SISCR
UW - 2018

Introduction to Clinical Trials – Day 1

Session 1 - Background

Presented July 23, 2018

Susanne J May
Department of Biostatistics
University of Washington

Daniel L Gillen
Department of Statistics
University of California, Irvine

General comments

- Organizational…
  - UW netID: to be announced
  - Password: to be announced
- Assumed prior knowledge
  - Basis statistics
  - Basis study designs
- PLEASE ask questions at any time
- Includes slides/content from Scott Emerson, Tom Fleming

Overview – Introduction to Clinical Trials

- Primary Design Issues
  - Introduction / Background
  - Screening Studies
  - Fundamentals of Trial Design
  - Statistical Tasks in Trial Design
  - Elements of a Clinical Protocol
- Conduct and Implementation
  - Randomization strategies
  - Blinding
  - Specification of secondary endpoints
  - Handling of missing data
  - Conduct and monitoring of the study
  - Independent data monitoring committee
  - Reporting of the result

Clinical Trials

- Clinical trials in the context of other studies
- Observational epidemiology of disease, risk
- Preclinical experiments
  - Laboratory, animal studies of mechanisms, toxicology
- Clinical trials
  - Safety for further investigations / dose
  - Safety of therapy
  - Measures of efficacy
  - Confirmation of efficacy / effectiveness
- Synthesis and quantification of evidence
- Adoption of new treatment indication

Start at the end….

- Where do we want to be?
  - Find a new treatment that improves health of individuals
  - Find a new treatment that improves health of the population
    - Treatments administered to a community
    - Treatments tested on a population

Treatment Indication

- Disease
  - Putative cause vs signs / symptoms
    - May involve method of diagnosis, response to therapies
- Population
  - Restrict by risk of AEs or actual prior experience
- Treatment or treatment strategy
  - Formulation, administration, dose, frequency, duration, ancillary therapies
- Outcome
  - Clinical vs surrogate; timeframe; measurement

Disease

- A moving target heavily influenced by treatment
  - Then: “fevers”
  - Now: “MRSA-related pneumonia”
- Trends over place and time in definition because
  - Symptoms
    - Cultural / geographic effects, earlier recognition, symptomatic treatments, comorbidities
  - Signs
    - New diagnostic modalities, other prevention strategies (e.g., TB vaccine) and treatments
  - Unmet need
    - Effective treatment already discovered for subset

Definition of Disease

- Specify the disease targeted by the therapy
  - Scientifically
    - Putative cause of constellation of symptoms
    - Symptoms / signs from multiple causes
  - Clinically
    - Diagnostic criteria
      - Incident vs prevalent
      - Symptoms
        - Intensity, frequency, duration, response to treatment
      - Signs
        - Method of measurement
        - Magnitude, reproducibility
    - Prior treatment history

Population

- Treatment indications may be restricted to a specific population
  - Demographics: age, sex
  - Genetics: drug metabolism
  - Comorbid conditions
    - Drug metabolism: renal, liver disease
    - Drug side effects: cardiovascular disease, bleeding
  - Prior treatment history: resistance to alternatives
  - Vulnerable populations
    - Pediatrics, pregnancy

Definition of Treatments

- Full description
  - Formulation of treatment
  - Dose, administration, frequency, duration
    - Rules for responsive dosing (e.g., insulin)
    - Include plans for
      - Treatment of adverse events
      - Dose reduction
      - Dose discontinuation
  - Ancillary treatments
    - Prescribed vs allowed vs prohibited

Outcomes

- The desired beneficial response from the treatment
  - Clinical outcomes
    - Prolonged survival
    - Quality of life
  - Surrogate outcomes
    - Improvement in some risk factor believed to be predictive of a good clinical outcome
- Definition
  - Method of measurement
  - Timeframe

Diagnostic Test "Indication"

- Disease
  - Putative cause vs eventual outcomes
- Population
  - Identify risk factors (prevalence)
  - Eliminate known false positives, false negatives
- Test or testing strategy
  - Formulation, administration, method of measurement
- Outcome
  - Sensitivity, specificity
  - Predictive value of positive, negative

Prognostic Test "Indication"

- Disease
  - Clinical diagnosis
- Population
  - Identify risk factors (prevalence)
  - Eliminate known false positives, false negatives
- Test or testing strategy
  - Formulation, administration, method of measurement
- Outcome: Clinical event or survival
  - Sensitivity, specificity
  - Predictive value of positive, negative

Clinical Decision for Diagnosis

- What test provides the best information for a patient
  - Based on what we know about the patient?
  - Based on what we know about the test?

Clinical Decision for Treatment

- What is the best treatment to give a patient
  - Based on what we know about the patient?
  - Based on what we know about the treatment?

Second Consideration

- Synthesize and evaluate evidence for a therapy
  - Analysis and interpretation of clinical studies

- Evidence Based Medicine (PICO)
  - Patient population
    - Disease and population characteristics
  - Intervention
    - Precise description of treatment strategy
  - Comparator
    - Alternatives in the absence of the new treatment
  - Outcome
    - Both beneficial and adverse

Third Consideration

- Where do we get the data to be synthesized?
  - Well designed clinical interventional studies

- Clinical trials
  - Experimentation in human volunteers
  - Investigates a new treatment/preventive agent
    - Safety:
      - Do adverse effects outweigh potential benefit?
    - Efficacy:
      - Does treatment beneficially alter the disease process
    - Effectiveness:
      - Would adoption of the treatment improve morbidity / mortality in the population?

Scientific Setting

- The goal of medical science is to produce the evidence that can be used to
  - Gain approval of new treatments and diagnostic tests
  - Provide evidence to be used in applying those treatments and tests

Level of Evidence

- U.S. Preventive Services Task Force
  - **Level I:** At least one properly designed RCT
  - **Level II:**
    - **II-1:** Well-designed, nonrandomized CT
    - **II-2:** Well-designed, multicenter cohort/case-control
    - **II-3:** Multiple time series with/without intervention; Dramatic results from uncontrolled trial
  - **Level III:** Opinions of respected authorities = *Eminence based (not their wording!)*

Legal Requirements for Good Science

- Wiley Act (1906)
  - Labeling
- Food, Drug, and Cosmetics Act of 1938
  - Safety
- Kefauver – Harris Amendment (1962)
  - Efficacy / effectiveness
    - " [If] there is a lack of substantial evidence that the drug will have the effect ... shall issue an order refusing to approve the application. "
    - "...The term 'substantial evidence' means evidence consisting of **adequate and well-controlled investigations, including clinical investigations**, by experts qualified by scientific training"
- FDA Amendments Act (2007)
  - Registration of RCTs, Pediatrics, Risk Evaluation and Mitigation Strategies (REMS)

Kefauver-Harris Amendments (1962)

- required that manufacturers prove the effectiveness of drug products before they go on the market, and afterwards report any serious side effects
- required that evidence of effectiveness be based on adequate and well-controlled clinical studies conducted by qualified experts
- mandated that FDA conduct a retrospective evaluation of the effectiveness of drugs approved for safety—but not for effectiveness—between 1938 and 1962
- transferred to FDA control of prescription drug advertising, which would have to include accurate information about side effects

Typical Chronology

- Observational epidemiology of disease, risk
- Preclinical experiments
  - Laboratory, animal studies of mechanisms, toxicology
- Clinical trials
  - Safety for further investigations / dose
  - Safety of therapy
  - Measures of efficacy
  - Confirmation of efficacy / effectivenesss
- Synthesis and quantification of evidence
- Adoption of new treatment indication

Types of Studies - 1

- Anecdotal observations
  - Case report
  - Case series
  - Hypothesis generation

That's not an experiment you have there, that's an experience.
Sir Ronald A. Fisher (1890 - 1962)

The plural of anecdote is not data.
Roger Brinner

Types of Studies - 2

- Designed observational study: Case - control
  - Sample diseased and nondiseased
  - Examine rates of exposures
  - Efficient for rare diseases
  - Can look at multiple risk factors
  - Limitation: Cannot infer cause and effect
    - Correlations with other factors
    - Protopathic associations

Types of Studies - 3

- Designed observational study: Cohort study
  - Sample exposed and nonexposed
  - Examine rates of disease
  - Efficient for common diseases
  - Can look at multiple diseases
  - Can identify “retrospective cohort”
  - Limitation: Cannot infer cause and effect
    - Correlations with other factors
    - Protopathic associations

Types of Studies - 4

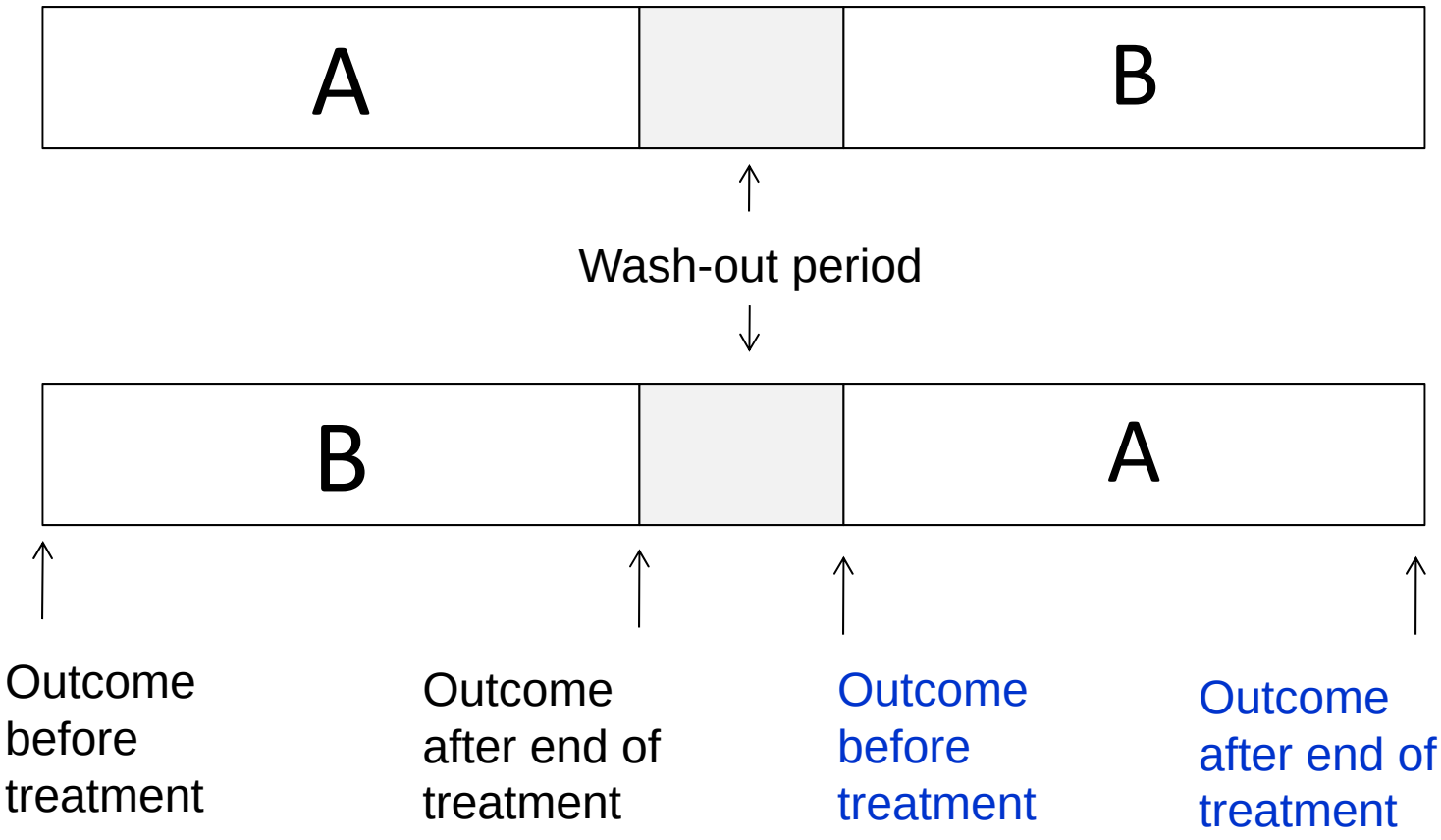
- Designed interventional study: Clinical trial
  - Assign subjects to treatments
  - Examine outcomes
  - Can look at multiple diseases
  - Can infer cause and effect

Example – Design/Statistics

- Example
  - Double-blind placebo controlled cross-over trial
  - Completed
  - Needed help with statistical analysis and write-up

Example: Cross-over trial

- Typical design for two treatments

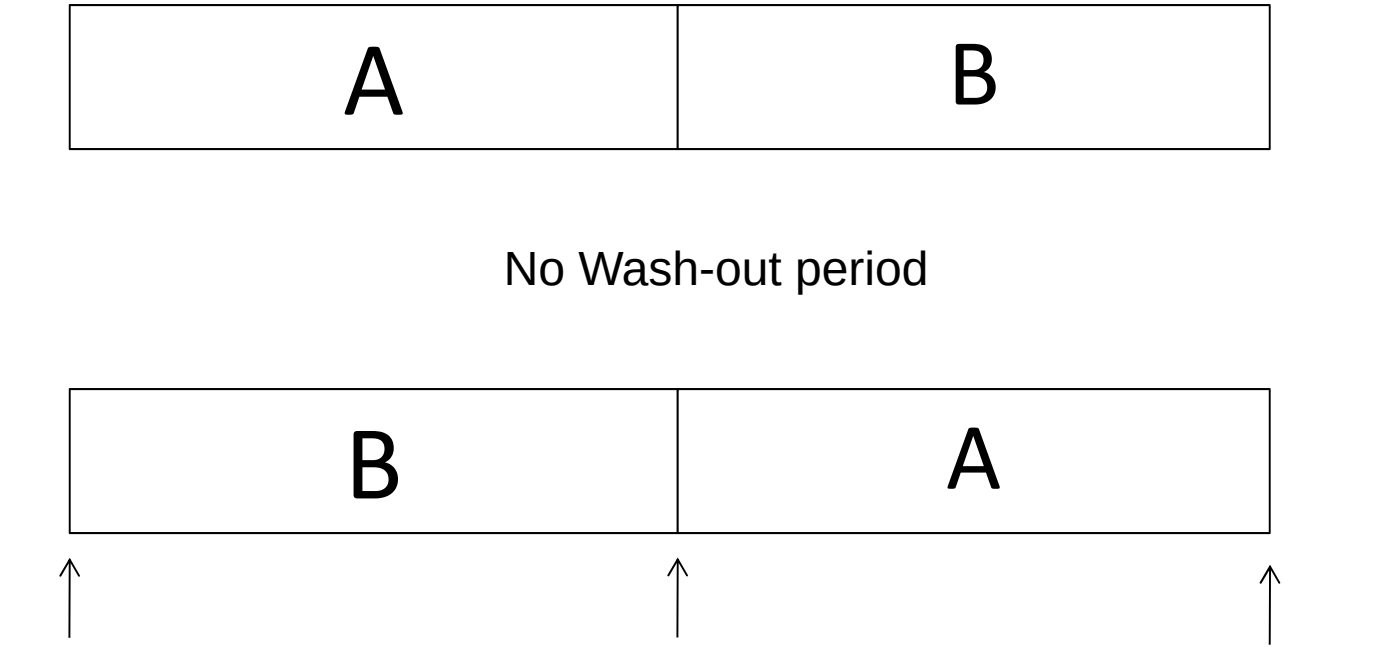


Example: Cross-over trial

- Design for this study
  - Primary outcome
    - Hemoglobin level
  - Secondary outcome
    - Frequency and intensity of nose bleeds
  - Time points of evaluation
    - Before treatment initiation
    - 3 months
    - 6 months

Example: Cross-over trial

- Design for this study



Example: Cross-over trial

- Potential issues
  - Last observation from first phase
    = first observations last phase
  - No wash-out period … justifiable?

