Module 1:

Introduction to Clinical Trials

Scott S. Emerson, M.D., Ph.D.
Professor Emeritus of Biostatistics
University of Washington

Seattle Institute for Statistics in Clinical Research
July 22, 2019

1

Abstract

This module provides an overview of the clinical and scientific issues that must be considered by the many disciplines that collaborate on a clinical trial. In this module, we consider the clinical, scientific, and regulatory setting of clinical trials, describing the phased approach to "treatment discovery" in which the safety, efficacy, and effectiveness of candidate treatments is investigated. In particular, we discuss the issues surrounding identification of the target population, definition of the treatment, choice of clinical outcomes, and choice of comparators. It will include introductory discussion of the choice of randomization strategies, blinding, conduct and monitoring of the study, and plans for reporting of the result. The importance of a well-defined study protocol is emphasized.

2

Where Am I Going?

Overview of Use of RCTs in Clinical Research

Organization of the Course

1. Scientific / ethical setting: "Treatment Discovery"
2. Clinical trial goal: Estimands / endpoints
3. Materials and methods:
   a) Patient eligibility criteria
   b) Randomization / blinding
   c) Treatments / study procedures
   d) Data collection / management / analysis
4. Documentation:
   a) Protocol
   b) Statistical Analysis Plan

3

Scientific Setting: Clinical Research

Experimentation in Human Volunteers

"Treatment Discovery"

Where am I going?

Improvements in population health are the result of a long history of scientific research into diagnosis, treatment and (ideally) prevention of disease

A succession of scientific studies and experiments
- "The result of a scientific study is another scientific study"

Knowledge solidified through clinical experiments
- Rigorous science
- Ethical treatment of individuals and populations

4

Medical Science and Statistics

- Statistics is about science
  - (Science in the broadest sense of the word)
- Science is about proving things to people
  - (The validity of any proof rests solely on the willingness of the audience to believe it)
- Medical science is about improving the health of people, but

"The physician must ... have two special objects in view with regard to disease, namely, to do good or to do no harm"

- Epidemics I, Hippocratic Corpus, c. 410 BC

"…it is only the second law of therapeutics to do good, its first law being this – not to do harm"

- Elisha Bartlett, 1844

5

Goals of Medical Research

- Identify methods to diagnose disease
- Identify risk factors for disease
- Identify treatments for disease
- Identify methods for disease prognosis
- Identify strategies for prevention of disease
- Basic science

6

Scientific Method

- Observation
- Form hypotheses
- Designed study
  - Overall goal
  - Specific aims
  - Materials and Methods
  - Data collection
  - Results (data analysis)
  - Interpretation and conclusions
- Repeat (new, refined hypotheses)

7

Bioethics in Clinical Research

- Classic principles
  - Autonomy
  - Beneficence
  - Non-maleficence
  - Justice
- Selected additional principles
  - Authority
  - Principle of double effect
  - Confidentiality
  - Equality
  - Equity
  - Mercy
  - Ownership
  - Privacy

8

Implementation in Clinical Research

- Governmental regulatory authorities
  - FDA (US), EMA (Europe), PMDA (Japan), MHRA (UK), etc.
  - International Council on Harmonisation (ICH)
- Institutional Review Boards (IRBs)
  - Institutional Ethics Commiteess, Research Ethics Boards, etc.
- Data Monitoring Committees
- Informed Consent Forms

9

U.S. Regulation of Drugs / Biologics

- 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)

10

Medical Devices

- Medical Devices Regulation Act of 1976
  - Class I: General controls for lowest risk
  - Class II: Special controls for medium risk - 510(k)
  - Class III: Pre marketing approval (PMA) for highest risk
    - "…valid scientific evidence for the purpose of determining the safety or effectiveness of a particular device … adequate to support a determination that there is reasonable assurance that the device is safe and effective for its conditions of use…"
    - "Valid scientific evidence is evidence from well-controlled investigations, partially controlled studies, studies and objective trials without matched controls, well-documented case histories conducted by qualified experts, and reports of significant human experience with a marketed device, from which it can fairly and responsibly be concluded by qualified experts that there is reasonable assurance of the safety and effectiveness…"
- Safe Medical Devices Act of 1990
  - Tightened requirements for Class 3 devices

11

Clinical Trials

- Experimentation in human volunteers
- Investigates a new treatment, preventive agent, or diagnostic method
- Safety:
  - Are there adverse effects that clearly outweigh any potential benefit?
- Efficacy:
  - Can it alter the disease process in a beneficial way?
- Effectiveness:
  - Would its adoption as a standard affect morbidity / mortality in the population?

12

The Enemy

"Let's start at the very beginning, a very good place to start…"

- Maria von Trapp

(as quoted by Rodgers and Hammerstein)

13

Overall Goal

- "Drug discovery"
- More generally
  - a therapy or preventive strategy
    - drug, biologic, device, behavior
  - for some disease
  - in some population of patients
- Medical diagnosis / prognosis
  - Evaluation of methods

14

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

15

First: Where Do We Want To Be?

- Perform some innovative experiment?
- Find a use for some proprietary drug / biologic / device?
  - "Obtain a significant p value"
- Find an efficacious treatment?
  - "Efficacy" seems to benefit some people somehow
- Find a new treatment that improves health of individuals
  - "Personalized medicine" (fixed effect or random effect?)
- Find effective treatment that improves health of the population
  - Treatments administered to a community
  - Treatments tested on a population then administered correctly

16

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

17

Disease

- A moving target heavily influenced by treatment
  - Then: "fevers"
  - Now: "MRSA-related pneumonia"
- Trends over place and time in definition because
  - Symptoms
    - Cultural 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

18

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

19

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

20

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

21

Outcomes

- The desired beneficial reponse from the treatment as can be defined for ***every*** patient
  - Clinical outcomes
    - Prolonged survival
    - Quality of life
  - Surrogate outcomes believed to be prognostic of clinical outcome
    - Cardiovascular: blood pressure, cholesterol, arrhythmias
    - Cancer: tumor size, tumor progression
    - Diabetes: blood glucose, glycosylated hemoglobin
    - Infectious diseases: vaccine titers
- Definition
  - Method of measurement
  - Timeframe

22

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

23

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

24

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?

25

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?

26

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

27

Observational Studies - 1

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

28

Observational 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

29

Observational 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
  - What is time 0?

30

Why Not Observational Studies?

- Effects of arrhythmias post MI on survival
  - Observational studies: high risk for death
  - CAST: Specific anti-arrhythmics have higher mortality
- Effects of beta-carotene on lung CA and survival
  - Observational studies: high dietary beta carotene has lower cancer incidence and longer survival
  - CARET: beta carotene supplementation in smokers leads to higher lung CA incidence and lower survival
- Effects of hormone therapy on cardiac events
  - Observational studies: HT has lower cardiac morbidity and mortality
  - WHI: HT in post menopausal women leads to higher cardiac mortality

31

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 that outweigh potential benefit?
    - Efficacy:
      - Does treatment beneficially alter the disease process
    - Effectiveness:
      - Would adoption of the treatment improve morbidity / mortality in the population?

32

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-cntrl
    - **II-3:** Multiple time series with/without intervention; Dramatic results from uncontrolled tral
  - **Level III:** Opinions of respected authorities

33

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)

34

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
- Hypothesized refinements to treatment indication
  - More clinical trials

35

Interventional Studies

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

36

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

37

Safety

- Safety to conduct trials
  - No serious adverse effect appears in first few subjects treated
- Safety in practice
  - Manageable adverse reaction profile
  - Severity: mild or moderate
  - Serious: rarely leads to hospitalization, disability, death, congenital abnormalities
  - Risk factors: patients at risk for adverse reactions can be identified before lasting clinical harm occurs
    - Errors of commission: treatment causes lasting disability
    - Errors of omission: lack of tolerability causes delay in better therapy
- Risk / benefit tradeoffs
  - Lack of efficacy in presence of any adverse reactions may be a safety problem

38

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

39

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

40

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

41

Disease

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

42

Target Population

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

43

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

44

Control Treatment

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

45

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?

46

Intervention

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

47

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)

48

Measurement of Outcome

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

49

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

50

Which: Efficacy or Effectiveness

- Factors leading to efficacy trials
  - "Knowledge is good"
  - As screening studies before confirmatory registrational studies
  - 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, $$)

51

Typical Scientific Hypotheses

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

{ better than, worse than, or about the same as } { an absolute standard, or what it would have been with some other treatment }

52

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

53

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

54

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

55

First Scientific Refinement

- Determine whether the group that received the treatment will have outcome measurements that are

{ higher than, / lower than, or / about the same as } { an absolute standard, or / measurements in an otherwise comparable group (that did not receive the treatment) }

56

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)

57

Second Scientific Refinement

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

{ higher than, / lower than, / about the same as } { an absolute standard, / measurements in an otherwise comparable group (that did not receive the treatment) }

58

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
- Post approval of indication
  - Phase IV: Post-marketing surveillance, REMS

59

Choice of Scientific Hypotheses

- Hypotheses addressed in a particular clinical trial must take into account the ethics related to current state of knowledge
  - First-in-human studies of a new experimental compound
  - New indication for previously approved drug
  - Comparative effectiveness of two approved therapies
- Hypotheses must also consider the (cost) efficient use of resources as evidence for effectiveness is accumulated
  - Rigor demanded in registrational confirmatory studies leads to very expensive RCT: $20M - $70M per study (2014 estimates)
  - Earlier phase studies might be used to screen out treatments that do not show effect on surrogate endpoints related to hypothesized mechanisms of action

60

Examples of Special RCT Settings

- First-in-human studies of a new experimental compound
  - Phamacokinetic / pharmacodynamic studies (PK / PD)
    - Healthy volunteers vs patients
    - Single dose vs multiple dose
    - Fasting vs with food
  - Dose finding
    - Cancer: Maximum tolerated dose through dose escalation
    - Others: Parallel dose groups for risk/benefit tradeoffs
      - Perhaps using surrogate biomarker for signs of effect
- Targeted safety studies
  - Thorough QT/QTc prolongation studies
  - Drug-drug interactions (DDI)
  - Cardiovascular outcome trials (CVOT)

61

Typical Regulatory Chronology

- Phase 1:
  - PK/PD; dose finding; dose limiting toxicities
- Phase 2:
  - Initial estimates of efficacy
    - Perhaps surrogate in enriched population over short timeframe
  - Initial assessment of safety
    - Adverse events (AEs), serious adverse events (SAEs)
    - Perhaps some protocolized safety measurements
- Phase 3: (“adequate and well-controlled investigations”)
  - Confirmatory efficacy (closer to effectiveness)
  - Detailed safety: protocolized as indicated and AEs, SAEs
- Phase 4: (post marketing)
  - Additional safety and special populations as indicated

62

Some Regulatory Jargon

- Not covered in depth here:
  - Pivotal studies
  - Noninferiority studies
  - Bioequivalence studies
  - Bridging studies

63

Clinical Trial Goals

Phases of Investigation

Estimands / Endpoints

Where am I going?

