SISCR
UW - 2016

Introduction to Clinical Trials – Day 1

Session 4 – Statistical Tasks in Trial Design

Presented July 25, 2016

Susanne J May
Department of Biostatistics
University of Washington

Daniel L Gillen
Department of Statistics
University of California, Irvine

Outline

- Refinement of hypotheses
- Probability model and summary measures
- Determination of sample size

Statistical Tasks in Clinical Trials

Clinicians perspective (some!)

Just one, right?

- How many people do I need?

Statistical Tasks in Clinical Trials

First question…

- What is your hypothesis?

Sample initial aims/hypothesis (1)

Specific Aim 1. To determine whether, among individuals with HIV-associated neurocognitive impairment (HNCI), antiretroviral therapy (ART) applied according to a CNS-targeted strategy (CNS-T) improves neurocognitive outcomes compared to a conventional (non-CNS-targeted) comparison strategy. All patients enrolled will be HIV-infected individuals with cognitive impairment eligible for new ART regimens according to contemporary consensus treatment guidelines. CNS-T will comprise three components: (1) optimizing the CNS-penetration of agents in the regimen; (2) augmenting the antiretroviral regimen if an interim assessment determines that viral load in cerebrospinal fluid (CSF) is not suppressed (CSF HIV RNA < 50 copies/mL); and (3) augmenting the regimen if an interim evaluation determines CSF drug concentrations to be subtherapeutic.

Hypothesis 1. Neurocognitive outcome in the CNS-T arm will be better than in the non-CNS-T arm.

Revised aim/hypothesis

Later round…

- **Specific Aim 1.** To evaluate the effectiveness of CNS-T as compared to non-CNS-T ART in treating HNCI globally and in different domains of functioning known to be affected by HIV.
- Hypothesis 1. Participants in the CNS-T arm will demonstrate greater improvement in NC functioning than participants in the non-CNS-T arm.

Sample initial aims/hypothesis (2)

- The **long-term goal** is to enhance the understanding of social and cultural pressures among women within the Ethiopian Community regarding HIV and to reduce (?) gender disparity in this population. [Should one of the goals of the analysis be to show that there is gender disparity regarding "seeking testing, treatment and counseling"?] The **overall objective** is to determine areas of misconception, misunderstanding and fear among Ethiopian women in this city regarding HIV infection and transmission [versus Ethiopian men or versus non-Ethiopian women or versus the general US population?].

Sample initial aims/hypothesis (2)

- (Observational study)
- Our **central hypothesis** is that although likely multi-factorial, gender disparity regarding knowledge about (?) HIV in the Ethiopian culture contributes to misconceptions regarding HIV prevention and transmission and possibly limits access to healthcare.

Sample initial aims/hypothesis (3)

- The goal of this project is to develop and evaluate the efficacy of an “Motivational Interview (MI) toolbox" to promote the adoption of risk reduction behaviors among newly infected HIV+ persons in enrolling sites, and to assess the efficacy of this intervention on HIV transmission behaviors and HIV incidence in discordant partnerships. In Year 1, an MI algorithm will be developed and piloted based on the sociodemographic and behavioral risk profile of the HIV+ participant.

Sample initial aims/hypothesis (3)

- This study is a multi-site study to determine whether the MI toolbox is associated with a decrease in HIV transmission behaviors.
- Compared to HIV-1 seroconverters in the control arm, HIV seroconverters randomized to receive the MI Toolbox will:
  (a) have a significantly lower proportion of unprotected sex acts with partners of unknown or negative HIV serostatus (***primary endpoint)***;

Statistical Refinements of Hypotheses

- Recall….
- Determine whether the group that received the treatment will tend to have outcome measurements that are

Better than
Worse than, or
About the same as

than

an absolute standard, or measurements in an otherwise comparable group that did not receive treatment

Choice of summary measure

- We need to refine scientific hypotheses about a clinical endpoint into testable statistical hypotheses about some summary measure of a distribution
- For Each Outcome Define "Tends To"
- In general, the space of all probability distributions is not totally ordered
  - There are an infinite number of ways we can define a tendency toward a "larger" outcome
  - This can be difficult to decide even when we have data on the entire population
    - Ex: Is the highest paid occupation in the US the one with
      - the higher mean?
      - the higher median?
      - the higher maximum?
      - the higher proportion making $1M per year?

Primary Endpoint: Statistical

- For a specific clinical endpoint, we still have to summarize its distribution
- Consider (in order of importance)
  - The most relevant summary measure of the distribution of the primary endpoint
  - The summary measurement the treatment is most likely to affect
  - The summary measure that can be assessed most accurately and precisely
- Statistical hypotheses are then stated in terms of the (single) summary measure

Marginal Summary Measures

- Many times, statistical hypotheses are stated in terms of summary measures for univariate (marginal) distributions
  - Means (arithmetic, geometric, harmonic, …)
  - Medians (or other quantiles)
  - Proportion exceeding some threshold
  - Odds of exceeding some threshold
  - Time averaged hazard function (instantaneous risk)
  - …
- What is most important scientifically?

