Facilitating Antibacterial Drug Development: Bayesian vs Frequentist Methods

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

The Brookings Institution
May 9, 2010

First: Where Do We Want To Be?

- Describe some innovative experiment?

- 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"

- Find a new treatment that improves health of the population
  - "Effectiveness"

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

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)

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

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 testing
  - Decrease average sample size = Maximize number of new drugs
- Distinctions without differences:
  - Every frequentist RCT design has a Bayesian interpretation
  - Every Bayesian RCT design has a frequentist interpretation

Phases of Investigation

- A “piecewise continuous” process
- During any individual clinical trial
  - Sequential monitoring, adaptation addresses issues of that trial
- “White space” between trials
  - More detailed analyses
  - Evaluation of multiple endpoints; cost/benefit tradeoffs
  - Exploratory analyses
  - Integration of results from other studies
  - Management decisions
  - Regulatory and ethical review
- Next RCT: May address different question or indication

Phase 3 Confirmatory Trials

- The major goal of a “registrational trial” is to confirm a result observed in some early phase study
  - Selection of “promising” early phase results introduces bias
  - The smaller the early phase trial, the greater the bias
- 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

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

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

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

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

Bayes' Rule

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

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

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

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 (frequentist type 1 error)
  - High probability that adopted drugs work (posterior probability)

Slightly Different Setting

- Usually we are interested in some continuous parameter
  - E.g., proportion of infections cured is $0 < p < 1$
- "Prevalence" is replaced by a probability distribution
  - Prior (subjective) probability of selecting a drug to test that cures proportion p of the population
- Sum over two hypotheses replaced by weighted average (by some subjective prior) over all possibilities

$$\Pr(p \mid \hat{p}) = \frac{\Pr(\hat{p} \mid p) \times \Pr(p)}{\int \Pr(\hat{p} \mid p) \times \Pr(p)\, dp}$$

$$= \frac{freq\ samp\ distn\ \times\ prior\ \ prob}{weighted\ \ average\ \ freq\ samp\ distn}$$

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”

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”

Frequentist vs Bayesian

- Frequentist and Bayesian inference truly complementary
  - Frequentist: Design an RCT so the same data is not likely to arise from both sets of hypotheses
  - 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

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

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

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

Recommended Best Practices

- Phased investigation
- Optimize process to maximize new drugs found with available patient resources
- Sequential sampling at each phase
  - Phase 2:
    - Choose type I error, power to increase prevalence (to ~50%?)
    - Best choice will depend on prior prevalence of “good ideas”
    - (Power of entire process depends on power at phase 2)
  - Phase 3:
    - Low type I error to ensure meet objective standards
    - High power to detect drugs that are clinically important
    - (False discovery rate depends on type I error at phase 3)

Comparisons: 10% Prior Prevalence

	RCT	Eff	Not	n
• Nonadaptive				
– Only Phase 3	2,000	160	45	500
– Homogeneous effect	2,047	165	5	1,181
– Homogeneous, 10% misleading	1,812	147	8	1,181
– Homogeneous, 20% misleading	1,627	132	12	1,181
– Inhomogeneous effect	2,123	99	5	1,181
• Adaptive subgroups: inflate error				
– Homogeneous effect	1,485	134	11	1,181
– Inhomogeneous effect	1,490	109	11	1,181
• Adaptive subgroups: control error				
– Homogeneous effect	1,707	139	4	1,277
– Inhomogeneous effect	1,720	105	4	1,277

Recommended Best Practices

- Examine scientific / statistical credibility using Bayesian analyses with a population of prior probabilities
  - Science is adversarial
  - Whom have we convinced?
- Priors should mainly consider beliefs before any testing
  - Update after studies
  - But consider bias introduced by selection of promising results
  - “Regression to the mean”

Final Comments

- Some aspects of RCT design can increase efficiency
  - Controlling / stratifying important factors, factorial designs, …
- Sequential sampling plans decrease average N
  - Increase number of drugs identified with fixed number of patients
  - May increase number of patients for any single trial
- Bayesian vs frequentist is an issue for inference
  - Every RCT design should (and does) allow either
  - Frequentist inference is “sufficient statistic” to allow others to perform Bayesian analyses that are relevant to their prior beliefs
- Any claim for greater efficiency in Bayesian inference merely reflects a change in standards
  - Incorporating prior information vs prior bias