Example: Cross-over trial

- What if there is an effect under A?
- Calculations
  - Change under A – Change under B

  AB sequence
  - hbA [t=0] – hbA [t=3] – (hbB [t=3] – hbB [t=6])

    9 – 14 – (14 – 9) = – 10

  BA sequence
  - hbA [t=6] – hbA [t=3] – (hbB [t=0] – hbB [t=3])

    9 – 14 – (9 – 9) = – 5

  - Under this calculation they should be the same!!!

Example: Cross-over trial issues

- Absence of carry-over effect – violated
- Basic premise underlying the design
- Misunderstanding – wash-out period
- “Wash-out” of drug in the system
- Sustained effect on hemoglobin

Example: Cross-over trial

- N = 106 patients interested and contacted
- N = 34 examined
- N = 22 started drug intake
- N = 2 non-compliant

- 9 Female, 11 Male
- 52 mean age (range 34 – 72)

Take home message

- "Bottom line" ?
- Statistician NEEDS to understand concepts of the application
- Investigator NEEDS to understand concepts of the statistical analysis

Phases of Investigation

- Series of studies support adoption of new treatment
  - Preclinical
    - Epidemiology including risk factors
    - Basic science: Physiologic mechanisms
    - Animal experiments: Toxicology
  - Clinical
    - Phase I: Initial safety / dose finding
    - Phase II: Preliminary efficacy / further safety
    - Phase III: Confirmatory efficacy / effectiveness
  - Approval of indication
    - (Phase IV: Post-marketing surveillance, REMS)

Phase III Clinical Trials

- Common scenarios
  - Establish efficacy / effectiveness of new treatment
    - superiority over no intervention
    - superiority over existing treatment
  - Establish equivalence with current treatment
    - Two-sided equivalence: bioequivalence
      - establish response not markedly higher or lower
    - One-sided equivalence: noninferiority
      - establish treatment not markedly worse
      - perhaps superior on secondary endpoint
  - Establish harm of existing treatment

Clinical Trials

- Experimentation in human volunteers
- Investigation of a new treatment or preventive agent
  - Safety: Do adverse effects outweigh any benefit?
  - Efficacy: Can treatment beneficially alter disease?
  - Effectiveness: Would adoption of the treatment help population's health?
- Investigation of existing treatments
  - Relative benefits: Is one treatment clearly superior?
  - Harm: Should a therapy currently in use be removed?
- Some questions cannot be answered by a clinical trial
  - E.g., establishing harm of a new substance

Example

- Random high bias

- Fleming (2010, Ann Intern Med, 153, 400)
  - Exploratory analyses
  - Misleading *P values*
  - Overestimate of effect sizes

Maternity Ward Example

- Fleming (2010)
  - Nursery, 22 infants
    2 boys
    20 girls

P-value: 0.0001

Maternity Ward Example

- Performing second study based on promising results

  (not part of the Fleming article)

  – Maternity unit 1, 22 infants (2 boys, 20 girls)
  – Maternity unit 2, 22 infants

    11 boys

    11 girls

Maternity Ward Example

- Combining study data
  - Maternity unit 1, 22 infants (2 boys, 20 girls)
  - Maternity unit 2, 22 infants (11 boys, 11 girls)

  Combined (13 vs 31)

  P-value: 0.0096

  Need to address the fact that these were two studies!

  Even then, carrying the problem forward!

Efficacy: A Moving Target

- Definition of efficacy can vary widely according to choice of endpoint and magnitude of importance
  - Basic science
    - Does treatment have any effect on the pathway
  - Clinical science
    - Does treatment have a sufficiently large effect on a clinically relevant endpoint in some subpopulation of the target population

Effectiveness: A Moving Target

- A treatment is “effective” if its introduction improves health in the population
  - Considers the net effect of safety and efficacy in the population as a whole
  - Takes into account such issues as
    - Noncompliance
    - Off-label use

Effectiveness vs Efficacy

- A treatment can be both efficacious and ineffective depending on such factors as
  - Target population
    - Restricted eligibility due to toxicity, compliance
  - Intervention
    - Training, quality control, compliance
  - Comparison treatment
    - No treatment, active treatment, ancillary treatments
  - Measurement of outcome(s)
    - Clinical disease vs subclinical markers
  - Summary measure of outcome distribution
    - Effects on mean, median, outliers

Disease

- Efficacy and effectiveness study populations may differ with respect to
  - Certainty of diagnosed disease
  - Subgroups with more (less) severe disease

Target Population

- Efficacy and effectiveness study populations may differ with respect to
  - Properly diagnosed disease
  - Subgroups with more (less) severe disease
  - Tolerance of treatment
  - Willingness to comply with treatment
  - Ancillary treatments
  - Different risk factors

Ex: Desensitization in Allergy

- Efficacy trial might consider
  - Patients with proven allergy who have shown "response" in open label study (perhaps due to genetic profile?)
  - Exclusion criteria for safety in trial
    - Cannot tolerate oral food challenge
    - Patients likely to be noncompliant
  - Exclusion criteria to ensure adequate data

- Effectiveness populations might include
  - All patients with reported allergy

Control Treatment

- Efficacy and effectiveness study populations may differ with respect to
  - Use of existing alternative treatments
  - Allowed ancillary treatments

Ex: Control Treatment in Allergy

- Efficacy trial might consider
  - Placebo
  - Careful control of diet

- Effectiveness populations should be best current standard of care
  - Will patient's behavior differ when they know their treatment assignment?

Intervention

- Efficacy and effectiveness populations may differ with respect to
  - Dose
  - Administration
  - Duration
  - Training
  - Quality control

Ex: Insulin Dependent Diabetes

- Efficacy trial might consider
  - Glucose monitoring according to protocol
  - Lengthy training
  - Close monitoring and retraining when necessary

- Effectiveness trial should strive for realistic setting
  - What would instructions and training, monitoring be if treatment were efficacious
  - What if treatment fails (use another)

Measurement of Outcome

- Efficacy and effectiveness populations may differ with respect to
  - Clinical measurement
  - Timing of measurement

Ex: Hypercholesterolemia

- Efficacy trial might consider
  - Lowering of serum cholesterol
  - Means

- Effectiveness trial should strive for relevant outcome
  - Proportion exceeding acceptable thresholds
    - Normal cholesterol levels
  - Time of survival

Which: Efficacy or Effectiveness

- Factors leading to efficacy trials
  - "Knowledge is good"
  - As pilot studies before prevention studies

- Factors leading to effectiveness trials
  - Serious conditions
    - Patients generally want to get better
  - Short therapeutic window for treatment
  - Waiver of informed consent
    - Do not withhold beneficial treatments in order to establish mechanisms
  - High cost of clinical trials (time, people, $$)

Typical (?!) Scientific Hypotheses

- The treatment will cause an individual's outcome to be

Better than
Worse than, or
About the same as

than

an absolute standard, or

what it should have been with some other treatment

Counterfactual

- The statement of the hypotheses assumed that it is possible to know what would have happened under some other treatment
  - Generally we instead have to measure outcomes that are observed
    - in another place (patient),
    - at another time, and / or
    - under different circumstances

Causation vs Association

- Truly determining causation requires a suitable interventional study (experiment)
  - Comparisons tell us about associations
  - Associations in the presence of an appropriate experimental design allows us to infer causation
    - But even then, we need to be circumspect in identifying the true mechanistic cause
      - E.g., a treatment that causes headaches, and therefore aspirin use, may result in lower heart attack rates due entirely to the use of aspirin

Investigating the Unknown

- We must acknowledge that we might be wrong
  - It will be impossible to prove something that is not true
  - The treatment might not work as we had hoped

Review: Statistical Hypotheses Testing

- According to Karl Popper (Austrian Philosopher):
- We can NOT prove that a hypothesis is true
- We can ONLY falsify a hypothesis

- Thus, “if we want to show” that a treatment “works” compared to a control, we start out by assuming that it has the same effect as the control, and try to disprove it.

Review: Statistical Hypotheses Testing

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

	H_0 is true	H_A is true
We fail to reject H_0	No error Prob 1-α	Type II error Prob β
We reject H_0	Type I error Prob α	No error Prob 1-β

- Type I error: falsely rejecting H_0
- Type II error: falsely not rejecting H_0
- Yogi Berra (slightly misquoted): Don't make the wrong mistake!

(Yogi Berra said: "I made a wrong mistake")

First Statistical Refinement

- Determine whether the group that received the treatment will 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

Variation in Response

- There is, of course, usually variation in outcome measurements across repetitions of an experiment
  - Variation can be due to
    - Unmeasured (hidden) variables
      - In the process of scientific investigation, we investigate one “cause” in a setting where others are as yet undiscovered
      - E.g., mix of etiologies, duration of disease, comorbid conditions, genetics when studying new cancer therapies
    - Inherent randomness
      - (as dictated by quantum theory)

Second Statistical Refinement

- 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

Ethics and Roles of Oversight Committees

- IRBs (REBs) Institutional Review Board
- PRC Protocol Review Committee
- DMC / DSMB Data Monitoring Committee
  Data Safety Monitoring Board
- SMC Study Monitoring Committee

Equipoise

- A state of equilibrium
- A state of balance
- Counterbalance

- Individual / collective / on average ?????

IRB – Protection of Human Subjects

- Review and approval by the Institutional Review Board (IRB) or the Human Subjects Division (HSD) is required before starting research involving human subjects.
- Authority to determine:
  - activity IS or IS NOT research and/or involving human subjects, or
  - that research qualifies for exemption
  - approve
  - conditionally Approve
  - require Modifications
  - defer
  - disapprove
  - terminate or suspend some or all parts of a study

PRC – Protocol Review Committee

- Typically, established for network studies otherwise not peer reviewed
- tasked with assessing the scientific and design merit of each protocol including:
- Importance of the question to be addressed
- Merit of experimental design, including appropriate controls
- Availability of adequate resources
- Adequacy of patient population and number of patients, including appropriate representation of minorities and women

PRC – Protocol Review Committee

- Appropriate recruitment strategies
- Adequacy of proposed plans for data acquisition, transfer, management and analysis
- Adequacy of quality control of data collection and monitoring and overall coordination of protocol management
- Description of appropriate plans to train center personnel to accomplish proposed research goals

Once protocols are finalized through the PRC process, DSMB members may still subsequently vote for approving the protocol but their responsibilities are for data quality and safety assurance.

DMC / DSMB

- To protect the interests of the study participants
- To preserve trial integrity and credibility in a manner that will enable the clinical trial
- To provide timely and reliable insights to the broader clinical community

- This requires
  - Judgement
  - … well informed
  - … independent
  - … scientifically objective
- Motivates fundamental principles for membership and function

DMC / DSMB

- Multidisciplinary representation
  - Clinical trialist
  - (Bio)statistician
  - Ethics
  - Specific area(s) of research in question
- Freedom from apparent significant conflicts of interest… financial, professional, regulatory
- Sole Access to Interim results on relative efficacy & safety of interventions

NHLBI policy regarding blinded data

- Note that as a general rule, the representative of the investigators is not permitted to receive blinded data or participate in discussions of blinded data that are collected during the investigation. This is particularly important if the representative is involved with seeing participants or is otherwise involved in a major way with running a clinic.

