Hot Topics in Clinical Trials

Graybill Conference VII
Biopharmaceutical Statistics
June 11, 2008

Scott Evans, Lingling Li, LJ Wei, Marvin Zelen
Program for Quantitative Sciences in Medicine
Department of Biostatistics
Harvard University
Meta Analysis

LJ Wei
Harvard University

Introduction
What is Meta Analysis

- What is meant by the word “Meta-analysis”
  - Meta is Greek for “later in time”
  - Meta is now used to denote something that goes to a higher level or is more comprehensive.
  - How is an analysis made more comprehensive?
- In empirical research, there are often multiple studies addressing the same research question
  - A standard analysis attempts to reach a conclusion based on a single study without reference to any other studies.
  - A meta-analysis attempts to reach a conclusion based on a set of studies that address the same hypothesis.

Introduction
History of Meta Analysis

- In 1805, Legendre developed least squares to combine data on the orbits of comets from different observatories.
- In 1930’s, statisticians working in agricultural research developed methods for combining the results of studies. Most notable are Fisher and Cochrane.
- In 1960’s, Cohen popularized the notion of effect size for use in sample size determination in the social and behavioral sciences
  - Effect size measures the differences between null hypothesis and the truth
  - Effect size + sample size determines the power.
- In 1976, Grass published an article “Primary, secondary and meta-analysis of research”. This is when the term “meta-analysis” was first used.

Introduction
Why Meta Analysis

- There are several reasons for conducting a meta-analysis of the results of previous studies:
  - The increasingly large # of research studies
    - 40,000 journals for the sciences
    - 1 article every 30 seconds
  - Unsystematic expert reviews of an area of research are often biased or years behind the current research.
  - Systematic and quantitative reviews are needed to summarize findings in a timely manner without bias.
Principles of Meta-analysis

- Meta-analysis is typically a two stage process
  1) Summary statistic calculated from each study.
     - For controlled trials, these values describe the treatment effects observed in each trial.
  2) Pooled effect is calculated by combining treatment effect estimates from individual studies.
     - Typically a weighted average of individual effects

- The combination of treatment effect estimates across studies may assume
  - **Fixed effects**: treatment effects the same across studies.
  - **Random effects**: treatment effects ~ a distribution across studies.

Formulating Hypothesis and Effect Measures

- Before conducting a meta analysis, it is important to decide the hypothesis or aim of the analysis. When formulating a hypothesis for meta-analysis, it is important to determine
  - the precise question the meta-analysis aims to address
  - whether the meta analysis is exploratory or hypothesis testing
    - **Hypothesis testing**: is the intent of the study to provide a definitive test (usually a test of average effect = 0)
    - **Exploratory**: are there variations in the treatment or characteristics of the studies that lead to better outcomes?
  - One also needs to select an appropriate **effect measure**

Type of Data and Effect Measures

- **Dichotomous or binary outcome**
  - when events of interest are rare
    - \( 1 - p_1 \approx 1 \) and \( 1 - p_0 \approx 1 \) \( \Rightarrow \) \( RR \approx OR \)
    - essential to assess the absolute risk in addition to relative risk
  - \( RR \) and \( OR = \infty \) if there are no events in the control group
  - \( RR \) and \( OR \) **not defined** if both groups have zero events
    - Standard procedures either exclude studies with 0/0 events or add 0.5 to empty cells.

- **Ordinal outcome**
  - Takes values \( y_1 < y_2 < \ldots < y_K \)
    - Example: “mild”, “moderate” and “severe”.
  - When the number of categories is large, such data are often analyzed as continuous data.
  - One may transform ordinal data into binary data by combining adjacent categories.
  - Proportional odds ratio under a proportional odds model
    - \( \text{logit} P(Y \leq y_k \mid \text{Tt}) = \alpha_k + \beta \cdot \text{Tt} \)
    - \( \beta \), the log-odds ratio, summarizes the treatment effect
Statistical Methods for Combining Results Across Studies

- What is the true effect? Depends on the underlying assumption about the study specific effect
  - Fixed Effects Assumption
    - Assumes that all studies have the same true effect
    - Variability only within each study
    - Precision depends mainly on study size
  - Random Effects Assumption
    - Studies allowed to have different underlying or true effects
    - Allows variation between studies as well as within studies
- Basic assumption: study results are independent.

Fixed Effects

Vote Count

- Consider # of studies in favor of the conclusion (say, reach the 0.05 level of significance) and examine if they are the majority
- This approach has been used a lot due to its simplicity, but has several drawbacks
  - Significance depends on study sample size and effect size.
  - Even if the null hypothesis is wrong and studies are not small, the percentage of trials reaching significance could still be less than 50% 
    - low power of detecting a treatment effect
  - Vote counts do not provide an estimate of effect size.

Fixed Effects

Pooling Using Effect Sizes

- Since Glass’s work in 1976, combining effect sizes has become the main form of meta-analysis.
- Suppose the estimated effect sizes are \( \{\hat{\beta}_i, i=1,...,S\} \)
- To ascertain the true underlying effect, a common approach is to consider a weighted average of the effect estimates from individual studies:
  \[
  \hat{\beta} = \frac{\sum_{i=1}^{S} w_i \hat{\beta}_i}{\sum_{i=1}^{S} w_i}
  \]

Fixed Effects

Pooling Using Effect Sizes

Binary Outcomes

- When the event rates are low or trial sizes are small, the standard error estimates used in the inverse variance method may be poor.
- Cochrane-Mantel-Hansel uses a different weighting scheme that depends upon the effect measure (eg RR, OR, RD).
- Cochrane-Mantel-Hansel pooled OR:
  \[
  \text{OR}_{MH} = \frac{\sum_{i=1}^{S} \hat{w}_i^{MH} \text{OR}_i}{\sum_{i=1}^{S} \hat{w}_i^{MH}}
  \]
  \[
  \hat{w}_i^{MH} = \frac{m_0(n_{ii} - m_{ii})}{n_{ii} + n_{0i}}
  \]

<table>
<thead>
<tr>
<th>Study</th>
<th>Event</th>
<th>No Event</th>
<th>Total</th>
<th>Event</th>
<th>No Event</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intervention</td>
<td>( m_i )</td>
<td>( n_i - m_i )</td>
<td>( n_i )</td>
<td>( m_i )</td>
<td>( n_i - m_i )</td>
<td>( n_i )</td>
</tr>
<tr>
<td>Control</td>
<td>( m_i )</td>
<td>( n_i - m_i )</td>
<td>( n_i )</td>
<td>( m_i )</td>
<td>( n_i - m_i )</td>
<td>( n_i )</td>
</tr>
</tbody>
</table>
Between Study Heterogeneity

- The key assumption of fixed effects meta analysis methods is that all primary studies are estimating the same underlying true effect.
- The underlying effects across studies may be heterogeneous.
  - Each study effect size $\hat{\beta}_i$ is estimating an individual population effect $\beta_i$.
  - As study sample size $N_i \to \infty$, $\hat{\beta}_i \to \beta_i$.
  - Some of the $\beta_i$ may be the same, but not all of them.

Testing for heterogeneity
- Cochran’s Q-test: $Q = \sum_{i=1}^{k} w_i (\hat{\beta}_{pooled} - \hat{\beta}_i)^2 \sim \chi^2_{k-1}$ under $H_0$.
  - Provides a measure of between study variation.
- Other descriptive measures of heterogeneity:
  - H statistic: $H = \sqrt{Q/(S-1)}$ has mean 1 under $H_0$.
  - Hoggins and Thompson (2002) suggested: $H > 1.5$ → caution regarding heterogeneity; $H < 1.2$ → little heterogeneity.
  - $I^2$ statistic = $(H^2 - 1)/H^2$
    - $I^2$ → % of total variability in effect size due to between study variation.
    - $I^2 \sim 0$ → little heterogeneity; $I^2 \sim 1$ → high heterogeneity.
    - termed the “inconsistency” of the trials included in meta-analysis and has become a preferred measure of heterogeneity.

Meta Regression

- The pooled effect size estimates the average effect across all studies.
- In the presence of heterogeneity:
  - the validity of such an average measure?
  - not a single population effect size that applies to all studies.
  - random effects pooling addresses heterogeneity to some extent.
- Meta regression:
  - provide an alternative approach that allows exploration of why studies have varied effect sizes.
  - one uses characteristics of the studies to explain the excess variation in effect sizes.
  - Thompson and Higgins (2002) reviewed several meta-regression methods.

Meta Regression

- The DerSimonian and Laird (1986) Model:
  $$\hat{Y}_i = \alpha + \beta_i + \zeta_i$$
- The Begg and Pilote (1991) Model:
  $$\hat{Y}_i = X_i \alpha + \beta_i + \zeta_i$$
  - This model allows the inclusion of single treatment historical controls as well as comparative trials in a treatment effect assessment.
Example

Effect of Rosiglitazone on MI or CVD Deaths

- Nissen and Wolski (2007) performed a meta analysis to examine whether Rosiglitazone (Avandia, GSK), a drug for treating type 2 diabetes mellitus, significantly increases the risk of MI or CVD related death.
- Avandia was introduced in 1999 and is widely used as monotherapy or in fixed-dose combinations with either Avandamet or Avandaryl. The original approval of Avandia was based on the ability of the drug to reduce blood glucose and glycated hemoglobin levels.
- Initial studies were not adequately powered to determine the effects of this agent on microvascular or macrovascular complications of diabetes, including cardiovascular morbidity and mortality.

### Exact Meta-Analysis Procedure
Combining Exact Intervals

- Questions:
  - Could we combine information across studies without excluding studies with 0 events or artificial imputation?
  - Could we make exact inference without relying on possibly inaccurate large sample approximations when the total number of studies is small, or the sample sizes of individual studies are small, or when the event rates are low.

---

**Exact Procedure for Combining 2x2 Tables for Rare Events**

**Example**

Effect of Rosiglitazone on MI or CVD Deaths

- Nissen and Wolski (2007) performed a meta analysis to examine whether Rosiglitazone (Avandia, GSK), a drug for treating type 2 diabetes mellitus, significantly increases the risk of MI or CVD related death.
- Avandia was introduced in 1999 and is widely used as monotherapy or in fixed-dose combinations with either Avandamet or Avandaryl. The original approval of Avandia was based on the ability of the drug to reduce blood glucose and glycated hemoglobin levels.
- Initial studies were not adequately powered to determine the effects of this agent on microvascular or macrovascular complications of diabetes, including cardiovascular morbidity and mortality.

### Exact Meta-Analysis Procedure
Combining Exact Intervals

- Questions:
  - Could we combine information across studies without excluding studies with 0 events or artificial imputation?
  - Could we make exact inference without relying on possibly inaccurate large sample approximations when the total number of studies is small, or the sample sizes of individual studies are small, or when the event rates are low.

---

**Exact Procedure for Combining 2x2 Tables for Rare Events**

**Example**

Effect of Rosiglitazone on MI or CVD Deaths

- Nissen and Wolski (2007) performed a meta analysis to examine whether Rosiglitazone (Avandia, GSK), a drug for treating type 2 diabetes mellitus, significantly increases the risk of MI or CVD related death.
- Avandia was introduced in 1999 and is widely used as monotherapy or in fixed-dose combinations with either Avandamet or Avandaryl. The original approval of Avandia was based on the ability of the drug to reduce blood glucose and glycated hemoglobin levels.
- Initial studies were not adequately powered to determine the effects of this agent on microvascular or macrovascular complications of diabetes, including cardiovascular morbidity and mortality.

### Exact Meta-Analysis Procedure
Combining Exact Intervals

- Questions:
  - Could we combine information across studies without excluding studies with 0 events or artificial imputation?
  - Could we make exact inference without relying on possibly inaccurate large sample approximations when the total number of studies is small, or the sample sizes of individual studies are small, or when the event rates are low.

---

**Exact Procedure for Combining 2x2 Tables for Rare Events**

**Example**

Effect of Rosiglitazone on MI or CVD Deaths

- Nissen and Wolski (2007) performed a meta analysis to examine whether Rosiglitazone (Avandia, GSK), a drug for treating type 2 diabetes mellitus, significantly increases the risk of MI or CVD related death.
- Avandia was introduced in 1999 and is widely used as monotherapy or in fixed-dose combinations with either Avandamet or Avandaryl. The original approval of Avandia was based on the ability of the drug to reduce blood glucose and glycated hemoglobin levels.
- Initial studies were not adequately powered to determine the effects of this agent on microvascular or macrovascular complications of diabetes, including cardiovascular morbidity and mortality.