Comparisons Across Groups

- Comparisons across groups then use differences or ratios
  - Difference / ratio of means (arithmetic, geometric, …)
  - Difference / ratio of proportion exceeding some threshold
  - Difference / ratio of medians (or other quantiles)
  - Ratio of odds of exceeding some threshold
  - Ratio of hazard (averaged across time?)
  - …
- What is most important scientifically?

Statistical tasks

- While we claim that the choice of the definition for “tends to be larger” is primarily a scientific issue, statisticians do usually play an important role
  - Quantifying how different summary measures capture key features of a probability distribution
  - Ensuring that the statistical analysis model truly addresses the scientific goal

Criteria for Summary Measure

- Choose some summary measure of the probability distribution according to the following criteria (in order of importance)
  - Scientifically (clinically) relevant
    - Also reflects current state of knowledge
  - Is likely to vary across levels of the factor of interest
    - Ability to detect variety of changes
  - Statistical precision
    - Only relevant if all other things are equal

Common Practice

- The overwhelming majority of statistical inference is based on means
  - Means of continuous random variables
    - t test, linear regression
  - Proportions (means of binary random variables)
    - chi square test (t test)
  - Rates (means) for count data
    - Poisson analyses

Use of the Mean

- Rationale
  - Scientific relevance
    - Measure of “central tendency” or “location”
    - Related to totals, e.g. total health care costs
  - Plausibility that it would differ across groups
    - Sensitive to many patterns of differences in distributions (especially in tails of distributions)
  - Statistical properties
    - Distributional theory known
    - Optimal (most precise) for many distributions
    - (Ease of interpretation?)

When Not to Use the Mean

- Lack of scientific relevance
  - The mean is not defined for nominal data
  - The mean is sensitive to differences that occur only in the tail of the distribution
    - E.g., increasing the jackpot in Lotto makes one person richer, but most people still lose
  - Small differences may not be of scientific interest
    - Extend life expectancy by 24 hours
    - Decrease average cholesterol in patients with familial hypercholesterolemia by 20 mg/dl

Wilcoxon Rank Sum Test

- Common teaching:
  - A nonparametric alternative to the t test
  - Not too bad against normal data
  - Better than t test when data have heavy tails
  - (Some texts refer to it as a test of medians)
- More accurate guideline
  - In general, the t test and the Wilcoxon are not testing the same summary measure
  - Wilcoxon is not transitive (can allow A > B > C > A)
  - The summary measure tested does not allow determination of clinical importance
  - Efficiency theory derived when a shift model holds for some monotonic transformation
    - If propensity to outliers is different between groups, the t test may be better even with heavy tails

Comments

- In any case, the decision regarding which parameter to use as the basis for inference should be made prior to performing any analysis directly related to the question of interest
  - Basing decisions regarding choice of analysis method on the observed data will tend to inflate the type I error
    - Decrease our confidence in our statistical conclusions

Probability Model and Summary Measures

- The scientific question posed by a clinical trial is typically translated into a statistical comparison of probability distributions
  - Unadjusted or adjusted comparison of summary measures

- We will need to describe the statistical implications of any randomization strategy in the context of statistical analysis model
  - Notation for regression on means, odds, or hazards

Summary Measures

- The measures commonly used to summarize and compare distributions vary according to the types of data
  - Means: binary; quantitative
  - Medians: ordered; quantitative; censored
  - Proportions: binary; nominal
  - Odds: binary; nominal
  - Hazards: censored
    - hazard = instantaneous rate of failure

Everything is Regression

- The most commonly used two sample tests are special cases of regression
  - Regression with a binary predictor
    - Linear → t test
    - Logistic → chi square (score test)
    - Proportional hazards → logrank (score test)

- General notation for variables and parameter
  - Y_i Response measured on *i*-th subject
  - X_i Value of predictor of interest for *i*-th subject
  - $W_{1i}, W_{2i} \ldots$ Value of adjustment variables for *i*-th subject
  - Θ_l Parameter of distribution of Y_i
  - The parameter might be the mean, geometric mean, odds, rate, instantaneous risk of an event (hazard), etc.