Evidence necessary to adopt a new treatment, prevention strategy, diagnostic method is accumulated gradually across multiple pre-planned studies

The goal of any individual study is ultimately defined by the "estimand" targeted by the study

Precise definition of the estimand in clinical research must account for the ethical and logistical issues inherent in human experimentation

64

Scientific Basis

- A clinical trial is planned to detect the effect of a treatment on some outcome
- Statement of the outcome is a fundamental part of the scientific hypothesis

65

Ethical Basis

- Generally, subjects participating in a clinical trial are hoping that they will benefit in some way from the trial
- Clinical endpoints are therefore of more interest than purely biological endpoints

66

Statistics and Game Theory

- 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

67

Ex: Level 0.05 per Decision

- Experiment-wise Error Rate

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

68

Primary Endpoint: Clinical

- Consider (in order)
  - The most relevant clinical endpoint
    - Survival, quality of life ("feels, functions, or survives")
  - The endpoint the treatment is most likely to affect
  - The endpoint that can be assessed most accurately and precisely

69

Additional Endpoints

- Other outcomes are then relegated to a "secondary" status
  - Supportive and confirmatory
  - Safety
- Some outcomes are considered "exploratory"
  - Subgroup effects
  - Effect modification

70

Primary Endpoint: Clinical

- Consider (in order of importance)
  - The phase of study: What is current burden of proof?
  - The most relevant clinical endpoint
    - Survival, quality of life
    - Proven surrogates for the above
      - But how can we be sure?
  - The endpoint the treatment is most likely to affect
    - Therapies directed toward improving survival
    - Therapies directed toward decreasing AEs
  - The endpoint that can be assessed most accurately and precisely
    - Avoid unnecessarily highly invasive measurements
    - Avoid poorly reproducible endpoints

71

Primary Endpoints: Missing Data

- The primary endpoint must be chosen so that it is measurable on every subject
- The method and timeframe of measurement should not be influenced by “post-randomization” variables
- Missing data on subjects can lead to uninterpretable results
  - Define how deaths will contribute to primary endpoints
  - Anticipate treatment discontinuation in some subjects and continue to collect their outcome data
  - Take steps to prevent loss to follow-up
  - Try to minimize patients withdrawing consent entirely

72

Multiple Endpoints

- Sometimes we must consider multiple endpoints
- We then control experimentwise error
- Possible methods
  - Composite endpoint
    - AND: Individual success must satisfy all
    - OR: Individual success must only satisfy one
    - AVERAGE: Sum of individual scores
    - EARLIEST: Event-free survival
      - Progression free survival(PFS): progression or death
      - Major Adverse Cardiovascular Events (MACE): MI, stroke, cardiovascular death
  - Co-primary endpoints
    - Must show improvement in treatment group on all endpoints
      - No guarantee that the same subjects are experiencing the improvement

73

Composite Endpoints: Caveats

- Sometimes investigators try to deal with missing data through composite endpoints
  - E.g., in cancer study count patient discontinuation from study drug as a “treatment failure” along with progression free survival
- Such an approach is not in keeping with the true goals of treatment discovery
  - Equates treatment d/c with cancer progression or death
  - If the treatment is not known to be effective, there is no reason to presume that d/c of treatment is in any way harmful
  - Furthermore, patients may have reasons to continue on experimental treatments due to off-target effects
    - Famous examples include minoxidil and sildenafil

74

Competing Risks

- Occurrence of some "nuisance" event precludes observation of the event of greatest interest, because
  - Further observation impossible
    - E.g., death from CVD in cancer study
  - Further observation irrelevant
    - E.g., patient advances to other therapy (transplant)
- Methods
  - Event free survival: time to earliest event
  - Time to progression: censor competing risks
  - "U statistics": define ranking based on both events
  - More recent missing data methods somehow imputing results to populations without competing risks

75

Competing Risks: Caveats

- Competing risks produce missing data on the event of greatest interest
- As with all missing data problems, there is nothing in your data that can tell you whether your actions are appropriate
  - Are subjects with competing risk more or less likely to have event of interest?
  - (the term "competing risk" has become shorthand for a setting in which your results are in doubt)

76

Surrogate Outcomes

Motivation and Examples

Where am I going?

- The goal of a RCT is to find effective treatment indications
- Statistical and logistical constraints often lead to the desire for surrogate outcomes
  - But these have led us astray in the past

77

Goal of Clinical Trial

- Establish whether an experimental treatment will prevent a particular clinical outcome
  - Incidence of disease
  - Decreased quality of life
  - Mortality

78

Problems

- Relevant clinical outcomes are often relatively rare events that occur after a significant delay
  - Believe that earlier interventions have greater chance of benefit
- Difficulty in measuring clinical outcome
  - Quality of life needs to be assessed over a sufficiently long period of time

79

Impact on Clinical Trial Design

- Large sample size required to assess treatment effect on rare events
- Long period of follow-up needed to assess endpoints
- Isn't there something else that we can do?

80

Motivation for Surrogate Endpoints

- Hypothesized role of surrogate endpoints
  - Find a biological endpoint which
    - can be measured in a shorter timeframe,
    - can be measured precisely, and
    - is prognostic for the clinical outcome
- Use of such an endpoint as the primary measure of treatment effect will result in more efficient trials

81

Identifying Potential Surrogates

- Typically use observational data to find prognostic risk factors for clinical outcome
- Treatments attempt to intervene on those risk factors
- Surrogate endpoint for the treatment effect is then a change in the risk factor

82

Examples

- Colon cancer prevention
  - Patients with adenomatous colon polyps have two-fold increase in risk of colon cancer over the general population
    - (It is believed that all colon cancer will first arise in colon polyps, and lifetime incidence of colon polyps is about 50%)
  - Prevention directed toward preventing colon polyps
  - Treatment effect measured by decreased incidence of colon polyps
  - True clinical outcome is preventing mortality

83

Examples

- AIDS
  - HIV leads to suppression of CD4 cells
  - Decreased CD4 levels correlates with development of AIDS
  - Treatment effects measured by following CD4 counts
  - True clinical outcome is prevention of morbidity and mortality

84

Examples

- Coronary heart disease
  - Poor prognosis in patients with arrhythmias following heart attack
  - Therapies directed toward preventing arrhythmias
  - Treatment effects measured by prevention of arrhythmias
  - True clinical outcome is prevention of mortality

85

Examples

- Liver failure
  - Poor prognosis in patients who develop renal failure
  - Therapies directed toward treating renal failure (dialysis)
  - Treatment effects measured by creatinine, BUN
  - True clinical outcome is prevention of mortality

86

Examples

- Other surrogate endpoints used historically
  - Cancer: tumor shrinkage
  - Coronary heart disease: cholesterol, nonfatal MI, blood pressure
  - Congestive heart failure: cardiac output
  - Arrhythmia: atrial fibrillation
  - Osteoporosis: bone mineral density
- Future surrogates?
  - Gene expression
  - Proteomics

87

Problem

- Establishing biologic activity does not always translate into effects on the clinical outcome
- May be treating the symptom, not the disease
  - Examples
    - Concorde: ZDV improves CD_4, not survival
    - CAST: encainide, flecainide prevents arrhythmias, worsens survival
    - Laromustine in AML: improves complete remissions, worsens survival
- May be missing effect through other pathways
  - Example
    - Intl CGD group: Gamma-INF no affect on biomarkers, decreases serious infections

88

Example: Concorde Trial

- (Lancet, April 3, 1993)
  - Asymptomatic HIV positive patients
  - Randomize to
    - Immediate ZDV (n = 877)
    - Placebo then progression to ZDV (n = 872)
  - Mean follow-up: 3 years

89

Concorde Trial: Surrogate Results

- Surrogate results on CD4 changes
  - 3 mos relative to baseline
    - Immediate ZDV: +20 cells
    - Placebo: -10 cells
  - Difference between treatment arms
    - 3 mos: 30 cells (P < .0001)
    - 6 mos: 35 cells (P < .0001)
    - 9 mos: 32 cells (P < .0001)
- Clinical results on AIDS/Death composite and all cause death

	ZDV (n = 877)	Placebo (n = 872)
AIDS / Death	175	171
Death	95	76
3 year survival	92%	93%

90

Concorde Trial: Conclusions

- "Results cast doubt on the value of using changes over time in CD4 count as a predictive measure for effects of antiviral therapy on disease progression and survival."

- Review of ZDV, ddI and ddC on Surrogate Markers and Clinical Endpoints (later meta-analysis)
  - 16 trials reviewed by NIAID S.O.T.A. Panel, Jun 93

		AIDS/Death		Survival			
		+	-	+	-	--	?
CD4	+	7	6	2	6	3	2
Effect	-	1	2	2	1	0	0

91

Example: CAST

- Cardiac Arrhythmia Suppression Trial
- Arrhythmia a risk factor for sudden death following a myocardial infarction
- Antiarrhythmic drugs (encainide and flecainide) successfully decrease incidence of arrhythmias
- CAST
  - Placebo controlled trial using mortality as outcome
  - Encainide and flecainide TRIPLE the death rate

92

Example: CGD

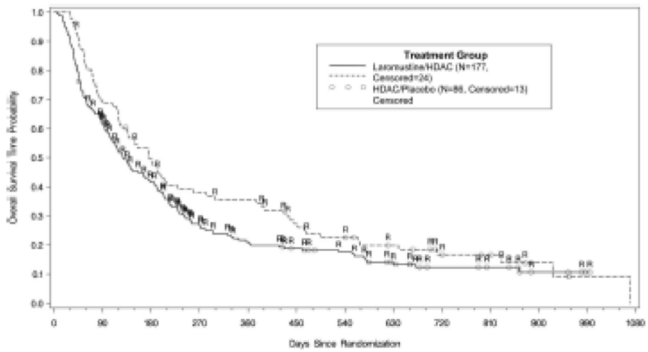
- Chronic Granulomatous Disease (CGD) leads to recurrent serious infections
- Gamma interferon increases bacterial killing and superoxide production?
- International CGD Study Group Trial of Gamma-INF
  - 70% reduction in recurrent serious infections
  - Essentially no effect on biological markers

93

Example: Laromustine in AML

- Laromustine + HDAC vs Placebo + HDAC
  - Surrogate outcome: Complete response 35% vs 19%, p= 0.005
  - Clinical outcome: Overall survival HR = 1.82, p= 0.004

BLOOD, 5 NOVEMBER 2009 • VOLUME 114, NUMBER 19 HDAC ± LAROMUSTINE IN FIRST RELAPSE AML 4031



Note: Patients alive as of the date of last follow-up have their survival time censored at last date of known mortality status.
R = patient achieved a CR or CRp.

Figure 2. Kaplan-Meier estimate of overall survival time.

94

Surrogate Outcomes

Possible Mechanisms

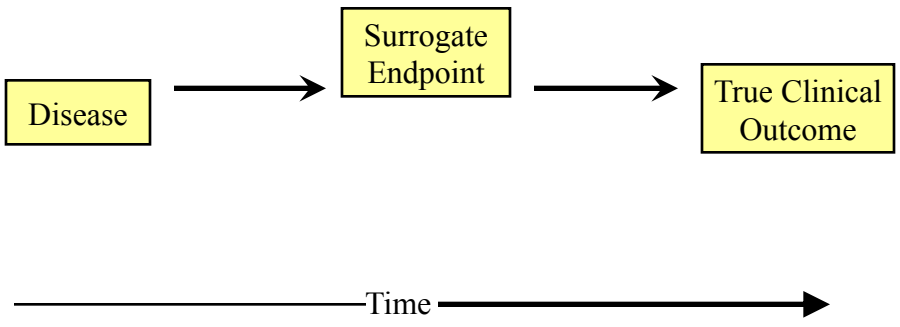
Where am I going?