### Exact Meta-Analysis Procedure
Combining Exact Intervals

- Questions:
  - Could we combine information across studies without excluding studies with 0 events or artificial imputation?
  - Could we make exact inference without relying on possibly inaccurate large sample approximations when the total number of studies is small, or the sample sizes of individual studies are small, or when the event rates are low.

---

**Exact Procedure for Combining 2x2 Tables for Rare Events**

**Example**

Effect of Rosiglitazone on MI or CVD Deaths

- Nissen and Wolski (2007) performed a meta analysis to examine whether Rosiglitazone (Avandia, GSK), a drug for treating type 2 diabetes mellitus, significantly increases the risk of MI or CVD related death.
- Avandia was introduced in 1999 and is widely used as monotherapy or in fixed-dose combinations with either Avandamet or Avandaryl. The original approval of Avandia was based on the ability of the drug to reduce blood glucose and glycated hemoglobin levels.
- Initial studies were not adequately powered to determine the effects of this agent on microvascular or macrovascular complications of diabetes, including cardiovascular morbidity and mortality.

### Exact Meta-Analysis Procedure
Combining Exact Intervals

- Questions:
  - Could we combine information across studies without excluding studies with 0 events or artificial imputation?
  - Could we make exact inference without relying on possibly inaccurate large sample approximations when the total number of studies is small, or the sample sizes of individual studies are small, or when the event rates are low.

---

**Exact Procedure for Combining 2x2 Tables for Rare Events**

**Example**

Effect of Rosiglitazone on MI or CVD Deaths

- Nissen and Wolski (2007) performed a meta analysis to examine whether Rosiglitazone (Avandia, GSK), a drug for treating type 2 diabetes mellitus, significantly increases the risk of MI or CVD related death.
- Avandia was introduced in 1999 and is widely used as monotherapy or in fixed-dose combinations with either Avandamet or Avandaryl. The original approval of Avandia was based on the ability of the drug to reduce blood glucose and glycated hemoglobin levels.
- Initial studies were not adequately powered to determine the effects of this agent on microvascular or macrovascular complications of diabetes, including cardiovascular morbidity and mortality.

### Exact Meta-Analysis Procedure
Combining Exact Intervals

- Questions:
  - Could we combine information across studies without excluding studies with 0 events or artificial imputation?
  - Could we make exact inference without relying on possibly inaccurate large sample approximations when the total number of studies is small, or the sample sizes of individual studies are small, or when the event rates are low.

---

**Exact Procedure for Combining 2x2 Tables for Rare Events**

**Example**

Effect of Rosiglitazone on MI or CVD Deaths

- Nissen and Wolski (2007) performed a meta analysis to examine whether Rosiglitazone (Avandia, GSK), a drug for treating type 2 diabetes mellitus, significantly increases the risk of MI or CVD related death.
- Avandia was introduced in 1999 and is widely used as monotherapy or in fixed-dose combinations with either Avandamet or Avandaryl. The original approval of Avandia was based on the ability of the drug to reduce blood glucose and glycated hemoglobin levels.
- Initial studies were not adequately powered to determine the effects of this agent on microvascular or macrovascular complications of diabetes, including cardiovascular morbidity and mortality.

### Exact Meta-Analysis Procedure
Combining Exact Intervals

- Questions:
  - Could we combine information across studies without excluding studies with 0 events or artificial imputation?
  - Could we make exact inference without relying on possibly inaccurate large sample approximations when the total number of studies is small, or the sample sizes of individual studies are small, or when the event rates are low.

---

**Exact Procedure for Combining 2x2 Tables for Rare Events**

**Example**

Effect of Rosiglitazone on MI or CVD Deaths

- Nissen and Wolski (2007) performed a meta analysis to examine whether Rosiglitazone (Avandia, GSK), a drug for treating type 2 diabetes mellitus, significantly increases the risk of MI or CVD related death.
- Avandia was introduced in 1999 and is widely used as monotherapy or in fixed-dose combinations with either Avandamet or Avandaryl. The original approval of Avandia was based on the ability of the drug to reduce blood glucose and glycated hemoglobin levels.
- Initial studies were not adequately powered to determine the effects of this agent on microvascular or macrovascular complications of diabetes, including cardiovascular morbidity and mortality.

### Exact Meta-Analysis Procedure
Combining Exact Intervals

- Questions:
  - Could we combine information across studies without excluding studies with 0 events or artificial imputation?
  - Could we make exact inference without relying on possibly inaccurate large sample approximations when the total number of studies is small, or the sample sizes of individual studies are small, or when the event rates are low.

---

**Exact Procedure for Combining 2x2 Tables for Rare Events**

**Example**

Effect of Rosiglitazone on MI or CVD Deaths

- Nissen and Wolski (2007) performed a meta analysis to examine whether Rosiglitazone (Avandia, GSK), a drug for treating type 2 diabetes mellitus, significantly increases the risk of MI or CVD related death.
- Avandia was introduced in 1999 and is widely used as monotherapy or in fixed-dose combinations with either Avandamet or Avandaryl. The original approval of Avandia was based on the ability of the drug to reduce blood glucose and glycated hemoglobin levels.
- Initial studies were not adequately powered to determine the effects of this agent on microvascular or macrovascular complications of diabetes, including cardiovascular morbidity and mortality.

### Exact Meta-Analysis Procedure
Combining Exact Intervals

- Questions:
  - Could we combine information across studies without excluding studies with 0 events or artificial imputation?
  - Could we make exact inference without relying on possibly inaccurate large sample approximations when the total number of studies is small, or the sample sizes of individual studies are small, or when the event rates are low.
Exact Meta-Analysis Procedure
Combining Exact Intervals

- For a given confidence level \( \eta \), one may obtain \( S \) study specific one-sided \( \eta \)-level confidence intervals for the risk difference.
  - each interval is constructed based on the data only from its corresponding study.

- For any given value of \( \Delta \), we examine whether \( \Delta \) is the true value. If \( \Delta = \Delta_0 \), then by the definition of \( \eta \)-level confidence intervals
  - any given such interval should contain \( \Delta \) with probability \( \eta \)
  - on average \( \Delta \) should belong to at least 100\( \eta \)% of the above \( S \) independent intervals.

Thus, we propose to include \( \Delta \) in the 100(1 – \( \alpha \))% level confidence interval \((a, \infty)\) if
\[
t(\Delta, \eta) = \sum_{i=1}^{S} w_i \{ y_i (\Delta, \eta) - \eta \} \geq c
\]
- where \( w_i \) is a study specific positive weight (e.g. sample size)
- \( c \) is chosen such that \( P(T(\eta) < c) \leq \alpha \).
  - \( T(\eta) = \sum_{i=1}^{S} w_i (B_i - \eta) \) is the null counterpart of \( t(\Delta, \eta) \)
  - \( \{B_i, i = 1, \ldots, S\} \) are \( n \) independent Bernoulli random variable with “success” probability of \( \eta \).

We repeat this process will all other possible values for \( \Delta \) and obtain the final interval.

Example
Effect of Rosiglitazone on MI or CVD Deaths

- For RR or OR effect measures
  - unless prior information about the underlying event rates is available, it is not clear how to utilize studies with zero events without continuity correction.

- RD may be used as an alternative effect measure
  - appealing interpretation
  - exact inference may be used

We examine the effect of Rosiglitazone on MI or CVD deaths based on \( \Delta = \text{RD} \) (Rosiglitazone – Control).
Non-parametric Inference for the Random Effects Distribution in Meta Analysis

Random Effects Meta Analysis
Non-parametric Estimation of the Median

- Wang et al (2008) proposed interval estimation procedures for the quantiles of $\beta_i$ without requiring the number of studies to be large.

- Suppose we are interested in estimating the median of $\beta_i$ denoted by $\mu_0$.
  - If $\beta_i$ known, exact confidence interval for $\mu_0$ can be obtained by inverting a sign test:
    \[ T(\mu) = \sum_{i=1}^{S} \{ I(\beta_i < \mu) - 0.5 \} \]
    The null distribution of $T(\mu_0) + S/2$ is a Binomial$(S, 0.5)$. 
Random Effects Meta Analysis
Non-parametric Estimation of the Median

- Alternatively, one may replace \( I(\hat{\beta}_i < \mu) \) with a measure of likelihood for the event \( \beta_i < \mu \)
- Example: the observed coverage level of the interval \((\infty, \mu)\) for \( \beta_i \) which is \( \Phi\{((\mu - \hat{\beta}_i) / \hat{\sigma}_i) \}
- The test statistic based on the coverage level is
\[
\hat{T}(\mu) = \sum_{i=1}^{S} \Phi\{((\mu - \hat{\beta}_i) / \hat{\sigma}_i) - 0.5\]
- Studies with data that are more informative for the event \( \beta_i < \mu \) would yield coverage level closer to either 0 or 1 and thus carry more weight in the test statistic.

Random Effects Meta Analysis
Non-parametric Estimation of the 100\(p^{th}\) Percentile

- For the 100\(p^{th}\) percentile, the test statistic is
\[
\hat{T}_p(\mu) = \sum_{i=1}^{S} [\Phi\{((\mu - \hat{\beta}_i) / \hat{\sigma}_i) - 0.5\}
- Unconditionally \( \hat{T}_p(\mu) \) is asymptotically Binomial(S,p)
- The null distribution of \( \hat{T}_p(\mu) \) can be approximated by
\[
T_p^*(\mu) = \sum_{i=1}^{S} [\Phi\{((\mu - \hat{\beta}_i) / \hat{\sigma}_i) - 0.5\} (2\varepsilon_i - 1)
- \{\varepsilon_i, i=1,\ldots,S\} \sim Bernoulli(p) \) independent of data

Example
Effect of ESA on the Risk Mortality

<table>
<thead>
<tr>
<th>95% Confidence Intervals for the Hazard Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Median</td>
</tr>
<tr>
<td>---</td>
</tr>
<tr>
<td>DL</td>
</tr>
<tr>
<td>(\tilde{T}(\cdot)) with (I(\cdot))</td>
</tr>
<tr>
<td>(\hat{T}(\cdot)) with (\Phi(\cdot))</td>
</tr>
</tbody>
</table>
Data Monitoring in Clinical Trials Using Prediction

Scott Evans
Harvard University
Graybill Conference
June 11, 2008

Practical Questions

- Should the trial be stopped?
  - For efficacy?
  - For futility?

- Are there non-eficacious arms that should be dropped?
  - E.g., if there are several arms (e.g., doses)

Motivation

- Answering these questions has:
  - Ethical attractiveness
    - Fewer patients generally exposed to inefficacious and potentially harmful therapies
  - Economical advantages
    - Smaller expected sample sizes and shorter expected duration than designs without interim analyses
      - Saving time, money, and other resources

Limitations of Many Traditional Methods

- Do not provide
  - Estimates of effect or associated precision (only test statistics, p-values, and decision rules)
    - Cannot evaluate “clinical relevance”
  - Information regarding the reasons for:
    - High p-values (or test statistics):
      - Negligible effect vs. insufficient data
    - Low p-values (or test statistics):
      - Large effect vs. lots of data
Repeated Confidence Intervals

- Simultaneous coverage of all sequential CIs
- Provides estimates of effects sizes
- Allows for flexibility in decision making
  - Can terminate trial based on current CI, but overall confidence level is guaranteed regardless of way that the decision to stop is made

Limitations of Repeated CIs

- At the interim, we wish to weigh the options of stopping vs. continuing
- But we do not know the ramifications of continuing
- Methods do not answer “What will happen if the trial continues?”
  - What effect size estimates and associated precision will be observed if the trial continues?
- Thus how do we weigh the options?

