Biost 536 / Epi 536
Categorical Data Analysis in Epidemiology

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

Lecture 8:
Exploratory vs Screening vs Confirmatory Studies

October 30, 2014

1

General Message

"Make new friends, but don't forget the old.
One is silver, and the other is gold."

- Children's song

2

Possible Perception

"You're going the wrong way. All y'all
are going the wrong way…"

- Wade
(Robert Altman's *Nashville,* 1975)

3

Apologies in Advance

"If there is anyone here whom I have
not offended, I sincerely apologize"

- Johannes Brahms

4

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)

5

Biomedical Science

- Basic science
  - Understanding mechanisms of cell biology, physiology, and pathophysiology
  - "Knowledge is good" (Emil Faber)
- Clinical science
  - Identifying treatment and preventive strategies that benefit an individual (most often ill or at high risk for disease)
- Public health science
  - Identifying strategies that promote the well-being of the population (on average)

6

Classification of Scientific Goals

- Exploratory hypothesis generation
  - What question should we study next?
  - CI, P values might rank the hypotheses in order of importance, but scientific issues are also important
- Screening studies
  - Does the preliminary data suggest continuing to investigate this hypothesis?
  - P values screen for the hypotheses that "clear the hurdle"
  - May involve principled refinement of hypothesis
- Confirmatory studies
  - Statistical inference about a well-defined question

7

Phases of Investigation

- Science has always been a sequential, adaptive process
  - Series of studies support adoption of new hypothesis
  - "The proper result of a scientific study is another scientific study."
- Preclinical
  - Epidemiology including risk factors
  - Basic science:
    - Biochemistry, physiologic mechanisms, physics / engineering
  - Animal experiments: Toxicology / safety
- Clinical
  - Designed observational studies
  - Interventional
    - Phase I: Initial safety / dose finding
    - Phase II: Preliminary efficacy / further safety
    - Phase III: Confirmatory efficacy / effectiveness

8

Major Issues

- We want a scientific process that extends our knowledge by formulating and testing hypotheses
- Basic tools at our disposal are well-conducted studies
- Observation
  - Exploring data and formulating hypotheses
- Designed studies
  - First choice: Randomized interventional experiments
  - Second choice: Analysis of observational data in a carefully chosen dataset that is independent of the data that generated the hypothesis

9

Observational vs Interventional Statistics

- Ultimately, the same statistical methods are used on both
  - Observational data that is potentially confounded, and
  - Interventional data that is better protected from confounding
- In randomized experiments, we can imply causation from the experimental design, not the statistical techniques
  - We do need to make sure that the statistical methods respect the design (i.e., comparisons are made across randomized groups)
  - Even then, we may not understand the mechanisms of action
- In observational studies, we can never be sure of causation
  - But we should generally try to use methods that would be appropriate if a cause-effect relationship exists

10

Modeling Decisions

- Associations (variable importance) vs prediction
- Proportion vs odds vs (average) hazard
- Difference vs ratio
- Modeling predictor of interest (exposure)
- Modeling effect modification
- Covariates for adjustment: confounders, precision
- Method of adjustment:
  - Stratification (weighting?)
  - Dummy variables
  - Linear
  - Transformed linear
  - Splines / smooths

11

Relevance to Biost 536

- We will need to make many decisions
  - We cannot make an informed decision without understanding the distinctions among our many choices
    - Scientific interpretation
    - Statistical behavior
- The validity of our scientific generalizations will be greatly affected by when we make the decisions
  - Exploratory data-driven analyses
    - Attempts to model the data generation process
    - Prone to high false discovery rate
  - Confirmatory hypothesis driven analyses
    - Attempts to model our scientific question

12

Breslow (*Int. Stat. Rev*, **67**, pp. 252-255)

"As a medical statistician, I am appalled by the large number of irreproducible results published in the medical literature. There is a general, and likely correct, perception that this problem is associated more with statistical, as opposed to laboratory, research. I am convinced, however, that results of clinical and epidemiological investigations could become more reproducible if only the investigators would apply more rigorous statistical thinking and adhere more closely to well established principles of the scientific method. While I agree that the investigative cycle is an iterative process, I believe that it works best when it is hypothesis driven."