- Understanding the pitfalls of surrogate outcomes requirese thinking about the mechanisms of treatments

95

Scenario 1: The Ideal

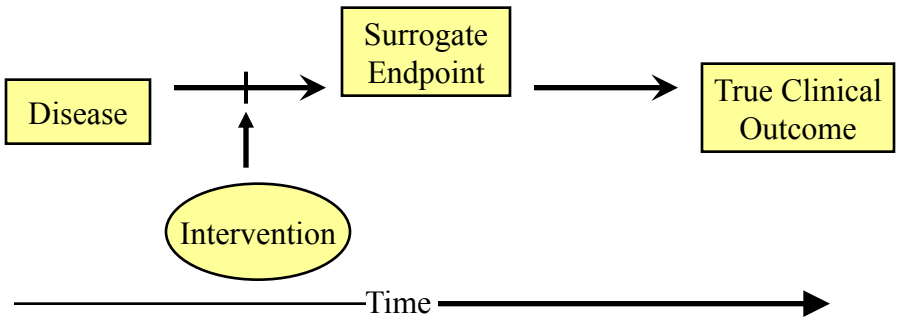
- Disease progresses to Clinical Outcome only through the Surrogate Endpoint



96

Scenario 1a: Ideal Surrogate Use

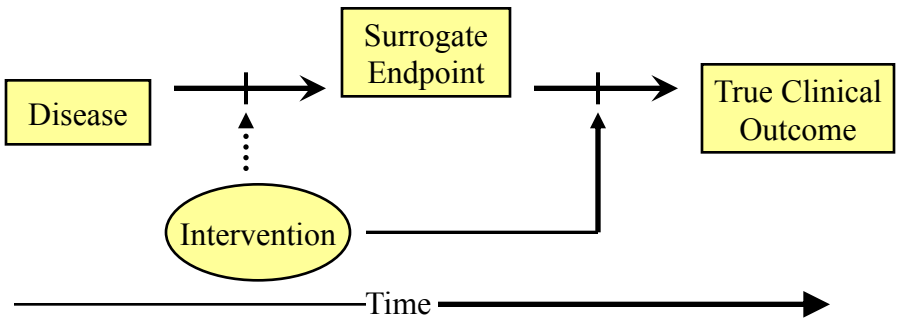
- The intervention's effect on the Surrogate Endpoint accurately reflects its effect on the Clinical Outcome



97

Scenario 1b: Inefficient Surrogate

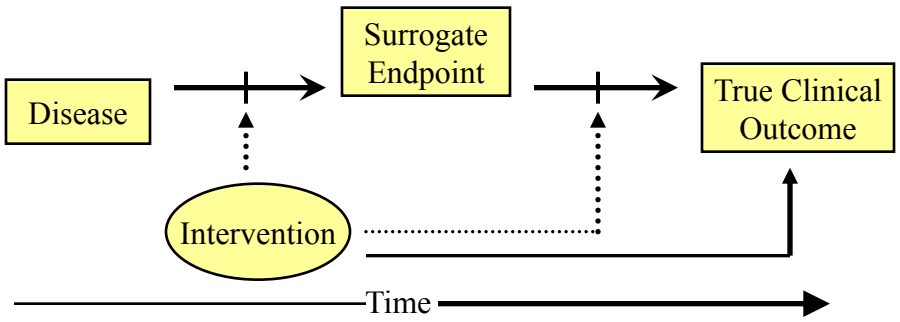
- The intervention's effect on the Surrogate Endpoint understates its effect on the Clinical Outcome



98

Scenario 1d: Dangerous Surrogate

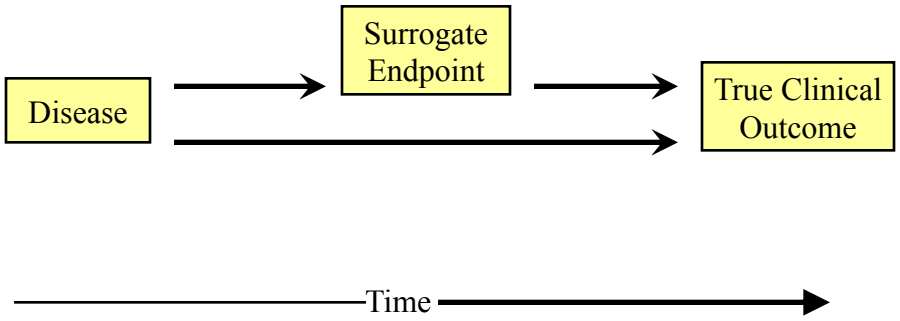
- Effect on the Surrogate Endpoint may overstate its effect on the Clinical Outcome (which may actually be harmful)



99

Scenario 2: Alternate Pathways

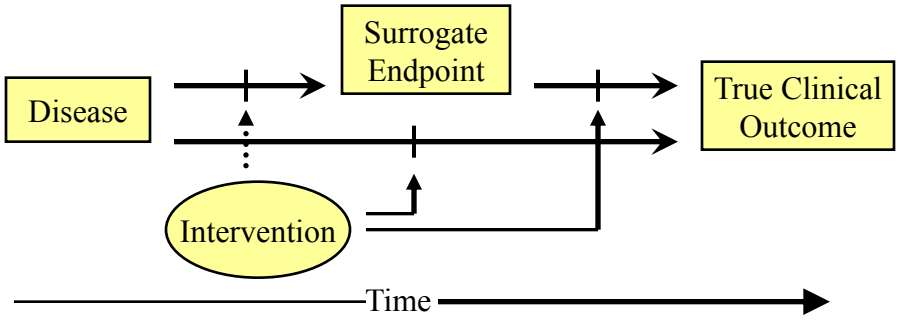
- Disease progresses directly to Clinical Outcome as well as through Surrogate Endpoint



100

Scenario 2b: Inefficient Surrogate

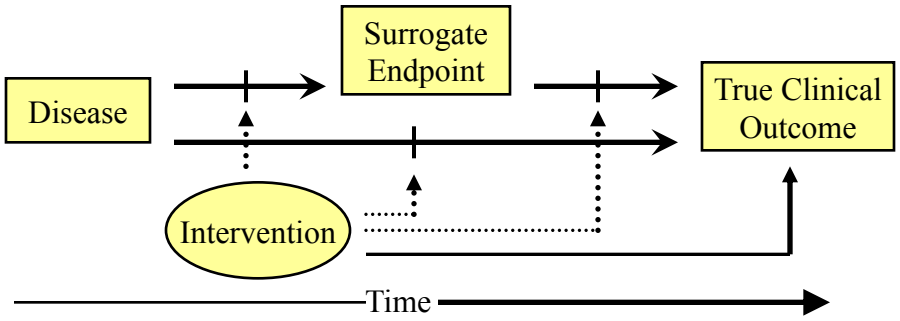
- Treatments' effect on Clinical Outcome is greater than is reflected by Surrogate Endpoint
  - E.g., Gamma interferon in CGD



101

Scenario 2d: Dangerous Surrogate

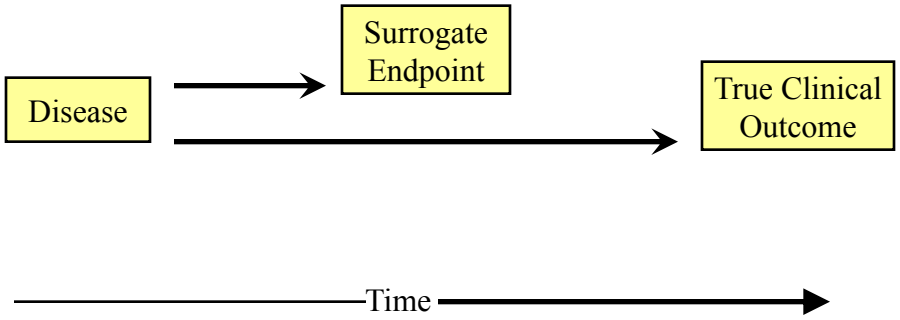
- The effect on the Surrogate Endpoint may overstate its effect on the Clinical Outcome (which may actually be harmful)



102

Scenario 3: Marker

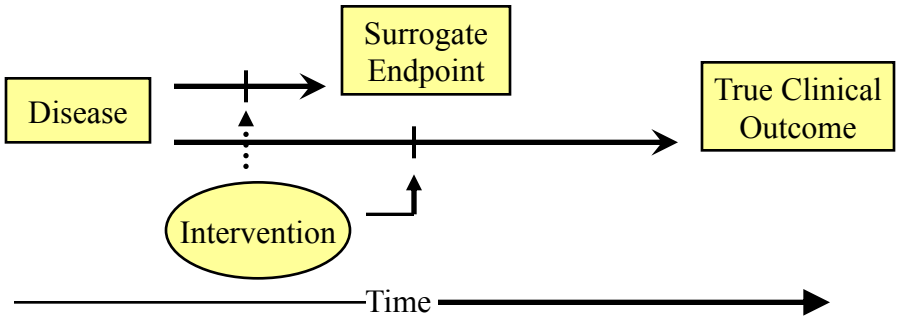
- Disease causes Surrogate Endpoint and Clinical Outcome via different mechanisms



103

Scenario 3b: Inefficient Marker

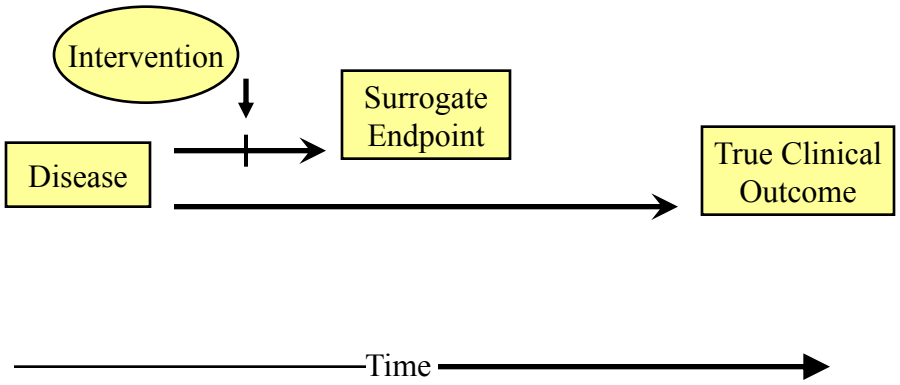
- Treatments' effect on Clinical Outcome is greater than is reflected by Surrogate Endpoint



104

Scenario 3c: Misleading Surrogate

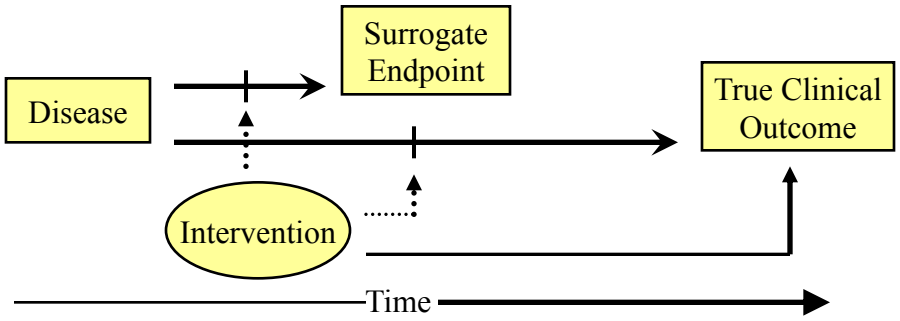
- Effect on Surrogate Endpoint does not reflect lack of effect on Clinical Outcome



105

Scenario 3d: Dangerous Surrogate

- Effect on the Surrogate Endpoint may overstate its effect on the Clinical Outcome (which may actually be harmful)
  - E.g., anti-arrhythmics?, laromustine in AML?



106

Can We Validate a Surrogate Endpoint?