DMC / DSMB – more specifics

- monitor the data from the clinical trial regularly, review and assess the safety and performance of its operations, safeguard the interests of study participants,
- And make recommendations with respect to:
  - Participant Safety
  - Efficacy of the study intervention
  - Benefit/risk ratio of procedures and participant burden
  - Selection, recruitment, and retention of participants
  - Adherence to protocol requirements

DMC / DSMB – more specifics continued

- And make recommendations with respect to:
  - Data and Statistical Analysis plan
  - Adequacy of measured and collected data
  - Possible amendments to the study protocol and consent forms
  - Performance of individual centers and core labs
  - Impact of proposed ancillary studies and sub-studies on participant burden and overall achievement of the main study goals
- Open and closed (and closed closed) sessions
- Summary of deliberations are reported to IRBs

- ⇨ Module 5 (tomorrow)

Relative Responsibilities/Relationships

- Sponsors, Investigators, Care Givers
  - Decision making responsibilities for design, conduct, & analysis of the trial
  - Primary patient care responsibilities
- IRBs and Regulatory Authorities
  - Approval of Ethics/Science of the Trial Design
  - Real time Monitoring of Safety (SAEs)
- Data Monitoring Committees
  - Sole access during conduct of the clinical trial to: aggregated efficacy/safety data across the trial unblinded by treatment group

Study Monitoring Committee

- Internally reviewing the conduct of studies with respect to
  - quality assurance of protocol implementation and data collection
  - proposing standards for acceptable adherence to protocols and data collection/data entry
  - development of performance reports
  - regular review of the reports of study progress and (site) performance
  - interacting with sites necessary to develop plans for addressing any particular areas of concern.
  - reporting to investigators etc. on study progress, adherence to study protocols, and adherence to standards for data completeness, timeliness, and quality
  - Potentially advise on sites readiness for starting enrollment on RCTs

Elements of a Clinical Protocol

1. Background
2. Objectives
3. Study Design
4. Materials and Methods
5. Assessment of Safety
6. Investigator Requirements
7. Human Subjects Concerns

Elements of a Clinical Protocol

1. Background
   - 1.1 Introduction
   - 1.2 Investigational treatment A
     - 1.2.1 Description of treatment
     - 1.2.2 Prior clinical experience
       - 1.2.2.1 Safety profile
       - 1.2.2.2 Evidence of effectiveness
   - 1.3 Investigational treatment B
     - …
   - 1.4 Combined therapy A & B
     - …
       - 1.4.2.1 Safety profile
         - Subadditive, additive, supraadditive effects
       - 1.4.2.1 Evidence of effectiveness
         - Subadditive, additive, supraadditive effects

Elements of a Clinical Protocol

2. Objectives

2.1 Primary Objectives

State in terms of
- targeted population
- treatments being compared
- the clinical measure of outcome
- summary measure of that outcome

2.2 Secondary Objectives

List of each secondary objective stated similar to above

Elements of a Clinical Protocol

3. Objectives

   3.1 Description of the study

   3.2 Rationale for the study design

   3.3 Outcome measures

   - Primary effectiveness outcome measure
   - Secondary effectiveness outcome measure
   - Safety outcome measures
   - summary measure of that outcome

   3.4 Safety Plan

   3.5 Ethical Issues

   3.6 Administrative Structure

   3.7 Regulatory oversight

Elements of a Clinical Protocol

4. Materials and Methods
   - 4.1 Patients
     - 4.1.1 Patient selection
     - 4.1.2 Inclusion criteria
     - 4.1.3 Exclusion criteria
   - 4.2 Method of treatment assignment and blinding
   - 4.3 Study Treatment(s)
   - 4.4 Concomitant and excluded therapies
   - 4.5 Study assessments
   - 4.6 Patient withdrawal from study
   - 4.7 Study termination
   - 4.8 Statistical Methods

Elements of a Clinical Protocol

4. Materials and Methods…

4.8 Statistical Methods

4.8.1 Timing of analyses

4.8.2 Monitoring of study conduct

4.8.3 Analysis populations

4.8.3.1 Safety population

4.8.3.2 Primary effectiveness population

4.8.3.3 Secondary effectiveness population

4.8.4 Assessing comparability of treatment arms

4.8.5 Efficacy Analyses

4.8.6 Safety Analysis

4.8.7 Pre-specified exploratory analyses

4.8.8 Determination of sample size

4.8.9 Safety monitoring plan

Elements of a Clinical Protocol

5. Assessment of safety
   - Any further information on specific safety variables
   - Definition of methods and timing for assessing adverse events, serious adverse events
   - Reporting requirements for SAEs
   - Data safety monitoring board

6. Investigator requirements
   6.1 Study initiation
   6.2 During conduct of the study
   6.3 Study completion

7. Human Subjects Concerns

Summary

- Devil is in the details…. !!!
- Have to think through all aspects of project development – taylored designs

Introduction to Clinical Trials - Day 1

Session 2 - Screening Studies

Presented July 23, 2018

Susanne J. May
Department of Biostatistics
University of Washington

Daniel L. Gillen
Department of Statistics
University of California, Irvine

Screening Trials

Need for exploratory science

- Before we can do a large scale, confirmatory Phase III trial, we must have
  - A hypothesized treatment indication to confirm
    - Disease
    - Patient population
    - Treatment strategy
    - Outcome
  - Comfort with the safety / ethics of human experimentation
- In “drug discovery", in particular, we will not have much experience with the intervention

Phases of investigation

- Preclinical
  - Epidemiology including risk factors
  - Basic science: Physiologic mechanisms
  - Animal experiments: Toxicology
- Clinical
  - Phase I: Initial safety / dose finding
  - Phase II: Preliminary efficacy / further safety
  - Phase III: Confirmatory efficacy / effectiveness
- Approval of indication based on total evidence to date
  - Evidence based medicine
  - (Phase IV: Post-marketing surveillance, REMS)

Case Study: Selenium for cancer prevention

Epidemiologic findings (Clark LC, *Nutr Cancer*. 1984;6(1):13-21)

- Case-control study: Plasma selenium and skin neoplasms:
  - 142 cases (basal cell epithelioma or squamous cell carcinoma); 103 noncancer controls.
    - Odds ratio = 4.39:
      lowest vs highest selenium decile (cases vs controls)

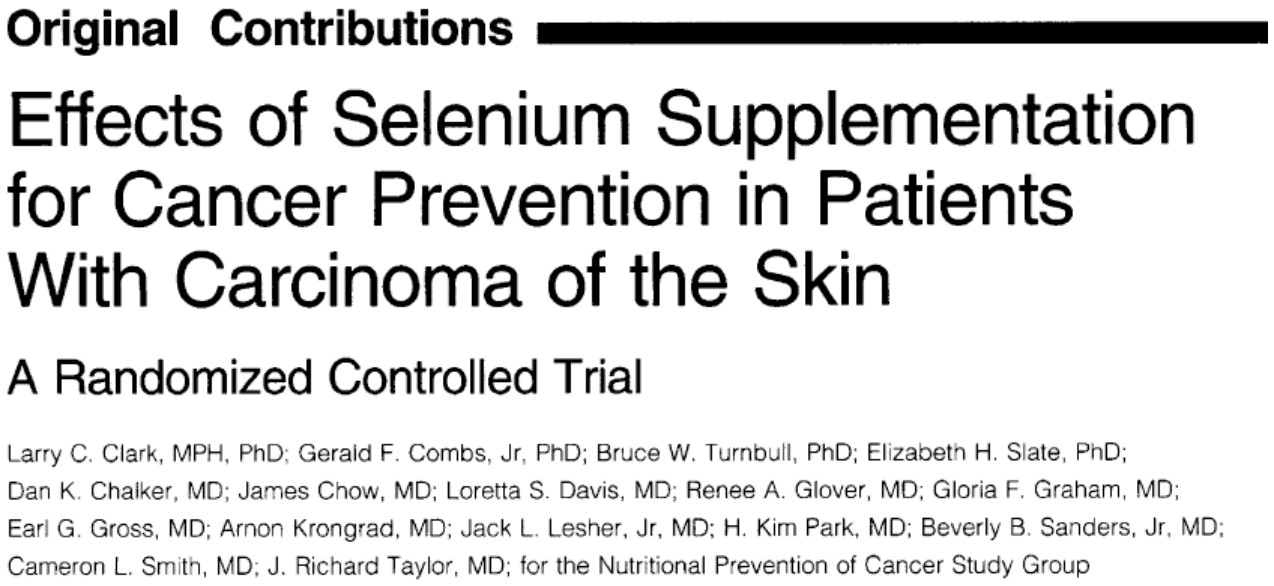
Abstract

Although experimental studies in animals show that selenium may prevent cancer, case-control studies of internal human cancers have been difficult to interpret because neoplastic tissue sequesters selenium. We therefore conducted a case-control study to examine the association between plasma selenium level and skin cancer, a neoplasm with minimal tumor mass at the time of diagnosis. The mean selenium level among patients with either basal cell epithelioma (N = 142), squamous cell carcinoma (N = 48), or both (N = 50), was 0.141 micrograms/g. This was significantly lower than the mean plasma selenium level of the 103 control subjects, which was 0.155 micrograms/g. The noncancer control groups were drawn from current clinic patients and past clinic patients. The logistic estimate of the odds ratio for the lowest versus the highest decile of selenium for all cases combined versus the group of current patient controls was 4.39; for all cases combined versus the past patient controls, the logistic estimate of the odds ratio was 5.81.

Case Study: Selenium for cancer prevention

Follow-up clinical trial (Clark, JAMA 1996; 276:1957-1963)

- Design: RCT (double-blind placebo-controlled; 1983-1991)
  - Dietary supplement: oral selenium (200μg) vs placebo
  - Patients with history of basal or squamous cell skin cancer
  - 1312 patents in seven dermatology clinics in eastern US



Original Contributions

Effects of Selenium Supplementation for Cancer Prevention in Patients With Carcinoma of the Skin

A Randomized Controlled Trial

Larry C. Clark, MPH, PhD; Gerald F. Combs, Jr, PhD; Bruce W. Turnbull, PhD; Elizabeth H. Slate, PhD; Dan K. Chalker, MD; James Chow, MD; Loretta S. Davis, MD; Renee A. Glover, MD; Gloria F. Graham, MD; Earl G. Gross, MD; Arnon Krongrad, MD; Jack L. Lesher, Jr, MD; H. Kim Park, MD; Beverly B. Sanders, Jr, MD; Cameron L. Smith, MD; J. Richard Taylor, MD; for the Nutritional Prevention of Cancer Study Group

Case Study: Selenium for cancer prevention

Clark trial results

- Lung cancer results (incident cases):
  - Selenium: 17 cases; Placebo: 31 cases
  - RR: 0.54 (95%CI: 0.30-0.98; p = 0.04)
- Prostate cancer results (incident cases):
  - Selenium: 13 cases; Placebo: 35 cases
  - RR: 0.37 (95%CI: 0.18-0.71; p = 0.002)

Table 3.—Cancer Incidence by Treatment Group

Cancer Sites, No.	Selenium	Placebo	RR (95% CI)*	*P* Value	HR (95% CI)†	*P* Value
Lung‡	17	31	0.54 (0.30-0.98)	.04	0.56 (0.31-1.01)	.05
Prostate‡	13	35	0.37 (0.18-0.71)	.002	0.35 (0.18-0.65)	.001
Colorectal‡	8	19	0.42 (0.18-0.95)	.03	0.39 (0.17-0.90)	.03
Head and neck	6	8	0.74 (0.21-2.43)	.58	0.77 (0.27-2.24)	.64
Bladder	8	6	1.32 (0.40-4.61)	.62	1.27 (0.44-3.67)	.66
Esophageal	2	6	0.33 (0.03-1.84)	.15	0.30 (0.06-1.49)	.14
Breast	9	3	2.88 (0.72-16.5)	.09	2.95 (0.80-10.9)	.11
Other specific carcinomas	5	9	0.55 (0.14-1.82)	.27	0.54 (0.18-1.62)	.27
Total carcinomas‡§	59	104	0.55 (0.40-0.77)	<.001	0.54 (0.39-0.75)	<.001
Melanomas	8	8	0.97 (0.32-2.96)	.91	0.92 (0.34-2.45)	.87
Leukemia/lymphomas	8	5	1.58 (0.46-6.14)	.41	1.50 (0.49-4.60)	.48
Other specific noncarcinomas	3	3	0.99 (0.13-7.37)	.98	0.99 (0.20-4.94)	.99
Total noncarcinomas	19	16	1.17 (0.57-2.44)	.65	1.16 (0.60-2.27)	.65
Total cancer‡§	77	119	0.63 (0.47-0.85)	.001	0.61 (0.46-0.82)	<.001

*RR indicates relative risk; and CI, confidence interval. P values derived from log rank tests.

Case Study: Selenium for lung cancer prevention

ECOG 5597

Phase III Chemoprevention Trial of Selenium Supplementation in Persons with Resected Stage I Non-Small Cell Lung Cancer

- Design: RCT (double-blind placebo-controlled)
  - Dietary supplement: oral selenium ($200\mu g$) vs placebo
  - Patients with resected stage I NSC lung cancer
  - 1522 patients from ECOG-participating clinics from 2000-2009.
- Results (interim analysis in 2009):
  - 5-year risk of recurrence or death: Selenium: 72%; Placebo 78%
  - Trial stopped early: “not an effective chemoprevention agent."

Case Study: Selenium for prostate cancer prevention

SELECT trial; JAMA. 2009 301(1): 39-51

- Randomized 35,533 men to 4 treatment groups (2×2 factorial:

  Selenium + Vit E placebo
  Selenium placebo + Vit E
  Selenium + Vit E
  Selenium placebo + Vit E placebo

- Follow-up for 4.17-7.33 years over 12 years

Effect of Selenium and Vitamin E on Risk of Prostate Cancer and Other Cancers
The Selenium and Vitamin E Cancer Prevention Trial (SELECT)

Case Study: Selenium for prostate cancer prevention

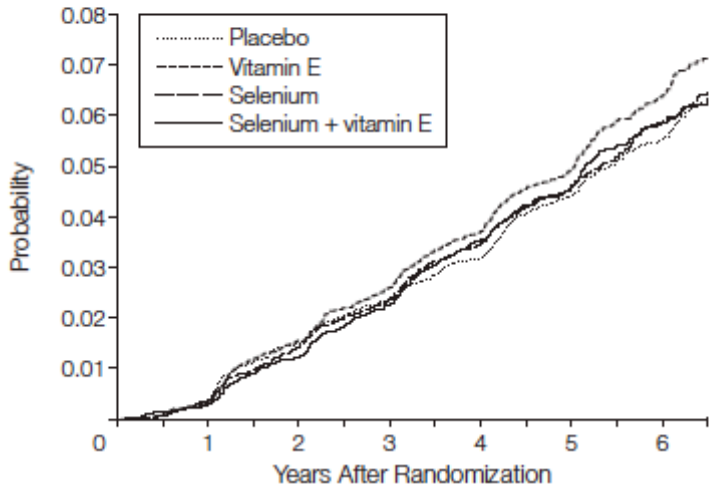
Follow-up clinical trial (Clark, JAMA 1996; 276:1957-1963)

SELECT trial results

Hazard ratios and 99% CI for prostate cancer:

Vit E:	1.13 (0.95 to 1.35)
Selenium:	1.04 (0.87 to 1.24)
Selenium + Vit E:	1.05 (0.88 to 1.25)

Figure 2. Cumulative Incidence of Prostate Cancer Detected Each Year by Intervention Group



No. at risk							
Placebo	8689	8553	8328	8039	7389	4892	2516
Vitamin E	8732	8610	8373	8098	7401	4867	2537
Selenium	8750	8597	8341	8083	7393	4848	2558
Selenium + vitamin E	8700	8585	8371	8097	7428	4894	2580

Compared with placebo, there was a statistically nonsignificant increase in prostate cancer in the vitamin E group (P=.06) and not in the selenium + vitamin E group (P=.52) or the selenium group (P=.62).