13

Breslow (*Int. Stat. Rev*, **67**, pp. 252-255)

"The epidemiology literature is replete with irreproducible results stemming from the failure to clearly distinguish between analyses that were specified in the protocol and that test the a priori hypotheses whose specification was needed to secure funding, and those that were performed post-hoc as part of a serendipitous process of data exploration."

14

This Lecture

- In observational epidemiologic studies we are often trying to discern which factors are associated with greater or lesser risk of disease
- From a clinical or public health standpoint, such studies will ideally lead to adoption of new treatments or behaviors that will
  - Better treat an individual (medicine)
  - Lead to a healthier population (public health)
- I will use the process of "treatment discovery" to illustrate some of the common pitfalls of unorganized scientific research
  - But everywhere I use "drug discovery" we could just as easily state "discovery of a risk factor"

15

Overall Goal

- "Drug discovery"
  - More generally
    - a therapy / preventive strategy or diagnostic / prognostic procedure
    - for some disease
    - in some population of patients
- A series of experiments to establish
  - Safety of investigations / dose
  - Safety of therapy
  - Measures of efficacy
    - Treatment, population, and outcomes
  - Confirmation of efficacy
  - Confirmation of effectiveness

16

Topic for Today: Optimizing the Process

- How do we maximize the number of drugs adopted while
  - Ensuring effectiveness of adopted drugs
  - Ensuring availability of information needed to use drugs wisely
  - Minimizing the use of resources
    - Patient volunteers
    - Sponsor finances
    - Calendar time
- The primary tool at our disposal: Sequential sampling
  - Decrease average sample size used for each drug
  - Maximize number of new drugs using limited resources

17

Phases of Investigation

- A sequential, adaptive process
  - But only "piecewise continuous"
- During any individual clinical trial
  - Sequential monitoring, adaptation addresses that trial's issues
- "White space" between trials: Detailed and exploratory analyses
  - Evaluation of multiple endpoints; cost/benefit tradeoffs
  - Exploratory analyses
  - Integration of results from other studies
  - Management decisions
  - Regulatory and ethical review

18

The Enemy

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

- Maria von Trapp

(as quoted by Rodgers and Hammerstein)

19

First

- Where do we want to be?
  - Describe some innovative experiment?
  - Establish that our hypothesis is true?
  - Find a use for some proprietary drug / biologic / device?
    - "Obtain a significant p value"
  - Find a new treatment that improves health of some individuals
    - "Efficacy" seems to benefit some people somehow
  - Find a new treatment that improves health of the population
    - "Effectiveness" requires proper use of a treatment that modifies an important clinical endpoint

20

Which: Efficacy or Effectiveness

- Factors leading to efficacy trials
  - "Knowledge is good"
  - As pilot studies before prevention studies
  - Inability to perform experiment under realistic conditions
- 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, $$)

21

Treatment "Indication"

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

22

Evidence Based Medicine

- Decisions about treatments should consider PICO
  - Patient (population)
  - Intervention
  - Comparators
  - Outcome
- There is a need for estimates of safety, effect

23

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

24

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)

25

Diagnostic Medicine: Evaluating a Test

- **We condition on diagnoses** (from gold standard)
  - Frequentist criteria: We condition on what is unknown in practice
- **Sensitivity: Do diseased people have positive test?**
  - Denominator: Diseased individuals
  - Numerator: Individuals with a positive test among denominator
- **Specificity: Do healthy people have negative test?**
  - Denominator: Healthy individuals
  - Numerator: Individuals with a negative test among denominator

26

Diagnostic Medicine: Using a Test

- **We condition on test results**
  - Bayesian criteria: We condition on what is known in practice
- **Pred Val Pos: Are positive people diseased?**
  - Denominator: Individuals with positive test result
  - Numerator: Individuals with disease among denominator
- **Pred Val Neg: Are negative people healthy?**
  - Denominator: Individuals with negative test result
  - Numerator: Individuals who are healthy among denominator