- Is there a way to validate a surrogate endpoint by establishing which causal pathway holds?

- What does not work:
  - It is not sufficient to establish that the surrogate endpoint predicts the clinical outcome in each treatment group separately
  - Treatment can affect the distribution of the surrogate endpoint while increasing mortality in every level

107

Hypothetical Example

- Treatment leads to shift toward lower values on surrogate outcome but higher death rates within each post randomization stratum
- Overall higher death rate on treatment

	Treatment		Control	
Surrogate	n	% die	n	% die
Low	30	50%	10	30%
Medium	40	60%	30	40%
High	30	70%	60	50%
Total	100	60%	100	45%

108

Example: CARET

- Beta-carotene supplementation for prevention of cancer in smokers
- Treatment group had excess cancer incidence and death
- Within each group, subjects having higher beta-carotene levels in their diet had better survival

109

Prentice's Criteria

- A surrogate endpoint must be correlated with the clinical outcome
- A surrogate endpoint must fully capture the net effect of treatment on the clinical outcome: Mediator analysis
  - After adjustment for the surrogate endpoint, there must be no treatment effect on the clinical outcome

110

However…

- The validity of a surrogate endpoint is dependent upon
  - the disease
  - the clinical outcome
  - the treatment
- Thus it is not possible to validate a surrogate endpoint for every combination of treatment and disease without doing a trial looking at the clinical outcome

111

Hence…

- When considering a number of treatments that can be presumed to act in a similar manner, meta-analyses of clinical trial results can sometimes be used to establish the suitability of a surrogate endpoint for other treatments in that class
  - Even then, we must watch for outliers within such a meta-analysis
  - Such outliers suggest that the presumption of similar action is violated

112

Nonetheless…

- Surrogate endpoints have a place in screening trials where the major interest is identifying treatments which have little chance of working
  - In cancer therapy, the premise is often that a treatment that does not lead to less progression cannot lead to better survival
- But for confirmatory trials meant to establish beneficial clinical effects of treatments, use of surrogate endpoints can (AND HAS) led to the introduction of harmful treatments

113

Nonetheless…

- Surrogate endpoints have a place in screening trials designed to identify treatments which have little chance of working
  - In cancer therapy, the premise is often that a treatment that does not lead to less progression cannot lead to better survival
- A very few surrogate endpoints are seeing some wide use
  - Hypertension is now generally regarded as a clinical disease
  - Blood glucose in diabetes (but often need CVOT)
  - Accelerated approval in cancer based on ORR, PFS
    - *JAMA Int Med* 2019: only 19% of accelerated approval later with proved clinical benefit
- But for confirmatory trials meant to establish beneficial clinical effects of treatments, use of surrogate endpoints can (AND HAS) led to the introduction of harmful treatments

114

Materials and Methods

Patient Volunteers

Eligibility Criteria

Where am I going?

Patients are the fundamental "material" used in clinical research

Definition of the patients typically involves lists of
- Inclusion criteria
- Exclusion criteria

Such criteria define the targeted disease / population, as well as
- Enriching the population to best demonstrate efficacy
- Avoiding safety issues related to other medical conditions

115

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

116

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

117

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
  - Exception: might enrich for a patient population especially likely to benefit, if safety of a later expanded indication will be obtained
- 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

118

Documentation

- Precise definition of target patient population is crucial
  - Scientific:
    - Materials and methods of scientific experiment
  - Clincal:
    - Generalization of safety outcomes
    - Generalization of efficacy outcomes

119

Inclusion / Exclusion Criteria

- 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

120

Conceptual Framework

- Population of patients with disease
  - Definition of disease by cause vs signs / symptoms
- Subpopulation with disease targeted by intervention
  - I argue "disease" is really 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

121

Ideal

- 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

122

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
    - Liver, renal, heart disease
    - Elderly

123

Safety Considerations

- Generalizing study results: Efficacy vs effectiveness
  - Treatment may have to be delivered to a population larger than studied
    - Diagnostic procedures after approval may be less rigorous
      - Time requirements: Definition of gram negative sepsis
  - "Diagnostic creep"
    - If some disease has no treatment, then there may be tendency to diagnose a disease that does
      - Gram negative sepsis, non VT/VT cardiac arrest
  - Off-label use

124

Inclusion / Exclusion Criteria

- Inclusion criteria:
  - Definition of ultimate target population
- Exclusion criteria:
  - Exceptions required for clinical trial setting
- Above definitions based on my ideal.
  - In fact, the safety and efficacy of the investigation treatment will only have been established in patients meeting both inclusion and exclusion criteria

125

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

126

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

127

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

128

Exclusion Criteria

- 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

129

Comments re Specification

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

130

Materials and Methods

Randomization / Blinding

Where am I going?

Scientifically, we desire to establish cause and effect relationships

Scientific demands that we ensure that the interventions being compared are not confounded by the many other factors that also influence patient outcomes

- Randomization precludes confounding on average
- Blinding (when possible) ensures subject and investigator prejudice does not bias our conclusions

131

Common Pitfalls of Studies

- Data driven hypotheses
  - Multiple comparisons
  - Over-fitting of data
- Poor selection of subjects, outcomes
- Noncomparability of treatment groups

132

Issues: Bias

- 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 (lack consistency)
- N.B.: The definition of bias is very much dependent upon what we wish we were estimating (the estimand)
  - How are we going to generalize our results?

133

Sources of Bias

- Attributing an observed difference to a particular treatment
  - Disease
    - Misclassification, overly restrictive
  - Patients
    - Insufficiently 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

134

Confounding Bias

- The treatment groups being compared differ with respect to other important (measured or unmeasured) variables that are prognostic for 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

135

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)

136

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

137

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

138

Statistics and Game Theory

- 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

139

Ex: Level 0.05 per Decision

- Experiment-wise Error Rate
-

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

140

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?

141

Statistical Issues

- Need to choose a primary summary measure or multiple comparison issues result

- Example: Type I error with normal data
  - 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

142

Statistical Issues

- Need to choose a primary summary measure or multiple comparison issues result

- Example: Type I error with lognormal data
  - Any single test: 0.050
  - Mean, geometric mean 0.074
  - Mean, Wilcoxon 0.077
  - Mean, geom mean, Wilcoxon 0.082
  - Above plus median 0.107
  - Above plus Pr (Y > 1) 0.152
  - Above plus Pr (Y > 1.645) 0.192

143

Issues: Variability

- Even when unbiased, studies that are conducted with low precision present a problem

- Decreased power leads to decreased positive predictive value of statistically significant results

- The same number of patients spread across multiple small studies increases the number of statistically significant studies
  - 10,000 pts in 10 studies: Expect 0.25 false positive
  - 10,000 pts in 400 studies: Expect 10 false positive

144

Statistical Design Issues

- Variability of measurements decreased by
  - Homogeneity of patient population
  - Precise definition of treatment(s)
  - Appropriate choice of clinical, statistical endpoints
  - High precision in measurements
  - Appropriate sampling strategy
- NB: But first and foremost, the RCT must be relevant

145

Blinding

- In studies with concurrent comparison groups, blinding of treatment assignment can minimize some types of bias
  - Participant or investigator prejudices for or against treatments
  - Some sources of ascertainment bias
- Single blind experiments:
  - Participant is unaware of treatment assignment
- Observer blind experiments
  - Investigator making assessments is unaware of treatment
- Double blind experiments:
  - Neither participant nor provider know treatment assignment

146

Goals

- Minimize "placebo effect": A participant being treated does better than one not being treated, irrespective of the actual treatment
  - This should be distinguished from secular trends in outcome that might happen over time
  - To detect a placebo effect, you compare a group that is unknowingly receiving a placebo to a group that is receiving nothing
- Minimize investigator bias in assessing
  - adverse events
  - treatment outcomes

147

Logistical Issues

- Blinding is not always possible
  - Placebo not always possible to be identical in appearance
    - weight of fiber, hardness of calcium,
  - Side effects of treatment may be noticeable
    - skin discoloration with beta-carotene
  - Burden of treatment may not be ethical
    - surgery, hospitalizations

148

Blinded Evaluation

- When blinding of participants and investigators is not possible, blinded evaluation may be
  - E.g., blinded radiologist, pathologist, clinical assessor
- Must still ensure similar schedule of assessment
  - side effects might lead to more frequent monitoring
- Competing risks (e.g., death from other causes) still a problem

149

Issues

- Issues that must be addressed with blinded experiments
  - Appearance of treatments
  - Dosage, administration schedules
  - Coding and dispensing treatments
  - When and how to unblind
    - Emergent situations
    - Only unblind when treatment of toxicities differs between therapies
  - Assessing how well the blind was maintained

150

When Blinding is Unnecessary

- Blinding less of an issue with harder endpoints
  - The more objective the measurement of outcome, the less important blinding is to the scientific credibility of a clinical trial.
  - (Of course, the ideal is a blinded experiment with solidly objective endpoints.)

151

Cause and Effect

- Necessary conditions for establishing cause and effect of a treatment
  - The treatment should precede the effect
    - Beware protopathic signs
      - Marijuana and risk of MI within 3 hours
  - When comparing groups differing in their treatment, the groups should be comparable in every other way
    - At baseline
    - And in how the exact same presentation will receive future treatment

152

Major Scientific Tool

- Randomization is the major way in which cause and effect is established
  - Ensures comparability of populations
    - Each treatment group drawn from same population
    - Differences in other prognostic factors will only differ by random sampling
      - Provides balance on the total effect of all other prognostic factors
      - May not provide balance on each individual factor
- NB: Sequential allocation of patients is not randomization
  - Possible time trends in recruitment, treatments, etc.

153

Points Meriting Repeated Emphasis

- Randomization is our friend…
  - If we randomize, we do not (on average) need to worry about differences between the treatment groups with respect to factors present at time of randomization
    - Any difference in outcomes can be attributed to treatment
      - Again, recognize that treatment can lead to differential use of other ancillary treatments, however
- But like all friends, we must treat it with respect.
  - We must analyze our data in groups defined at the time of randomization
    - Discarding or missing data on randomized subjects may lead to bias
      - It certainly leads to diminished scientific credibility

154

Points for Further Elucidation

- Confounding not an issue (on average)
  - P value measures probability of observed effects occurring due only to randomization imbalance
- Gain precision if
  - Control important predictors, or
  - Adjust for stratification variables
- Subgroup analyses
  - If effect modification is concern

155

Randomization Strategies

- Complete randomization (CRD)
- Blocked randomization
  - Ensure balance after every *k* patients
  - Ensure closer adherence to randomization ratio
  - Undisclosed block sizes to prevent bias
- Stratified randomization
  - Separately within strata defined by strong risk factors
    - Lessens chance of randomization imbalance
  - Need to consider how many variables can be used
- Dynamic randomization
  - Adaptive randomization to achieve best balance on marginal distribution of covariates

156

Complete Randomization (CRD)

- With each accrued subject a (possibly biased) coin is tossed to determine which arm
  - Probability of treatment arm = *r / (r + 1)*
  - Independence of successive randomizations
- Issues
  - No bias (on average)
  - Face validity if some covariates appear imbalanced
  - Precision lower due to variability in covariate distribution and sample sizes