Case Study: Selenium for cancer prevention

Summary remarks

The selenium story represents:

- Excellent demonstration of careful evaluation of a hypothesis illustrating:
  - Interplay between careful epidemiology and clinical trials in a range of diseases:
    1. Epidemiology as foundation for major intervention trials.
    2. Demonstrates the importance of confirmatory trials for subgroup effects in large trials.
    3. Large RCT's of the same hypothesis in multiple diseases
    4. The question has been answered??

Screening Trials

Phases of investigation

- Preclinical
  - Epidemiology including risk factors
  - Basic science: Physiologic mechanisms
  - Animal experiments: Toxicology
- Clinical
  - Phase I: Initial safety / dose finding
  - Phase II: Preliminary efficacy / further safety
  - Phase III: Confirmatory efficacy / effectiveness
- Approval of indication based on total evidence to date
  - Evidence based medicine
  - (Phase IV: Post-marketing surveillance, REMS)

Phase II clinical trials : Screening

- Phase II clinical trials seek to establish preliminary evidence of efficacy
- Goals:
  - Screening for any evidence of treatment efficacy
  - Incidence of major adverse effects
  - Decide if worth studying in larger samples
    - Gain information about best chance to establish efficacy
    - Choose population, treatment, outcomes
- This initial screening is essential for achieving the following public health objectives...

Public Health Objective for Clinical Research

Formulating the public health objective

- Ultimate objectives:
  - Discover things that are true
  - Develop the science in order to provide public health benefit (therapies, prevention, etc...)
  - Want high prevalence of truly beneficial therapies/practices among all things (therapies or public health recommendations) that are adopted in practice.
- These objectives are quantified as the *positive predictive value* (PPV) of clinical research
  - Medical studies as diagnostic tests
  - Review of PPV: cervical cancer screening
  - PPV in clinical trials
    - Illustration of practices to increase (or decrease) PPV.

Public Health Objective for Clinical Research

Medical studies as diagnostic tests

- Clinical testing of a new treatment or preventive agent is analogous to using laboratory or clinical tests to diagnose a disease
  - Goal is to find a procedure that identifies truly beneficial interventions
- Not surprisingly, the issues that arise when screening for disease apply to clinical trials
  - Predictive value of a positive test is best when prevalence is high
  - Use screening trials to increase prevalence of beneficial treatments

Public Health Objective for Clinical Research

Diagnostic testing

- We most often characterize the *sensitivity* and *specificity* of a diagnostic/screening test
  - Sensitivity of test: Probability of positive in diseased
    - Sample a cohort of subjects with the disease
    - Estimate the proportion who have a positive test result:

$$\text{Sensitivity} = \Pr[+|D]$$

    - 1 - False Negative Rate
  - Specificity of test: Probability of negative in healthy
    - Sample a cohort of healthy (non-diseased) subjects
    - Estimate the proportion who have a negative test result:

$$\text{Specificity} = \Pr[-|\bar{D}]$$

    - 1 - False Positive Rate

Public Health Objective for Clinical Research

Diagnostic testing

- We are actually interested in the diagnostic utility of the test:
  - Predictive value of a positive test: Probability of disease when test is positive

$$\text{PPV} = \Pr[D|+]$$

  - Predictive value of a negative test: Probability of not diseased when test is negative

$$\text{NPV} = \Pr[\bar{D}|-]$$

Public Health Objective for Clinical Research

Diagnostic testing

- We can compute the predictive value of positive and negative tests using *Bayes rule*:

$$\Pr[D|+] = \frac{\Pr[+|D]\Pr[D]}{\Pr[+|D]\Pr[D] + \Pr[+|\bar{D}]\Pr[\bar{D}]}$$

$$\Pr[\bar{D}|-] = \frac{\Pr[-|D]\Pr[D]}{\Pr[-|D]\Pr[D] + \Pr[-|\bar{D}]\Pr[\bar{D}]}$$

Public Health Objective for Clinical Research

Diagnostic testing

- Key property: Positive and Negative predictive value depends upon sensitivity, specificity, AND prevalence of disease

$$\Pr[D|+] = \frac{\Pr[+|D]\Pr[D]}{\Pr[+|D]\Pr[D] + \Pr[+|\bar{D}]\Pr[\bar{D}]}$$

$$\text{PPV} = \frac{\text{Sens} \times \text{Prev}}{\text{Sens} \times \text{Prev} + (\text{1-Spec}) \times (\text{1-Prev})}$$

$$\Pr[\bar{D}|-] = \frac{\Pr[-|\bar{D}]\Pr[\bar{D}]}{\Pr[-|D]\Pr[D] + \Pr[-|\bar{D}]\Pr[\bar{D}]}$$

$$\text{NPV} = \frac{\text{Spec} \times (\text{1-Prev})}{(\text{1-Sens}) \times \text{Prev} + \text{Spec} \times (\text{1-Prev})}$$

PPV Example: Cervical cancer screening in New Zealand

New Zealand National Cervical Screening Program (NCSP)

- Established in 1991: credited with reducing cervical cancer incidence and mortality.
- Over 70% participation
- Two screening tests (circa 2000)
  - Pap smear ($\sim$\$5): funded by NCSP
  - ThinPrep ($\sim$\$20): offered by some physicians for \$15 fee

ThinPrep versus Pap (Stein 2003)

- Pap smear (Papanicolaou test)
  - Cervical swab on slide for pathologist evaluation
  - Sensitivity $\sim$ 50% (up to 68?%)
  - Specificity $\sim$ 98% (up to 79%)
- “ThinPrep": liquid-based cytology screening test
  - Cervical swab rinsed in tube with liquid preservative
  - Sensitivity $\sim$ 80%
  - Specificity $\sim$ 90%

PPV Example: Cervical cancer screening in New Zealand

NCSP and equitably (*circa 2000*)

- Lower SES communities unable to pay for ThinPrep
- *IF* superior should NCSP adopt ThinPrep?
- Key questions:
  1. Is ThinPrep really more accurate than pap?
  2. What are the potential cost impacts?
     - ThinPrep costs $15 more
     - A positive screening test is referred for colposcopy ($200).
     - Lower specificity might overwhelm budget with unnecessary colposcopies.

PPV Example: Cervical cancer screening in New Zealand

Impact of sensitivity and specificity on the NCSP

- Suppose :
  - 1,000,000 women are screened
  - Prevalence of high grade lesions is 1%:
    - 10,000 with high grade lesion
    - 990,000 without high grade lesion
  - Each positive test is sent for colposcopy

PPV Example: Cervical cancer screening in New Zealand

Impact of sensitivity and specificity on the NCSP

- Suppose NCSP uses pap:
  - Sensitivity = 50%
  - Specificity = 98%

- Results of screening:
  - Number of positive tests:

$$\begin{aligned} \text{True positive tests:} \quad & 10,000 \times 0.50 = 5000 \\ \text{False positive tests:} \quad & 990,000 \times 0.02 = 19,800 \\ \text{PPV}: \quad & \frac{5000}{24,800} = 0.20 \end{aligned}$$

  - Cost:

$$\begin{aligned} \text{Cost of tests:} \quad & \$5.00M \\ \text{Cost of colposcopy:} \quad & \$4.96M \\ \text{Total:} \quad & \sim \$10M \end{aligned}$$

PPV Example: Cervical cancer screening in New Zealand

Impact of sensitivity and specificity on the NCSP

- Suppose NCSP uses ThinPrep:
  - Sensitivity = 80%
  - Specificity = 95%

- Results of screening:

  Number of positive tests:

$$\begin{aligned} \text{True positive tests:} \quad & 10,000 \times 0.80 = 8000 \\ \text{False positive tests:} \quad & 990,000 \times 0.05 = 49,500 \\ \text{PPV}: \quad & \frac{8000}{57,500} = 0.14 \end{aligned}$$

  Cost:

$$\begin{aligned} \text{Cost of tests:} \quad & \$20M \\ \text{Cost of colposcopy:} \quad & \$11.5M \\ \text{Total:} \quad & \sim \$31.6M \end{aligned}$$

PPV Example: Cervical cancer screening in New Zealand

Impact of sensitivity and specificity on the NCSP

- Suppose NCSP uses ThinPrep:
  - Sensitivity = 80%
  - Specificity = 90%

- Results of screening:

  Number of positive tests:

$$\begin{aligned} \text{True positive tests:} \quad & 10,000 \times 0.80 = 8000 \\ \text{False positive tests:} \quad & 990,000 \times 0.1 = 99,000 \\ \text{PPV}: \quad & \frac{8000}{99,900} = 0.075 \end{aligned}$$

  Cost:

$$\begin{aligned} \text{Cost of tests:} \quad & \$20.0M \\ \text{Cost of colposcopy:} \quad & \$21.4M \\ \text{Total:} \quad & \sim \$41.4M \end{aligned}$$

PPV Example: Cervical cancer screening in New Zealand

Summary remarks: public health objective

- Rare diseases:
  - High risk for false positive
  - Important to control *specificity*
- Consequences of a false positive
  - Costs to healthcare system
  - Anxiety costs for women
- Clearly:
  - Weigh costs against risk/consequences of false negative
- Public health objective:
  - Highest PPV for lowest total cost

Formulating the public health objective

PPV as the objective in public health research

- So what is the right answer?
  - Diagnostic testing
    - Identify people with disease who can benefit from care
    - Identify people who should not be treated
  - Public health research?
    - Identify hypotheses that are in fact true
    - Identify hypotheses that are not worthy of further exploration
- What are the consequences of a wrong answer?
  - Diagnostic testing?
    - People do not receive beneficial treatment
    - People receive non-beneficial treatment
  - Public health research?
    - Populations do not receive beneficial practice/care
    - Populations receive non-beneficial practice/care
- Objective:
  - Maximize the proportion of right answers (PPV)

Positive predictive value of research

PPV in research

- A Statistical hypothesis test can be viewed as a test for beneficial treatments.
  - α-level: probability of observing a positive (statistically significant) test in absence of a true treatment effect:
    - Level of significance is 1 - specificity.
    - Choosing $\alpha = 0.05$ gives 95% specificity.
  - Statistical power (β): Probability of observing a positive (statistically significant) test when there is a true treatment effect:
    - Power is sensitivity.
    - Common choice of 80% sensitivity (not usually recommended by me).
  - Prevalence (π_0): the percentage of effective treatments among all tested treatments.

Positive predictive value of research

PPV in research

- Positive predictive value: probability that a statistically significant trial indicates a truly effective treatment.

$$PPV = \frac{\beta\pi_0}{\beta\pi_0 + \alpha(1-\pi_0)}$$

- The probability that our public health recommendation is in fact beneficial.

PPV in biomedical research

Example: The Amgen experience

- CG Begley and LM Ellis:
  " Raise the standards for preclinical cancer research"
  Nature 483:531-533; 2012
  - Over the past decade Amgen scientists tried to confirm the results of 53 'landmark' studies
  - Only 6/53 (11%) of these results were confirmed
  - *"The scientific process demands the highest standards of quality, ethics, and rigour."*
- All true:
  - High standards are an absolute requirement.
  - Also need to note that lack of reproducibility is not surprising if initial false-positive risk is high

The Public Health Objective

Clinical trials as diagnostic tests

- We routinely consider power (β = sensitivity) and type I error (α = 1 - specificity).
- What is the prevalence (π_0)?
  - 9.6% of treatments entering phase I trials are positive in subsequent phase III trials (Biomedtracker, 2016)
    - Lowest in oncology (7.0%)
    - Highest in infectious diseases (17%) (Nature Biotech, 2014)
  - Results of NCI-sponsored trials 1955-2006 (Djulbegovic, 2008)
    - 743 randomized comparisons, 176 (24%) are significant
    - 116 (15%) discover 'breakthrough interventions'.
  - Results of phase II cancer trials (Hay, Nature Biotech, 2016)
    - 28% led to phase III.
  - Prevalence of truly beneficial treatments entering phase II trials is probably less than 10%.

The Public Health Objective
How do clinical trials determine PPV?

Example: Phase II studies as screening tests

- Consider the following approaches to evaluating new treatments:
  1. Study every treatment in a large definitive experiment.
  2. Perform small screening tests, and perform large definitive experiments only in those treatments that pass the screening tests.

- Suppose that we want to evaluate the efficiency of these strategies. Assume:
  - 10% of all treatments actually work.
  - Level of significance = 0.05 (specificity = 0.95).
  - 1,000,000 subjects are available for clinical trials.
  - Power for a clinically important difference:

$$\begin{aligned} 1000 \text{ subjects} &\rightarrow 97.5\% \text{ power} \\ 500 \text{ subjects} &\rightarrow 80\% \text{ power} \\ 50 \text{ subjects} &\rightarrow 15\% \text{ power} \end{aligned}$$

The Public Health Objective
How do clinical trials determine PPV?

Example: Phase II studies as screening tests

- Scenario 1 (only large trials):
  - Suppose we evaluate 1000 new treatments (100 effective and 900 ineffective) with 1000 subjects per trial.
  - On average we have positive tests for:
    - 98 of the 100 effective treatments ($0.975 \times 100 \approx 98$).
    - 45 of the 900 ineffective treatments ($0.05 \times 900 = 45$).
  - PPV: $98/(45 + 98) = 0.69$; that is, only 69% of the 143 treatments identified actually work.