Regression

- General notation for simple regression model
  - $g(\Theta_l) = \beta_0 + \beta_1 X_i + \beta_2 W_{1i} + \beta_3 W_{2i} + \ldots$
  - $g(\,)$ link function used for modeling
  - β_0 *"intercept"*
  - β_1 *"slope"* for predictor of interest X
  - β_j *"slope"* for covariate W_{j-1}

  The link function is usually either none (means) or log (geom mean, odds, hazard)

Comparisons

- Define a comparison across groups to use when answering scientific question
  - If straight line relationship in parameter, slope for POI is difference in parameter between groups differing by 1 unit in X when all other covariates in model are equal
  - If nonlinear relationship in parameter, slope is average difference in parameter between groups differing by 1 unit in X "holding covariates constant"
    - Statistical jargon: a "contrast" across the groups

Regression Models

- According to the parameter compared across groups
  - Means ⇨ Linear regression
  - Geom Means ⇨ Linear regression on logs
  - Odds ⇨ Logistic regression
  - Rates ⇨ Poisson regression
  - Hazards ⇨ Proportional Hazards regr

Comparison of Models

- The major difference between regression models is interpretation of the parameters
  - Summary: Mean, geometric mean, odds, hazards
  - Comparison of groups: Difference, ratio
- Issues related to inclusion of covariates remain the same
  - Address the scientific question
    - Predictor of interest; Effect modifiers
  - Address confounding
  - Increase precision

Interpretation of Parameters

- Intercept
  - Corresponds to a population with all modeled covariates equal to zero
    - Most often outside range of data; quite often impossible; very rarely of interest by itself
- Slope
  - A comparison between groups differing by 1 unit in corresponding covariate, but agreeing on all other modeled covariates
    - Sometimes impossible to use this definition when modeling interactions or complex curves

Randomization versus (?) Adjustment

- The fundamental statistical distinctions between unadjusted and adjusted regression models are central to the goals of randomization

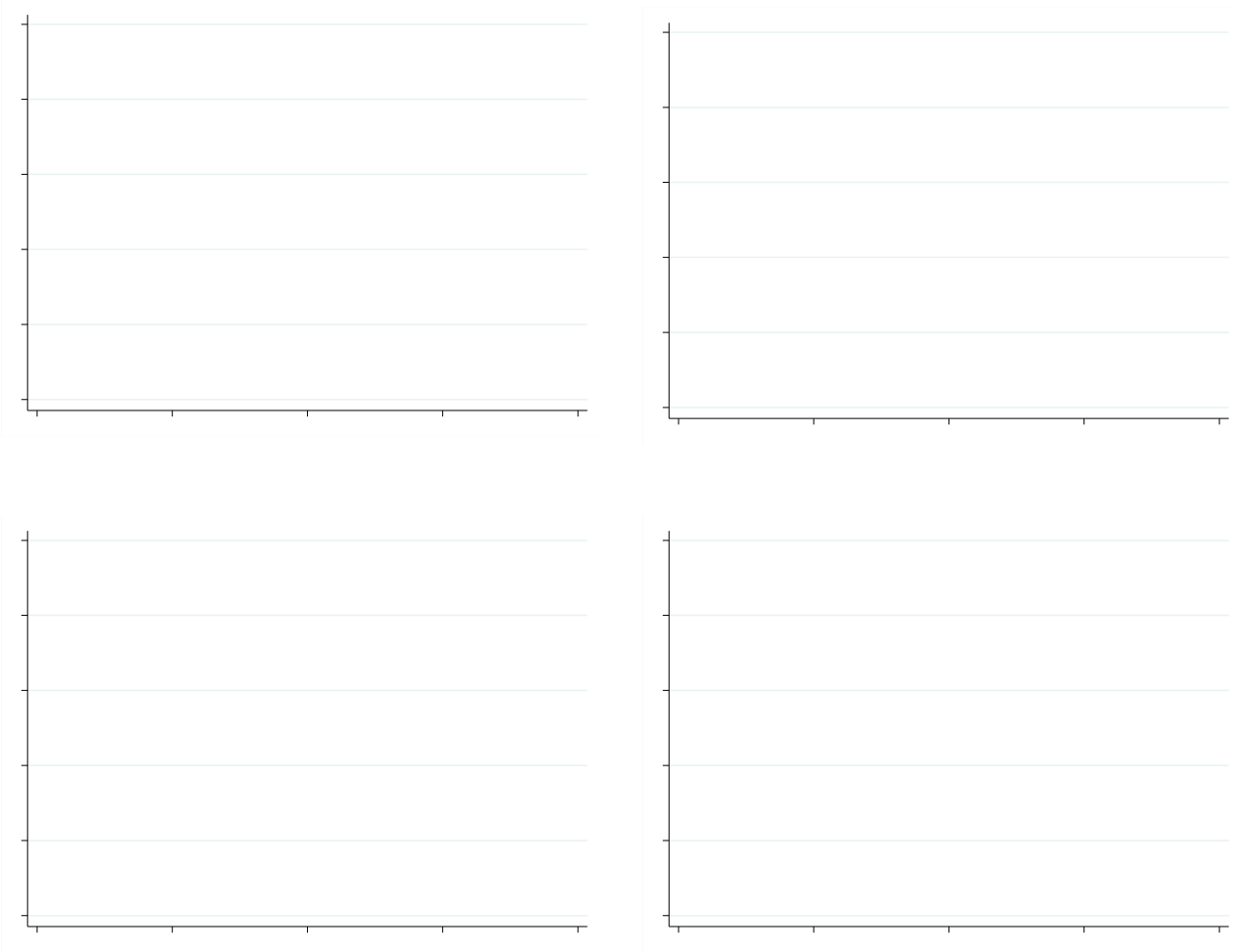
- We thus want to be able to consider the relationships between
  - unadjusted and adjusted parameters, and
  - the standard errors of the two parameter estimates.