27

Points Meriting Special Emphasis

- Discover / evaluate tests using frequentist methods
  - Sensitivity, specificity
- Consider Bayesian methods when interpreting results for a given patient
  - Predictive value of positive, predictive value of negative
- Possible rationale for our practices
  - Ease of study: Efficiency of case-control sampling
  - Generalizability across patient populations
    - Belief that sensitivity and specificity might be
    - Knowledge that PPV and NPV are not
  - Ability to use sensitivity and specificity to get PPV and NPV
    - But not necessarily vice versa

28

Bayes' Rule

- Allows computation of "reversed" conditional probability
- Can compute PPV and NPV from sensitivity, specificity
  - **BUT: Must know prevalence of disease**

$$PPV = \frac{sensitivity \times prevalence}{sens \times prevalence \;+\; (1-spec)\times(1-prevalence)}$$

$$NPV = \frac{specificity \times (1-prevalence)}{spec \times (1-prevalence) \;+\; (1-sens)\times prevalence}$$

29

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

30

Frequentist Inference

- Control type 1 error: False positive rate
  - Based on specificity of our methods
- Maximize statistical power: True positve rate
  - Sensitivity to detect specified effect
- Provide unbiased (or consistent) estimates of effect
- Standard errors: Estimate reproducibility of experiments
- Confidence intervals
- Criticism: Compute probability of data already observed
  - "A precise answer to the wrong question"

31

Bayesian Inference

- Hypothesize prior prevalence of "good" ideas
  - Subjective probability
- Using prior prevalence and frequentist sampling distribution
  - Condition on observed data
  - Compute probability that some hypothesis is true
    - "Posterior probability"
  - Estimates based on summaries of posterior distribution
- Criticism: Which presumed prior distribution is relevant?
  - "A vague answer to the right question"

32

Frequentist vs Bayesian

- 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:

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

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

$$posterior\ \ odds = Bayes\ Factor \times prior\ \ odds$$

33

Topic for Today: Optimizing the Process

- How do we maximize the number of drugs adopted while
  - Ensuring effectiveness of adopted drugs
  - Ensuring availability of information needed to use drugs wisely
  - Minimizing the use of resources
    - Patient volunteers
    - Sponsor finances
    - Calendar time
- The primary tool at our disposal: Sequential sampling
  - Decrease average sample size used for each drug
  - Maximize number of new drugs using limited resources

34

Phases of Investigation

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

35

Phase III Confirmatory Trials

- The major goal of a "registrational trial" is to confirm a result observed in some early phase study
- Rigorous science: Well defined confirmatory studies
  - Eligibility criteria
  - Comparability of groups through randomization
  - Clearly defined treatment strategy
  - Clearly defined clinical outcomes (methods, timing, etc.)
  - Unbiased ascertainment of outcomes (blinding)
  - Prespecified primary analysis
    - Population analyzed as randomized
    - Summary measure of distribution (mean, proportion, etc.)
    - Adjustment for covariates

36

Why Confirmation: Real-life Examples

- 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

37

Multiple Comparisons in Biomedicine

- Observational studies
  - Observe many outcomes
  - Observe many exposures
  - Perform many alternative analyses
    - Summary of outcome distribution, adjustment for covariates
  - Consequently: Many apparent associations
    - May be type I errors
    - But even when valid, may be poorly understood due to confounding
- Interventional experiments
  - Exploratory analyses ("Drug discovery")
    - Modification of analysis methods
    - Multiple endpoints
    - Restriction to subgroups

38

Statistics and Game Theory

- Multiple comparison issues
  - Type I error for each endpoint – subgroup combination
    - In absence of treatment effect, will still decide a benefit exists with probability, say, .025 in each such combination
- Multiple endpoints and subgroups 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
- Impact of increased type I error on Bayes factor is huge
  - Ratio of power to type I error means multiplicative effects

39

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)

40

U.S. Regulation of 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

41

Phase III Clinical Trials

- Confirmation of efficacy / effectiveness
  - Goals:
    - Obtain measure of treatment's efficacy on disease process
    - Incidence of major adverse effects
    - Therapeutic index
    - Modify clinical practice (obtain regulatory approval)
  - Methods
    - Relatively large number of participants from true target population (almost)
    - Clinically relevant outcome