The Public Health Objective
How do clinical trials determine PPV?

Example: Phase II studies as screening tests

- Scenario 2 (preliminary screening trials):
  - (a) Suppose we first screen 12,500 new treatments (1,250 effective and 11,250 ineffective).
    - Using 50 subjects in the screening trials (625,000 total) with 15% power.
    - On average the screening trials give positive tests for:
      - 187 of the 1,250 effective treatments ($0.15 \times 1250 \approx 187$).
      - 562 of the 11,250 ineffective treatments ($0.05 \times 11250 \approx 562$).
    - PV+ for the screening phase: $187/(187 + 562) = 0.25$.
  - (b) Now evaluate the 749 treatments (187 effective and 562 ineffective) from the screening trials.
    - Using 500 subjects per trial (374,500 total) with 80% power.
    - On average these confirmatory trials give positive tests for:
      - 150 of the 187 effective treatments ($0.8 \times 187 \approx 150$).
      - 28 of the 562 ineffective treatments ($0.05 \times 562 \approx 28$).
    - PV+ for confirmatory trials: $150/178 = 0.84$.

The Public Health Objective
How do clinical trials determine PPV?

Example: Phase II studies as screening tests

- Comparison of scenarios:
  - Scenario 1 (large trials only):
    - Use 1,000,000 subjects
    - Screen 1,000 new treatments
    - Adopt 98 effective treatments
    - Adopt 45 ineffective treatments
    - PPV = 98/143 = 0.69
  - Scenario 2 (screening studies followed by large trials):
    - Use 999,500 subjects
    - Screen 12,500 new treatments
    - Adopt 150 effective treatments
    - Adopt 28 ineffective treatments
    - PPV = 150/178 = 0.84

The Public Health Objective
How do clinical trials determine PPV?

Example: Phase II studies as screening tests

- Bottom line:
  - Using the same number of subjects, phase II studies increase the predictive value of a positive study. A greater number of effective treatments are identified due in part to the greater number of treatments screened.
  - (Different choices of statistical power in screening and confirmatory trials can be used to optimize the strategy for a particular setting.)

The Public Health Objective
How do clinical trials determine PPV?

PPV is increased through good experimental practice

- How do we increase PPV in the clinical trials paradigm?

$$PPV = \frac{\beta\pi_0}{\beta\pi_0 + \alpha(1-\pi_0)}$$

1. Increase π_0:
   - Careful planning of preliminary studies (choice of endpoints; patient population)
   - Avoid "novel" and "innovative" ideas
   - Careful specification of hypothesis-driven research (avoid “science by hunch")

The Public Health Objective
How do clinical trials determine PPV?

Sensitivity to π_0 (how likely is it that the new treatment works?)

1a. Trial of an 'incremental' advance for a known compound:

- $\pi_0 = 0.20$; $\alpha_2 = 0.05$; $\beta_2 = 0.15$; $\alpha_3 = 0.05$; $\beta_3 = 0.80$
- Results:

	Trials	True Pos	False Pos	PPV
Phase 2	11765	353	471	0.43
Phase 3	824	282	24	0.92

1b. Trial of a novel and innovative therapy:

- $\pi_0 = 0.01$; $\alpha_2 = 0.05$; $\beta_2 = 0.15$; $\alpha_3 = 0.05$; $\beta_3 = 0.80$
- Results:

	Trials	True Pos	False Pos	PPV
Phase 2	13245	20	656	0.029
Phase 3	675	16	33	0.33

The Public Health Objective
How do clinical trials determine PPV?

PPV is increased through good experimental practice

- How do we increase PPV in the clinical trials paradigm?

$$PPV = \frac{\beta\pi_0}{\beta\pi_0 + \alpha(1-\pi_0)}$$

2. Increase β:

    - Good practice (no missing data, low variation in outcome assessment, good adherence, etc.)
    - Increase sample size.

The Public Health Objective
How do clinical trials determine PPV?

Sensitivity to β_3 (ultimate sensitivity for effective therapies)

2a. Sufficiently powered phase III ($\beta_3 = 0.975$)

- $\pi_0 = 0.10$; $\alpha_2 = 0.05$; $\beta_2 = 0.15$; $\alpha_3 = 0.05$; $\beta_3 = 0.975$
- Results:

	Trials	True Pos	False Pos	PPV
Phase 2	9091	136	409	0.25
Phase 3	545	133	20	0.87

2b. Underpowered phase III:

- $\pi_0 = 0.10$; $\alpha_2 = 0.05$; $\beta_2 = 0.15$; $\alpha_3 = 0.05$; $\beta_3 = 0.50$
- Results:

	Trials	True Pos	False Pos	PPV
Phase 2	15385	231	692	0.25
Phase 3	923	115	35	0.77

The Public Health Objective
How do clinical trials determine PPV?

PPV is increased through good experimental practice

- How do we increase PPV in the clinical trials paradigm?

$$PPV = \frac{\beta\pi_0}{\beta\pi_0 + \alpha(1-\pi_0)}$$

3. Reduce α:

   - Pre-specify outcomes
   - Pre-specify all analyses
   - Avoid multiple comparisons
   - Avoid surrogate outcomes
   - Avoid subgroups

The Public Health Objective
How do clinical trials determine PPV?

Sensitivity to α (false positive risk; specificity)

3a. Relax phase II alpha ($\alpha_2 = 0.20$)

- $\pi_0 = 0.10$; $\alpha_2 = 0.20$; $\beta_2 = 0.15$; $\alpha_3 = 0.05$; $\beta_3 = 0.80$
- Results:

	Trials	True Pos	False Pos	PPV
Phase 2	6780	102	1220	0.077
Phase 3	1322	81	61	0.571

3b. Relax both phase II and III alpha ($\alpha_2 = 0.2$, $\alpha_3 = 0.10$):

- $\pi_0 = 0.10$; $\alpha_2 = 0.20$; $\beta_2 = 0.15$; $\alpha_3 = 0.10$; $\beta_3 = 0.80$
- Results:

	Trials	True Pos	False Pos	PPV
Phase 2	6780	102	1220	0.077
Phase 3	1322	81	122	0.40

The Public Health Objective
How do clinical trials determine PPV?

Summary: PPV as a function of π_0, α, and β

Scenario	π_0	α_2	β_2	α_3	β_3	Drugs Evaluated	True Pos	False Pos	PPV
1	0.10	*	*	0.05	0.800	1000	98	45	0.685
2	0.10	0.05	0.15	0.05	0.800	12500	150	28	0.842
3	0.20	0.05	0.15	0.05	0.800	11765	282	24	0.923
4	0.01	0.05	0.15	0.05	0.800	13265	16	33	0.327
5	0.10	0.05	0.15	0.05	0.975	9091	133	20	0.867
6	0.10	0.05	0.15	0.05	0.500	15385	115	35	0.769
7	0.10	0.20	0.15	0.05	0.800	6780	81	61	0.571
8	0.10	0.20	0.15	0.10	0.800	6780	81	122	0.400

The Public Health Objective
PPV and good science?

Summary remarks: How to get high PPV with fewer trials

- Design for scientifically informative negative trials
  - All trials (positive or negative) must reduce the number of viable hypotheses.
- Accept that *no means no*
  - *(Never give up)*: Avoid inflating α with
    - multiple endpoints
    - subgroup analyses
    - surrogate endpoints
  - *(Try try again)*: Avoid recycling ideas

The Public Health Objective
PPV and good science?

Summary remarks: How to get high PPV with fewer trials

- Assure power (β)
  - Good practice reduces variability
  - Good recruitment/retention
  - Adequate sample size
- Avoid development programs with low pre-test probability (π_0)
  - “Novel" and “innovative" approaches have low π_0.

The Public Health Objective
PPV and good science?

Summary remarks: How to get high PPV with fewer trials

- Financial analysts are even quick to pick up on this! (Grainger, Forbes, 2015)
  - "More insidiously, the pivotal trials often adopt a different end-point, agreed with the regulators, to the Phase 2 trials (where a 'surrogate' end-point was used to predict whether the regulatory end-point is likely to be met). Unless this surrogate is perfect (and few are), some agents that are positive against the surrogate will be ineffective against the regulatory end-point."
  - "Similarly, the pivotal trials need to be performed in less selected patient populations. If the drug is more effective in the defined subset studied in Phase 2 than in the broader population, unexpected failure will again result."

For a discussion across all biomedical studies, see "Why Most Published Research Findings Are False", J. Ioannidis, *PLOS Medicine*, August 2005

Introduction to Clinical Trials - Day 1

Session 3 - Fundamentals of Trial Design

Presented July 23, 2018

Susanne J. May
Department of Biostatistics
University of Washington

Daniel L. Gillen
Department of Statistics
University of California, Irvine

Goals of Clinical Trial Design

Clinical trials

- Experimentation in human volunteers
- Investigation of a new treatment or preventive agent
  - *Safety* : Are there adverse effects that clearly outweigh any potential benefit?
  - *Efficacy* : Can the treatment alter the disease process in a beneficial way?
  - *Effectiveness* : Would adoption of the treatment as a standard effect morbidity in the population?

Goals of Clinical Trial Design

A trial must meet minimum scientific standards

- It must address a meaningful question
  - Discriminate between viable hypotheses (Science)
- Trial results must be credible to the scientific community
  - Valid materials, methods (Science, Statistics)
  - Valid measurement of experimental outcome (Science, Clinical, Statistics)
  - Valid quantification of uncertainty in experimental procedure (Statistics)

Goals of Clinical Trial Design

Individual Ethics

- Conducted in human volunteers, the clinical trial must be ethical for participants on the trial
  - Minimize harm and maximize benefit for participants in clinical trial
  - Avoid giving trial participants a harmful treatment
  - Do not unnecessarily give trial participants a less effective treatment

Goals of Clinical Trial Design

Group Ethics

- The clinical trial must ethically address the needs of the greater population of potential recipients of the treatment
  - Approve new beneficial treatments as rapidly as possible
  - Avoid approving ineffective or (even worse) harmful treatments
  - Do not unnecessarily delay the new treatment discovery process

Goals of Clinical Trial Design

Optimality criteria

- A good procedure will

1. Minimize “false positives"
   - Any treatment recommended for adoption will have a high probability of being a truly effective therapy
2. Minimize “false negatives"
   - Any truly effective therapy will have a high probability of being recommended for adoption
3. Be highly safe and ethical
   - Minimize the number of patients exposed to inferior treatments while investigations proceed
4. Be efficient
   - Minimize costs (patients, calendar time, money)

Goals of Clinical Trial Design

Role of Statistics

- Answering scientific questions in presence of variable response
- Scientific questions often reduce to comparing the magnitude of some measurement across groups
- Outcome measures are rarely constant
  - Inherent randomness
  - Hidden (unmeasured) variables
- Use of probability models for describing variability in the real world
  - Distribution of measurements
  - Summary measure (functional) for scientific tendency
  - Quantification of uncertainty in (contrast of) functional(s) (Signal and noise)

Goals of Clinical Trial Design

Common statistical approach

- Optimality criteria (1) and (2) speak directly to the need for achieving high PPV and low NPV
- Design an RCT to answer relevant question
  - Treatment, patient population, intervention, comparator, outcome
    - There is an underlying probability of our hypotheses being correct: “Prevalence of effective therapies"
- Fix probability of making wrong decisions
  - Erroneously decide against status quo < 2.5%
  - But: erroneously decide against status quo 2.5%
- Design trial to fix sensitivity of study
  - Power: High probability to detect beneficial treatment

Positive predictive value in research

- Relationship to type I error, power, and prevalence of truly effective therapies

$$PPV = \frac{\text{Power} \times \text{Prev}}{\text{Power} \times \text{Prev} + (\text{Type I Error}) \times (\text{1-Prev})}$$

$$NPV = \frac{(\text{1-Type I Error}) \times \text{1-Prev}}{(\text{1-Type I Error}) \times \text{1-Prev} + (\text{1-Power}) \times \text{Prev}}$$

Goals of Clinical Trial Design

Predictive value of statistically significant result depends on

1. Probability hypothesis is true to begin with (start with “good ideas")
   - Fixed when hypothesis is formulated
2. Type I error (Specificity)
   - Fixed by level of significance
3. Power (Sensitivity)
   - Statistical power made as high as possible by design

Goals of Clinical Trial Design

The later two elements are improved by

1. Minimizing bias
   - Remove confounding and account for effect modification
2. Decreasing variability of measurements
   - Homogeneity of population, appropriate endpoints, appropriate sampling strategy, more precise measuring device

Goals of Clinical Trial Design

Common pitfalls of studies

- Common pitfalls of experimentation are:
  - Data driven hypotheses (↑ Type I error)
  - Multiple comparisons (↑ Type I error)
  - Poor selection of subjects (↓ Power)
  - Over-fitting of data (↑ Type I error, (↓ Power)
  - Poor selection of subjects, outcomes (↓ Power)
  - Noncomparability of treatment groups (↑ Type I error)
- Each of these pitfalls leads to increases in variability and/or bias in clinical trials...

Variability

How does variability arise?

- Intuitively, if the same experiment is performed several times, the observed results will differ each time
- This variability in observed response depends on several factors including:
  1. The homogeneity of trial participants
  2. How consistently treatment is administered
  3. How consistently the response is measured
  4. Sample size

Reducing Variability

Increasing homogeneity of trial participants

- Inclusion/exclusion criteria to identify a population for whom
  - A new treatment is needed
  - Experimental treatment is likely to work
    - Expected to work equally well in all subgroups
  - All patients likely to eventually use the new treatment are represented (safety)
- Ex: A patient which allows only patients with limited disease and a ECOG score of 0-1 will be much less variable thane one which allows any extent of disease and a ECOG score of 0-4

Reducing Variability

Increased reliability of response measurement

- Objective response measurements and consistent reproducible measurements are critical
- Reducing the subjectivity of response assessment (inter- and intra-rater reliability) will decrease variability
- For biomarkers, use of a single assay analyzed as a single laboratory will decrease measurement error

Reducing Variability

Adequate sample size

- Statistical information is heavily dependent up the number of independent sampling units
- A larger number of patients will lead to reduced variability
- The result is a more precise estimate of treatment effect
- Note: Increasing the number of measurements on a given patient does not contribute the same amount of information as increasing the number of independent patients

Bias

What is bias?