157

Face Validity: Table 1

	Methotrexate Arm		Placebo Arm	
	n	Mean (SD; Min – Max)	n	Mean (SD; Min – Max)
Age (yrs)	132	50.4 (8.5; 32 - 69)	133	52.2 (8.5; 26 - 67)
Female	132	92.4%	133	92.5%
Pruritus score	116	7.7 (3.8; 4 - 16)	124	6.9 (3.8; 4 - 20)
Splenomegaly	131	8.4%	133	10.5%
Telangiectasia	132	4.6%	133	11.3%
Edema	132	6.1%	133	3.0%
Alkaline phosphatase	132	242.6 (145.9; 53 - 933)	133	245.0 (187.6; 66 - 1130)
ALT	131	54.5 (41.7; 12 - 202)	132	50.6 (41.4; 12 - 311)
Total bilirubin	132	0.7 (0.4; 0.1 - 2.7)	133	0.7 (0.4; 0.1 - 2.4)
Albumin	132	4.0 (0.3; 3.1 - 6.0)	133	4.0 (0.3; 3.0 - 4.8)
Prothrombin time INR	124	1.0 (0.1; 0.7 - 1.3)	132	1.0 (0.1; 0.7 - 1.3)
Mayo score	128	3.8 (0.8; 1.6 - 6.3)	133	3.9 (0.8; 1.6 - 6.1)
Avg stage	128	2.2 (0.9; 1.0 - 4.0)	128	2.3 (0.9; 1.0 - 4.0)
Avg fibrosis	128	1.2 (0.8; 0.0 - 3.0)	128	1.3 (0.9; 0.0 - 3.0)

158

Face Validity

- Spurious results due to covariate imbalance?
  - Unconditionally: Unbiased so no problem
  - Conditional on obtained randomization:
    - IF covariates are strongly predictive of outcome, then covariate imbalance is predictive of type I error
    - But need to consider that combined effect of other measured and unmeasured covariates may provide balance
- Ultimately, however, we need to have credible results
  - We do not always get to choose what others believe

159

Precision

- Impact of completely randomized design on precision of inference
  - Impact of imbalance in sample sizes
    - The number accrued to each arm is random
  - Impact of imbalance in covariates
    - "One statistician's mean is another statistician's variance"

160

Randomization Ratio

- Most efficient
  - When test statistics involve a sum, choose ratio equal to ratio of standard deviations
- Most ethical for patients on study
  - Assign more patients to best treatment
    - Many sponsors / patients presume new treatment
    - (Adaptive randomization: Play the winner)
- Most ethical for general patient population
  - Whatever is most efficient (generally not adaptive)
- Other goals
  - Attaining sufficient patients exposed to new treatment
  - Maintaining DSMB blind

161

CRD: Improved Performance

- If we adjust for important covariates, we will often gain back nearly all the precision lost by imbalance in covariates
  - Face validity in Table 1 if readers recognize that adjustment accounts for any observed imbalance
- Caveats:
  - If covariate imbalance by arm, model misspecification can be an issue re conditional bias
  - If covariate imbalance by arm, lack of effect can be an issue re variance inflation
  - If adjustment not TOTALLY prespecified, “intent to cheat” analysis can be an issue
    - Not too much loss of precision from imperfect model

162

Issues with CRD

- Imbalance across arms in sample sizes
  - Not much of an issue with large sample sizes
  - Could be problematic with sequential sampling
    - Interim analyses of data early in the study
- Imbalance across arms in time trends
  - Outcome may be associated with time of accrual
- Imbalance across arms in distribution of prognostic covariates
  - Loss of face validity
  - Loss of precision
  - May impact important subgroup analyses

163

Mechanisms Leading to Time Trends

- Patients accrued early may differ from those accrued later, because
  - Backlog of eligible patients
  - Startup of new clinical sites
  - Pressure to increase accrual
  - Secular trends in beliefs about intervention
    - (Made much worse if any interim results leak out)
  - Secular trends in diagnostic tools used for eligibility
  - Secular trends in ancillary treatments

164

Blocked Randomization

- Within every sequence of *k* patients, the ratio of treatment to control is exactly *r : 1*
  - Within each "block" ordering of treatments is random
- Important caveats:
  - Investigators must not know block size
    - Otherwise, decisions to enroll patients might be affected by knowledge of next assignment
  - Hence, often use "concealed blocks of varying sizes"

165

Stratified Randomization

- Strata are defined based on values of important covariates
  - E.g., sex, age, disease severity, clinical site
- Within each stratum defined by a unique combination of stratification variables, CRD or blocked randomization
- Important caveats:
  - Number of strata is exponential in number of stratification variables
    - E.g., 4 two level stratification variables ➔ 16 strata
  - Gains in precision depend upon pre-specified adjustment for the stratification variables in the analyses

166

Dynamic Randomization

- Subjects are assigned to the treatment arm that will achieve best balance
- "Minimization": minimize the difference between the distribution of covariate effects between arms
  - Define a "distance" between arms for covariate vectors
  - Probability of assignment depends upon arm that would provide smallest difference
- Advantages / Disadvantages
  - Typically improved face validity (but no guarantees)
  - Can handle an arbitrary number of covariates
  - Logistically more involved
  - Decreased credibility if too deterministic
  - Does not necessarily facilitate subgroup analyses

167

Minimize Subjects on Inferior Arms

- Most commonly used
  - Sequential sampling
    - Interim analyses of data
    - Terminate trials when credible decisions can be made
- Also proposed
  - Response adaptive randomization
    - Change randomization probabilities as evidence accumulates that one treatment might be best
      - "Play the winner"
      - Bayesian response adaptive randomization
    - Complicates standard analyses and may lead to worse ethics re the entire population with disease
    - (Not considered further here)

168

Overall Recommendations

- Blind as much as possible
- Randomize
  - 1:1 randomization unless really good reason
- Stratify on
  - Clinical center if possible
  - No more than 4 important variables (including center)
- Use hidden blocks of varying sizes
  - Block sizes 4, 6, or 8 in 1:1 randomization
- Prespecify an analysis model that includes the covariates that are thought *a priori* to be the most highly prognostic
  - (Less of an issue with logistic regression)

169

Materials and Methods

Treatments / Study Procedures

Where am I going?

In order to be able to generalize study results all treatments and measurement methods need to be well-defined

Treatments must anticipate situations that commonly arise in clinical practice and that might affect formulation, dose, administration, prophylaxis for adverse effects, concomitant therapies, criteria for treatment discontinuation

Study measurements must consider timing, technology

- Tradeoffs between standardization and generalizability

170

Treatment Strategies

- In human experimentation, 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
    - Patients may discontinue a therapy due to headache
    - If forced to continue, those patients may have CVA
- 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

171

Definition of Treatments

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

172

Special Issues

- Ultimately, the scientific credibility of the clinical trial stems from our ability to assign a treatment to the participants
- Ideally we do this in a random fashion
- At a given point in time, we can only assign a strategy
  - Competing risks may make further treatment impossible
  - Intervening events may change indications
  - Informed consent can be withdrawn
- We must avoid ruining the comparisons of strategies
  - Naïve attempts to compare "treatment" may ruin our ability to assess what really can be tested

173

Ramifications

- Multiple actions are possible after progression
- Stay the course
  - "Progression" dichotomizes a continuous process
  - Treatment may be delaying that process
- Advance to other therapies
  - Ideally the same for both treatment arms
- Cross-over to other arm
  - Sometimes motivated to increase sample treated
  - A huge scientific mistake but
    - Ethics sometimes demands it
      - PA catheterization vs central line
      - Pemetrexed vs docetaxel

174

Can There Be Noncompliance?

- Experimentally: NO
  - By definition, all patients are following intent to treat
    - Clearly addresses effectiveness questions
    - If efficacy had been our goal:
      - Exclude noncompliant patients as much as possible
      - Increase sample size to deal with attenuation
- Safety: MAYBE
  - We do have to worry that adherence to treatment strategy may change after reporting efficacy
    - We will only have tested safety under the compliance actually achieved
  - Measuring compliance is important for interpretation

175

Ramifications

- Important distinctions need to be made
- "Stopping study drug"
  - This may happen due to
    - Adverse events
    - Progression
    - Study burden
  - While we hope for high compliance
    - Badgering patients to remain on therapy can lead to worse adverse events or the quitting the study
  - In the event of stopping study drug, all follow-up of primary outcomes should proceed as planned
- "Withdrawing consent"
  - No further data will be available

176

What is our Target?: Estimands

- Clinical issues:
  - The indication: Disease, Population, Treatment, Outcome
  - Clinical importance of distributional summary measure
  - Clinical importance of stratification
- Regulatory issues:
  - Efficacy, safety, effectiveness over time and populations
- Scientific issues:
  - Avoid bias (through randomization and blinding)
- Statistical issues:
  - Summarizing the outcome distribution
    - Mean, geom mean, median, proportion, odds hazard
  - Covariate adjustment (precision)
  - Per randomization analyses

177

Scientific Efficacy / Safety Estimands

- What is impact among patients who follow protocol?
  - No matter what: An interesting basic science question
  - Clinically may be used to explore mechanism of action
- Patients who don't follow protocol may be irrelevant for chronic tx
  - Patients who do not follow directions
  - Patients who have intolerable adverse reactions
    - Perhaps "intolerable" only because uncertain of efficacy, or
    - Perhaps leading to serious consequences with continued therapy
  - Patients with real or perceived lack of efficacy
    - Early clinical course is discouraging, or
    - Definitive progression to serious condition prior to primary endpoint
  - Development of contraindication to treatment (e.g., pregnancy)
  - Patients with early evidence of cure
  - Patients with competing risk that prevents measurement

178

Impact of Estimand on RCT

- The patients who are "relevant" differ according to the estimand of interest
- The primary goal should be to devise an experiment that only randomizes patients who are relevant to the estimand
  - This is often difficult
  - It may mean using more than one RCT, answering different aspects of the safety/effectiveness profile in different studies
    - Randomize all patients to assess short term safety
    - Randomized withdrawal among tolerators, responders, compliers to assess safety and efficacy with long term use
- Sometimes, however, a counterfactual estimand is of greatest scientific interest
  - In these cases, all results are subjective

179

Materials and Methods

Statistical Precision

Where am I going?

We design a clinical trial to discriminate between two plausible, important hypotheses.