NARC 009

- Randomized, double-blind, placebo-controlled, multicenter, dose-ranging study of prosaptide (PRO) for the treatment of HIV-associated neuropathic pain
- Subjects randomized to 2, 4, 8, 16 mg/d PRO or placebo administered via subcutaneous injection
- Primary endpoint:
  - 6 week change from baseline in weekly average of random daily Gracely pain scale prompts using an electronic diary

Interim Analysis Results: NARC 009

<table>
<thead>
<tr>
<th>Treatment</th>
<th>N</th>
<th>95% CI for Mean Change</th>
<th>95% CI for Diff¹</th>
<th>95% PI for Diff²</th>
<th>95% PI for Diff³</th>
<th>Required Diff⁴</th>
</tr>
</thead>
<tbody>
<tr>
<td>Placebo</td>
<td>31</td>
<td>(-0.35, -0.11)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2 mg</td>
<td>34</td>
<td>(-0.21, -0.04)</td>
<td>(-0.04, 0.25)</td>
<td>(-0.01, 0.21)</td>
<td>(-0.16, 0.06)</td>
<td>-0.54</td>
</tr>
<tr>
<td>4 mg</td>
<td>34</td>
<td>(-0.38, -0.12)</td>
<td>(-0.19, 0.16)</td>
<td>(-0.14, 0.10)</td>
<td>(-0.23, 0.01)</td>
<td>-0.45</td>
</tr>
<tr>
<td>8 mg</td>
<td>32</td>
<td>(-0.18, -0.02)</td>
<td>(-0.01, 0.28)</td>
<td>(0.03, 0.23)</td>
<td>(-0.15, 0.05)</td>
<td>-0.56</td>
</tr>
<tr>
<td>16 mg</td>
<td>36</td>
<td>(-0.34, -0.09)</td>
<td>(-0.16, 0.19)</td>
<td>(-0.11, 0.14)</td>
<td>(-0.21, 0.04)</td>
<td>-0.54</td>
</tr>
</tbody>
</table>

1: 95% CI for the difference in mean changes vs. placebo
2: 95% PI for the difference in mean changes vs. placebo assuming full enrollment, assuming current trend
3: 95% PI for the difference in mean changes vs. placebo assuming full enrollment, assuming per protocol, \( \mu_{\text{placebo}} = -0.17 \) and \( \mu_{\text{drug}} = -0.34 \)
4: Difference in mean changes needed in the remaining patients for the CI for the difference in mean changes to exclude zero (in favor of active treatment) at the end of the trial
Predicted Intervals

- Intuitive
- Practical
  - Provides information on effect size and associated precision (clinical relevance)
  - Invariant to study design (superiority vs. noninferiority)
  - Allows flexibility in decision process
  - Can use for binary, continuous, or time-to-event endpoints
    - 1 and 2 sample problems

PIPs: Schema for Construction

1. Simulate future outcome
2. A simulated complete dataset at the end of the trial
3. Calculate the "final" result
4. Assumption of HR
5. Adjusted PI
6. Simulate many times and get many adjusted PIs

Predicted Interval Plot

18M after the 2nd interim analysis (4Y from the start) Assumed HR = 0.975.

Current Interval [0.88, 1.08]
Length: 0.20

In favor of Comb
In favor of Mono therapy
Sensitivity Analyses

- Perform sensitivity analyses to evaluate the effect of the uncertainty of the model assumption by creating PIPs for various assumed models
  - E.g., observed trend is true, $H_A$ is true, $H_0$ is true, optimistic/pessimistic case scenarios, etc.

CONSORT Diagram

Mean Difference and 95% CI in Response Rates

AE Summary
Should we re-think event-driven trials?

- **Time-to-event endpoints**
  - Power determined by number of events (numerator) due to estimation of relative risk

- **Absolute risk**
  - Powering on absolute risk may be more efficient because it uses denominator data
  - Presentation of absolute risk may be more interpretable than relative risk
  - Can still calculate relative risk

If you were a DMC member, what is your recommendation? Continue? Stop?

**Case #1**
At interim analysis
Relative Risk
[0.3, 158]

HINT: Very wide

**Case #2**
At interim analysis
Absolute Risk Diff.
[-0.04%, 0.12%]

HINT: Very tight. The difference is at most 0.12%
Randomization

Each patient has the same opportunity, as any other patient, of receiving any of the treatments under study.

Allocation of treatments to patients is generally carried out using a chance mechanism so that neither the patient nor the physician know in advance which therapy will be assigned.

Randomized clinical trials are regarded as the best way to carry out clinical trials.

Disadvantages of Randomization

- Patient or physician may not care to participate in experiment involving a chance mechanism to decide treatment.

- May interfere with physician-patient relationship
3. Conditioning on Sample Size

**Aim:** Randomize experimental units to two treatments \((A, B)\).

**Notation:**

\[
N = 2n \quad \text{number of experimental units evaluable}
\]

\[
Y_i : \quad \text{outcome for } i^{th} \text{ experimental unit (not random)}
\]

\[
\delta_i = \begin{cases} 
1 & \text{if } i^{th} \text{ unit assigned to } A \\
0 & \text{otherwise}
\end{cases}
\]

---

### Two Methods of Randomization

<table>
<thead>
<tr>
<th>{\delta_i}</th>
<th>I</th>
<th>II</th>
</tr>
</thead>
<tbody>
<tr>
<td>(P{\delta_i = 1})</td>
<td>independent</td>
<td>dependent</td>
</tr>
<tr>
<td>(\text{var } \delta_i)</td>
<td>(\frac{1}{2})</td>
<td>(\frac{1}{4})</td>
</tr>
<tr>
<td>(\text{cov} (\delta_i, \delta_j, i \neq j))</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>(E(S_a))</td>
<td>(NY/2)</td>
<td>(NY/2)</td>
</tr>
<tr>
<td>(V(S_a))</td>
<td>(\frac{1}{4} \sum \limits_{i=1}^{N} Y_i^2)</td>
<td>(\frac{1}{4} \sum \limits_{i=1}^{N} \frac{(Y_i - \bar{Y})^2}{(N - 1)})</td>
</tr>
</tbody>
</table>

\[
\frac{N}{4} \sum \limits_{i=1}^{N} \frac{(Y_i - \bar{Y})^2}{N - 1} \leq \frac{1}{4} \sum \limits_{i=1}^{N} Y_i^2 \quad \text{for nearly all } N
\]

---

### Cluster Sampling

Up to now we have considered randomization in which an individual experimental unit (patients) are randomized to treatments.

In some situations it is more feasible to randomize treatments to groups of experimental units; i.e., groups may be patients in a hospital, families, geographic regions. Cluster randomization or cluster sampling is used to describe this kind of random assignment.

We will show that cluster and individual randomization are equivalent if the outcomes have low probabilities of occurring.
Is asymptotic $N(0, 1)$

Where $p = \frac{n_a}{N}$, $q = 1 - p$, $S = \Sigma Y_i$

Alternatively, suppose $S_a, S_b$ are independent Poisson random variables with

$$E(S_a) = N_a \lambda, \quad E(S_b) = N_b \lambda.$$  

Then the distribution of $S_a$, conditional on $S = S_a + S_b$ is binomial with

$$E(S_a | S) = pS, \quad V(S_a | S) = pqS$$

**Conclusion**: Cluster Sampling from a Poisson distribution is equivalent to individual randomization.

---

### Idealized Clinical Trial Process

- **Population with Disease**
- **Random Sample of Patients**
- **Randomization**

**Idealized Clinical Trial Process**

- Nearly all clinical trials do not have a random sample of patients.
- Only a local inference can be made unless additional assumptions are assumed.
- The randomization process can serve as the basis for making a local inference.
- In most situations power will be increased as the inference is more narrow.

---

**Investigations of Local Inference : Power**

- Outcome variables: continuous, binomial, survival
- Randomized multi center trials (two treatments)
- Methodology conditions on the number of patients randomized within a center (conditions on ancillary statistic)
- Study design is permuted blocks of size four:
  - Sample size for each treatment is equalized for every fourth patient entered on trial.
- Group sequential and fixed sample size.
Table 2: Power Comparison for Survival Outcomes (N = 4)

<table>
<thead>
<tr>
<th>Sample Size</th>
<th>120</th>
<th>360</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. of Institutions</td>
<td>10</td>
<td>20</td>
</tr>
<tr>
<td>Conditional test (Gehan)</td>
<td>.38</td>
<td>.33</td>
</tr>
<tr>
<td>Stratified Gehan</td>
<td>.19</td>
<td>.14</td>
</tr>
<tr>
<td>Conditional test (Logrank)</td>
<td>.37</td>
<td>.32</td>
</tr>
<tr>
<td>Stratified Logrank</td>
<td>.20</td>
<td>.14</td>
</tr>
</tbody>
</table>

NOTE: α = .05, two-sided test.

Data are generated from exponential distribution. Ratio of two hazards is 2. Percentage of censored observations is 18.5%.
Causal Inference in Clinical Trials

Lingling Li
Department of Ambulatory Care and Prevention
Harvard Medical School
June 11, 2008

Motivating Example I

- The Study to Understand Prognoses and Preferences for Outcomes and Risks of Treatments (SUPPORT)
- Examine the effect of the use of right heart catheterization (RHC) on survival time up to 30 days
- Observational study with 5735 critically ill adult patients
- 2184 patients received RHC

Motivating Example I

For those 5735 patients, data available on
- A dichotomous treatment indicator variable \( A \)
  - \( A = 1 \) indicates receiving RHC during the initial 24 hours of an ICU stay (2184)
- Response variable \( Y \): survival time up to 30 days
- A vector of baseline covariates \( L \) including age, years of education, income, weight, blood pressure etc.

Motivating Example I

- Confounding: decision to use or withhold RHC was left to the discretion of the physician based on patient characteristics
- Analyzed using different methods in several published papers (Connors et al., 1996; Lin et al., 1998; Tan, Z., 2006)
- Lack of evidence suggesting beneficial effects of RHC on the 30-day survival
- The appearance of a harmful effect could be explained away by unmeasured confounding
Motivation Example II

- The AIDS Clinical Trial Group (ACTG) study 021, a double-blind randomized clinical trial
- Compare the effect of bactrim versus aerosolized pentamidine (AP) as prophylaxis therapy for pneumocystis pneumonia (PCP) in AIDS patients
- Primary endpoint: time to recurrent PCP
- Cross over to the other treatment allowed if PCP developed
- Secondary endpoint: time to death

310 patients accrued, 94 died, the remaining 216 patients either alive at the end of the trial or dropped out in the middle
- Of the 94 deaths, 21 crossed over, 37 stopped all prophylactic therapy (21 for non-medical reasons and 16 for medical indications)
- Suppose we are interested in the survival status at the end of the trial:
  - $Y = 1$ if dead at the end of the trial
  - $Y = 0$ if alive at the end of the trial

Outline

- Introduction to Causality
- Confounding in Observational Studies
- Dependent Censoring in Randomized Trials
- Sensitivity Analysis for Nonignorable Censoring
Outline

- Introduction to Causality
- Confounding in Observational Studies
- Dependent Censoring in Randomized Trials
- Sensitivity Analysis for Nonignorable Censoring

Introduction

There are generally two notions of causation:

- Cause of an effect: first observe an event/outcome, and subsequently identify the causes or events that lead to the observed outcome
- Effect of a cause: assess the effect of a well defined exposure or intervention. e.g. does smoking cause lung cancer? does azidothymidine (AZT) prevent the advent of AIDS among HIV infected patients?

Introduction

An example of (1):
- In the 80s, unusual high number of patients dying from a combination of syndromes including a rare Kaposi’s skin cancer and pneumonia
- Later, HIV found to be the cause

Limited to (2) in this lecture with our focus on challenges in clinical trials

Introduction

Why do we need a formal theory of causation?

- Make it explicit what we mean by "causal effect", i.e., what is the quantity/estimand we seek?
- Give explicit assumptions under which "association is causation", and therefore standard statistical methods may be used
- Give explicit assumptions needed for the identification of causal effects even when "association is not causation"
- Allow for the derivation of new statistical methods when standard and familiar methods fail
Counterfactuals

- Suppose you are contemplating taking an aspirin for your headache, and the outcome $Y$ denotes whether or not you are headache free within say the next hour.
- As a thought experiment, think of two potential outcomes
  
  $Y_0$ : headache outcome after not taking aspirin
  $Y_1$ : headache outcome after taking aspirin
- Note everything else remains exactly the same
- Either of the two will be observed depending on whether or not you decide to take the aspirin, but never both

Counterfactuals

- $Y_3$ is the outcome that you would observe if, possibly contrary to fact, you followed treatment $a \in \{0, 1\}$.
- The English sentence "aspirin has no causal effect on my headache" is the mathematical statement about my potential outcomes:
  
  $Y_1 = Y_0$.
- Suppose larger $Y$ indicates better outcome, an individual has a beneficial causal effect of aspirin if $Y_1 > Y_0$; or has a harmful causal effect of aspirin if $Y_1 < Y_0$.