Unadjusted vs Adjusted Models

- Adjustment for covariates changes the scientific question
  - Unadjusted models
    - Slope compares parameters across groups differing by 1 unit in the modeled predictor
      - Groups may also differ with respect to other variables
  - Adjusted models
    - Slope compares parameters across groups differing by 1 unit in the modeled predictor but similar with respect to other modeled covariates

Linear regression

- When are estimated parameters for *X* the same in adjusted and unadjusted models?



Adjustment in clinical trials

- When are estimated parameters for *X* the same in adjusted and unadjusted models?
- Answer…

- Consequence regarding presenting p-values in Table 1…

Adjustment in clinical trials

- Precision variable
  - *W* is associated with *Y* after adjustment for *X*
  - *X* and *W* are uncorrelated (no association in means)
    - Randomization !!!
- Confounding variable
  - *W* is associated with *Y* after adjustment for *X*
  - *X* and *W* are correlated (difference in means)
- If stratified randomization ⇨ adjust for stratification variable (e.g. site)

Summary Measures

- Other considerations….
- Typically: power study for (single) primary hypothesis
- Potentially: show power for important secondary hypotheses
- Pre-specify analysis for primary hypothesis in detail
- Including how to deal with missing values etc. (more tomorrow)

Sample Size Considerations

- At the end of the study, we analyze our data in order to be able to make an informed decision about the effectiveness of a new treatment

- We choose a sample size for our study in order to have sufficient precision to make such inference

Sample Size Considerations

- Hypothesis testing

The truth can only be: either H_0 true, or H_A true

	H_0 true	H_A true
We do not reject H_0	No error Prob = $1 - \alpha$	Type II error Prob = β
We reject H_0	Type I error Prob = α	No error Prob = $1 - \beta$

Type I error: falsely rejecting H_0 Probability: α
Type II error: falsely not rejecting H_0 Probability: β

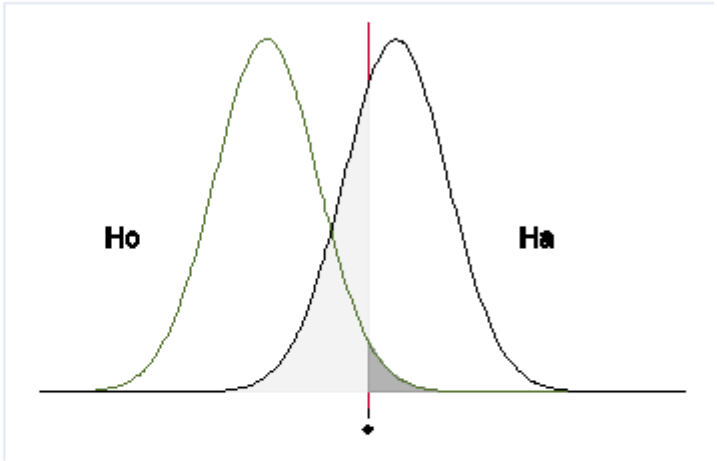
$1 - \beta$ = Power of the test = Probability of rejecting H_0 when it is false.
(more on Power later)

Sample Size Considerations

- Main goals of power / sample size calculations
  - Avoid sample size that is TOO small
  - Avoid sample size that is TOO large
  - Ethical issues
  - Financial issues

Sample Size Considerations

Power of a test:
(Assume for now, we know μ_0, μ_a and σ)