- In statistics, bias is a tendency of a statistical estimate to deviate in one direction from a “true value"
  - What defines the “truth" is dictated by the scientific goal
- A biased study is one that will systematically tend to estimate a treatment effect that is not correct
  - across replicated experiments (frequentist bias), or
  - with a large sample size (consistency)
- As in the statistical definition, the definition of a biased study is very much dependent upon what we wish we were estimating
  - How are we going to generalize our results?

Sources of Bias

- Attributing an observed difference to a particular treatment
  - Disease
    - Misclassification, overly restrictive
  - Patients
    - Insufficiently selected or overly restrictive
  - Intervention
    - Administered incorrectly, improper restriction of ancillary treatments
  - Comparator
    - Irrelevant comparator, treatment groups not similar
  - Outcomes
    - Irrelevant outcome, measurements differ by group

Confounding Bias

- The treatment groups being compared differ with respect to other important (measured or unmeasured) variables that are predictive of outcome
  - Systematic confounding
    - Process of assigning treatments tends to create groups that are dissimilar
    - Patient or provider preference
    - Time trends in diagnosis, treatment
  - Stochastic (conditional) confounding
    - No systematic trends, but we got unlucky this time

Ascertainment Bias

- Assessment of outcomes differs across treatment groups
  - Method of measurement
    - Clinical versus subclinical triggers for assessment
  - Frequency of measurement
    - Adverse events leading to higher surveillance
    - Impact on minima, maxima, time to event
  - Misclassification
    - Accuracy and/or precision of measurement affected by treatment (e.g., tumor growth vs inflammation)

Effect Modification Bias

- Treatment effect varies across subgroups
  - Can lead to appearance of confounding if subgroup membership differs across treatment groups
  - Also leads to problems in generalizing effectiveness to eventual treated population

Reporting Bias

- Tendency to report results agreeing with preconceived notions
  - Publication bias in literature
  - Selection of historical results to get most favorable outcomes
  - Multiple comparison issues in selecting primary outcomes
  - Multiple comparison issues in selecting summary of outcome distributions
- Increases type I error substantially

Multiple Comparisons

Multiple comparisons

- Multiple comparison issues
  - Type I error for each endpoint
    - In absence of treatment effect, will still decide a benefit exists with probability, say, .025
- Multiple endpoints increase the chance of deciding an ineffective treatment should be adopted
  - This problem exists with either frequentist or Bayesian criteria for evidence
  - The actual inflation of the type I error depends
    - the number of multiple comparisons, and
    - the correlation between the endpoints

Multiple Comparisons

Multiple comparisons

“When you go looking for something specific, your chances of finding it are very bad, because of all the things in the world, you’re only looking for one of them.

“When you go looking for anything at all, your chances of finding it are very good, because of all the things in the world, you’re sure to find some of them."

- Darryl Zero in “The Zero Effect"

Multiple Comparisons

Multiple comparisons

In Statistics-Speak “When you go looking for something specific, your chances of finding [a spurious association by chance] are very bad, because of all the things in the world, you’re only looking for one of them.

“When you go looking for anything at all, your chances of finding [a spurious association by chance] are very good, because of all the things in the world, you’re sure to find some of them."

Multiple comparisons

- Goal is to achieve reproducible scientific evidence, but multiple comparisons lead to
  - Inflation of type I error rates
  - Spurious associations
- Consider the experiment-wise type I error rate as a function of the number of comparisons and the correlation between endpoints

Number Compared	Worst Case	Correlation 0.00	0.30	0.50	0.75	0.90
1	.050	.050	.050	.050	.050	.050
2	.100	.098	.095	.090	.081	.070
3	.150	.143	.137	.126	.104	.084
5	.250	.226	.208	.184	.138	.101
10	.500	.401	.353	.284	.193	.127
20	1.000	.642	.540	.420	.258	.154
50	1.000	.923	.806	.624	.353	.193

Multiple comparisons

- Some believe that this problem only exists when testing different outcomes
- However, the issue also exists when testing multiple summary measures for the same outcome!
- As an example, consider the type I error for a two group comparison of a normally distributed outcome

Any single test:	0.050
Mean, geometric mean:	0.057
Mean, Wilcoxon:	0.061
Mean, geom mean, Wilcoxon:	0.066
Above plus median:	0.085
Above plus Pr (Y > 1 sd):	0.127
Above plus Pr (Y > 1.645 sd):	0.169

- Bottom line: Need to specify a primary summary measure or multiple comparison issues result!

Summary Remarks

Essentials of trial design

- A good procedure will

1. Minimize “false positives"
   - Any treatment recommended for adoption will have a high probability of being a truly effective therapy
2. Minimize “false negatives"
   - Any truly effective therapy will have a high probability of being recommended for adoption
3. Be highly safe and ethical
   - Minimize the number of patients exposed to inferior treatments while investigations proceed
4. Be efficient
   - Minimize costs (patients, calendar time, money)

Summary Remarks

Essentials of trial design

- First and foremost, the trial must be relevant
  - Conducted in appropriate patient population (new treatment needed and likely to work)
  - Testing appropriate hypothesis
- Predictive value of trial results is increased by
  - Decreasing variability:
    - Homogeneity of patient population
    - Precise definition of treatment(s)
    - Appropriate choice of clinical, statistical endpoints
    - High precision in measurements
    - Appropriate sampling strategy
  - Minimizing bias:
    - Use of appropriate comparison group
    - Blinding
    - Use of randomization
    - Avoiding multiple comparisons

Summary Remarks

We're not alone...

- International Conference on Harmonisation (ICH: www.ich.org):
  - Launched in 1990: a harmonization of requirements for pharmaceutical registration in US, Europe, and Japan.
  - An excellent resource for current best practice.
- ICH Part E9 - Statistical Principles.
- CONSORT guidelines
  - An agreement between major journals on standards of evidence.
    * The Revised CONSORT Statement for Reporting Randomized Trials: Explanation and Elaboration.
    * Douglas G. Altman, DSc; Kenneth F. Schulz, PhD; David Moher, MSc; Matthias Egger, MD; Frank Davidoff, MD; Diana Elbourne, PhD; Peter C. Gotzsche, MD; and Thomas Lang, MA, for the CONSORT Group.
    * Ann Intern Med. 2001;134:663-694.
    * http://www.consort-statement.org

Defining the Target Population

Inclusion/Exclusion Criteria

- Patients are the fundamental “sampling units" of our scientific experiment
- We thus want to be able to
  - have a clear definition of the disease we are targeting,
  - exclude patients for whom the likelihood of successfully completing the RCT is low

Defining the Target Population

Scientific basis

- A patient population for whom
  - An improved treatment is desired
  - There is no contraindication to the use of the investigational treatment
  - The investigational treatment might reasonably be expected to work
    - Furthermore: the degree of benefit is expected to be nearly the same for all subgroups of patients that can be identified beforehand

Defining the Target Population

Clinical basis

- For clinical utility, the definition of the target population must be based on information commonly available prior to start of treatment
  - Definitions based on diagnostic criteria available only after some delay should be avoided
    - e.g., bacterial culture is often only available 24 hours after start of therapy
  - Definitions based on diagnostic tests that are not routinely available should be avoided
    - genetic profile?
    - clinical utility versus basic science

Defining the Target Population

Target population

- Patient population should generally reflect clinical basis as closely as possible
  - Exception: when it is ethical to conduct a clinical trial to answer a basic science question
- Additional concerns in clinical trial setting
  - Clinical equipoise among choice of all possible treatment assignments
  - Conservatism in using untested treatments
  - Patients' compliance with heightened surveillance in a clinical study

Defining the Target Population

Inclusion/exclusion criteria

- Precise definition of target patient population is crucial
  - Scientific:
    - Materials and methods of scientific experiment
  - Clinical:
    - Generalization of safety outcomes
    - Generalization of efficacy outcomes
- Inclusion / exclusion criteria define target population
- Source of patients also of great interest for generalizability
  - Primary care versus tertiary care centers' patient populations
  - Regional differences in possible effect modifiers
    - environmental exposures
    - genetic factors

Defining the Target Population

Conceptual framework

- Population of patients with disease
  - Definition of disease by cause vs signs / symptoms
- Subpopulation with disease targeted by intervention
  - Defined by treatment?
- Subpopulation eligible for study accrual
  - Restricted due to general clinical trial setting
- Eligible patients from which sampled
  - Restricted due to specific clinical trial (location, time)
- Study sample
  - Restricted due to willingness to participate (efficacy vs. effectiveness)

Defining the Target Population

Ideal study sample

- The study sample should look like a random sample from the subpopulation of all diseased patients who would ultimately be judged suitable for the intervention.
  - Negligible impact of restrictions due to clinical trial procedures
  - Negligible impact of restrictions due to locale of clinical trial
  - High participation rate by eligible patients

Defining the Target Population

Safety considerations

- In conduct of clinical trial may want to exclude some patients
  - Need to consider whether at-risk patients should be exposed to unproven therapy
    - Pregnancy, children, co-morbidities, elderly
- Generalizing study results: Efficacy vs effectiveness
  - Self-selection into trial
  - Treatment may have to be delivered to a population larger than studied
    - Diagnostic procedures after approval may be less rigorous (eg. time requirements in definition of gram negative sepsis)
  - Off-label use

Defining the Target Population

Inclusion/exclusion criteria

- Inclusion criteria:
  - Definition of ultimate target population
- Exclusion criteria:
  - Exceptions required for clinical trial setting
- The safety and efficacy of the investigation treatment will only have been established in patients meeting both inclusion and exclusion criteria

Defining the Target Population

Inclusion criteria

- Objective criteria of disease
  - Strive for common clinical definitions
  - Minimize subjective criteria
- Measures of severity of disease that might preclude inclusion in target population
  - Mild disease might not be of interest
  - Severe disease might not be ethical

Defining the Target Population

Inclusion criteria

- Subgroups of interest
  - E.g., age: adult vs children (though avoid unnecessary restriction)
  - E.g., not candidate for surgery or having failed other treatments
  - E.g., genetic subtype
- Contraindications to treatment
  - Ideally, only if ultimate labeling of treatment would include such contraindications
  - E.g., liver disease, renal disease, diabetes

Defining the Target Population

Exclusion criteria

- Contraindications to treatments in clinical trial setting
  - E.g., safety concerns with new drug that might lead to compliance issues with unproven efficacy
  - E.g., contraindication to comparison treatment
  - E.g., language barriers
- Requirements for evaluation of treatment outcome
  - E.g., lack of measurable disease
  - E.g., inability to make clinic visits
  - E.g., simultaneous participation in other clinical trials
- Requirements for compliance to protocol
  - E.g., not passing a run-in period
  - (but need to avoid lessening generalizability)
- Requirements for ethical investigation
  - unwillingness or inability to provide informed consent

Defining the Target Population

Specification of inclusion/exclusion criteria

- Criteria for inclusion / exclusion should consider
  - Methods of measurement
  - Need for and impact of multiple measurements
    - effect of more frequent surveillance
    - possible contradictory measurements
  - Timeframes for all criteria
    - usually stated relative to randomization

Defining the Intervention

Treatments vs. treatment strategies

- The trial will ultimately compare outcomes across populations receiving different treatments
- In a clinical trial, we never test a treatment
  - We may not ethically force people to continue a therapy
  - It may not be medically advisable to even want a patient to continue
- Instead we test a treatment strategy
  - We prescribe an initial treatment
  - Patients may also receive ancillary treatments
    - These may be precipitated by experimental therapy
  - Patients may progress to other therapies

Defining the Intervention

Definition of treatments

As such, a full description of the treatment is necessary

- Formulation of treatment
- Dose, administration, frequency, duration
  - Rules for responsive dosing (e.g., insulin)
  - Include plans for
    - Treatment of adverse events
    - Dose reduction
    - Dose discontinuation
- Ancillary treatments
  - Prescribed vs allowed vs prohibited
    - (Distinguish safety issues from efficacy issues)

Choice of Outcome

Clinical outcomes

- Goal of a clinical trial is to establish if an experimental treatment will prevent a particular clinical outcome:
  - Development of disease
  - Decreased quality of life
  - Mortality
- Essential to define relevant outcome and summary measure
  - Probability of mortality within 28 days
  - Number of days alive and out of ICU
  - Mean 6-minute walk distance
  - Median survival

Choice of Outcome

Clinical outcomes

- A common problem is that the clinical outcomes are rare or occur after a long time.
- This has an impact on trial design:
  - Larger sample sizes are required to detect treatment effects on rare events.
  - Long periods of follow-up may be needed to assess clinical endpoints.

Choice of Outcome

Surrogate outcomes

- A surrogate outcome is a biological endpoint which:
  - Can be measured in a shorter time frame
  - Can be measured precisely
  - Is predictive of the clinical outcome.
- Use of a surrogate may increase trial efficiency.
  - Assume that treatment effect on the surrogate is a good indication of its effect on the clinical outcome

Choice of Outcome

Examples of surrogate outcomes

- AIDS:
  - HIV leads to depression of CD4 cells
  - Increased viral load correlates with development of AIDS
  - Surrogate endpoint: viral load
  - Clinical endpoint: morbidity and/or mortality

- Coronary heart disease:
  - People with arrhythmia following heart attack (MI) have poor survival.
  - Therapies have been developed toward preventing arrhythmia.
  - Surrogate endpoint: arrhythmia
  - Clinical endpoint: mortality.

Choice of Outcome

Clinical endpoints

- Clinical versus biological (surrogate) endpoints
  - Typically, subjects participating in a trial are hoping that they will benefit in some way from the trial
  - Clinical endpoints are therefore of more interest than purely biological endpoints
  - For late stage trials, how well does the proposed surrogate correlate with the targeted clinical endpoint?
  - Often there is great potential for being led astray by a surrogate outcome which may pose safety issues
- More later (and Day 2!)