Counterfactuals

**Consistency assumption:** the observed outcome $Y$ satisfies

$$ Y = A Y_1 + (1 - A) Y_0 $$

<table>
<thead>
<tr>
<th>ID</th>
<th>A</th>
<th>Y</th>
<th>$Y_0$</th>
<th>$Y_1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>?</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>?</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>1</td>
<td>?</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>1</td>
<td>?</td>
<td>1</td>
</tr>
</tbody>
</table>

- Note that since both outcomes are never simultaneously observed, it is impossible to evaluate individual causal effects.

Counterfactuals

**Remark:** the mere definition of the potential variable $Y_3$ carries the so-called assumption of no-interference between units (i.e., Stable unit treatment value assumption). Under this assumption, the value of the outcome of one subject who receives treatment $a$, is unaffected by the treatments received by the other subjects in the population.
- As an example, the potential outcome would be ill-defined if $A = 0$ was placebo and $A = 1$ was treatment with a vaccine for a highly contagious disease. Obviously, if the vaccine is effective, the value of $Y_0$ would depend on whether or not the contacts of the person get the vaccine.
Association vs Causal Measures in the Population

- The variables we can hope to observe are \((A, Y)\).

<table>
<thead>
<tr>
<th>ID</th>
<th>A</th>
<th>Y</th>
<th>Y_0</th>
<th>Y_1</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

- **Average causal effect** (a causal effect measure in the population)

\[
\psi \equiv E(Y_1) - E(Y_0) = 3/4 - 3/4 = 0
\]

- **Difference of conditional expectations** (an association measure)

\[
\zeta = E(Y|A=1) - E(Y|A=0) = 2/2 - 1/2 = 1/2
\]

Counterfactuals

- **Technical interlude**: in the expressions \(E(Y|A=1), E(Y|A=0), E(Y_1), E(Y_0)\), the random variables \(Y, (Y_1, Y_0)\) are to be understood as the outcome and potential outcomes of a person chosen at random from the finite population of four individuals. The expectations, of course, coincide with the averages of the values of the variables in the four members of the finite population. Later on, we will regard these variables as arising from one random draw from an infinite population and the expectations will be understood as the average of the values of the variables in the infinite members of the population.

Randomization

- Suppose you randomize the population of patients to either aspirin with probability \(p > 0\) or to no aspirin with probability \(1 - p > 0\). Then, with \(\perp\) denoting independence, it holds that

\[Y_a \perp A, \text{ for } a = 0, 1\]

because, \(Y_1\) and \(Y_0\) are, like gender and age, pretreatment variables.

- In such case we have for \(a = 0, 1\), that

\[
P(Y_a = 1) \overset{\text{by randomization}}{=} P(Y_a = 1|A = a) \overset{\text{by consistency}}{=} P(Y = 1|A = a)
\]

\(0 < p < 1\)
Randomization

- Thus, the probability distribution of the counterfactuals $Y_a$, $a = 0, 1$, can be written in terms of the distribution of the observed data $(Y, A)$ and hence it is identified.
- Note that randomization does not imply $Y \perp A$ since $Y = AY_1 + (1 - A)Y_0$ is determined by treatment and therefore is a post-treatment variable.
- In fact, under randomization,

$$Y \perp A \iff Y_1 \overset{D}{=} Y_0 \ (Y_1 \text{ and } Y_0 \text{ equal in distribution})$$

Motivating Example I

- The Study to Understand Prognoses and Preferences for Outcomes and Risks of Treatments (SUPPORT)
- Examine the effect of the use of right heart catheterization (RHC) on survival time up to 30 days
- Observational study with 5735 critically ill adult patients
- 2184 patients received RHC

Motivating Example I

For those 5735 patients, data available on

- A dichotomous treatment indicator variable $A$
  - $A = 1$ indicates receiving RHC during the initial 24 hours of an ICU stay (2184)
- Response variable $Y$: survival time up to 30 days
- A vector of baseline covariates $L$ including age, years of education, income, weight, blood pressure etc.
- Decision to use or withhold RHC was left to the discretion of the physician based on patient characteristics

Outline

- Introduction to Causality
- Confounding in Observational Studies
- Dependent Censoring in Randomized Trials
- Sensitivity Analysis for Nonignorable Censoring
Observational Study

- Randomization not applicable due to ethical and practical reasons
- Observed data comes from a cross-sectional observational study
- A random sample from \((L, A, Y)\) where \(L\) is a vector of pre-treatment covariates.

- **No unmeasured confounders assumption (NUCA):**
  \[ Y_a \perp A \mid L, \quad \text{for } a = 0, 1 \]
  
  i.e. \(Y_a\) and \(A\) are conditionally independent given \(L\).

The G-formula.

- **Theorem 1:** if (i) consistency, (ii) NUCA and (iii) (Positivity) \(P(A = a \mid L) > 0\) (CNP) hold w.p. 1, the distribution of \(Y_a\) is identified and it satisfies

  \[
  p^a(y) = \int p(y \mid l, a) \, dF(l)
  \]

- \(p^a \equiv p_{Y_a}(y)\) : the density (with respect to some measure \(\mu\)) of \(Y_a\) at \(y\)
- \(p(y \mid l, a) \equiv p_{Y \mid L, A}(y \mid l, a)\) : the conditional density of \(Y\) (at \(y\)) given \(L = l\) and \(A = a\)
- \(F(l)\) : the c.d.f. of \(L\) at \(l\)

The G-formula.

- **Proof:** for \(a = 0, 1\)
  \[
  p^a(y) \equiv p_{Y_a}(y) = \int p_{Y _a \mid L}(y \mid l) \, dF(l)
  \]
  
  \[
  = \int p_{Y _a \mid L, A}(y \mid l, a) \, dF(l) \quad \text{by randomization and positivity}
  \]
  
  \[
  = \int p_{Y _a \mid L, A}(y \mid l, a) \, dF(l) \quad \text{by consistency}
  \]
Crude vs Counterfactual Means

- In general, the crude mean \( E(Y|A=a) \) is not equal to the counterfactual mean \( E(Y_a) \).
  - \( E(Y_a) \): the mean of the outcome \( Y \) if, contrary to the fact, everyone in the population was forced to take treatment \( A=a \).
  - \( E(Y|A=a) \): the mean of the outcome \( Y \) among the sub-population who chose to take treatment \( A=a \).

\[
E(Y_a) \equiv E[E(Y|A=a,L)] \tag{1}
\]
\[
E[E(Y|A=a,L)] \equiv E(E(Y|A=a,L)) \tag{2}
\]
\[
E\left(\frac{I(A=a)}{\pi(a,L)}Y\right) \tag{3}
\]

where \( \pi(a,L) \equiv p(A=a|L) \), the so-called propensity score (PS).

Propensity Score Adjustments

Equity (3)

\[
E\left(E(Y|A=a,\pi(a,L))\right) \equiv E\left(\frac{I(A=a)}{\pi(a,L)}Y\right)
\]
holds since

\[
E\left(\frac{I(A=a)}{\pi(a,L)}Y\right) = E\left(E\left(\frac{I(A=a)}{\pi(a,L)}Y|A=a,\pi(a,L)\right)\right)
\]
\[
= E\left(E\left(\frac{I(A=a)}{\pi(a,L)}Y|A=a,\pi(a,L)\right)\right)
\]
\[
= E\left(E\left(\frac{I(A=a)}{\pi(a,L)}\right)Y|A=a,\pi(a,L)\right)
\]
\[
= E\left(E\left(\frac{I(A=a)}{\pi(a,L)}\right)Y|A=a,\pi(a,L)\right)
\]
\[
= E\left(E\left(E(Y|A=a,\pi(a,L))\right)\right)
\]
Propensity Score Adjustments

Equity (2)

\[ E( Y \mid A = a, L) \mid \pi \mid \pi (a, L) \]

holds since equity (3) holds and

\[
E \left( \frac{I (A = a)}{\pi (a, L)} Y \right) = E \left( \frac{E[I (A = a)]}{\pi(a, L)} Y \mid A = a, L \right) \\
= E \left( \frac{E[I (A = a)]}{\pi(a, L)} \right) E \left[ Y \mid A = a, L \right] \\
= E \left( \frac{E[I (A = a)]}{\pi(a, L)} \right) E \left[ Y \mid A = a, L \right] \\
= E \left( E \left[ Y \mid A = a, L \right] \right)
\]

\( Y_a \perp A \mid L \Rightarrow Y_a \perp \pi (a, L) \) (Rosenbaum and Rubin, 1984)

Stratification by the Propensity Score

- Exploiting the formula

\[ E(Y_a) = E \left[ E \left( Y \mid \pi (a, L), A = a \right) \right] \]

Rosenbaum, Rubin and co-authors in a number of papers dating back to the mid-80’s have proposed the following estimator of \( E(Y_a) \).

1. Specify a model \( \pi (L; \alpha) = \pi (1, L; \alpha) \) for the conditional probability of taking treatment \( p_{A \mid L} (1) \), say \( \pi (L; \alpha) = \left( 1 + \exp \left( -\alpha^T L \right) \right)^{-1} \) indexed by a finite dimensional vector \( \alpha \). Estimate \( \alpha \) with its ML estimator \( \hat{\alpha} \) using data \( (A_j, L_j) \) \( i = 1, \ldots, n \).
2. Stratify the units into, say quintiles, of the estimated propensity scores \( \pi (L; \hat{\alpha}) \), \( i = 1, \ldots, n \).
3. Estimate the mean of \( Y \) among those having \( A = a \), separately for each stratum.
4. Compute the weighted average of the means estimated in 3) with weights equal to the proportion of subjects of the sample in each stratum.

Inverse Probability Weighting (IPW) Estimators

- Exploiting the formula

\[ E(Y_a) = E \left[ \frac{I(A = a)}{p_{A \mid L}(a \mid L)} Y \right] \]

we can construct the following estimator

1. Specify a model \( \pi (a, L; \alpha) \) for the conditional probability of taking treatment \( p_{A \mid L} (a \mid L) \), say \( \pi (a, L; \alpha) = \left( 1 + \exp \left( -\alpha^T L \right) \right)^{-1} \) if \( a \in \{ 0, 1 \} \). Estimate \( \alpha \) with its ML estimator \( \hat{\alpha} \) using data \( (A_j, L_j) \) \( i = 1, \ldots, n \).
2. Construct an estimator of \( E(Y_a) \) using \( \mathbb{P}_n \left\{ \frac{I(A = a)}{\pi(a, L; \alpha)} Y \right\} \) where \( \mathbb{P}_n (Z) \equiv n^{-1} \sum_{i = 1}^n Z_i \) for any \( n \) iid copies \( Z_1, \ldots, Z_n \) of \( Z \).
3. Naturally, \( \hat{\psi}_{ipw} \equiv \mathbb{P}_n \left\{ \frac{I(A = 1)}{\pi(1, L; \alpha)} Y \right\} - \mathbb{P}_n \left\{ \frac{I(A = 0)}{\pi(0, L; \alpha)} Y \right\} \) is a consistent estimator of \( \psi = E(Y_{a=1}) - E(Y_{a=0}) \).
IPW Estimator: A Heuristic Explanation

Suppose that $L$ is a binary baseline covariate, with $L = 1$ indicating bad health.