$$* = \mu_0 + z_{1-\alpha/2}\left(\sigma/\sqrt{n}\right)$$

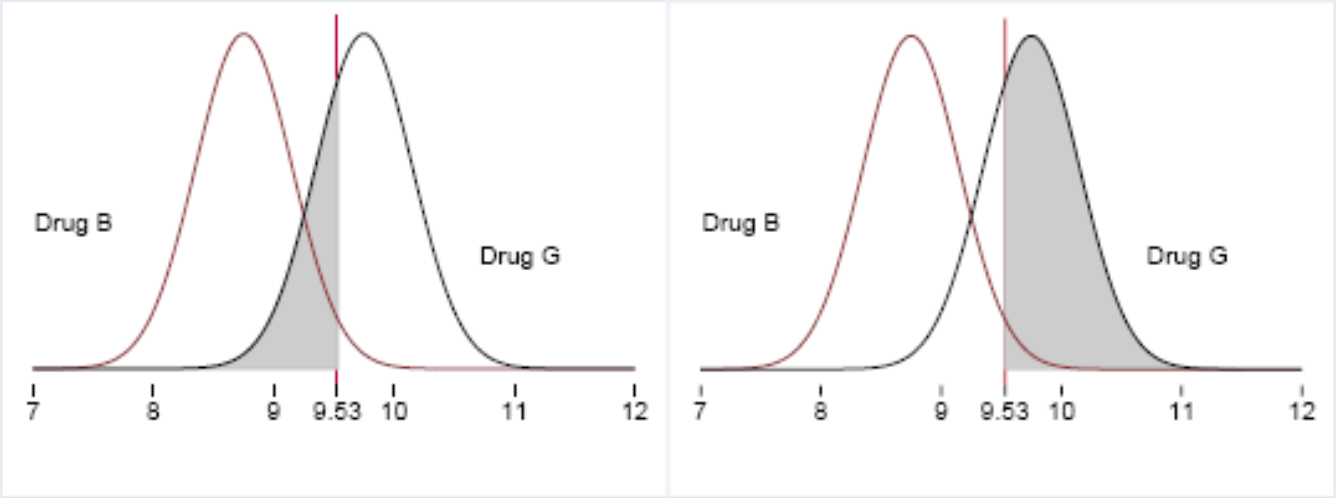
also: $$* = \mu_a - z_{1-\beta}\left(\sigma/\sqrt{n}\right) \qquad \text{(note: } z_\beta = -z_{1-\beta}\text{)}$$

Thus, $$\mu_0 + z_{1-\alpha/2}\left(\sigma/\sqrt{n}\right) = \mu_a - z_{1-\beta}\left(\sigma/\sqrt{n}\right)$$

Sample Size Considerations

- Normally distributed outcome

Shaded area represents β,
the probability of type II error



Shaded area represents $1-\beta$,
the power of the test.

Sections

Sample Size Considerations

- Can express this as ….

With, $\mu_0 + z_{1-\alpha/2}\left(\sigma/\sqrt{n}\right) = \mu_a - z_{1-\beta}\left(\sigma/\sqrt{n}\right)$

In terms of:

⇨ $\mu_a - \mu_0 = \left(z_{1-\alpha/2} + z_{1-\beta}\right)\dfrac{\sigma}{\sqrt{n}}$ magnitude of the difference

⇨ $z_{1-\beta} = \dfrac{\sqrt{n}\left(\mu_a - \mu_0\right)}{\sigma} - z_{1-\alpha/2}$ power

⇨ $n = \sigma^2 \dfrac{\left(z_{1-\alpha/2} + z_{1-\beta}\right)^2}{\left(\mu_a - \mu_0\right)^2}$ sample size (total, $n = n_1 + n_2$)

If σ is estimated, use s_p instead of σ, use t-distribution instead of normal.

Reporting Inference

- At the end of the study analyze the data
- Report three measures (four numbers)
  - Point estimate
  - Interval estimate
  - Quantification of confidence / belief in hypotheses

Reporting (Frequentist) Inference

- Three measures (four numbers)
  - Consider whether the observed data might reasonably be expected to be obtained under particular hypotheses
    - Point estimate: minimal bias? MSE?
    - Confidence interval: all hypotheses for which the data might reasonably be observed
    - P value: probability such extreme data would have been obtained under the null hypothesis
      - Binary decision: Reject or do not reject the null according to whether the P value is low

Reporting Bayesian Inference

- Three measures (four numbers)
  - Consider the probability distribution of the parameter conditional on the observed data
    - Point estimate: Posterior mean, median, mode
    - Credible interval: The “central” 95% of the posterior distribution
    - Posterior probability: probability of a particular hypothesis conditional on the data
      - Binary decision: Reject or do not reject the null according to whether the posterior probability is low

Parallels Between Tests, CIs

- If the null hypothesis not in CI, reject null
  - (Using same level of confidence)
- Relative advantages
  - Test only requires sampling distn under null
  - CI requires sampling distn under alternatives
  - CI provides interpretation when null is not rejected

Scientific Information

- "Rejection" uses a single level of significance
  - Different settings might demand different criteria
- P value communicates statistical evidence, not scientific importance
- Only confidence interval allows you to interpret failure to reject the null:
  - Distinguish between
    - Inadequate precision (sample size)
    - Strong evidence for null

Hypothetical Example

- Clinical trials of treatments for hypertension
  - Screening trials for four candidate drugs
    - Measure of treatment effect is the difference in average SBP at the end of six months treatment
    - Drugs may differ in
      - Treatment effect (goal is to find best)
      - Variability of blood pressure
    - Clinical trials may differ in conditions
      - Sample size, etc.