Comparison Groups

Need for Comparison Groups

- Clinical trials can utilize:
  - No comparison group
  - Historical controls
  - Concurrent comparison group(s)
- Having a comparison groups is important when
  - Deciding whether a proposed treatment is effective
  - Deciding among the alternatives when treating a single patient

Comparison Groups

No comparison group

- Appropriate when an absolute criterion for treatment exists
- Single arm clinical trial
  - Cohort design
  - Includes “pre-post" designs
- Rarely do such absolute criterion exist. Instead, we are really invoking the the results from previous investigations
  - Ex: Pearl Index of 2.0 for evaluation of oral contraceptives

Comparison Groups

Historical controls

- An attempt to make more efficient use of limited research resources
- Single arm clinical trial
- Compare results to
  - Absolute criterion derived from historical trials
    - Dishonest : Treat historical estimates as known and use only one-fourth the sample size compared to a 2-arm study
  - Sample from historical clinical trial (better)
    - More honest : Account for variability in historical control estimate save only half the sample size relative to a 2-arm study

Sample size requirements

- Consider generic sample size formulae for the no control, historical control, and concurrent comparison groups
- Sample size requirements in a single arm study to detect a mean outcome greater than μ_0

$$n = \frac{(z_{1-\alpha/2} + z_\beta)^2 \sigma^2}{(\mu_1 - \mu_0)^2}$$

- Sample size requirements on experimental arm in a two arm study to detect a mean outcome greater than μ_0

$$n_1 = \frac{(z_{1-\alpha/2} + z_\beta)^2 \left(1 + \frac{n_1}{n_0}\right) \sigma^2}{(\mu_1 - \mu_0)^2}$$

Sample size requirements

- Sample size requirements on experimental arm when using historical controls in a study to detect a mean outcome greater than μ_0

$$n_1 = \frac{(z_{1-\alpha/2} + z_\beta)^2 \left(1 + \frac{n_1}{n_0}\right) \sigma^2}{(\mu_1 - \mu_0)^2}$$

  - n_0 historical controls are presumably already available

Comparison Groups

Use of historical controls

- Thus, compared to a two arm study of a new treatment and a historical treatment, use of historical controls can save time and money
  - Use of historical control sample obviates the need for one arm; thus only half the subjects when 1:1 randomization utilized
  - Using the estimates from a historical clinical trial as if they were known treatment effects decreases sample size requirements even further:
    - Only one-fourth the number of subjects are required
    - However, we are pretending that we have an infinite number of relevant historical controls (no variability)!

Use of historical controls

- However, the validity of such methods is heavily dependent upon the historical trial being comparable in every way
  - No changes in comparison treatment
  - No changes in definition of study population
  - No changes in ancillary treatments
  - No changes in measurement of treatment outcome
- Pocock (J Chronic Disease, 1976) described conditions for acceptability of historical control group

Comparison Groups

Pocock conditions for use of historical controls

1. Such a group must have received a precisely defined standard treatment
   - relevance of standard treatment must remain
   - measurement of treatment parameters must be the same
   - ancillary treatments must not have changed
2. Group must have been a part of a recent clinical study containing the same requirements for patient eligibility
   - measurement methods used in eligibility must be the same
   - clinical trial setting must have same selection pressures on patient participation
3. Methods of treatment evaluation must be the same
   - same criteria (schedule) for performing evaluations
   - same criteria for judging outcomes

Comparison Groups

Pocock conditions for use of historical controls (cont'd)

4. Distributions of important patient characteristics should be comparable
   - same univariate distributions of risk factors (within range dictated by eligibility criteria)
   - same correlations among risk factors
   - must hold for both measured/unmeasured risk factors of
     - disease, adverse outcomes, and competing risks
5. Previous study must have been performed in the same organization with largely the same clinical investigators
   - must control any subjective aspects of definition of eligibility, treatments, outcome
   - must control for unique patient populations due to location and/or referral patterns
6. There must be no other indications leading one to expect differing results

Additional criteria for use of historical controls

- The analysis should reflect the variability in the original data, not just the estimates of treatment effect
  - It is “cheating" to pretend there was no variability in assessing the outcome from the historical comparison group.
  - Ideally: use the exact distribution of the covariates
    - Nonlinearities of effects of covariates on outcome and interactions among the covariates might alter the inference

Comparison Groups

Statistical remedies for meeting these criteria?

- Attempts to circumvent some of these requirements using statistical methods
  - Clearly, the above conditions are rarely, if ever, satisfied.
  - Attempts have been made to use statistical models to adjust for differences between the historical control group and a current treatment group.
    1. Adjustment for covariates
    2. Propensity score analysis

Comparison Groups

Adjustment for covariates

- Analysis with adjustment for confounding due to dissimilarities between treatment groups
  - Adjust for important predictors of treatment outcome
  - E.g., analyze treatment effect in a regression model including indicator of treatment
  - include as covariates those prognostic variables that differ between the groups

Comparison Groups

Propensity score analyses

- Propensity score analyses attempt to mimic randomization; does not worry about prognostic capability for outcome
  - Confounding = association between covariate and treatment AND association between covariate and outcome
- Creates a “propensity score" measuring the propensity for an individual with specific covariates to be in the new treatment group
- Perform an analysis adjusting for propensity scores
  - In each stratum, there is no association between covariate and treatment

Comparison Groups

Statistical remedies for meeting these criteria?

- Both approaches suffer from drawbacks noted by Byar (Biometrics, 1980) and Simon (Ca Treat Rep, 1982):
  - The variables that are measured and properly recorded typically explain only a small percentage in the variability in treatment group membership and treatment outcome.
    - That is, the regression models used have a very low R^2, thus our ability to have properly matched groups is rather low.
- Furthermore, progress in diagnostic methods and therapeutic strategies means that few measurements made in the past are exactly comparable to those made now
  - Laboratory and imaging techniques lead to improved diagnosis and staging of disease
    - E.g., earlier diagnosis of disease
    - E.g., detection of metastases at earlier stages causes trends toward milder disease being diagnosed as Stage IV
  - Supportive measures may improve outcomes

Internal controls

- Each subject serves as his/her own control
  - Different treatments at different times
  - Different treatments for different parts of body (eg. eye diseases, skin diseases)
- Note: This does not include “pre-post" designs looking at the change from baseline in a single arm study
  - These would be uncontrolled experiments

Comparison Groups

Concurrent controls

- Two or more treatment arms
  - Placebo or standard therapy
    - “If it is ethical to use a placebo, it is not ethical not to." -Lloyd Fisher
  - Active treatments
    - Sometimes consider equivalence
  - Multiple levels of same treatment
    - Evidence of dose-response
    - Identification of optimal dose

Blinding

What is blinding and how does it differ from concealed allocation?

- *Blinding (or masking)* is when neither the the study subject (single blind) nor the study investigator (double-blind) have knowledge of the treatment being received or delivered.
- *Concealed allocation* is when the study investigator (personnel) do not know the allocation sequence.

Blinding

What is blinding?

- ICH guidelines (part E9):

  Blinding or masking is intended to limit the occurrence of conscious and unconscious bias in the conduct and interpretation of a clinical trial arising from the influence which the knowledge of treatment may have on the recruitment and allocation of subjects, their subsequent care, the attitudes of subjects to the treatments, the assessment of end-points, the handling of withdrawals, the exclusion of data from analysis, and so on. The essential aim is to prevent identification of the treatments until all such opportunities for bias have passed.

Blinding

Types of blinding

- Participant and investigator bias can be (and have been) a major source of bias in RCTs
  - Such bias generally stems from knowledge of the type of treatment a participant is assigned in the trial
- In studies with concurrent comparison groups, blinding of treatment assignment can minimize bias
  - Single blind experiments : Participant is unaware of treatment assignment
  - Double blind experiments : Neither the participant nor treatment provider know treatment assignment
  - Triple blind experiments : Monitoring committee also blinded

Goals of Blinding

Consider the scientific objective

- ICH guidelines (www.ich.org) part E9 Statistical Principles

  "The most important design techniques for avoiding bias in clinical trials are blinding and randomisation, and these should be normal features of most controlled clinical trials intended to be included in a marketing application."

- Similar criteria are required in the CONSORT guidelines.

Goals of Blinding

Blinding can serve to

1. Minimize “placebo effect", wherein a participant being treated does better than one not treated, irrespective of the actual treatment
   - This is distinguished from secular trends in outcome that might occur over time (cohort effects)
   - To detect a placebo effect, one can compare a group that unknowingly received placebo to a group that received nothing

2. Minimize investigator bias in assessing
   - Adverse events
   - Treatment outcomes (consider subjective assessments such as time to hemostasis or time to tumor response)

3. Minimize bias due to missing data
   - Patients with chronic disease where multiple competing trials are ongoing may be less likely to continue in a given study with knowledge that they are receiving placebo

Goals of Blinding

Concealed allocation can serve to

- Prevent selection bias attributable to

  1. the participants
  2. the investigator

- “Allocation concealment seeks to prevent selection bias, protects the assignment sequence before and until allocation, and can always be successfully implemented. In contrast, blinding seeks to prevent ascertainment bias, protects the sequence after allocation, and cannot always be implemented."

  (Schultz, JAMA; 1995; 274(18)1456:1458)

Goals of Blinding

Does this really matter?

- Noseworthy (1994).*Neurology* 1994;44:16-20.
  - All patients examined and response judged by both a blinded and unblinded neurologist.

Table 1. *p* Value* of between-treatment comparison of proportion of subjects improved, stable, or worse

Assessment (no. pts)	IV cyclo versus placebo		PLEX versus placebo	
	Blinded	Unblinded	Blinded	Unblinded
6 Months (165)	0.159	0.069	0.246	0.047
12 Months (144)	0.295	0.084	0.086	0.004
18 Months (108)	0.418	0.255	0.106	0.072
24 Months (91)	0.088†	0.152	0.201	0.031
Final (mean, 30.4 months; 165)	0.290†	0.490	0.990	0.590

* Derived from chi-square test of the 2 (treatment) × 3 (improved, stable, worse) frequency table at each assessment point.
† Trend favoring placebo; all other comparisons favor active therapy.
IV cyclo Intravenous cyclophosphamide group (group I).
PLEX Plasma exchange group (group II).

Goals of Blinding

Does this really matter?

- Wright, Am Heart J (1948) 36:801-815.
- Odd/even day allocation in a trial of anticoagulants in MI gave 589 patients in the active treatment arm and 442 patients in the control arm.
  * 57% (589/1031) assigned to active treatment (95% CI for assignment probability: 0.54 to 0.60).
  * Clearly Biased allocation; cannot rule out differences between treatment groups.

Goals of Blinding

Does this really matter?

- Schulz, JAMA (1995) 273(5):408-412.
  * Meta-analysis of 250 trials from Cochrane pregnancy and childbirth database.
  * In trials with inadequate concealment of treatment allocation, odds ratios for treatment benefit were 41% larger (i.e., 41% better) than in trials with adequate concealment:

Table 2.—Odds Ratios in the Unclearly and Inadequately Concealed Trials Compared With Those in Adequately Concealed Trials*

Level of Allocation Concealment	Ratio of Odds Ratios (95% Confidence Interval)	χ^2 (*df*)	*P*
Adequate	1.00 (referent)		
Unclear	0.67 (0.60-0.75)	57.9 (2)	<.001
Inadequate	0.59 (0.48-0.73)		

*Multiple logistic regression model with the dependent variable being binary outcome measures from each meta-analysis. The independent variables included a binary variable for treatment group (experimental vs control); indicator variables to control for the effects of each of the 250 trials; terms for the "meta-analysis by treatment group" interaction to control for the different summary odds ratios for the treatment effects in the 33 meta-analyses; and the "allocation concealment by treatment" interaction terms displayed in this table to analyze their associations with estimates of treatment effects. Model deviance=434.2; *df*=215.

Issues With Blinding

Blinding is not always possible

- Placebo not always possible to be identical in appearance
  - Weight of fiber, viscosity of fluid for injections
- Side effects of treatment may be noticeable
  - Skin discoloration with beta-carotene
  - Injection site reactions
- Burden of treatment may not be ethical
  - Surgery, hospitalizations, repeated radiation exposure from CT scans

Issues With Blinding

Other issue need to be considered

- Appearance of treatments
- Dosage, administration schedules
- Coding and dispensing of treatments
- When and how to unblind
  - Emergent situations
  - Only unblind when treatment of toxicities differs between therapies
- Assessing how well the blind was maintained

Issues With Blinding

Blinded evaluation

- When blinding of participants is not possible, blinded evaluation may be
- Must still ensure a similar schedule of assessments
  - Side effects might lead to more frequent monitoring
- Competing risks (eg. death from other causes) still a problem

Issues With Blinding

Blinded evaluation

- In some cases, use of a blinded independent review committee may be mandated
  - Ex: Progression of disease in the setting of follicular non-Hodgkins lymphoma
  - Investigators at each perform measurable lesion assessments based on CT scans and physical examination to determine response and progression
  - Blinded independent radiology review committee retrospectively read and interpret all CT scans for response evaluation and progression
  - Primary response based upon independent review committee
- Bias and monitoring issues can still arise (cf. Dodd et al, JCO (2008), Brummel and Gillen, OJS (2013))

Blinding

When is blinding unnecessary?

- Blinding is less of an issue with harder endpoints (eg. survival)
- The more objective the measurement of outcome the less important blinding is to the scientific credibility of a clinical trial

Blinding

Subjective outcomes

- In cases where blinding is not possible it is important to make outcome assessments as objective as possible
  - Ex: Hemostatic agents for cessation of minor to moderate bleeding during surgery
  - Control: Sponge; Treatment: Powder
  - Not possible to blind surgeon
  - Surgeon responsible for determining when hemostasis has occurred
  - How to define hemostasis?

Treatment allocation

Objective, need, requirements

- Objectives:
  - Treatment groups must be comparable so that differences between groups are due to treatment.
  - Assure against confounding (by both measurable and unmeasurable differences):
    - We might be able to adjust for confounders that can be measured.
    - We cannot adjust for unmeasured differences.
    - To measure confounders we would have to know them *a priori*.
- Requirement:
  - Randomization assures that on average all treatment groups are comparable.