1. Suppose there are 100 subjects with $L = 1$ and another 100 subjects with $L = 0$

2. 80 subjects with $L = 1$ assigned to $A = 1$ and $E(Y|L = 1, A = 1) = 30$; the remaining 20 assigned to $A = 0$ and $E(Y|L = 1, A = 0) = 20$

3. 30 subjects with $L = 0$ assigned to $A = 1$ and $E(Y|L = 0, A = 1) = 70$; the remaining 70 assigned to $A = 0$ and $E(Y|L = 0, A = 0) = 60$

Under $Y_2 \perp A|L$, to estimate $E(Y_{3=1})$, each of the 80 subjects with $L = 1$ and $A = 1$ counts for themselves and for the other 20 subjects with $L = 1$ but assigned to $A = 0$ ($1 + 20/80 = \frac{1}{3}$); similarly, each of the 30 subjects with $L = 0$ and $A = 1$ counts for themselves and for the other 70 subjects with $L = 0$ and $A = 0$ ($1 + 70/30 = \frac{1}{3}$)

Ratio Estimators

$E(Y_a) = E\left(\frac{I(A=a)}{\hat{p}_{A|L}(a|L)}Y\right)$ can be re-written as

$$E(Y_a) = \left[ E\left(\frac{I(A=a)}{\hat{p}_{A|L}(a|L)}\right)\right]^{-1} E\left(\frac{I(A=a)}{\hat{p}_{A|L}(a|L)}Y\right)$$

Suppose $\pi(a, L; \hat{\alpha})$ is correctly specified, $E(Y_a)$ can be estimated by

$$\hat{E}(Y_a) = \hat{p}_n \left\{ \frac{I(A = a)Y}{\pi(a, L; \hat{\alpha})} \right\} \left(\hat{p}_n \left\{ \frac{I(A = a)}{\pi(a, L; \hat{\alpha})} \right\}\right)^{-1}$$

$\hat{\psi}_{ipw2} \equiv \hat{E}(Y_1) - \hat{E}(Y_0)$

$\hat{\psi}_{ipw2}$ is generally more efficient than $\hat{\psi}_{ipw1}$

Known as a ratio estimator in the sampling literature (Horvitz and Thompson, 1952)

IPW Estimators with Augmentations

In fact, in the model that a parametric form $\pi(a, L; \alpha)$ for $\hat{p}_{A|L}(a|L)$ is assumed, there exists a class of IPW estimators (with augmentations) of the form

$$\hat{E}(Y_a)_{ipw, d} = \frac{\hat{p}_n \left\{ \frac{I(A = a)}{\pi(a, L; \alpha)}Y \right\}}{\hat{p}_n \left\{ \frac{I(A = a)}{\pi(a, L; \alpha)} \right\} + d(L)}$$

$\hat{\alpha}$ is the ML estimator
$d(\cdot)$ is an arbitrary function of $L$
When $\pi(a, L; \alpha)$ is correct, the augmentation term converges to 0 in probability for any $d$.

Remarks

The IPW (augmented) estimators, and the estimator based on PS stratification will tend to give highly variable and unreliable estimators when the propensity scores are close to 1 or to 0 for some values of $L$.

This problem is often referred to in the literature as the problem of poor overlap of the propensity scores between the treated and untreated.
Remarks

- The regression estimator based on a parametric model, $\phi(a, L; \eta)$, for $E(Y|A = a, L)$ does not suffer from this problem because it extrapolates from the fitted model $\hat{E}(Y|A = a, L) = \phi(a, L; \hat{\eta})$ to produce estimates of $E(Y|A = a, I)$ for values of $I$ with $\pi(a, I) \approx 0$.
- However, the estimation now strongly depend on the validity of the regression model $\phi(a, L; \eta)$.

Propensity Score Matching

- Rosenbaum and Rubin (1983) first proposed this approach to estimate the average treatment effect on the treated, i.e.,
  \[ E(Y_{a=1}|A = 1) - E(Y_{a=0}|A = 1). \]
- Specify a model $\pi(L; \alpha) = \pi(1, L; \alpha)$ for the conditional probability of taking treatment $P_{A|L}(1|L)$, say $\pi(L; \alpha) = (1 + \exp(-\alpha^T L))^{-1}$ indexed by a finite dimensional vector $\alpha$. Estimate $\alpha$ with its ML estimator $\hat{\alpha}$ using data $(A_i, L_i) i = 1, ..., n$, and $\bar{Y}_i = \pi(L_i; \hat{\alpha})$
- Match each treated subject to one or more untreated subjects on propensity score based on certain distance (Rosenbaum, 2002)
- The estimator can be constructed as
  \[ \hat{\psi}_{PSM} = \frac{\sum_i I(A_i = 1) \{Y_i - \bar{Y}_{m,i}\}}{\sum_i I(A_i = 1)}, \]
  where $\bar{Y}_{m,i}$ is the average outcome of a few subjects in the untreated group (the "controls") with propensity scores close to $\bar{Y}_i$.

Propensity Score Matching

- Suppose $\hat{\alpha}$ converges in probability to $\alpha^*$
- Suppose model $\pi(L; \alpha)$ is correctly specified, $\hat{\psi}_{PSM}$ converges in probability to
  \[ E\{[E(Y|A = 1, \pi(L; \alpha^*)) - E(Y|A = 0, \pi(L; \alpha^*))]|A = 1\} \]
  \[ = E\left\{\left[\frac{E(Y_{a=1}|A = 1, \pi(L; \alpha^*))}{E(Y_{a=0}|A = 1, \pi(L; \alpha^*))}\right]|A = 1\right\} \]
  \[ = E(Y_{a=1} - Y_{a=0}|A = 1) \]

Confounding in Observation Studies

- Identify causation from association
- Standard regression
- Propensity Score (PS) adjustments
  - PS stratification
  - Inverse probability weighting (IPW)
  - PS matching
Outline

- Introduction to Causality
- Confounding in Observational Studies
- Dependent Censoring in Randomized Trials
- Sensitivity Analysis for Nonignorable Censoring

Motivation Example II

- The AIDS Clinical Trial Group (ACTG) study 021, a double-blind randomized clinical trial
- Compare the effect of bactrim versus aerosolized pentamidine (AP) as prophylaxis therapy for pneumocystis pneumonia (PCP) in AIDS patients
- Our endpoint of interest: the survival status at the end of the trial

Motivation Example II

- Informative drop-out: patients stopped prophylactic therapies for non-medical reasons and medical indications
- Nonrandom nonadherence: cross over to the other treatment arm if PCP developed
- Causal effect: the effect of two prophylaxis therapies on the survival rate at the end of the trial, if everyone followed the assigned therapy and didn’t stop the therapy unless for toxicity
- (Artificially) regard subjects as dependently censored at the first time a subject stops therapy due to reasons other than toxicity or switches therapy

Motivation Example II

- Appropriate palliative therapies available to combat the toxicity
- Causal effect: the effect of two prophylaxis therapies on the survival rate at the end of the trial, if everyone followed the assigned therapy and never stopped the therapy
- (Artificially) regard subjects as dependently censored at the first time a subject stops therapy due to any reason or switches therapy
A Randomized Clinical Trial with Missing Data

- Suppose we have i.i.d. observations \( O_i = (R_i, Y_i, A_i, L_i) \) \( i = 1, 2, \ldots n \).
- \( L_i \) : a vector of baseline covariates
- \( A_i \) : a dichotomous variable indicating the randomly assigned treatment arm, \( A_i \in \{0, 1\} \)
- \( Y_i \) : continuous outcome measured at the end of a fixed follow-up period
- \( R_i \) : missing indicator, \( Y_i \) observed if \( R_i = 1 \)

Missing Completely at Random (MCAR)

- MCAR:
  \[ p(R = 1|Y, A, L) = p(R = 1|A, L) \iff R \perp Y|A = a, L \]
  \[ p_{Y|R=1,A,L}(y|R=1, a, l) = p_{Y|A,L}(y|a, l) \]
- Further, as \( A \perp L \) and \( p_{Y_a}(y) = p_{Y|A=a}(y) \) by randomization,
  \[ E[E(Y|R=1, A=a, L)] \]
  \[ = \int p(Y|A=a, l) p_{L}(l) \, d\mu(l) \]
  \[ = \int p(Y|A=a, l) A_{L|A}(l|a) \, d\mu(l) \]
  \[ = E(Y|A=a) \text{ rand.} = E(Y_a) \]
- Standard complete case analysis is valid

Motivation Example II

- Per-protocol analysis valid if the probability of being censored is independent of the outcome given the treatment assignment and some pre-treatment covariates
- Apparently NOT true
  - stopping therapy due to medication indications and other reasons
  - switching therapy if PCP developed
  - PCP status and the medication indications likely to affect the survival status

A Randomized Clinical Trial with Missing Data

- Suppose we have i.i.d. observations \( O_i = (R_i, Y_i, A_i, L_i, W_i) \) \( i = 1, 2, \ldots n \).
- \( L_i \) : a vector of baseline covariates
- \( A_i \) : a dichotomous variable indicating the randomly assigned treatment arm, \( A_i \in \{0, 1\} \)
- \( W_i \) : a vector of post-treatment covariates correlated with \( Y_i \) and always observed
- \( Y_i \) : continuous outcome measured at the end of a fixed follow-up period
- \( R_i \) : missing indicator, \( Y_i \) observed if \( R_i = 1 \)
Missing at Random (MAR)

- MAR:
  \[ p(R = 1|Y, A = a, L, W) = p(R = 1|A = a, L, W) \]

- Positivity: \( \Pr (R = 1|A = a, L, W) > 0 \) w.p. 1

- However, in general \( \tilde{W}_1 \equiv (L, W) \) II A DOES NOT hold.

  Equivalently,
  \[
  E \left[ E(Y|R = 1, A = a, \tilde{W}_1) \right] = \int p_{Y|A, \tilde{W}_1}(y|a, \tilde{w}_1) p_{\tilde{W}_1}(\tilde{w}_1) d\mu(\tilde{w}_1) \\
  \neq \int p_{Y|A, \tilde{W}_1}(y|a, \tilde{w}_1) p_{\tilde{W}_1}(\tilde{w}_1|a) d\mu(\tilde{w}_1) = E(Y_a)
  \]

- Standard regression using complete cases will be biased.

A "Missing Data" Problem

- As thoroughly discussed in Robins et al. (1994), the estimation of average treatment effect \( \psi = E(Y_{a=1}) - E(Y_{a=0}) \) can be viewed as a "missing data" problem.

- Suppose we want to estimate \( E(Y_{a=1}) \)
  - "full data": \( (Y_{a=1}, A, L) \)
  - \( Y_{a=1} \) is observed for the subgroup with \( A = 1 \), but missing for the other subgroup with \( A = 0 \)
  - Under NUCA (i.e., \( Y_a \equiv A|L \)), \( p(A = 1|Y_{a=1}, L) = p(A = 1|L) \)

- Similarly, for \( E(Y_{a=0}) \)
  - "full data": \( (Y_{a=0}, A, L) \)
  - \( Y_{a=0} \) is observed for the subgroup with \( A = 0 \), but missing for the other subgroup with \( A = 1 \)

Inverse Probability Censoring Weighting (IPCW)

- Suppose we are interested in \( \tau_1 = E(Y_{a=1}) \)

- By randomization, \( p_{Y_{a=1}}(y) = p_{Y|A=1}(y) \)

- \( p_{Y|A,R}(y|A = 1, R = 1) \neq p_{Y|A,R}(y|A = 1, R = 0) \)

- \( p_{Y|A,R,\tilde{W}_1}(y|A = 1, R = 1, \tilde{w}_1) = p_{Y|A,R}(y|A = 1, R = 0, \tilde{w}_1) \)

- Within each level of \( \tilde{W}_1 \), by assumptions, the distribution of \( Y \) among those with \( R = 0 \), though not observed in our data, equals the distribution of \( Y \) among those with \( R = 1 \).
Inverse Probability Weighting (IPW) Estimators

- Exploiting the formula: \( E(Y_a) = E \left( \frac{f(A=a)}{p_{A|L}(a|L)} Y \right) \)
- \( \hat{E}_1(Y_a) \equiv \mathbb{P}_n \left\{ \frac{f(A=1)}{p_{A|L}(a|L)} Y \right\} \)
- \( \hat{E}_2(Y_a) = \mathbb{P}_n \left\{ \frac{f(A=a)Y}{p_{A|L}(a|L,a)} \right\}^{-1} \) as \( E \left( \frac{f(A=a)}{p_{A|L}(a,\tilde{a})} \right) = 1 \)
- ...