Reporting P values

Study	P value
A	0.1974
B	0.1974
C	0.0099
D	0.0099

Reporting point estimates

Study	SBP Diff
A	27.16
B	0.27
C	27.16
D	0.27

Reporting P values & point estimates

Study	SBP Diff	P value
A	27.16	0.1974
B	0.27	0.1974
C	27.16	0.0099
D	0.27	0.0099

Reporting Confidence Intervals

Study	SBP Diff	95% CI	P value
A	27.16	-14.14, 68.46	0.1974
B	0.27	-0.14, 0.68	0.1974
C	27.16	6.51, 47.81	0.0099
D	0.27	0.06, 0.47	0.0099

- Interpreting non-significance
- Studies A and B are both “nonsignificant”
  - Only study B ruled out clinically important differences
  - The results of study A might reasonably have been obtained if the treatment truly lowered SBP by as much as 68 mm Hg

Reporting Confidence Intervals

Study	SBP Diff	95% CI	P value
A	27.16	-14.14, 68.46	0.1974
B	0.27	-0.14, 0.68	0.1974
C	27.16	6.51, 47.81	0.0099
D	0.27	0.06, 0.47	0.0099

- Interpreting Significance:
- Studies C and D are both statistically significant results
  - Only study C demonstrated clinically important differences
  - The results of study D are only frequently obtained if the treatment truly lowered SBP by 0.47 mm Hg or less

Reporting

- If ink/space is not in short supply, there is no reason not to give point estimates, CI, and P value

- If ink/space is in short supply, the confidence interval provides most information
  - (but sometimes a confidence interval cannot be easily obtained, because the sampling distribution is unknown under the null)

General Comments

- What alternative to use?
  - Minimal clinically important difference (MCID)
    - To detect?
    - To declare significant?
- What level of significance?
  - "Standard": one-sided 0.025, two-sided 0.05
  - "Pivotal": one-sided 0.005?
    - Do we want to be extremely confident of an effect, or confident of an extreme effect
- What power?
  - Science: 97.5% (unless MCID for significance ⇨ ~50%)
  - Subterfuge: 80% or 90%
- Adjustment for sequential monitoring…

Role of Secondary Analyses

- We choose a primary outcome to avoid multiple comparison problems
  - That primary outcome may be a composite of several clinical outcomes, but there will only be one CI, test

- We select a few secondary outcomes to provide supporting evidence or confirmation of mechanisms
  - Those secondary outcomes may be
    - alternative clinical measures and/or
    - different summary measures of the primary clinical endpoint

Secondary Analysis Models

- Selection of statistical models for secondary analyses should generally adhere to same principles as for primary outcome, including intent to treat

- Some exceptions:
  - Exploratory analyses based on dose actually taken may be undertaken to generate hypotheses about dose response
  - Exploratory cause specific time to event analyses may be used to investigate hypothesized mechanisms

Safety Outcomes

- During the conduct of the trial, patients are monitored for adverse events (AEs) and serious adverse events (SAEs)
  - We do not typically demand statistical significance before we worry about the safety profile
    - We must consider the severity of the AE / SAE
  - If we perform statistical tests, it is imperative that we not use overly conservative procedures
    - When looking for rare events, Fisher's Exact Test is far too conservative
      - Safety criteria based on nonsignificance of FET is a license to kill
    - Unconditional exact tests provide much better power

Sample Size Considerations

- We can only choose one sample size
  - Secondary and safety outcomes may be under- or over-powered

- With safety outcomes in particular, we should consider our information about rare, devastating outcomes (e.g., fulminant liver failure in a generally healthy population)
  - The “three over N” rule pertains here
  - A minimal number of treated individuals should be assured
    - Control groups are not as important here, if the event is truly rare

Interpreting a "Negative Study"

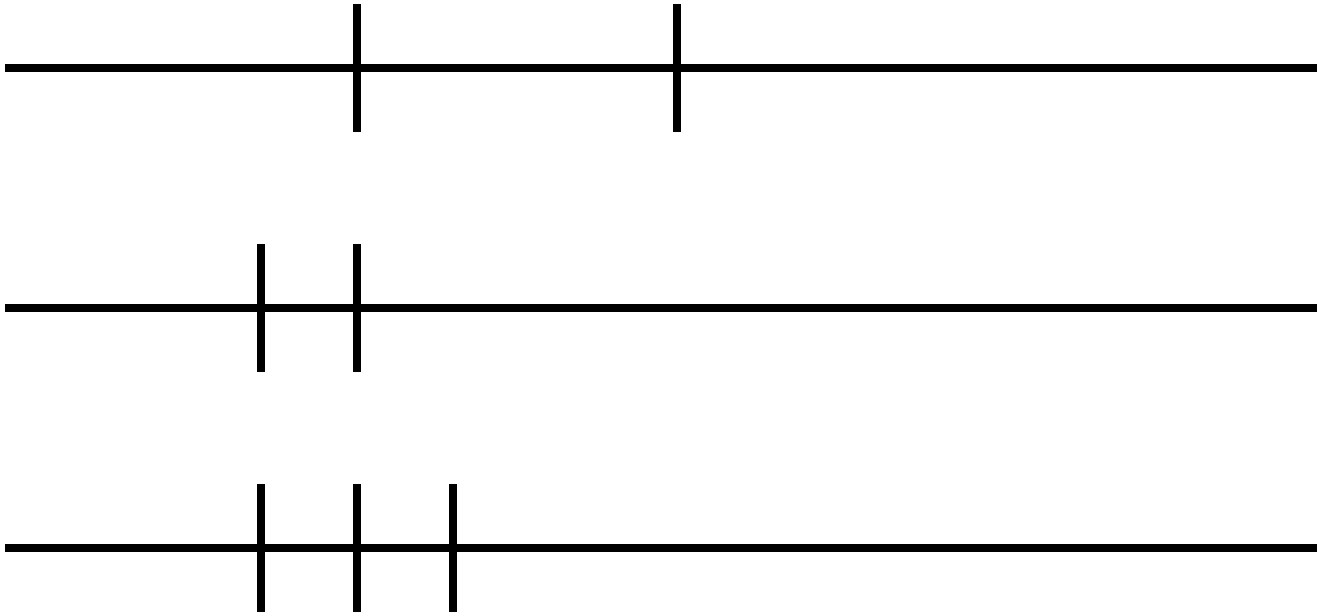
- Possible explanations for no statistically significant difference in estimate of θ
  - There is no true difference in the distribution of response across groups
  - There is a difference in the distribution of response across groups, but the value of θ is the same for both groups
    - (i.e., the distributions differ in some other way)
  - There is a difference in the value of θ between the groups, but our study was not precise enough
    - A "type II error" from low "statistical power"