42

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

43

Phase II Clinical Trials

- Preliminary evidence of efficacy
  - Goals:
    - Screening for any evidence of treatment efficacy
    - Incidence of major adverse effects
    - Decide if worth studying in larger samples
      - Gain information about best chance to establish efficacy
        - » Choose population, treatment, outcomes
  - Methods
    - Relatively small number of participants
    - Participants closer to true target population
    - Outcome often a surrogate
    - Sometimes no comparison group (especially in cancer)

44

Screening Studies as Diagnostic Tests

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

45

Preliminary Studies in Screening

- In cancer less than 5% of treatments studied in clinical trials are adopted
- NCI drug development program 1970 - 1985
  - 350,000 unique chemical structures studied
  - 83 pass preclinical and phase I testing
  - 24 pass phase II tests for biological activity

46

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

47

Scenario 1: Only Phase III

- Only large trials using 1,000,000 subjects
  - 10% of drugs being investigated truly work
  - Level of significance .025, .025, or 0.05
  - Sample size / power
    - 979 subjects, α=0.025, 97.5% power ➔ 1,021 RCT
    - 500 subjects, α=0.025, 80.0% power ➔ 2,000 RCT
    - 394 subjects, α=0.050, 80.0% power ➔ 2,538 RCT
  - Results
    - N= 979: 99 effective / 23 ineffective (PV+ = .81)
    - N= 500: 160 effective / 45 ineffective (PV+ = .78)
    - N= 394: 202 effective / 114 ineffective (PV+ = .64)

48

Scenario 2a: Screening Phase II

- Use 700,000 subjects in Phase II studies
  - 10% of drugs being investigated truly work
  - Level of significance .025
  - Sample size / power
    - 100 subjects provide 24% power ➔ 7,000 RCT
  - Results
    - N= 100: 168 effective / 158 ineffective (PV+ = .52)
- Use 300,000 subjects in confirmatory Phase III studies
  - 52% of drugs being investigated truly work
  - Level of significance .025
  - Sample size / power
    - 921 subjects provide 96.7% power ➔ 326 RCT
  - Results
    - N= 921: 162 effective / 4 ineffective (PV+ = .98)

49

Scenario 2b: Screening Phase II

- Use 700,000 subjects in Phase II studies
  - 10% of drugs being investigated truly work
  - Level of significance .10
  - Sample size / power
    - 342 subjects provide 85% power ➔ 2,047 RCT
  - Results
    - N= 342: 173 effective / 184 ineffective (PV+ = .49)
- Use 300,000 subjects in confirmatory Phase III studies
  - 49% of drugs being investigated truly work
  - Level of significance .025
  - Sample size / power
    - 839 subjects provide 95% power ➔ 357 RCT
  - Results
    - N= 839: 165 effective / 5 ineffective (PV+ = .97)

50

Summary

		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

Summary: Phase 2

		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

Summary: Phase 3

		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
    - 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?

54

Summary: "Drug Discovery"

		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**

Coherent vs Incoherent Bayes

- The previous results were based on "staying the course"
- *A priori* we presumed a certain treatment effect
  - Phase 2 studies powered for that treatment effect
  - When progress to phase 3, still power for that treatment effect
- The problem: Phase 2 results are used to decide to go to phase 3
  - Results from "promising phase 2 trials" are biased

56

The Problem of Small Studies

- Using 700,000 patients
  - Small sample size ➔ Big bias of "positive" studies

			Null: Δ =0			Alt: Δ= .125		
N per RCT	RCTs	Crit Value	Prob Sig	N Sig RCT	Expected Estimate	Prob Sig	N Sig RCT	Expected Estimate
7000	**100**	**0.0234**	0.025	2	0.028	1.000	100	0.125
3500	**200**	**0.0331**	0.025	5	0.039	1.000	200	0.125
700	**1000**	**0.0741**	0.025	25	0.089	0.912	912	0.132
350	**2000**	**0.1048**	0.025	50	0.125	0.649	1,298	0.156
70	**10000**	**0.2343**	0.025	250	0.280	0.180	1,801	0.299
35	**20000**	**0.3313**	0.025	500	0.390	0.114	2,271	0.407