IP(T)W Estimators with Augmentations

Recall the class of IP(T)W estimators (with augmentations)
\[
\hat{E}(Y_{a_{ipw,d}}) = \frac{\mathbb{P}_n \left\{ \frac{f(A=a)}{p_{A|L}(a|L)} Y \right\}}{\mathbb{P}_n \left\{ \frac{f(A=a)}{p_{A|L}(a|L,a)} \right\}}
\]

- The parameter of interest \( E(Y_a) = E \left( \frac{f(A=a)}{p_{A|L}(a|L)} Y \right) \) under CNP
- \( \pi \left( a, L; \tilde{a} \right) \) is an assumed parametric form for \( p_{A|L}(a|L) \) and \( \tilde{a} \) is the ML estimator using \( (A_i, L_i) \) \( i = 1, ..., n \)
- \( d(\cdot) \) is an arbitrary function of \( L \)
- When \( \pi \left( a, L; \tilde{a} \right) \) is correct, the augmentation term converges to 0 in probability for any \( d \).

IPCW Estimators with Augmentations

- Similarly, a class of IPCW estimators with augmentation terms
  \[
  \hat{H}_{1,pcw,d} = \frac{\mathbb{P}_n \left\{ \frac{R}{\pi_1(W_1;\tilde{a})} Y \right\}}{\mathbb{P}_n \left\{ \frac{R}{\pi_1(W_1;\tilde{a})} \right\} - \mathbb{P}_n \left\{ \frac{R}{\pi_1(W_1;\tilde{a})} - 1 \right\} \}
  \]
- The parameter of interest
  \( E(Y_{a-1}) = E \left( \frac{R}{\pi_1(W_1;\tilde{a})} Y \right) \) under MAR and positivity
- \( \pi_1 \left( W_1; \tilde{a} \right) \) is an assumed parametric form for \( \text{Pr} \left( R = 1 | W_1, A = 1 \right) \)
- \( \tilde{a} \) is the ML estimator using realizations of \( (R, W_1) \) in the treated group only
- \( d(\cdot) \) is an arbitrary function of \( W_1 \)
- When \( \pi_1 \left( W_1; \tilde{a} \right) \) is correct, the augmentation term converges to 0 in probability for any \( d \).
Dependent Censoring in a Longitudinal Study

- Suppose the surrogate variables \( W \) are to be measured at fixed times \( k = 1, 2, \ldots, K - 1 \).
- For any \( k \leq K - 1 \), let \( \mathcal{W}_{k,i} = (W_{0,i} = L_i, W_{1,i}, \ldots, W_{k-1,i}) \), the \( W \)–history for subject \( i \) up to but not including the \( k \)th occasion.
- The outcome \( Y_i \) measured at time \( K \), the end of follow-up.
- \( R_{k,i} = 1 \) if subject \( i \) remains in the follow-up at time \( k \); and 0 otherwise.
- Monotone missing pattern, i.e., \( R_{k,i} = 0 \Rightarrow R_{s,i} = 0 \) for any \( s \geq k \).

For any \( 1 \leq k \leq K \),
- Sequentially ignorable missingness (SIM):
  \[
  \Pr(R_k = 1|R_{k-1} = 1, \mathcal{W}_k, A, Y) = \Pr(R_k = 1|R_{k-1} = 1, A, \mathcal{W}_k)
  \]
- Positivity: with probability 1,
  \[
  \lambda_k (a, \mathcal{W}_k) \equiv \Pr(R_k = 1|R_{k-1} = 1, \mathcal{W}_k, A = a) \geq \sigma > 0
  \]

\[
\pi (a, \mathcal{W}_K) \equiv (R_K = 1|\mathcal{W}_K, A = a) = \prod_{k=1}^{K} \lambda_k (a, \mathcal{W}_k).
\]

- \( E(Y_a) = E(Y_a|A = a) = E \left( \frac{R}{\pi(a, \mathcal{W}_K)} Y|A = a \right) \) under (i)
  Consistency, (ii) SIM, and (iii) Positivity

Outline

- Introduction to Causality
- Confounding in Observational Studies
- Dependent Censoring in Randomized Trials
- Sensitivity Analysis for Nonignorable Censoring
Back to the MAR Setting

- Suppose we have i.i.d. observations $O_i = (R_i, Y_i, A_i, \bar{W}_{1,i} = (L_i, \bar{W}_i))$ for $i = 1, 2, \ldots, n$.
- $L_i$: a vector of baseline covariates
- $W_i$: a vector of post-treatment covariates correlated with $Y_i$ and always observed
- $A_i$: a dichotomous variable indicating the randomly assigned treatment arm, $A_i \in [0, 1]$.
- $Y_i$: continuous outcome measured at the end of a fixed follow-up period
- $R_i$: missing indicator, $Y_i$ observed if $R_i = 1$

Nonignorable Censoring

- Ignorable censoring: $p(R = 1|Y, A = a, \bar{W}_1) = p(R = 1|A = a, \bar{W}_1)$
- Nonignorable censoring: $p(R = 1|Y, A = a, \bar{W}_1) \neq p(R = 1|A = a, \bar{W}_1)$
- Because of ignorable, the likelihood for one individual is given by

$$
\begin{align*}
\mathcal{L} &= \left\{ \left[ p_{\bar{Y}|R, \bar{W}_1, A = a} (y|Y = 1, \bar{W}_1) \right]^{R} p_{\bar{W}_1|A = a} (\bar{W}_1) \right\} \\
& \quad \times \left( \pi (a, \bar{W}_1; \alpha) \right)^{R} (1 - \pi (a, \bar{W}_1; \alpha))^{1-R}
\end{align*}
$$

- The parameter of interest $\tau = E(Y|A = a)$ depends on $\eta_1$ and $\eta_2$.

Nonignorable Censoring

- $p(R = 1|Y, A = a, \bar{W}_1) = \pi(a, \bar{W}_1, Y; a)$ instead of $\pi(a, \bar{W}_1; a)$
- For example, $\pi(a, \bar{W}_1, Y; a) = \exp \{ a^T \bar{W}_1 + \gamma Y \}$
- $\gamma = 0 \iff$ ignorable censoring
- $\gamma$: a measure of the correlation between the censoring indicator $R$ and the outcome $Y$ after adjusting for $\bar{W}_1$
- Larger value of $\gamma$ implies more hidden bias

Sensitivity Analysis

For any given value $\gamma$,

- ML estimator $\hat{\alpha}_{ML}$ not obtainable as the following likelihood depends on the unobserved outcome

$$
L = \prod_{i=1}^{n_2} \frac{e^{(a^T \bar{W}_{1,i} + \gamma Y_i) R_i}}{1 + e^{a^T \bar{W}_{1,i} + \gamma Y_i}}
$$

- a CAN estimator $\hat{\alpha}$ still obtainable by solving the following estimating equations

$$
\sum_{i=1}^{n_2} \left( \frac{R_i}{\pi(a, \bar{W}_{1,i}; Y_i; \alpha) - 1} \right) d(\bar{W}_{1,i}) = 0
$$

where $d(\cdot)$ is a vector of arbitrary functions of $\bar{W}_{1,i}$, and has the same dimension as $\alpha$. 
Sensitivity Analysis

- The IPCW estimator is robust if insensitive to different values of $\gamma$.
- More flexible models for the missing mechanism:
  \[ \pi(a, \tilde{W}_1, Y; \alpha) = \exp\{\alpha^T \tilde{W}_1 + \gamma q_\alpha(Y, \tilde{W}_1)\} \]
  \[ \text{with } q_\alpha(Y, \tilde{W}_1) \]
  being a known function of $Y$ and $\tilde{W}_1$.
- The choice of $d(\cdot)$ in the estimating equations affects the efficiency.
- See Rotnitzky et al. (1998, 2001) and Scharfstein et al. (1999) for detailed discussions.

Conclusions (I): the framework of counterfactuals can be useful

- Make it explicit what we mean by "causal effect", i.e., what is the quantity/estimand we seek?
- Give explicit assumptions under which "association is causation", and therefore standard statistical methods may be used.
- Give explicit assumptions needed for the identification of causal effects even when "association is not causation".
- Allow for the derivation of new statistical methods when standard and familiar methods fail.

Conclusions (II): methods need to be used with caution

- The counterfactual framework may not apply to some settings.
- Make sure the estimand has the right meaning and interpretation.
- The validity of associated assumptions, some assumptions not even testable, e.g., NUCA.
- Expert knowledge can be used to enhance the plausibility of the assumptions.

References

References


References


References


References

Noninferiority Trials

Scott Evans
Harvard University

Graybill Conference
June 11, 2008

Outline

• Noninferiority Trials
  – Design Issues
    • Fixed Margin
    • Combination Approach
    • Precision-Based Approach
    • Synthesis Method
  – Analyses and Reporting
    • Interim Analyses
    • CONSORT
    • ITT vs. PP
    • Missing Data
    • Switching between Noninferiority and Superiority

Background on Scientific Designs to Evaluate an Intervention

• Single arm study
  – Administer an intervention to a group of patients and see if they improve or are cured

Single Arm Study

• Limitations:
  – Cannot control for “natural history”
    • People may have improved anyway
  – Placebo-effect (or Hawthorne effect)
    • Patients/clinicians believe that they are getting better because of treatment; manifesting itself in better outcomes
      – E.g., pain
  – Miss a “good result” when you observe no change but patients would have gotten worse if left untreated.
**Placebo Controlled Trial**

- Randomize to one of two arms
  - New intervention or placebo
- Blind the intervention to patients and investigators when possible
- Analysis compares the two arms with respect to response

---

**Absence of Evidence is Not Evidence of Absence**

- Hypothesis testing is analogous to a court trial:
  - People are assumed innocent until proven guilty
    - $H_0$: Innocent
    - $H_A$: Guilty
  - If VERDICT = Not Guilty (i.e., do not reject $H_0$), then:
    - We cannot say that we have proven innocence
    - We say that we failed to find enough evidence to prove guilt
      - There is a subtle but important difference between the two

---

**Noninferiority Trial**

- Randomize to one of two arms
  - New intervention or active-control
- Blind the intervention to patients and investigators when possible
- Analysis compares the two arms with respect to response
- Need to show new intervention is “no worse” than the active control

---

**Superiority**

$H_0$: $p_1 - p_2 = 0$

$H_A$: $p_1 - p_2 \neq 0$

$p_1 = \text{efficacy of new treatment}$

$p_2 = \text{efficacy of control group}$
Noninferiority Trial

- Decide on a NI margin, M (more later)

- Analysis
  - Get a confidence interval (CI) for the difference between arms (new intervention minus active control with respect to efficacy) and note if lower bound of CI is within the NI region

Statistical vs. Clinical Significance

- Not the same thing!

Example: ACTG 116A

- DDI vs. placebo is not ethical with availability of AZT

- Endpoint
  - Time to AIDS defining event or death

- Noninferiority
  - If upper bound of CI for HR (DDI vs. AZT) < 1.6
    - DDI (500mg): HR=1.02 90% CI = (0.79, 1.33)
    - DDI (750mg): HR=1.04 90% CI = (0.80, 1.34)

Example: FDA SGE Experience

- A randomized, double-blind, multicenter study comparing the efficacy and safety of Piperacillin/Tazobactam (PT, 4G/500MG) and Imipenem/Cilastatin (IC, 500MG/500MG) administered intravenously every six hours to treat nosocomial pneumonia in hospitalized patients
Example: FDA SGE Experience

• Lower bound of 95% CI for the difference in response rates (PT-IC) is –0.066 (> –0.20)
  – Was a margin of 20% too large?
  – Noninferiority would be shown for a margin as small as 7%

• Result
  – PT was noninferior to IC
  – Approved by the FDA

Noninferiority Trial

It is impossible to show that two treatments have identical efficacy. Instead we choose a NI margin, M, and seek to prove that \( \beta_{TC} \) is less than M. This is done by testing the null hypothesis:

\[
H_0: \beta_{TC} \geq M
\]

gainst the one-sided alternative

\[
H_0: \beta_{TC} < M
\]

Note: The traditional roles of the hypotheses are reversed.