Interpreting a "Positive Study"

- Analogous interpretations when we do find a statistically significant difference in estimate of θ
  - There is a true difference in the value of θ
  - There is no true difference in θ, but we were unlucky and observed spuriously high or low results
    - Random chance leading to a "type I error"
      - The p value tells us how unlucky we would have had to have been
    - (Used a statistic that allows other differences in the distn to be misinterpreted as a difference in θ
      - E.g., different variances causing significant t test)

Goal of Clinical Trial

- Establish evidence (typically) for
  - Superiority
  - Noninferiority
  - Equivalence
- Technically… confidence intervals…



Criteria for Selection

- Fundamental criteria for choosing among these types of trials
  - Under what conditions will we change our current practice by
    - Adopting a new treatment
    - Discarding an existing treatment

Conditions for Change in Treatment

- Adopting a new treatment
  - Better than using no treatment (efficacious)
  - Equal to some existing efficacious treatment
  - Better than some existing efficacious treatment

- Discarding an existing treatment
  - Worse than using no treatment (harmful)
  - (? Equivalent to using no treatment)
  - Not as efficacious as another treatment

Ethical Issues

- When is it ethical to establish efficacy by comparing a treatment to no treatment?

- When is it ethical to establish harm by comparing a treatment to no treatment?

Scientific Issues

- How to define scientific hypotheses when trying to establish
  - efficacy by comparing a new treatment to no treatment
  - efficacy by comparing a new treatment to an existing efficacious treatment
  - superiority of one treatment over another

- How to choose the comparison group when trying to establish efficacy by comparing a new treatment to an existing efficacious treatment

Statistical Issues

- How to choose sample size to discriminate between scientific hypotheses
  - To establish difference between treatments
  - To establish equivalence between treatments

Goals of Equivalence Studies

- Interplay of ethical, scientific, and statistical issues
  - Ethics often demands establishing efficacy by comparing new treatment to an active therapy
  - Scientifically the relevant hypothesis is then one of equivalence
  - Statistically it takes an infinite sample size to prove exact equivalence

Superiority over No Treatment

- Desire to establish that a new treatment is better than nothing (efficacious)
  - New treatment will be added to some standard therapy if shown to be efficacious
  - Placebo controlled if possible

Superiority over Existing Treatment

- Desire to establish that a new treatment is better than some existing treatment
  - An efficacious treatment already in use
  - New treatment will replace that efficacious treatment if shown to be superior
  - Not ethical or of interest to merely prove efficacy
  - Active control group

Common to Both

- In either case, the goal of superiority trials is to rule out equality between two treatments
  - And thus also rule out inferiority of the new treatment

Noninferiority Trials

- Desire to establish that a new treatment is not so much worse than some other treatment as to be nonefficacious
  - Show new treatment is efficacious
    - New treatment will be made available if it provides benefit
    - An efficacious treatment already in use
    - Not ethical to compare new treatment to no treatment
    - Active control group
      - But, we need not be superior to the active group, nor ostensibly even at the same level of efficacy
      - Define a “Noninferiority Margin” as the level of decrease in efficacy relative to active control that is “unacceptably inferior”

Use of Noninferiority Trials

- Noninferiority trials of use when
  - Trying to adopt a new treatment without the expense of proving superiority
    - Often the sponsor actually believes it is superior
  - Trying to improve secondary endpoints without removing efficacy on primary endpoint
    - E.g., in cancer chemotherapy, adverse events often correlated with efficacy

Major Issues with Noninferiority Trials

- Presumption that active control would be efficacious in the current trial
  - And the need to quantify that level of efficacy
- Establishing the noninferiority margin
  - How much of a decrease in efficacy is “unacceptably inferior”?
  - How certain do we have to be that we have not exceeded that limit?