RCT design involves maximizing statistical precision through appropriate use of

- Trial structure
- Statistical analysis model
- Sample size

180

Statistical Planning

- Satisfy collaborators as much as possible
  - Discriminate between relevant scientific hypotheses
    - Scientific and statistical credibility
  - Protect economic interests of sponsor
    - Efficient designs
    - Economically important estimates
  - Protect interests of patients on trial
    - Stop if unsafe or unethical
    - Stop when credible decision can be made
  - Promote rapid discovery of new beneficial treatments

181

Sample Size Calculation

- Traditional approach
  - Sample size to provide high power to "detect" a particular alternative that is
    - Clinically important improvement
    - Plausible
- Decision theoretic approach
  - Sample size to discriminate between hypotheses
    - "Discriminate" based on interval estimate
    - Standard for interval estimate: 95%
      - Equivalent to traditional approach with 97.5% power

182

Issues

- Summary measure
  - Mean, geometric mean, median, proportion, hazard…
- Structure of trial
  - One arm, two arms, k arms
  - Independent groups vs cross over
  - Cluster vs individual randomization
  - Randomization ratio
- Statistic
  - Parametric, semi-parametric, nonparametric
  - Adjustment for covariates

183

Refining Scientific Hypotheses

- Scientific hypotheses are typically refined into statistical hypotheses by identifying some parameter θ measuring difference in distribution of response
  - Difference/ratio of means
  - Ratio of geometric means
  - Difference/ratio of medians
  - Difference/ratio of proportions
  - Odds ratio
  - Hazard ratio

184

Inference

- Generalizations from sample to population
- Estimation
  - Point estimates
  - Interval estimates
- Decision analysis (testing)
  - Quantifying strength of evidence

185

Measures of Precision

- Estimators are less variable across studies
  - Standard errors are smaller
- Estimators typical of fewer hypotheses
  - Confidence intervals are narrower
- Able to statistically reject false hypotheses
  - Z statistic is higher under alternatives

186

Statistics to Address Variability

- At the end of the study we perform frequentist and/or Bayesian data analysis to assess the credibility of clinical trial results

  (Frequentist statistic)

  - Estimate of the treatment effect
    - Single best estimate (consistent or unbiased)
    - Precision of estimates (confidence interval)
  - Decision for or against hypotheses
    - Binary decision (whether to reject null)
    - Quantification of strength of evidence (p value)

187

Frequentist Criteria for Precision

- Standard error
- Width of confidence interval
- Statistical power: Probability of rejecting the null hypothesis
  - Select level of significance
    - Standard: One-sided 0.025; two-sided 0.05
    - Pivotal: One-sided 0.005; two-sided 0.01
  - Select "design alternative"
    - Minimal clinically important difference
      - To detect versus declaring significant
      - May consider what is feasible
    - Minimal plausible difference
  - Select desired power
    - High power for a decision theoretic approach

188

Sample Size Determination

- Based on sampling plan, statistical analysis plan, and estimates of variability, compute
  - Sample size that discriminates hypotheses with desired power,

    OR
  - Hypothesis that is discriminated from null with desired power when sample size is as specified, or

    OR
  - Power to detect the specific alternative when sample size is as specified

189

Ideal Results

- Goals of "drug discovery" are similar to those of diagnostic testing in clinical medicine
- We want a "drug discovery" process in which there is
  - A low probability of adopting ineffective drugs
    - High specificity (low type I error)
  - A high probability of adopting truly effective drugs
    - High sensitivity (low type II error; high power)
  - A high probability that adopted drugs are truly effective
    - High positive predictive value
    - Will depend on prevalence of "good ideas" among our ideas

190

My Perspective

- Public Health is primary goal
  - Improve average health of population
  - Patients willing to participate in RCT are a public resource
- Must be aware of economic reality
  - Most “drug discovery” RCT funded by industry
  - Government funded research for orphan or high risk areas
- Must be aware of conflicts of interest
  - Academic researchers’ “pet” hypotheses that might bring fame
  - Industry with proprietary drugs that might bring wealth
  - Political / public pressure on Government funders and regulators

191

Design: Distinctions without Differences

- There is no such thing as a “Bayesian design”
- Every RCT design has a Bayesian interpretation
  - (And each person may have a different such interpretation)
- Every RCT design has a frequentist interpretation
  - (In poorly designed trials, this may not be known exactly)
- In this talk I focus on the use of both interpretations
  - Phase 2: Bayesian probability space
  - Phase 3: Frequentist probability space
  - Entire process: Both Bayesian and frequentist optimality criteria

192

Application to Drug Discovery

- We consider a population of candidate drugs
- We use RCT to "diagnose" truly beneficial drugs
- Use both frequentist and Bayesian optimality criteria
  - Sponsor:
    - High probability of adopting a beneficial drug (frequentist power)
  - Regulatory:
    - Low probability of adopting ineffective drug (freq type 1 error)
    - High probability that adopted drugs work (posterior probability)
  - Public Health (frequentist sample space, Bayes criteria)
    - Maximize the number of good drugs adopted
    - Minimize the number of ineffective drugs adopted

193

Frequentist vs Bayesian: Bayes Factor

- Frequentist and Bayesian inference truly complementary
  - Frequentist: Design so the same data not likely from null / alt
  - Bayesian: Explore updated beliefs based on a range of priors
- Bayes rule tells us that we can parameterize the positive predictive value by the type I error and prevalence
  - Maximize new information by maximizing Bayes factor
  - With simple hypotheses:

$$PPV = \frac{power \times prevalence}{power \times prevalence \quad + \quad type\ I\ err \times (1 - prevalence)}$$

$$\frac{PPV}{1-PPV} = \frac{\boldsymbol{power}}{\boldsymbol{type\ I\ err}} \times \frac{prevalence}{1-prevalence}$$

194

$$posterior\ \ odds = \boldsymbol{Bayes\ Factor} \times prior\ \ odds$$

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

195

Preliminary Studies in Screening

- Two general approaches to studying new treatments
- Scenario 1:
  - Study every treatment in a large definitive experiment
    - Only do Phase III studies
      - Level of significance 0.025, high power
    - (Ignore, for now, the safety / ethics of this)
- Scenario 2:
  - Perform small screening trials, with confirmatory trials of promising treatments passing early tests
    - Phase II studies
      - Level of significance, power (sample size) to be determined
    - Confirmatory
      - Level of significance 0.025, high power

196

Apples with Apples

- To explore impact of different approached, we consider
  - Only large trials using 1,000,000 subjects
  - 10% of drugs being investigated at start of phase 2 truly work
  - 500 patients provide 80% power in a fixed sample study
  - 1,000,000 patients are willing to be randomized
- We evaluate
  - Number of drugs investigated in Phase 2
  - Number of drugs investigated in Phase 3
  - Number of drugs adopted after Phase 3
    - Truly effective drugs adopted (work for at least some)
    - Truly ineffective drugs adopted
  - Available information for safety on adopted drugs

197

Summary: Screening Trials

		Scenario 1	Scenario 2a	Scenario 2b
Phase 2	Number RCT	2,000 (10% eff)	7,000 (10% eff)	2,047 (10% eff)
	N per RCT	0	100	342
	Type 1 err; Pwr		0.025; 24%	0.100; 85%
	"Positive" RCT		168 eff; 158 not	173 eff; 184 not
Confirmtaory Phase 3	Number RCT	2,000 (10% eff)	326 (52% eff)	357 (49% eff)
	N per RCT	500	921	839
	Type 1 err, Pwr	0.025; 80%	0.025; 97%	0.025; 95%
	# Effctve Adopt	**160**	**162**	**165**
	# Ineff Adopt	**45**	**4**	**5**
	Pred Val Pos	78%	98%	97%
	N per Adopt	500	1,021	1,181

Screening Phase II: Bottom Line

- Pilot studies increase the predictive value of a positive study while using the same number of subjects.
  - Screening parameters can be optimized based on prevalence
    - Proportion of subjects in Phase II vs Phase III
    - Type I error at Phase II
    - Power at Phase II
- Additional considerations when choosing among screening parameters
  - Will we have same prevalence of "good" ideas when we screen 2,000 drugs vs 7,000 drugs?
  - Holding predictive value of positive constant, which strategy provides more information about safety and secondary endpoints for the treatments eventually adopted?

199

Burden of Larger Phase II Studies?

- It appears to be advantageous to use larger Phase 2 studies than is typical currently in cancer research
- BUT: Ethical and efficiency concerns can be addressed through sequential sampling
  - During the conduct of the study, data are analyzed at periodic intervals and reviewed by the DMC
  - Using interim estimates of treatment effect decide whether to continue the trial
  - If continuing, decide on any modifications to
    - scientific / statistical hypotheses and/or
    - sampling scheme

200

General Classification of Approaches

- What aspects of the RCT are modified?
  - *Statistical*: Modify only the sample size to be accrued
  - *Scientific*: Possibly modify the hypotheses related to patient population, treatment, outcomes
- Are all planned modifications described at design?
  - *"Prespecified adaptive rules":* Investigators describe
    - Conditions under which trial will be modified and
    - What those modification will consist of
  - *"Fully adaptive":* At each analysis, investigators are free to use current data to modify future conduct of the study

201

Statistical Design Issues

- Under what conditions should we use fewer subjects?
  - Ethical treatment of patients
  - Efficient use of resources (time, money, patients)
  - Scientifically meaningful results
  - Statistically credible results
  - Minimal number of subjects for regulatory agencies
- How do we control false positive rate?
  - Repeated analysis of accruing data involves multiple comparisons

202

Potential Benefits of Stopping Rules

- Group sequential sampling
  - Aggressive early stopping for futility: Pocock boundaries
    - Greatest efficiency (or nearly so)
  - Conservative early stopping for efficacy: O'Brien-Fleming
    - Burden of proof, other endpoints
- Type I error, power maintained exactly at each phase
  - Worst case maximum sample size increases
- Average sample size requirements assuming 10% truly effective drugs at start of Phase II
  - Only large studies : 58.5% of fixed sample
  - Pilot scenario 2a : 56.0%
  - Pilot scenario 2b : 61.0%

203

What Can Go Wrong?

- We ignore bias / lack of precision and believe our phase 2 results
  - Phase 3 study will be poorly designed
- We fish through phase 2 data to change outcome
  - We might inflate type 1 error without sufficient increase in power
- We use a surrogate at phase 2 that is not sensitive/specific for true clinical outcome used at phase 3
  - We inflate "effective" type 1 error without increasing power
- We fish through phase 2 data to change eligibility criteria or dose
  - We might inflate type 1 error without sufficient increase in power

204

Phase 3 Sample Size from Phase 2

- Phase 2: type 1 error 0.10 with N= 342 provides 85% power
  - Estimated treatment effects: mn (sd; min – max)
    - Ineffective drugs 0.095 (0.022; 0.069 - 0.313)
    - Effective drugs 0.140 (0.043; 0.069 - 0.396)
- Phase 3 using Phase 2 results
  - Estimated N for 95% power:
    - Ineffective drugs 1665 (610; 134 – 2745)
    - Effective drugs 893 (571; 84 – 2745)
  - Type 1 error 0.025, avg power 86%
- Screen 1,759 drugs with 1,000,000 patients
  - End of phase 2: 150 effective, 159 ineffective (PVP= 0.49)
  - End of phase 3: 129 effective, 4 ineffective (PVP= 0.97)

205

Scientific Adaptation

- We also use phase 2 data to “refine” our indication
  - Revising outcomes to reflect the most promising results
    - Statistical summary measure vs clinical endpoint
  - Revising eligibility criteria based on subgroup analyses
  - Changing from surrogate efficacy to effectiveness endpoints
    - “Treating the symptom not the disease”
- But, we must be concerned about data dredging (“data mining”) leading to inflation of type 1 error
  - Easily triple or quadruple the type 1 error
  - Cannot increase the power by as much
  - Lower Bayes Factor ➔ lower positive predictive value

206

Examples without Error Control

- Consideration of multiple summary measures
  - Mean, geometric mean, Wilcoxon, median, two proportions
  - Type 1 error 0.10 ➔ 0.23
  - Power 0.85 ➔ 0.92
- Consideration of subgroups
  - Overall sample and equal subgroups defined by three variables
  - Type 1 error 0.10 ➔ 0.33
  - Power 0.85 ➔ 0.95
- Consideration of change of endpoint between phase 2 and 3
  - Phase 2: potential surrogate and Phase 3: clinical outcome
  - Type 1 error 0.10 ➔ 0.19 (10%) ➔ 0.27 (20% misleading)
  - Power 0.85 ➔ 0.85

207

Control Error (Inhomogeneous Effects)

- With inhomogeneous effects across subgroups, we also need to consider additional errors
- A “True Positive” would be adoption of a new treatment in exactly the population that benefits
- “False Positives” might include drugs with too broad an indication
  - It does not work in part of the population
- “False Negatives” might include a drug that has omitted part of the population that would truly benefit