T is Often Better than C in Other Ways

• Better safety profile
• Less expensive
• More convenient to administer
  – Less invasive
  – Fewer pills
• Easier to comply
• Shorter treatment duration

Assumption: Constancy

• Historical data:
  – C showed superiority to P in historical trial (\( \beta_{CP,H} > 0 \))

• Constancy Assumption
  – \( \beta_{CP,H} = \beta_{CP,NI} \) if placebo was present in the NI trial
    • May not be the case in the presence of resistance development or with differing trial conduct (e.g., administration of treatment, differences in populations or endpoints, etc.)
  – Not verifiable in current trial (without placebo)
  – Implication: to conduct the NI trial in the same manner as the trial that established \( \beta_{CP} > 0 \)
Assumption: Assay Sensitivity

- Trial is able to detect differences between treatments if they exist
- Need a sensitive enough instrument to measure response and detect differences if they exist
  - Otherwise, all interventions will display similar responses due to the insensitivity of the instrument

Choice of Active Control

- Must have clinical efficacy
  - Of substantial magnitude
  - That is precisely estimated
  - With estimates that are relevant to the setting in which the NI trial is being conducted
    - Constancy assumption
  - Preferably measured by multiple controlled trials
- Regulatory approval does not necessarily imply that an intervention can be used as an active control in an NI study
  - Superiority to placebo must be reliably established

Choice of the Active Control

- Must have assurance that the active control would be superior to placebo if a placebo arm was employed in the trial

Biocreep

- The tendency for recently demonstrated noninferior treatments to be active controls in new NI trials even though they are slightly inferior to historically proven active controls vs. placebo (D’Agostino, SIM, 2003)
- If A is noninferior to B and B is noninferior to C, then it does NOT necessarily follow that A is noninferior to C
Choice of Active Control

• Example: OHARA is developing a treatment trial for oral candidiasis (OC) in Africa within the ACTG system
  – Fluconazole is not readily available in Africa (too expensive)
  – Nystatin is used in many places as the standard of care
  – Gentian Violet (GV, an inexpensive topical agent) showed excellent in vitro activity
  – A NI trial of GV compared to Nystatin was proposed

Choice of Active Control

• Similar issues have arisen with treatments that were once shown to be effective but may no longer be effective because of the development of resistance
  – Example: MRSA

MRSA

• Methicillin-resistant Staphylococcus aureus
  – Bacterial infection that has become resistant to antibiotics (such as penicillin, amoxicillin, methicillin)
  – Lives on the skin
  – Community acquired
    • Spread through contact

DMC Discussion

• ID clinicians insist antimicrobials are "highly effective" in treating skin infections, but seem unable to site evidence to back this claim other than anecdotal (and then claim it is "unethical" to do any other trial design other than NI).

• Without the data to support one of the drugs as the "control" based upon reliable and reproducible evidence of the magnitude of the benefit of the control compared to placebo, the results of an NI trial are not meaningful
Choosing the Noninferiority Margin (M)

Traditional Approach

- Subjective but structured
  - Combination of statistical reasoning and clinical judgment (CHMP Guidance)

- Concepts
  - “Maximum treatment difference that is clinically irrelevant”
  - “Largest treatment difference that is acceptable in order to gain other advantages of the experimental intervention”

- Design parameter is not present in superiority trials

Choosing the Noninferiority Margin (M)

Investigators frequently set $M$ equal to half of the estimated effect of the active control ($C$) relative to placebo ($P$) from the historical evidence. That is, set

$$M = \frac{\hat{\beta}_{C,P}}{2}$$

This is known as “preserving a fraction of the effect”. Note that the method does not consider the fact that the estimate from historical data is subject to uncertainty.

Tamoxifen vs. Placebo: NSABP P1 Trial

Subset of Women $\geq 50$ years old

Favors Placebo  Favors Tamoxifen

Relative Risk for Invasive Br Ca: Placebo / Tamoxifen

Interpretation: $P$ increases the rate of invasive breast cancer incidence compared to Tam by 112% (CI: 52% to 303%)

RR = 2.12 (1.52 - 3.03)

STAR Trial: Claiming NI

- Thus the upper bound of the 95% CI estimate for the relative risk needs to be less than 1.56
Tamoxifen vs. Placebo: NSABP P1 Trial
Subset of Women ≥ 50 years old

Favors Placebo  Favors Tamoxifen

RR = 2.12 (1.52 - 3.03)

NI margin: Lower bound of the 95% CI estimate for the RR of Tamoxifen vs. placebo

Relative Risk for Invasive Br Ca: Placebo / Tamoxifen

Failed Noninferiority Trials
Poor Choice of Noninferiority Margin

In the SPORTIF trials, ximelegatran was compared to warfarin for stroke prevention in atrial fibrillation patients. The event rates in the warfarin group (control) were 2.3% (Sportif III) and 1.2% (Sportif V). Based on the historical evidence, the sponsor chose an absolute NI margin of 2%. Because of the low event rates in the control arm, this resulted in a NI margin that allowed the conclusion of NI even if the trial did not rule out a doubling of the event rate.

The common theme in these trials was that a NI margin based on the effect of the comparator drug was not consistent with the community standard for therapeutic equivalence.

Sample Size

- Still only get $\alpha=0.025$ on the one side
  - Using 0.05 lowers the level of evidence for drawing conclusions vs. accepted practice in superiority trials

- General wisdom (including regulatory agencies) is to use 2-sided 95% CIs to evaluate both superiority and NI

- Important to power for per protocol analyses too
  - More on this later

NI Trials May Not Be Appropriate When:

- Constancy assumption is in doubt
- Assay sensitivity is in doubt
- Historical data of active control effect over placebo are thin
- Variability of active control effect is very large
Fixed Margin Method

- Define NI margin $M$
  - Using data from historical trials
    - Note: subject to bias and uncertainty
- Use the 2-sided 95% CI to rule out $M$
- Too conservative?

Testing for NonInferiority

To address $O_1$, noninferiority, we determine a NI margin, $M$, which, if met will allow us to conclude that $T$ is NI to $C$. $M$ should be chosen based on clinical considerations, not the historical data regarding the effect of $C$ relative to $P$.

Given the margin, $M$, we test the null hypothesis

$$H_{10}: \beta_{TC} = M$$

Against the alternative hypothesis

$$H_{1A}: \beta_{TC} < M$$

Using standard methods for NI trials.

Testing for Superiority to Placebo

Assuming assay constancy, we test the null hypothesis:

$$H_{20}: \beta_{TP} (= \beta_{TC} + \beta_{CP}) = 0$$

against the alternative $H_{1A}: \beta_{TP} < 0$ with the test statistic

$$\hat{T} = \frac{\beta_{TP}}{\sqrt{\hat{V}_{\beta_{TP}} + \hat{V}_{\beta_{CP}}}}$$

If we assume approximate normality of the estimates, this can be done with standard methods.

Design and Analysis

Since the goal of the NI design is to reject both $H_{01}$, the hypothesis of inferiority, and $H_{02}$, the hypothesis of no difference relative to placebo, we perform sample size calculations for each test and choose $N$ equal to the larger of the two sample sizes.

The analyses do not require two unrelated tests. Since the results from the historical trials are known at the time the NI trial is conducted, the two hypothesis tests can be reduced to two tests involving the estimated treatment effect of $T$ relative to $C$ in the NI trial. The two tests can be expressed as inequalities involving $\hat{T}$.

If the estimate satisfies the more stringent of the two inequalities, we reject both $H_{01}$ and $H_{02}$ and achieve both objectives.
Designs Based on Precision

- Based on principles of estimation rather than hypothesis testing
  - Do not specify explicit hypotheses
  - Avoid making the distinction between superiority and NI

- Specify the precision with which you want to estimate treatment differences (e.g., the maximum length of a CI for the treatment effect difference)
  - Power the study to estimate the effects with this precision
  - Interpret the CI as usual (ruling out effects in either direction as appropriate)
  - May wish to pre-specify a NI margin for regulatory purposes

Inconsistency with Preservation of Effect?

- New intervention could look better than active control (point estimates) but not meet the preservation of effect condition

- Two trials with different active controls have different standards for success

- If new intervention is better than active control, should the active control be withdrawn?

Synthesis Method

- Combines standard errors from historical trials and NI trial
  - Note: not the standard error from a randomized comparison

- NI is not considered
  - M is irrelevant

Issues with Synthesis Method

- Proper control of conditional type I error is not warranted

- Assumes no bias in historical point estimate
  - Often not the case with publication bias
  - Problematic when constancy assumption does not hold

- Cannot produce a fixed NI margin

- No consensus on role of across-trial error

- Too liberal?
**Noninferiority:**

\[ H_0: p_1 - p_2 > -\Delta \]
\[ H^*_A: p_1 - p_2 \geq -\Delta \]

**Interim Result: What Would You Recommend?**

**CLINICALLY INFERIOR**

A

B

C

D

E

**CLINICALLY NONINFERIOR**

**STATISTICALLY INFERIOR**

\(-\Delta\)

\(0\)

**STATISTICALLY SUPERIOR**

\(p_1 - p_2\)

**Analyses**

- 2-sided CIs are generally used
  - Consistent between significance testing and subsequent estimation
  - Consistent with ICH E-9
  - P-values are not appropriate

- Choice of NI margin plays a direct role and thus must be justified

**Interpretation Issues**

- If active control displays different efficacy than in prior trials vs. placebo, then the validity of the pre-defined NI margin may be suspect and the interpretation of treatment differences is challenging
  - Can we model the changes in the effect of \(C\) vs. \(P\)
    - May not have appropriate data as we need a comparison of \(C\) vs. \(P\) from multiple trials conducted at different times
**ITT vs. Per Protocol**

- Sensitivity analyses are important
  - ITT and Per Protocol should both be conducted
  - Need consistency of qualitative result (CPMP, 2000, points to consider)
- PP usually results in a larger effect size
  - ITT dilutes effects
- PP usually has wider CIs
  - Fewer patients than ITT

**Idea: Analysis of Missing Data**

- Consistent with the null hypothesis of the NI trial
  - E.g., continuous endpoint: impute reasonable expected value (i) for the active control and (i-M) for the new intervention
  - E.g., binary data: impute expected proportion (p) for active control and (p-M) for new intervention
  - Then analyze using analysis of means

**Switching from Superiority to Noninferiority**

- Generally not acceptable to go from failing to demonstrate superiority to NI
  - Unless a NI margin was pre-specified
  - Post-hoc definition of NI margin is difficult to justify
- Choice of NI margin needs to be independent of trial data

**Switching from Superiority to Noninferiority**

- May be feasible if:
  - Noninferiority margin was pre-specified (or can be justified, which is difficult to do)
    - Must be based on external information and not chosen to fit the data
  - ITT and PP show similar results
  - The trial was of high quality with few drop-outs and good adherence
  - The control group displayed similar efficacy to trials vs. placebo
  - The trial was sensitive enough to detect effects
Benefit:Risk

Scott Evans
Harvard University
Graybill Conference
June 11, 2008

Benefit:Risk

• Fundamental concept in clinical trials
  – Weighed by regulators in product approval decisions
  – Evaluated by sponsors to aid in development decisions
  – Assessed by data monitoring committees during interim analyses to make recommendations

Benefit:Risk

• September, 2006
  – Institutes of Medicine recommended that FDA develop and continually improve a systematic approach to benefit: risk

• December, 2006
  – European Committee for Proprietary Medicinal Products (CPMP) called for improved methodology leading to a more systematic approach to benefit: risk analysis

BRAT - PhRMA

• Goal
  – Develop a structured, transparent benefit: risk framework and integrate it into the regulatory review process
  • Quantitative and qualitative elements
  • Pilot studies being planned
BRAT - PhRMA

• Separate B:R balance by each indication
• Consider all data
• Incorporate uncertainty
• Evaluate temporal effects
• Transparency via systematic, consistent, reproducible model

Benefit: Risk

? May be acceptable with a very serious disease with no known cure

? May not be acceptable with a disease that is not life-threatening and other effective and safe treatment options are available

Challenges

• Clarity of objective
  – Benefit:risk of the population vs. the trial cohort vs. an individual patient

• Generalizing benefit:risk assessment in clinical trials to a population
  – Trials may not be representative of real patient usage

Challenges

• Disease-specific context
  – Toleration of toxicities for some indications but not others

• Patient-specific context
  – E.g., benefits and risks are different for old vs. young

• Population-specific context
  – Multinational trials with differing cultures, ethics, and availability of medicines
**MRSA (The “Superbug”)**

- Methicillin-resistant Staphylococcus aureus
  - Bacterial infection that has become resistant to antibiotics (such as penicillin, amoxicillin, methicillin)
  - Lives on the skin
  - Community acquired
    - Spread through contact

**Benefits and Risks Tradeoff**

- FDA Advisory Committee (CDRH)
  - “Substantial Equivalence”
    - Distinct from traditional concept of noninferiority
    - Simultaneous evaluation of benefit and risk
  - NeuroStar™ for the treatment of Major Depressive Disorder
    - Repetitive transcranial magnetic stimulation (rTMS)
    - Vs. Electroconvulsive Therapy (ECT)
      - Causes a seizure via electronic stimulation

**Substantial Equivalence**

- From my packet sent by the FDA:

  “the experimental device does not need to be as effective as the predicate device, if the clinical data demonstrated that any reduction in effectiveness was off-set by an improvement in patient safety/risk”

- What metric/weights?