57

Coherent vs Incoherent Bayes

- An alternative optimistic strategy
  - Phase 2 studies powered for presumed treatment effect
  - Phase 3 studies powered for observed phase 2 estimate
- An alternative Bayesian strategy
  - Phase 2 studies powered for presumed treatment effect
  - Phase 3 studies powered for posterior mean treatment effect
    - (up to some maximum)

58

Summary

		Scenario 2b	Optimistic	Mod. Bayes
Phase 2	Number RCT	2,047 (10% eff)	1,759 (10% eff)	1,959 (10% eff)
	N per RCT	342	342	342
	Type 1 err; Pwr	0.100; 85%	0.100; 85%	0.100; 85%
	"Positive" RCT	173 eff; 184 not	150 eff; 159 not	163 eff; 176 not
Confirmtaory Phase 3	Number RCT	357 (49% eff)	309 (49% eff)	339(48% eff)
	N per RCT	839	894 vs 1665	941 vs 998
	Type 1 err, Pwr	0.025; 95%	0.025; 95 vs 86%	0.025; 95%
	# Effctve Adopt	165	129	156
	# Ineff Adopt	5	4	4
	Pred Val Pos	97%	97%	97%
	N per Adopt	1,181	1,259	1,285

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

60

Ultimate Goal

- Modify the sample size accrued so that minimal number of subjects treated when
  - new treatment is harmful,
  - new treatment is minimally effective, or
  - new treatment is extremely effective
- Only proceed to maximal sample size when
  - not yet certain of treatment benefit, or
  - potential remains that results of clinical trial will eventually lead to modifying standard practice

61

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

62

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

63

Potential Benefits of Stopping Rules

- 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%

64

Furthermore

- Additional advantages of screening trials
  - Gathering more detailed preliminary safety data before embarking on expensive, large scale Phase 3 trials
  - Gathering preliminary efficacy data that allows fine tuning
  - Fine tune eligibility criteria
    - Include only susceptible patient populations
    - Exclude patients at high risk for AEs
  - Optimal treatment strategies
    - Fine tune formulation, dose, administration, frequency, duration
    - Develop dose modification strategies
    - Prophylactic treatments, rescue treatments for AEs
  - Optimal clinical endpoints
- Major disadvantage
  - "White space" (time delay) between phase 2 and phase 3
  - (Truly an issue for sponsors, rather than public health)

65

Inflation of the Type I Error

- Recall that in order to avoid inflation of type I error, we require confirmatory studies using prespecified
  - Patient population
  - Treatment
  - Primary clinical outcome
  - Statistical analysis
- Hence, we must be concerned about data dredging ("data mining") of the phase 2 data, because it may lead to differences between phase 2 and phase 3 due to
  - Revising outcomes to reflect the most promising results
  - Revising eligibility criteria based on subgroup analyses
  - Changing from surrogate efficacy to effectiveness endpoints
    - "Treating the symptom not the disease"

66

Mathematical Basis

- The multiple comparison problem is traced to a well known fact of probability

$$\Pr(A \text{ or } B) \geq \Pr(A)$$

$$\Pr(A \text{ or } B) \geq \Pr(B)$$

67

Statistics and Game Theory

- Multiple comparison issues
  - Type I error for each endpoint – subgroup combination
    - In absence of treatment effect, will still decide a benefit exists with probability, say, .025
- Multiple endpoints and subgroups 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

68

Data Dredging Examples: Endpoints

- We might look for the endpoint for which the treatment has the largest estimated effect
- Examples
  - Overall survival
    - Logrank test vs Wilcoxon logrank vs survival at fixed time …
  - Progression free survival
  - Major adverse cardiovascular events (MACE)
  - MACE plus hospitalization
  - …

69

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

70

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?