208

Comparisons

	RCT	Eff (TP)	Ptl FN	FP
Nonadaptive				
• Homogeneous effect	2,040	165(165)	0	5
• Homogeneous,10% misleading	1,812	147(147)	0	8
• Homogeneous,20% misleading	1,627	132(132)	0	12
• Inhomogeneous effect	2,130	97(0)	0	5
Adaptive subgroups: inflate error				
• Homogeneous effect	1,488	134(43)	91	11
• Inhomogeneous effect	1,493	122(88)	34	11
Adaptive subgroups: control error				
• Homogeneous effect	2,081	153(54)	99	5
• Inhomogeneous effect	2,110	132(99)	25	5

209

Comments

- Screening Phase II trials provide great protection
  - Ensure overwhelming majority of adopted drugs are truly effective
- Control of type I and II errors are of great importance at phase 2
  - But note that type 1 error of 0.025 not necessarily indicated
- Adaptive designs can help provide that control
  - But need to re-power the study to get greatest benefit
  - The added benefit over nonadaptive designs is not huge, but
    - Higher power and predictive value of the positive
    - More beneficial drugs identified with more safety data
  - It is not clear how far current practice is from what will be attained with wider use of formal adaptation at phase 2

210

Confirmatory Trials

- Adaptive trials cannot replace the need for confirmatory trials
- We need estimates of treatment effect independent of adaptation
  - Sampling distribution in adaptive trials depends on treatment effect in discarded doses, subgroups, etc.
- Regulatory agencies (and scientific community) need to understand "Intent to Cheat" analyses
  - It is not sufficient to understand how designs work ideally
  - Must also understand how biases can creep in

211

Seamless Phase 2 / 3

- In cases that no changes will be made between Phase 2 and Phase 3, can try to use same trial
  - Need to ensure that same level of evidence is provided as would be in two independent trials
    - Pivotal 0.005 vs 0.000625 in two independent trials?
    - One RCT setting vs two RCT settings (random effects)
- Such would eliminate "white space"
  - Nothing presented here specific to separate Phase 2 / Phase 3
  - But note that white space is truly an issue for those whose focus is on a particular agent
  - During "white space" other agents in the pipeline can be investigated
  - Eliminating "white space" limits scientific, regulatory, and ethical review of phase 2 results

212

Major Conclusions

- There is no substitute for planning a study in advance
  - At Phase 2, adaptive designs may be useful to better control parameters leading to Phase 3
    - Most importantly, learn to take "NO" for an answer
  - At Phase 3, there seems little to be gained from adaptive trials
    - We need to be able to do inference, and poorly designed adaptive trials can lead to some very perplexing estimation methods
- **"Opportunity is missed by most people because it is dressed in overalls and looks like work."** -- Thomas Edison
- In clinical science, it is the steady, incremental steps that are likely to have the greatest impact.

213

Really Bottom Line

"You better think (think)
about what you're
trying to do…"

-Aretha Franklin, "Think"

214

Materials and Methods

Data Collection / Management / Analysis

Where am I going?

Unless diligent attention is paid to the quality of the RCT data, no useful information will be obtained, and the RCT cannot possibly be an ethical experiment

Data collection must be timely and accurate.

- Missing data must be avoided as much as possible and handled appropriately when it cannot be avoided

Accruing data should be monitored for data quality and patient safety

Data analysis plans must be totally pre-specified

215

Data Collection

"The government are very keen on amassing statistics. They collect them, add them, raise them to the nth power, take the cube root and prepare wonderful diagrams. But you must never forget that every one of these figures comes in the first instance from the *chowky dar* (village watchman in India), who just puts down what he damn pleases."

- Sir Josiah Stamp

216

Ultimate Goal

- Reporting a scientific experiment
  - Overall goal
  - Specific aims
  - Materials and Methods
  - Results
  - Conclusions

217

Ultimate Goal

- Reporting a scientific experiment
  - Overall goal
  - Specific aims
  - Materials and Methods
    - Patients, dosing, adherence to monitoring
  - Results
    - Disposition, compliance, adverse events, outcomes
  - Conclusions

218

Overall Goal / Specific Aims

- Role of data
  - Goal / aims ideally determined prior to start of study
  - BUT, the question actually answered is specific to
    - the subjects actually sampled
    - the methods actually used
    - the data actually gathered
    - the analysis actually performed
  - Generalization of results depends on all of the above

219

Materials

- Role of data
  - Eligibility criteria are usually broad
  - Need to describe the population actually sampled
  - Need to describe how the sample might differ from the ultimate target population

220

Conceptual Framework

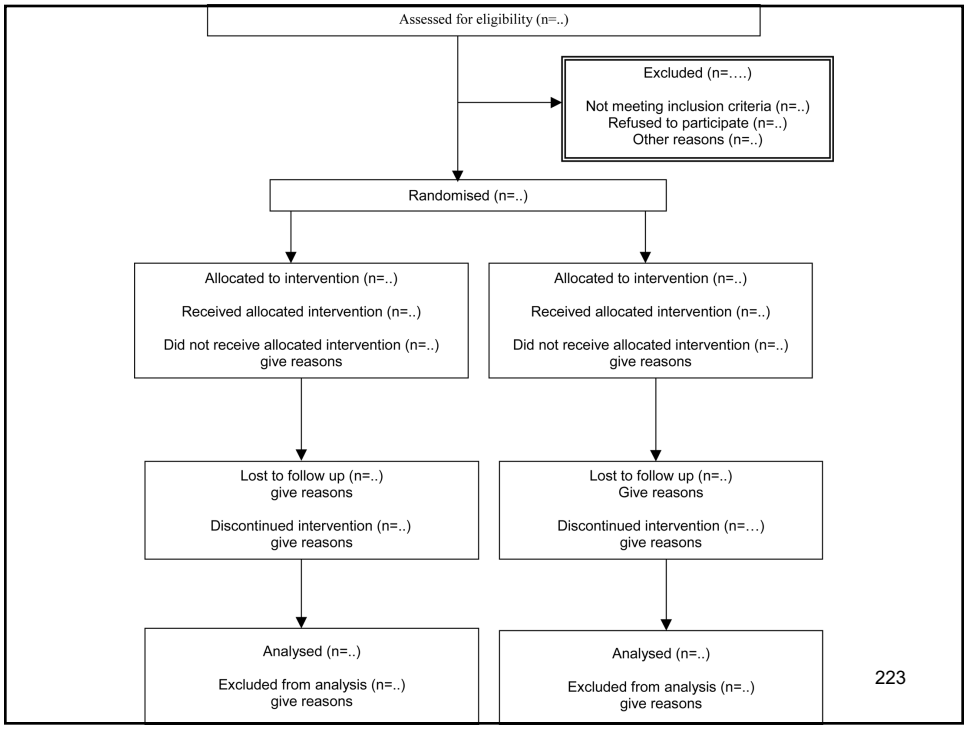
- Population of patients with disease
  - Definition of disease by cause vs signs / symptoms
- Subpopulation with disease targeted by intervention
  - I argue "disease" is really defined by treatment
- Subpopulation eligible for study accrual
  - Restricted due to general clinical trial setting
- Eligible patients from which sampled
  - Restricted due to specific clinical trial (location, time)
- Study sample
  - Restricted due to willingness to participate
- Analysis sample
  - Data collection

221

Generalizability

- CONSORT: Consolidated Standards of Reporting Trials
  - Evidence based, minimum standards
  - Report flow of patients from screening to collection of primary outcomes
    - Screened
    - Enrolled
    - Randomized
    - Completed

222



223

Initial Screening Data

- Source of screened patients
- Number screened
- Characteristics (may require consent)
  - Demographics
  - Disease characteristics
- Reasons for ineligibility
  - Inclusion criteria
  - Exclusion criteria
  - No participation
    - Unable to contact
    - Refused participation

224

Screening Visit(s) Data

- Consent for screening
- Contact information: Name, address, alternative contacts…
- Demographics: Sex, age, race, ethnicity…
- Disease characteristics: Duration, severity, …
- Prior and ongoing treatments
- Eligibility data
  - Inclusion criteria
  - Exclusion criteria
- Consent for randomization

225

Baseline Visit(s) Data

- Baseline data
  - Characterize patients
    - Severity of disease, concomitant disease…
  - Baseline measures of outcomes
    - Concomitant medications
    - Adverse events
    - Efficacy outcomes
      - E.g., initial SBP for reduction of HTN
      - E.g., tumor size for progression
- Note differing detail needed for screening vs baseline

226

Run-in Data

- Some clinical trials involve a run-in
  - Placebo: All patients take placebo
    - Washout vs assessing compliance
    - Patients may be blinded to existence of run-in
  - Active: All patients take experimental therapy
    - Allows randomized comparison of efficacy in patients actually taking drug
      - Randomized withdrawal of drug (among "responders"?)
      - Usually patients aware of run-in
    - Assess tolerability for AEs
    - Assess compliance

227

Randomization Data

- Documentation of eligibility
- Informed consent
- Stratification variables
- Variables needed for determination of dosing
  - Weight, BSA, renal function, severity of disease…
- Time, date of randomization
- Documentation of assigned group (blinded)
  - Cluster?
- Receipt of first treatment: time, date

228

Treatment Data: Why

- Intention to treat analysis is the standard for efficacy
  - Patients are analyzed in assigned group irrespective of their compliance
  - Compliance data is an outcome
    - Assess possible AEs
    - Assess possible mechanism for lack of effect
    - Describe realized exposure to treatment
    - Exploratory analyses for dose / response?
- Safety analyses are typically analyzed according to drug exposure
  - AEs / SAEs occurring within 28 days (?) of last dose

229

Treatment Data: What

- Initial assignment
  - Dose, administration, frequency, duration, ancillary treatments
- Protocol specified modifications
  - Dose reduction / escalation / holidays
    - Date, time
    - Reasons for change (AE, efficacy or lack of efficacy)
- Patient compliance
  - Dose, frequency, duration
  - Intermittent vs permanent change

230

Treatment Data: How

- Protocol specified modifications
  - Regularly scheduled visits
  - Interim visits
- Patient compliance
  - Patient diaries
  - Pill counts
  - Clocks on container lids
  - Biochemical measures: blood, biopsies

231

Patient Monitoring Data: Safety

- Protocol defined safety endpoints
  - Clinical events, subclinical laboratory measurements
- Adverse events (usually open-ended questions)
  - Review of interim AEs at regular visits
  - Undesirable clinical events that occur during the study
    - Treatment emergent: new or exacerbated
    - Classification (e.g. MEDRA), grade of severity
  - Treatment relatedness (but do not necessarily believe)
    - (if truly related, then an “Adverse Reaction”)
- Serious adverse events
  - Fatal, life-threatening, hospitalization or prolongation, birth defects
  - Expedited reporting if unexpected (SUSARs)

232

Patient Monitoring Data: Efficacy

- Protocol defined efficacy endpoints
  - Clinical events
    - Create patient symptoms
  - Quality of life
    - Patient reported outcomes (PRO)
  - Subclinical events
    - Signs thought to be indicative of clinical risk
    - Protocol specified monitoring schedules of
      - Patient performance (FEV, 6 minute walk, etc.)
      - Blood
      - Tissue biopsies
      - Radiology

233

Missing Data

- Ideal:

"Just say no."

(Nancy Reagan)

- Real life:

"Missing data happens"

(Bumper Sticker-
rough translation)

234

Types of Missing Data

- Ignorable
  - We can safely throw out the cases with missing data without biasing our results
- Nonignorable
  - Omitting cases with missing data leads to erroneous conclusions

235

Solutions?