Two-sided Equivalence Studies

- Desire to rule out all differences of clinical relevance
  - Show new treatment is approximately equivalent to existing treatment
    - New treatment will be made available if it provides approximately same level of benefit as existing treatment
    - Goal can be establishing efficacy or just establishing no harm
    - Key is in definition of “approximately equivalent” in a way to rule out the minimal clinically important differences

Statistical Consideration

- When confidence intervals are used as the criteria for statistical evidence
  - Superiority, noninferiority, and equivalence trials are distinguished only by
    - defining the hypotheses which you desire to discriminate
    - choosing sample sizes to ensure that confidence intervals will discriminate between those hypotheses

Sample Size

- Heretofore we have primarily considered randomization to two independent groups

- Sometimes we can gain efficiency by using more complex designs

Cluster Randomization

- When treatment cannot be administered on an individual level without contamination
  - E.g., smoking cessation programs
  - E.g., education strategies
  - E.g., out of hospital emergency response
- Subjects randomized to treatment or control in clusters
  - Often form matched sets of clusters to randomize in strata

Cluster Randomization

- Advantages
  - Allows investigation of community interventions
  - Intervention at clinic or village level may be perceived as more ethical
  - Logistical considerations for equipment, etc.
- Disadvantages
  - Sample size may be the number of clusters rather than the number of subjects
  - May lose substantial power over randomization by individual

Cross-over Trials

- Each subject receives every treatment
  - May gain precision because each subject serves as own control
  - Order of treatments should be randomized
    - **A pre/post design is not correctly termed a cross-over design**
  - Washout period to avoid carryover effects
    - Analyses should look for differences in treatment effect by order of administration
  - Not feasible with most time to event studies

Cross-over Trials

- Advantages
  - Greater statistical power in presence of high ratio of between subject to within subject variability in response
    - I.e., when high correlation between repeat measurements of response
- Disadvantages
  - Cannot be used in presence of
    - curative treatments
    - long carryover (and statistical power to detect carryover is usually low)

Factorial Designs

- Test two or more treatments simultaneously
  - Every subject gets either active or control for each treatment
  - Example: Two treatments: A vs PlcA and B vs PlcB
    - Four treatment groups
      - A and B; A and PlcB; PlcA and B; PlcA and PlcB
- Partial Factorial
  - Some subjects might only participate in one part of the trial
    - Additional treatment groups
      - A only; PlcA only; B only; PlcB only

Factorial Designs

- Advantages
  - Answer multiple questions with the same study
    - In absence of effect modification, same power as individual studies
    - Ability to address effect modification (but with low power)
- Disadvantages
  - Exclusion criteria must consider all treatments
  - One treatment may affect compliance on all treatments
  - AEs from one treatment may affect ascertainment bias on all treatments

Large Simple Trials

- Use many subjects and minimize amount of data collected
  - Definition of treatment must be straightforward
  - Definition of outcome must be straightforward
- Allows looking at smaller increments of benefit
- Must not sacrifice scientific rigor, however
  - Ability to assess mechanism of action
  - Ability to detect unexpected toxicity

Case Study

- From abstract:
- “This paper proves that the placebo group (saline) displays a tendency, as indicated by two statistical tests, towards a significant increase in the red blood cells lost in the 24 hours after the operation.”

Comments?

- **Reference:** Gray and Polakow, A study of Premarin intravenous and its influence on blood loss during transurethral prostatectomy, Journal of International Medical Research, 1979, 7(1) 96-99.

Case Study

From the same paper:

"… Once the patient had been operated upon and the exclusive pathology became known this disqualified the patient from the study retrospectively. The exclusions were:

- Coagulation disorders.
- Previous surgery to prostate.
- …
- …
- Severe pre-operative anaemia.
- Admission haemoglobin less than 11 grams %.
- History of salicylate, steroid or anti-inflammatory ingestion during the preceding six months.
- Prostatic carcinoma."

Case Study

- Out of “47 consecutive patients undergoing transurethral prostatectomy between 03/09/75 and 12/05/77 were studied”….

- Guess how many were excluded due to the above criteria?

- They did not report on how they handled potential exclusions in the power calculations

Cartoon

- From <www.CAUSEweb.org>

"We test thousands of new treatments each year, so to avoid multiple testing issues we always do a validation experiment to confirm our positive results".

Cartoon

- From <www.CAUSEweb.org>

How often do those work out?

Cartoon

- From <www.CAUSEweb.org>

About 5% of the time!

Questions?

References

- Friedman LM, Furberg CD and DeMets DL: *Fundamentals of Clinical Trials*
- Pocock SJ: *Clinical Trials: A Practical Approach*