71

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
- With lognormal data the type 1 error can be greater than 0.21

72

Data Dredging Examples: Modeling POI

- We might consider modeling a continuous predictor of interest
  - Untransformed linear
  - Log transformed linear
  - Dummy variables based on quartiles
  - Dummy variables based on scientific intervals
  - Linear splines

73

Statistical Issues

- Need to choose a primary statistcal analysis or multiple comparison issues result
- Example: Type 1 errors from simulations under the null
  - Untransformed linear 0.049
  - Log transformed linear 0.050
  - Dummy variables based on quartiles 0.043
  - Dummy variables based on 3 intervals 0.050
  - Linear splines on 3 intervals 0.049
  - Lowest p value from all of the above 0.131

74

Data Dredging Examples: Dose / Arms

- We might look for the dose group or treatment arm with largest effect
  - Treatment effect
  - Risk / benefit ratio
  - P value

75

Data Dredging Examples: Subgroups

- In phase 2 trials that are not significant, we search for subgroups that might show significant differences
  - If the results were significant overall, we use the overall results
- In phase 2 trials that are significant, we look for cases in which all the effect seems to be in a subgroup
  - Statistical significance in, say, males
  - Point estimate in wrong direction in females
- We look for the smallest p value among the overall comparison and several subgroups
  - We choose the indication with the smallest p value

76

Examples

- We can explore the impact of adaptive changes in RCT in several examples
  - Consideration of multiple summary measures
    - Mean, geometric mean, Wilcoxon, median, two proportions
  - Consideration of subgroups
    - Overall sample
    - Plus equal sized subgroups defined by three variables
  - Consideration of change of endpoint between phase 2 and 3
    - Phase 2: potential surrogate
    - Phase 3: clinical outcome
- We consider
  - Adaptations that do or do not control type I error
  - Treatment effect in all groups or only in one subgroup
  - Surrogates that do or do not always predict outcome

77

Homogeneous Effects, No Error Control

- First we consider treatments that are equally effective in all subjects
- Prevalence of beneficial treatments: 10%
- Possible adaptations
  - Adaptive choice of statistical summary measure
    - Mean, geometric mean, median, Wilcoxon, two thresholds
  - Look for subgroups having effects
    - Sex, Age (young vs old), BMI (normal vs obese)
    - Strategies
      - If significant overall, proceed with all, otherwise choose most significant subgroup
      - Choose subgroup if it is highly significant and opposite subgroup has estimated nil effect
      - Choose analysis with smallest p value

78

Summary (Homogeneous Effects)

		Scenario 2b	Alt Smry Meas	Subgroups
Phase 2	Number RCT	2,047 (10% eff)	1,695 (10% eff)	1,485 (10% eff)
	N per RCT	342	342	342
	Type 1 err; Pwr	0.100; 85%	0.227; 92%	0.334; 95%
	"Positive" RCT	173 eff; 184 not	155 eff; 346 not	141 eff; 446 not
Confirmtaory Phase 3	Number RCT	357 (49% eff)	501 (31% eff)	587 (24% eff)
	N per RCT	839	839	839
	Type 1 err, Pwr	0.025; 95%	0.025; 94%	0.025; 95%
	# Effctve Adopt	165	147	134
	# Ineff Adopt	5	9	11
	Pred Val Pos	97%	94%	92%
	N per Adopt	1,181	1,181	1,181

Inhomogeneous Effects, No Error Control

- Consider treatments effective only in females
- Prevalence of beneficial treatments: 10%
- Look for subgroups having effects
  - Sex, Age (young vs old), BMI (normal vs obese)
  - Strategies
    - If significant overall, proceed with all, otherwise choose most significant subgroup
    - Choose subgroup if it is highly significant and opposite subgroup has estimated nil effect
    - Choose analysis with smallest p value

80

Impact of Strategies for Subgroups

Analysis	Sig	Pref All	Choice	Min P
All	.64	.64	.40	.07
Females	.85	.20	.40	.60
Males	.10	.00	.00	.00
Young	.45	.02	.03	.06
Old	.45	.02	.03	.06
Norm Wt	.45	.02	.03	.06
Obese	.45	.02	.03	.06

81

Summary (Inhomogeneous Effects)