"If certain girls don't look at you
It means that they like you a lot
If other girls don't look at you
It just means they're ignoring you
How can you know, how can you know?
Which is which, who's doing what?
I guess that you can ask 'em
Which one are you baby?
Do you like me or are you ignoring me?"

Dan Bern, *"Tiger Woods"*

236

Sad Facts of Life

"Bloodsuckers hide beneath my bed"

"Eyepennies", Mark Linkous (Sparklehorse)

- Typically, nothing in your data can tell you whether missing data is ignorable or nonignorable
  - You just have to deal with what you worry about
  - At the time of study design, plans should be made
    - Sensitivity analyses?
      - Worst case for new treatment, best for control; vice versa
    - Imputation?
    - Ignore?

237

Missing Data: Efficacy

- Missing data is a pervasive problem in RCT
- Lack of training: Patients, investigators
  - Off study drug
  - Decline invasive procedures
  - Withdraw consent
- Poor endpoint definition
  - All randomized patients must have defined outcome
    - E.g, Quality of life after death, GFR in dialysis, symptom relief with noncompliance
- Sloppy conduct of RCT
  - Excessive loss of follow-up

238

Missing Data: Solutions

- Prevention of missing data is the only sure method
- Increasing enrollment to “replace” missing data just attains greater precision on a biased outcome
- Methods of modeling missing data are based on untestable assumptions
  - There is nothing in your data that can prove your assumptions are appropriate

239

Patient Monitoring Data: Compliance

- Patient adherence to measurement of outcomes
  - Clinic visits
    - Timing relative to window
  - Outcome assessments
    - Efficacy
      - Blood, tissue samples; radiology, special exams
      - Withdrawn consent for invasive procedures?
    - Adverse events
      - Periodic reports per protocol
      - Capture of interim SAEs

240

Patient Monitoring Data: Logistics

- Patient change of address
  - (sometimes schedule phone visits to maintain contact)
- Site compliance with timeliness completeness

241

End of Study

- Reason for stopping study
  - Completion per protocol
    - May be off study drug but still followed
    - Death
    - Withdrawn consent
      - Reasons
- Permission for further follow-up
  - Change of address
- Conjectured treatment assignment?

242

Sources of Data

- Subject self report
- Proxy for subject
- Clinic staff and study records
  - Standard medical care
  - Protocol specified procedures
- Medical records
- Laboratory, radiology, pathology
  - Local vs central labs
- Adjudication panels
- Public health records
  - Registries
  - National Death Index

243

Data Collection Issues

- Timeliness
  - Data collection and storage
- Completeness
  - Missing data
- Accuracy
  - Measurement methods, adjudication panels
- Precision
  - Measurement methods, adjudication panels

244

Data Collection Methods

- Forms
  - Abstracted from medical records
    - Indication bias
  - Completed by subject
  - Completed by proxy
  - Administered by study personnel
  - Completed by clinic staff, study personnel
  - Completed by adjudication panels
- Data files
  - E.g., laboratory, Medicare, National Death Index

245

Data Collection

- Development of forms
  - Administrative information
    - For follow-up, etc.
    - Often text
  - Scientific information
    - Needs to be appropriate for statistical analysis
      - Free text is difficult to analyze
      - Coding of response by person closest to the source

246

Data Collection

- Development of forms (cont.)
  - Format of forms should facilitate
    - Completion of form
      - Brief as possible
      - Make sure no portions overlooked
        - » “skip patterns”, two columns, back of page
      - Cover all cases (explicit “does not apply”)
    - Data entry
      - Coding on form

247

Data Collection

- Issues in form development
  - Number of distinct forms
  - Guidance to the subject, clinic staff on form
    - Study specific definitions
    - Indications for study procedures, other forms
    - Convenience versus increased length
  - Manual and training for form completion
  - Forms for subject vs proxy vs administered
  - Translations
  - Pretesting: subject, staff, investigators, statistician
  - Mapping between different versions over time

248

Data Management

- Planning for data management
  - Data to be collected: What? Why?
  - Methods of collection: Who? Where? When? How?
    - Forms development
  - Methods for data storage
    - Development of database
      - Administrative data: often dynamic
      - Scientific data: usually static
  - Methods for data entry
    - Distributed versus central

249

Data Management

- Handling of data
  - Collection
  - Data entry
  - Storage of forms, primary records
  - Data verification
  - Checking for errors
  - Data reporting
  - Data analysis
  - Final database

250

Data Management: Entry

- Data entry
  - Transcription of data from forms into computerized data base
    - Personnel often low-level clerical staff
    - Little scientific knowledge
  - Electronic data entry by research nurse may minimize data entry errors
- Computerized data checks
  - Screen for impossible values
  - Screen for inconsistencies within form
  - Double entry
- In any case, need to maintain source documents for verification

251

Data Management

- Storage of forms, primary records
  - Subject confidentiality is a major concern
  - Must ensure limited access to confidential information

252

Data Management

- Data verification and checking for errors
  - Data entry errors
  - Data collection errors
    - Audit clinics
    - Compare study data to medical records
  - Maintaining an audit trail
    - Changing database versus making corrections in separate files

253

Data Management

- Data reporting
  - Administrative analyses
    - Accrual rates
    - Timeliness of data collection
    - Completeness of data collection
  - Baseline characteristics
  - Event rates (combined treatment groups only)

254

Data Management

- Data analysis
  - The ultimate purpose of collecting the data
  - Much easier, more generalizable if all the previous stages conducted properly
  - Complete record of all analyses should be maintained
    - date of analysis
    - version of data base and software

255

Data Management

- Bottom line
  - The key qualification you should look for in the data management personnel is attention to detail

256

Monitoring the Trial

Data Monitoring Committees

Where am I going?

- In human experimentation, it is important that the trial be monitored for patient safety
- An independent DMC (DSMB) protects patient safety while avoiding investigator bias

257

Purpose

- During the course of a clinical trial, the accruing data is typically monitored
  - Need to ensure that ethical issues of human experimentation are addressed for
    - subjects on the trial
    - larger population of diseased patients
  - Need to ensure that the trial is conducted in a rigorously scientific manner

258

Issues in Monitoring

- Interpretation of interim results
  - unclean data
  - multiple comparison issues
  - effect of stopping rules on statistical inference
- Potential effect that early release of results can have on
  - continued accrual to the study
  - unblinding of the study

259

Issues

- Potential problems if interim monitoring of data (and decisions to terminate the trial) is performed by investigators
  - Enthusiasm of investigators may lead to overlooking safety issues that arise
  - Interim monitoring may compromise the blinding of study results
  - Interim monitoring may cause inappropriate changes to the study plan (early stopping)
    - Regulatory environment

260

DMC (DSMB)

- Data Monitoring Committee (Data and Safety Monitoring Board)
  - Panel of independent scientists who review the interim data
  - Advisory to the Steering Committee for the clinical trial
    - Make recommendations re
      - Modifying informed consent
      - Modifying protocol
      - Early termination of the study

261

Organization: Charge

- Charge of the Data Monitoring Committee
  - Single clinical trial
  - Multiple related clinical trials
    - by disease classification (e.g., cooperative groups)
    - by treatment (e.g., pharmaceutical industry)

262

Organization: Makeup

- Include experts in some or all of
  - Disease being studied
  - Investigational treatment
  - Measurement of primary outcome
  - Anticipated adverse events
  - Statistics / clinical trials
  - Ethics
- May include patient advocates

263

Organization: Makeup

- Do not include
  - Individuals with financial conflict of interest
    - employees of sponsor
    - stockholders
    - consultants to sponsor
  - Individuals with scientific conflict of interest
    - (Drive for fame can be just as corrupting as a drive for money)
  - Regulators (their role comes later)

264

Organization: Planning

- DMC members identified
  - Verify no conflict of interest
- Establish DMC charter
  - Charge to the DMC
  - Schedule / structure of meetings
  - Organizational chart
- DMC review of protocol
  - DMC role versus IRB role
  - Understand data available for monitoring
  - Understand data management

265

Organizational Meeting

- First DMC meeting
  - Finalize charter
  - Finalize meeting structure
    - open and closed sessions
  - Discuss data flow, timeframe for data reports
    - locking of database, analysis, reading
  - Data and format to be in interim reports
  - Stopping guidelines
    - Understanding intent of sponsor
    - Need to define stopping rule

266

Meeting Structure

- Open session
  - DMC, sponsor, data base management
  - Blinded reports
  - Typically no statistics by treatment group
- Closed session
  - DMC and unblinded data analysts
  - Statistics by treatment group
- Report to Steering Committee representative

267

Data Issues

- Need timely, accurate data
  - Ongoing data collection, QA monitoring
  - Locking of data base
  - Cleaning and analysis of data (1-2 months)
  - Report to DMC (1 week)
  - DMC meeting
- Tradeoff between current data and fully verified data

268

Reports and Tables

- Interim analysis reports and tables
  - Open (not by treatment) and closed (by treatment)
    - Accrual and retention status (by center)
    - Adherence to schedule of visits and data completion status (by center)
    - Missing data, errors in data
    - Adverse events
      - aggregate
      - by patient
    - Primary and secondary endpoints
      - stratified by important subgroups

269

DSMB Blinding

- Blinding of DMC to treatment group
  - As a general rule, the burden of proof is not the same for the two treatment arms
  - Typically, the same magnitude of difference between the arms should not always lead to symmetric actions
  - The DMC should have full access to the treatment assignment
    - (but I like to delay looking at it as long as possible)

270

Recommendations

- Report of DMC after interim analysis
  - Overall impression of study conduct
  - Overall impression of data quality
  - Suggested changes to
    - protocol
      - eligibility criteria; data collected; treatment (incl ancillary)
    - informed consent
    - monitoring schedule
  - Recommendations re early termination
    - Minimize unblinding in case Steering Committee decides otherwise
      - should not report actual treatment effect
      - report to subset of Steering Committee

271

Documentation

Prespecification of all Aspects of RCT

Where am I going?

Clinical trial standards have been developed for

- Protocol and amendments with review by regulatory authorities, IRBs, monitoring committees
- Statistical analysis plans that are fully defined prior to unblinding the data
- Registering clinical trials in advance in order to ensure pre-specification of goals, methods, etc.
- Reporting of clinical trial results

272

Purpose of Protocol

- We design an experiment to minimize the bias and variability of our measurement of treatment effect
- The protocol documents the ways in which we will conduct our experiment to achieve that goal

273

Study Protocol

- Clinical trial protocol
  - Formal definition of treatments, endpoints, hypotheses, eligibility, study procedures, etc.
  - Serves as
    - Documentation of rationale for study
    - Documentation of prior knowledge
    - Documentation of experimental method
    - Guide to development of manual of operations
    - Guide to development of Statistical Analysis Plan
- See FDA/NIH Template for clinical trials

  https://osp.od.nih.gov/clinical-research/clinical-trials/

274

Statistical Analysis Plan

- The ultimate reference for statistics reported at the end of RCT
- Formal pre-specification of all planned analyses including
  - Definition of datasets used for various analyses
    - Safety, Efficacy, ITT, PP, etc.
  - Definition of derived variables (including time conventions)
  - Description of tables, listings, figures to be produced
  - Detailed specification of analyses to address each endpoint
    - Inferential criteria (and any hierarchy for sequential testing and multiple comparisons)
    - Analysis model, covariates (and explicit form in model)
    - Estimation of standard errors and test statistic
    - Handling of missing data in primary analysis
    - Sensitivity analyses re missing data assumptions
  - Descriptive supportive analyses, exploration of subgroups
  - Etc.

275