{C_i}/n_i$ =null expected value of O_i, given marginals.

$V_i = p_i(1-p_i) = \frac{n_{C_i} n_{T_i}}{n_i^2}$ =null variance of O_i, given marginals.

General EZ Principle and Applications

- FUN FACT: Each $\hat{\delta}_i = (O_i - E_i)/V_i$ estimates the log hazard ratio and has variance $1/V_i$.
- Optimal weighted average of the $\hat{\delta}_i$ weights inversely proportional to variances, $w_i = 1/\mathrm{var}(\hat{\delta}_i) = V_i$.

$$\begin{aligned} \hat{\delta} &= \frac{\sum_{i=1}^d w_i \hat{\delta}_i}{\sum_{i=1}^d w_i} = \frac{\sum_{i=1}^d \cancel{V_i}\{(O_i - E_i)/\cancel{V_i}\}}{\sum_{i=1}^d V_i} \\ &= \frac{\sum_{i=1}^d (O_i - E_i)}{\sum_{i=1}^d V_i}. \end{aligned} \tag{4}$$

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

General EZ Principle and Applications

Logrank z-statistic is

$$Z = \frac{\hat{\delta}}{\sqrt{\mathrm{var}(\hat{\delta})}} = \frac{\sum_{i=1}^{d}(O_i - E_i)}{\sqrt{\sum_{i=1}^{d} V_i}} = \left(\frac{\sum_{i=1}^{d}(O_i - E_i)}{\sum_{i=1}^{d} V_i}\right)\sqrt{\sum_{i=1}^{d} V_i}$$

$$= \hat{\delta}\sqrt{\sum_{i=1}^{d} V_i}, \tag{5}$$

where $\hat{\delta}$ estimates the log hazard ratio. 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

$$\begin{aligned} Z &\approx \hat{\delta}\sqrt{d/4} \\ \mathrm{E}(Z) &\approx \delta\sqrt{d/4}, \end{aligned} \tag{6}$$

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

General EZ Principle and Applications

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

$$\begin{aligned} \mathrm{E}(Z) = \delta\sqrt{d/4} &= (z_{\alpha/2}+z_{\beta}) \text{ (EZ Principle)} \\ d &= \frac{4(z_{\alpha/2}+z_{\beta})^2}{\delta^2}, \end{aligned} \tag{7}$$

  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

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

$$\begin{aligned} \mathrm{E}(Z) &= \delta\sqrt{d/4} = (z_{\alpha/2} + z_{\beta}) = (1.96 + 1.04) = 3 \text{ (EZ Principle)} \\ d &= \frac{4(3)^2}{\delta^2}\frac{4(3)^2}{\{\ln(1.333)\}^2} \approx 436 \text{ events.} \end{aligned} \tag{8}$$

General EZ Principle and Applications

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

$$\begin{aligned} E(Z) &= \ln(1.333)\sqrt{350/4} = 2.689 \\ 2.689 &= (z_{\alpha/2} + z_\beta) = 1.96 + z_\beta \ \text{(EZ Principle)} \\ \Phi(2.689 - 1.96) &= \Phi(z_\beta) = 1 - \beta = \text{power.} \end{aligned} \tag{9}$$

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

General EZ Principle and Applications

- Suppose you get only 350 events (deaths).
- What hazard ratio can be detected with 85% power?,

$$
\begin{aligned}
E(Z) &= \ln(\lambda)\sqrt{350/4} = z_{\alpha/2} + z_{\beta} = 1.96 + 1.04 = 3 \text{ (EZ Principle)} \\
\lambda &= \exp\left(\frac{3}{\sqrt{350/4}}\right) \\
&= 1.378. \qquad (10)
\end{aligned}
$$

- 85% power for a hazard ratio of 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 is 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

$$Z = \frac{\hat{p}_S - \hat{p}_N + 0.10}{\sqrt{\frac{\hat{p}_S(1-\hat{p}_S)+\hat{p}_N(1-\hat{p}_N)}{n}}} \tag{11}$$

If $p_S = p_N$,

$$\begin{aligned} \mathrm{E}(Z) &\approx \frac{0.10}{\sqrt{\frac{2p(1-p)}{n}}} = (z_{\alpha/2} + z_{\beta}) \text{ (EZ Principle)} \\ \frac{0.10}{\sqrt{\frac{2p(1-p)}{n}}} &= 1.645 + 1.282 = 2.927. \end{aligned}$$

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

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 this represents the **true** treatment effect; i.e., $\mathrm{E}(Z) = 1.96$.
- If we repeat trial, power is, by EZ principle,

$$\begin{aligned} \mathrm{E}(Z) &= 1.96 + z_\beta \\ 1.96 &= 1.96 + z_\beta \\ 0 &= z_\beta \\ \Phi(0) &= \Phi(z_\beta) = 1 - \beta = \text{power} \\ 0.5 &= \text{power} \end{aligned} \tag{12}$$

Only a 50% chance of replicating the result!

Outline

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.
- **Understimating** 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
  - 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** $Y_{EOS} - Y_{BL}$. 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.”

Outline

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

$$n = \frac{(1.96)^2 p(1-p)}{(0.15)^2} = 341.4756 p(1-p). \tag{13}$$

- If we expect $p = 0.3$, substitute 0.3 into Equation (13) to get $n = 72$.
- Could also use $p = 0.5$ as a worst case scenario: Substituting 0.5 into (13) gives $n = 86$.

Outline

Sample Size: Safety

- In safety studies, want to know sample size needed to see at least one adverse event (AE) of a given probability.

  P(see at least one AE of probability p)

$$\begin{aligned} &= 1 - P(0 \text{ AEs of probability } p) \\ &= 1 - (1-p)^n. \end{aligned} \tag{14}$$

- 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

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

- For 2-sided $\alpha = 0.05$, $\mathrm{E}(Z)$ 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.

References I

Goodman, S. (1992). A comment on replication, p-values and evidence. Statistics in Medicine **11,** 875–879.

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.

Proschan, M. A. (2022). Statistical Thinking in Clinical Trials. Chapman and Hall/CRC.

Schoenfeld, D. (1981). The asymptotic properties of nonparametric tests for comparing survival distributions. Biometrika **68,** 316–319.