		Scenario 2b	Prefer All	Choose Subgrp
Phase 2	Number RCT	2,123 (10% eff)	1,490 (10% eff)	1,490 (10% eff)
	N per RCT	342	342	342
	Type 1 err; Pwr	0.100; 64%	0.334; 92%	0.334; 92%
	"Positive" RCT	136 eff; 191 not	137 eff; 448 not	137 eff; 448 not
Confirmtaory Phase 3	Number RCT	327 (42% eff)	584 (23% eff)	584 (23% eff)
	N per RCT	839	839	839
	Type 1 err, Pwr	0.025; 73%	0.025; 75%	0.025; 80%
	# Effctve Adopt	99	103	109
	# Ineff Adopt	5	11	11
	Pred Val Pos	95%	90%	91%
	N per Adopt	1,181	1,181	1,181

Adaptive Sampling Plans

- At each interim analysis, possibly modify
  - Conditions for early stopping
  - Schedule of analyses
  - Randomization ratios
  - Maximal statistical information
  - Statistical criteria for credible evidence
  - Scientific and statistical hypotheses of interest
    - Summary measures used to quantify treatment effect
      - Mean, median, etc.
    - Clinical endpoint
      - Objective response rate, progression, survival, etc.
    - Eligibility criteria
      - Restrict to a subgroup
    - Definition of treatment
      - Drop dose groups, change ancillary treatments, etc.

83

When Stopping Rules Not Pre-specified

- Methods to control the type I error have been described for fully adaptive designs
  - Most popular: Preserve conditional error function from some fixed sample or group sequential design
  - Can have loss of efficiency relative to prespecified plan
- Can choose revised sample size to maintain power
- Methods to compute bias adjusted estimates and confidence intervals not yet well-developed

84

"Partitioning Type 1 Error"

- When designing an adaptive design to look for alternative endpoints, subgroups, doses, we have to decide how to prioritize the different decisions
- This is akin to "spending type 1 error" in sequential trials
- We have to consider our relative beliefs in the treatment effect
  - Is it likely to be homogeneous across all subgroups examined?
  - Is it likely to be concentrated in some pre-specified subgroup?

85

Strategies for Subgroups: Type 1 Error

Example: Assuming independent covariates with 50-50 split

Analysis	Sig	Pref All	Choice	Min P
All	.023	.022	.021	.007
Females	.023	.013	.013	.015
Males	.023	.013	.013	.015
Young	.023	.013	.013	.015
Old	.023	.013	.013	.015
Norm Wt	.023	.013	.013	.015
Obese	.023	.013	.013	.015

86

Strategies for Subgroups: Alternatives

- Need to consider how we thing the overall treatment effect might differ from effects within subgroups
- Cases we have considered
  - Hypothesized treatment effect actually occurs only in subpopulation
    - Overall test is extremely underpowered: 45% instead of 85%
  - Slightly stronger hypothesized treatment effect only in subpopulation
  - Overall population's treatment effect as hypothesized
    - But one subgroup has double that effect and opposite subgroup has no effect

87

Generalizability

- We need to consider type 1 and type 2 errors relative to the ultimate result of the "drug discovery" process
- The previous results are dependent on
  - A mixture of 10% effective drugs and 90% ineffective drugs, *where "effectiveness" is defined based on the clinical outcome used in the phase 3 trial*
  - Phase 2 and phase 3 type I errors being controlled at the specified level *based on the phase 3 outcome*
  - Phase 2 and phase 3 power being controlled at the specified level *based on the phase 3 outcome*
- Many early phase RCT use alternative outcomes
  - "Surrogate endpoints" in restricted populations

88

Inhomogeneous Effects, Control Errors

- Consider treatments effective only in females
- Prevalence of beneficial treatments: 10%
- Look for subgroups having effects
  - Sex, Age (young vs old), BMI (normal vs obese)
  - Strategies as before
- Perform all tests using type I error of 0.023
  - Yields experimentwise type I error of 0.100
- Increase phase 2 sample size to obtain 0.85 experimentwise power

89

Control Error (Inhomogeneous Effects)