Treatment allocation

Methods

- Concealed allocation:
  - Study personnel cannot determine the treatment assignment before it occurs (or not until study completion in double blind trials).
  - Bias can occur with inadequate concealment.

Treatment allocation

Methods

- Approaches to Randomization
  - Completely randomized designs
  - Blocked randomization
  - Stratified randomization
  - Baseline-adaptive randomization
  - Response-adaptive randomization
  - Cluster randomization

Treatment allocation

Methods: Completely randomized designs

- Treatment assignment is made by randomly allocating a subject to one of the treatment groups without considering previous treatment allocations or the subject's covariates.
  - With equal probabilities of getting any one of the treatments (like flipping a coin).
  - With unequal probabilities of getting each of the treatments (like flipping a biased coin).
- Advantages:
  - Analysis is straightforward
  - Simple to implement
- Disadvantages:
  - In small trials this may result in loss of power and/or bias due to:
    - Unequal number of subjects on treatment arms.
    - Imbalances in the types of patients on different arms.
    - Time trends in non-study treatments or types of patients.

Methods: Blocked Randomization

- Random treatment allocation in (relatively small) blocks so that the desired number of subjects in each treatment is assured.
  - E.g., If you want 500 subjects in each of two treatment groups, then assign patients in 50 blocks of 20 patients so that in each block 10/20 are assigned to each treatment
- Advantages:
  - Potential for more power due to equal number of patients on each arm.
  - Better protection against time trends.
- Disadvantages (none really, but...):
  - (Analysis could account for blocking to attain higher power.)
  - (More complicated to implement and rarely done.)

Treatment allocation

Methods: Stratified Randomization

- Randomization in strata defined by important covariates:
  - E.g., To guarantee gender balance, randomize in small blocks in males and females separately (e.g., first 20 males are equally allocated between treatments A and B; first 20 females are equally allocated between treatments A and B).
  - Particularly useful in small trials with a few covariates that are strong predictors of outcome.
  - Difficult with small numbers in each strata (i.e., cannot have a large number of stratification variables).
- Advantages:
  - Guarantees balance on important covariates (reduces chance of confounding).
  - Reduces variation
- Disadvantages:
  - More difficult to implement
  - Analysis should account for stratification variables (adjust for stratification variables).

Treatment allocation

Methods: Baseline-Adaptive Randomization

- Adaptively modify the randomization procedure to ensure comparable frequency distributions of several covariates.
  - E.g., if there is currently an excess of males receiving treatment A, then the next male should be assigned to treatment B.
  - (Minimization) Each patient is allocated to minimize the imbalance between all important covariates.
- Advantages:
  - Same as for stratification
  - May work better in small samples
- Disadvantages:
  - Much more difficult to implement.
  - Analysis must account for the covariates that controlled the allocation.

Methods: Response-Adaptive Randomization

- “Play the winner" designs:
  - Modify the number of subjects assigned to each treatment according to outcomes of previous subjects.
  - You must have knowledge of previous outcomes by treatment group for each randomization.
- Advantages:
  - Decreases the number of subjects who receive an inferior treatment.
- Disadvantages:
  - May decrease power of the study (serious imbalances may result).
  - Increased chance for bias.
  - May not convince the scientific community.

Treatment allocation

Response-Adaptive Randomization (Example)

ECMO: Extracorporeal Membrane Oxygenation in neo-natal respiratory failure.

- Persistent pulmonary hypertension results in right to left shunt through the foramen ovale or ductus arteriosus causing hypoxemia. ECMO is used to maintain life until the condition resolves.
- Trial 1 (Play the winner absolutely): *Pediatrics* (1985) 76:479-487
  - First subject was randomized to conventional medical therapy (CMT); the infant died.
  - Second subject given ECMO; infant lived.
  - Next 8 subjects given ECMO; all lived.
  - Result:
    100% mortality with CMT
    0% with ECMO
    RR = 0.

Response-Adaptive Randomization (Example)

ECMO Example (con't):

- Trial 2 (Play the winner with higher probability): *Pediatrics* (1989) 84(6):957-63
  - Randomize until the 4th CMT death, then treat remainder with best approach.
  - 19 babies in first phase (4/10 die with CMT; 0/9 die with ECMO).
  - 20 babies on ECMO in second phase (1 death).
  - Result:
    40% (4/10) mortality with CMT;
    3% (1/29) with ECMO;
    RR = 0.086.

- Trial 3 (conventional RCT): *Pediatrics* (1998) 101(4):E1
  - Randomize 185 infants (92 to CMT, 93 to ECMO)
  - Result:
    59% (54/92) mortality with CMT;
    32% (30/93) with ECMO;
    RR = 0.55.

Response-Adaptive Randomization (Example)

ECMO Example (con't):

Implications of the ECMO example:

- ECMO looked better with response-adaptive randomization.
- Response-adaptive designs were not accepted as adequate justification for ECMO.
- Inadequate study designs can delay introduction of beneficial treatments or prolong use of inferior treatments.

"In fact, in the ECMO trial, the patient who failed on treatment B had the most extreme values on no fewer than four important covariates (Paneth & Wallenstein, 1985), and was clearly the sickest. In effect, the trial provides no information whatsoever regarding the treatment comparison. "

-Begg (1990)

Response-Adaptive Randomization (Example)

- The ECMO experience has tempered enthusiasm for randomized PTW
- This being said, there may be times were response-adaptive randomization will work, but
  - There needs to be a clear dilemma re individual ethics
  - There will tend to be decreased group ethics
  - It takes a lot of planning in order to obtain results that will be sufficiently credible

Treatment allocation

Methods: Cluster Randomization

- Subjects are randomized in groups (e.g., churches, schools, cities).
  - Useful when treatment cannot be administered on an individual level without contamination (e.g., smoking cessation studies).
  - Often clusters are matched and treatments are assigned within the matched pairs.
- Advantages:
  - Allows investigation of community interventions.
  - Eliminates contamination bias.
- Disadvantages:
  - Sample size is the number of clusters not the number of individuals.
  - May lose power over individual randomization.

Treatment allocation

Methods: Logistics of Randomization

(a) Completely randomized designs:
- Create column from 1 to $2N$.
- Create column of random numbers uniformly distributed between 0 and 1.
- If the random number is less than 0.5, then the subject receives active treatment, otherwise they receive placebo.

(b) Blocked randomization: For a block of size k with $k/2$ subjects in each of two groups:
- Create a column of $k/2$ A's and $k/2$ B's.
- Create column of random numbers uniformly distributed between 0 and 1.
- Sort the first column according to the second column.
- Repeat for as many blocks as desired.

(c) Stratified randomization: Repeat for each stratum.

Methods: Logistics of Randomization

- Where to perform randomization:
  - Central randomization:
    - Phone calls to the coordinating center.
    - Sequences can be determined at the start of the study (except with adaptive randomization).
  - Distributed randomization: Computer programs, envelopes, or lists at pharmacies.
- Important principles:
  - Strong quality assurance must be in place to ensure proper randomization.
  - Ensure adequate concealment/blinding.
  - Provide for emergency unblinding.
  - Exact randomization scheme must be known for analysis.

SISCR
UW - 2018

Introduction to Clinical Trials – Day 1

Session 4 – Statistical Tasks in Trial Design

Presented July 24, 2018

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
- Study designs

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

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}… 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

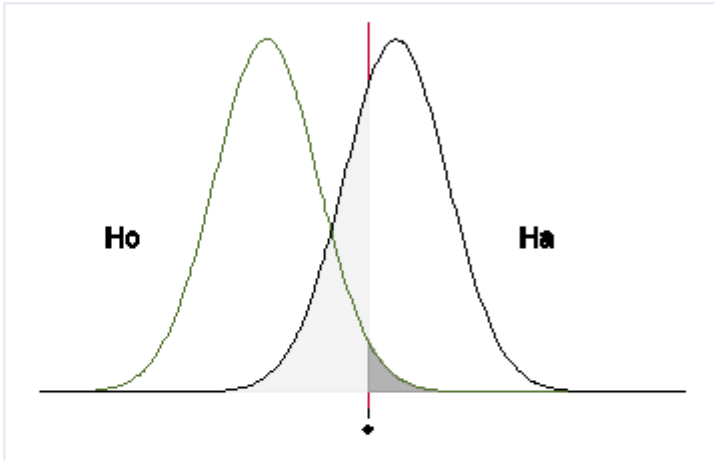
$1 - \beta$ = Power of the test = Probability of rejecting H_0 when it is false.
(more on Power later)

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)



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

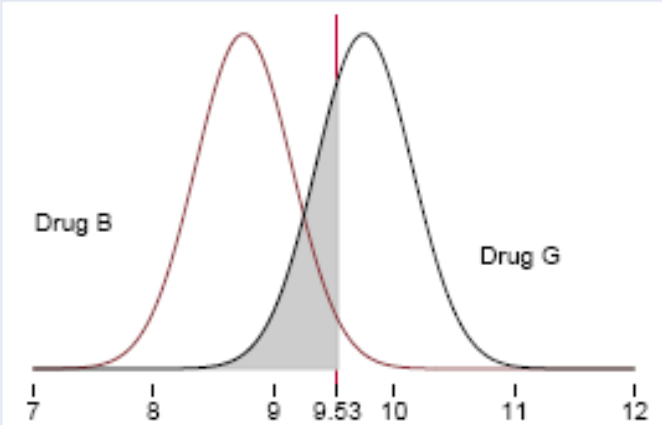
also: $$* = \mu_a - z_{1-\beta}\left(\sigma/\sqrt{n}\right) \qquad \text{(note: } z_\beta = -z_{1-\beta}\text{)}$$

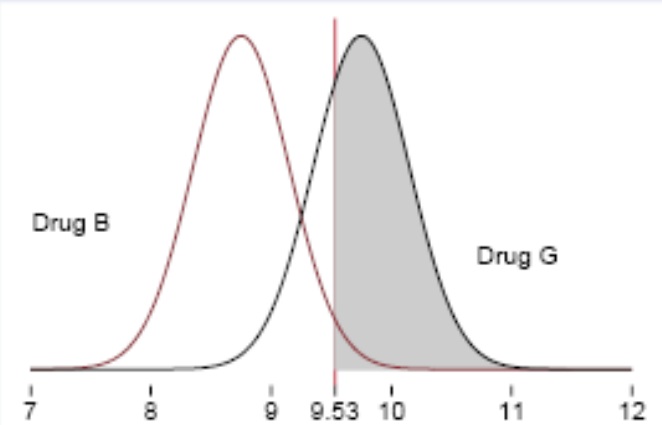
Thus, $$\mu_0 + z_{1-\alpha/2}\left(\sigma/\sqrt{n}\right) = \mu_a - z_{1-\beta}\left(\sigma/\sqrt{n}\right)$$

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.

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:

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

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

$\Rightarrow n = \sigma^2 \dfrac{\left(z_{1-\alpha/2} + z_{1-\beta}\right)^2}{\left(\mu_a - \mu_0\right)^2}$ 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 0.024, 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"

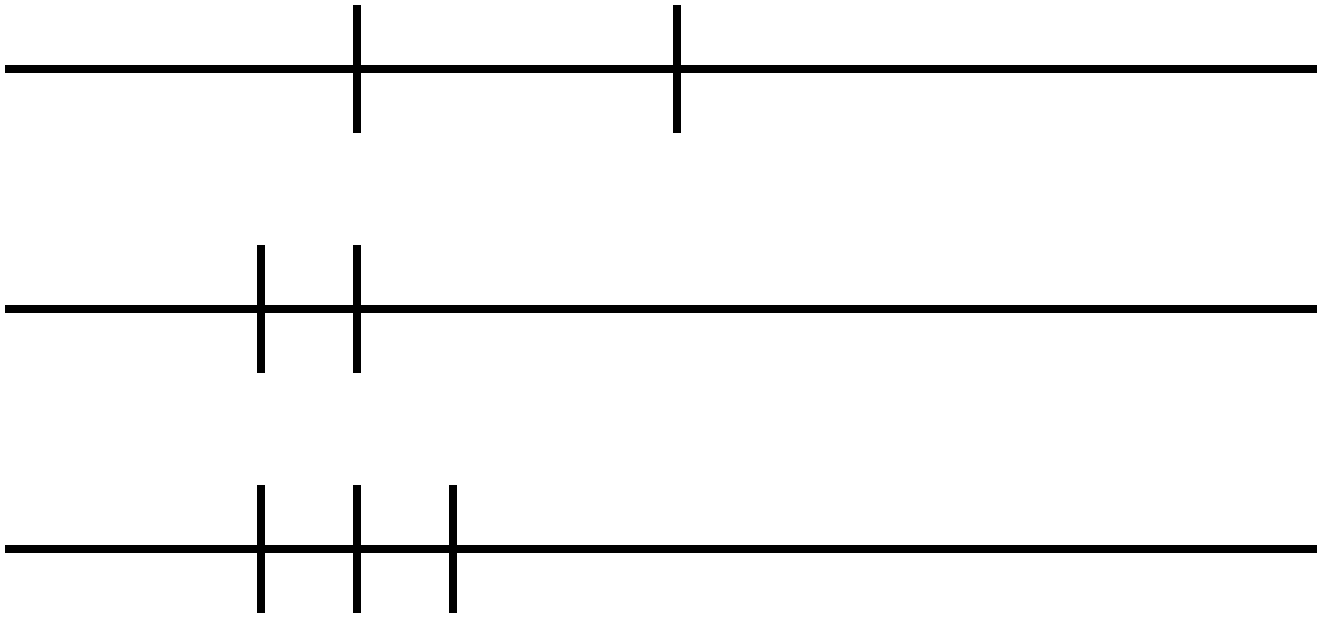
- Possible explanations for no statistically significant difference in estimate of θ
  - There is no true difference in the distribution of response across groups
  - There is a difference in the distribution of response across groups, but the value of θ is the same for both groups
    - (i.e., the distributions 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*