		Scenario 2b	Inflate Error	Control Error
Phase 2	Number RCT	2,123 (10% eff)	1,490 (10% eff)	2,081 (10% eff)
	N per RCT	342	342	342
	Type 1 err; Pwr	0.100; 64%	0.334; 92%	0.100; 77%
	"Positive" RCT	136 eff; 191 not	137 eff; 448 not	161 eff; 187 not
Confirmtaory Phase 3	Number RCT	327 (42% eff)	584 (23% eff)	348 (46% eff)
	N per RCT	839	839	829
	Type 1 err, Pwr	0.025; 73%	0.025; 80%	0.025; 95%
	# Effctve Adopt	99	109	153
	# Ineff Adopt	5	11	5
	Pred Val Pos	95%	91%	97%
	N per Adopt	1,181	1,181	1,177

Control Error (Inhomogeneous Effects)

- With inhomogeneous effects, 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

91

Homogeneous Effects, Surrogates

- We consider treatments that are equally effective in all subjects
- Prevalence of beneficial treatments: 10%
  - "Beneficial" defined based on phase 3 endpoint
- Prevalence of misleading treatments: 0%, 10%, or 20%
  - "Misleading" = efficacious on surrogate but not effective
  - 85% power to detect efficacy on surrogate
- No adaptations

92

Surrogates (Homogeneous Effects)

		0% Misleading	10% Misleading	20% Misleading
Phase 2	Number RCT	2,046 (10% eff)	1,812 (10% eff)	1,627 (10% eff)
	N per RCT	342	342	342
	Type 1 err; Pwr	0.100; 85%	0.100; 85%	0.100; 85%
	"Positive" RCT	174 eff; 184 not	154 eff; 337 not	138 eff; 494 not
Confirmtaory Phase 3	Number RCT	358 (49% eff)	491 (31% eff)	632 (22% eff)
	N per RCT	839	839	839
	Type 1 err, Pwr	0.025; 95%	0.025; 95%	0.025; 95%
	# Effctve Adopt	166	147	132
	# Ineff Adopt	5	8	12
	Pred Val Pos	97%	95%	91%
	N per Adopt	1,181	1,181	1,181

Comparisons

	RCT	Eff (TP)	Not(FP)	n
Nonadaptive				
• Homogeneous effect	2,040	165(165)	5	1,181
• Homogeneous,10% misleading	1,812	147(147)	8	1,181
• Homogeneous,20% misleading	1,627	132(132)	12	1,181
• Inhomogeneous effect	2,123	99(0)	5	1,181
Adaptive subgroups: inflate error				
• Homogeneous effect	1,488	134(43)	11	1,181
• Inhomogeneous effect	1,493	122(88)	11	1,181
Adaptive subgroups: control error				
• Homogeneous effect	2,040	153(56)	4	1,277
• Inhomogeneous effect	2,067	135(103)	4	1,277

94

Comments

- In a large, expensive study, it is well worth our time to carefully examine the ways we can best protect
  - Patients on the study
  - Patients who might be on the study
  - Patients who will not be on the study, but will benefit from new knowledge
  - Sponsor's economic interests in cost of trial
  - Eventual benefit to health care costs
- Adaptation to interim trial results introduces complications, but they can often be surmounted using methods that are currently well understood
  - It is not immediately clear how close we already are to optimality
  - (Multiple 0.023 tests yielded experimentwise 0.10)
- To get good results, we need to learn to take "NO" for an answer

95

Final Comments

- Though presented in the context of RCT, the results can be easily generalized to settings where RCT are unethical, unfeasible, or impossible
- In such settings, there will typically be more gradual progression from exploratory to screening to confirmation
  - Optimal strategies will depend on prevalence of "good" ideas
- And there will often be many confirmatory studies in as disparate settings as possible before a hypothesis is believed by a broader community
- And in those confirmatory studies, it is important not to keep changing the question
  - Stay the course re all aspects of sampling and statistical analyses

96

Final Comments

- Careful use of exploratory, screening, and confirmatory studies provide great protection
  - Ensure that overwhelming majority of "adopted" hypotheses are true
- However, control of type I and II errors are of great importance at every stage
  - Uncontrolled exploratory studies lessen prevalence of "good ideas" at the screening stage
  - But note that type 1 error of 0.025 not necessarily indicated
- Nothing can protect against false surrogates
  - It is probably best to only consider surrogates at the exploratory phase

97

Really Bottom Line

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

-Aretha Franklin, "Think"

98