**Censoring**

- Important safety data is often censored by the end of the study or by treatment discontinuation

- Important efficacy data is often censored by SAEs

- Safety data often gathered after approval, but not always the case for efficacy data
Asymmetry of Attention to Benefits and Risks

- Passive collection of some safety data
  - Report only if abnormal
  - Cannot distinguish between normal (indicating no safety issue) vs. missing data

ITT Principle

- Important with respect to safety data
- Need to follow patients that withdraw from treatment
  - Otherwise adverse outcomes may not be obtained as they can occur after withdrawal from exposure
  - Potential for informative censoring if subjects that withdraw are dissimilar to subjects that complete the study

Importance of Follow-up

- Carefully consider the length of $T_W$ in design to evaluate safety
- Count events in interval of $T_S + T_W$ to avoid informative censoring and avoid bias
  - More diligent follow-up needed on patients that discontinue the treatment early
    - Else can bias Kaplan-Meier estimates of cumulative incidence

Proposed Methods
# Combined Marginal Analyses

- **Current Practice**
  - Separate and marginal analyses of efficacy and safety
  - Then informal (usually non-quantitative) aggregation of the two marginal analyses are conducted as an assessment of benefit:risk

- **Meta-analyses of several trials**
  - E.g., Integrated Summary of Efficacy and Safety (ISES)

# Benefit:Risk Ratio

- **Advantages**
  - Easy to understand and communicate

- **Disadvantages**
  - Does not account for the relative timing of these events
  - Does not account for the censoring by competing events
  - Can be challenging to identify a threshold at which benefits and risks are neutralized

---

# NNTB and NNTR

- Let $\pi_A$ and $\pi_B$ ($\pi_B > \pi_A$) be the incidence rates of a beneficial outcome in treatment groups A (control) and B (experimental) respectively, then

$$\text{NNTB} = \frac{1}{\pi_B - \pi_A}$$

- E.g., if $\pi_B - \pi_A = 0.5$ then NNTB=2
  - Implying that on average if 4 patients are treated (2 on each arm), then we expect one more beneficial outcome on treatment B than treatment A

---

# Ratio based on NNT

- $R = \text{NNTB}/\text{NNTR}$
  - If the events of benefit and harm are of equal importance, then considering benefits and risks:
    - $R=1$, suggests that A and B are similar
    - $R>1$, suggests that B is better than A since B provides benefits at a faster rate than providing harms (relative to A)
    - $R<1$, suggests that B is worse than A since B provides benefits at a slower rate than providing harms (relative to A)
NNT and the Ratio based on NNT

- Statistical inference based on NNT or R is challenging
  - Lesaffre et.al., CCT, 1999.
  - Duncan & Olkin, SIM, 2005.
  - Alemayehu et.al., JBS, 2006.

NNT Extensions

- NNT for time-to-event data

- “Unqualified success” and “unmitigated failure”

Q-TWiST

- Quality-adjusted Time Without Symptoms or Toxicity (Gelber, et.al., Biometrics, 1989)
  - Proposed to evaluate adjuvant therapies for breast cancer
    - Has been used in other disease settings
      - HIV, melanoma, rectal cancer, lymphoma, prostate cancer, childhood AML.
    - Evaluates benefit:risk “trade-offs” in clinical trials

ECOG 1684: Q-TWiST

- Overall Survival
- Interferon Alfa-2b Observation
- P = 0.06
- Months from Randomization
ECOG 1684 Q-TWiST

Toxicity summary

- Grade 3 or worse toxicity occurred in 78% of patients
  - Dosing delays or reductions occurred in 62% of patients during the IV phase and for 51% during the SQ phase
  - Included constitutional, myelosuppression, hepatotoxicity and neurological

ECOG 1684: Q-TWiST

Step 1: Define clinical health states

- Toxicity
  - Time with grade 3 or worse treatment-related toxicity
    - The entire treatment duration was assigned to Tox if any grade 3 or worse toxicity was experienced
- Relapse
  - All time following disease relapse
- TWiST
  - All remaining survival time

ECOG 1684: Q-TWiST

Step 2: Partitioning overall survival

Interferon Alfa-2b

- Observation

ECOG 1684: Q-TWiST

Step 2: Partitioning overall survival

- Observation
### ECOG 1684: Q-TWiST

Step 3: Compare treatments using weighted components

<table>
<thead>
<tr>
<th>Outcome</th>
<th>Treatment Group</th>
<th>Obs</th>
<th>IFN</th>
<th>Difference</th>
<th>(95% C.I.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tox</td>
<td></td>
<td>0.0</td>
<td>5.8</td>
<td>5.8</td>
<td>(5.0 to 6.7)</td>
</tr>
<tr>
<td>TWiST</td>
<td></td>
<td>30.0</td>
<td>33.1</td>
<td>3.1</td>
<td>(-4.8 to 11.0)</td>
</tr>
<tr>
<td>Rel</td>
<td></td>
<td>12.4</td>
<td>10.4</td>
<td>-2.0</td>
<td>(-6.2 to 2.3)</td>
</tr>
<tr>
<td>Q-TWiST*</td>
<td></td>
<td>34.9</td>
<td>42.5</td>
<td>7.6</td>
<td>(0.0 to 15.1)</td>
</tr>
</tbody>
</table>

\[ Q-TWiST = u_{Tox} \times Tox + TWiST + u_{Rel} \times Rel \]

- **Threshold Utility Plot**
  - Fundamental outcome of a Q-TWiST analysis
    - Compares treatments across combinations of weights
  - Vertical axis
    - Weight for time with toxic effects
  - Horizontal axis
    - Weight for time after relapse
  - Plots contours of similar treatment effects

### Weight Coefficients

- 95 patients with melanoma provided their patient-specific weights for the clinical health states
- Applied to threshold analysis
- 16% had preferences in the region of significant Q-TWiST gain with interferon
- 84% had preferences in the region of non-significant Q-TWiST gain

Composite Endpoints

Desirable Endpoint Characteristics
- Clinically relevant
  - Addresses the scientific question
- Easy to interpret
- Easy to obtain
- Quantified/qualified in an unbiased manner
- Sensitive to changes induced by treatment
- Affordably obtained
- Results in a reasonable sample size
  - Note: continuous responses usually require smaller sample sizes than binary or time-to-event

Composite Endpoints: Advantages

- More complete characterization of treatment effect
  - May be interested in a variety of outcomes
    - E.g., Perspective of treating more than a single disease
- Reduces bias due to competing risks and informative censoring
- More events could imply more power
  - If expected effect size is unchanged
- Avoids multiplicity issue and avoids difficult selection between endpoints (ICH E-9)

Composite Endpoints: Disadvantages

- Competing risk issue
  - If censored patient has a different risk for event than a non-censored patient, then “informative censoring”
  - Occurs when an event (e.g., disease progression) is analyzed while ignoring circumstances (e.g., death) that preclude the occurrence of the event

Composite Endpoints: Considerations

- Are the components of similar importance?
- Do the components occur with similar frequency?
- Is the treatment effect similar across components?
Example: Composite Endpoint in HIV Trial
Time to “treatment failure”
- Time from randomization to:
  1. death due to any cause, or
  2. disease progression - defined as a new or recurrent AIDS-defining opportunistic infection or malignancy occurring after completion of 12 weeks of ARV treatment, or
  3. virologic failure defined as two successive measurements of plasma HIV-1 RNA ≥1000 copies/mL

Patient Level Measures
- Patients rate their overall experience with respect to perceived benefit and risk
- Possibly useful for therapies to treat symptoms
- Problematic when symptoms do not equate with risk
  - i.e., silent risk of abnormal labs (e.g., LFTs, bilirubin)

Global Benefit:Risk (GBR) Score
- Assign non-negative weights
  - Chosen to reflect the relative importance of each category when evaluating treatment
- Then compute summary measures and compare the distributions between treatment groups

Benefit-Less-Risk Measure
- Benefit is discounted for the presence of untoward safety events according to a prespecified algorithm
  - Similar discounting approach to Q-TWiST
Benefit-Less-Risk (BLR) Measure

- Step 2: Suppose benefit is measured on a single primary endpoint
  - Denote the benefit for subject $j$ as $S_j$
  - Discount $S_j$ by a multiple of $R_j$ using a “conversion factor”, $f$

$$BLR = S_j - f R_j$$

- Use BLR to compare treatments

Benefit:Risk Index

- Women’s Health Initiative
  - 15 year project sponsored by NHLBI
  - >161,000 postmenopausal women
  - 3 components
    - Randomized clinical trials
      - Enrolled > 68,000 women
      - Designed to evaluate effects of hormone replacement therapy (HRT), dietary modification, calcium/vitamin D on clinical endpoints
    - Observational study
    - Community prevention trial

Benefit:Risk Index

- HRT could have preventative effects and possible adverse effects (noted by DMC)
  - Freedman et.al. (CCT, 1996) proposed indices to combine the estimated treatment effects on the five endpoints
     - Estimate the effect of HRT on each endpoint
     - Combine the effects to form a composite HRT effect

WHI: PREMPRO

- Drug to help women manage symptoms of menopause
- Product label includes summary of benefits and risks
- Allows individual patients and their physicians to make patient-specific decisions based on benefit:risk
  - Patient-specific weighting and value judgment based on outcomes
Nonparametric Approach

- Useful when patients can be ranked with respect to each of benefit and risk (i.e., continuous or ordinal measures)
- Rank patients with respect to combined benefit and risk based on their benefit-specific and risk-specific ranks
- Compare the combined ranks between groups
- Issues
  - Ranking two patients, each higher on one scale (benefit or risk)
    - Subjective metric or weighting
    - Handling ties
    - Clinical relevance of effects?

Within Patient (Personalized Medicine)

- Define
  - Minimum tolerable benefit when risk = 0 ($b_1$)
  - Maximum tolerable risk when benefit = 10 ($r_1$)
  - Minimum tolerable benefit when risk = $r_1/2$ ($b_2$)
- Fit a smooth “tolerability curve” through these 3 points

Within Patient (Personalized Medicine)

- Analyses combines benefit and risk data within patient
- Allows for flexibility in analyses for varying trade-offs
  - By selecting different points that will define the tolerability curve
  - Assessment of treatment differences can be made on a personalized level

Ideas

- PK and biomarkers for benefit:risk evaluation?
  - If safety is dependent upon drug concentration and when predictive biomarkers are available
Absolute Risk

- More easily interpreted and may be more informative than relative risk
- More effectively summarizes medical impact
- Conveys absolute chance of event
- Describe the expected numbers of events (both beneficial and adverse) in a population of patients of a given size
  - For both the experimental treatment and the control
  - A comparison can be made by defining weights for the events

Interpretation of Graphic

- 100 patients given treatment for disease
  - 86 patients would have been disease free even if they had not received treatment
  - 9 patients (red faces) are patients that are not cured event with treatment
  - 5 patients (yellow faces) show a benefit
    - They would not have been cured without treatment but are cured when they receive treatment
    - NNTB = 20
    - It is not possible to identify which patients will benefit; thus all 100 patients need to be treated in order for 5 to benefit
  - Can similarly summarize NNTR

Benefit:Risk Reporting

- Remember CONSORT reporting principles
- Describe how benefit:risk varies across subgroups or disease characteristics
  - Often a single aggregate summary description of benefit:risk is not appropriate
  - E.g., by age, race, etc.
<table>
<thead>
<tr>
<th>Concluding Remarks</th>
</tr>
</thead>
</table>
| • More formal benefit:risk evaluation is needed to supplement traditional methods of analyses  
  - Estimates should include measures of variability  
  - Analyses should include sensitivity analyses to subjective weighting and other assumptions  
  - Consider analysis by time intervals |
| • Benefit:Risk is an ongoing process  
  - Continuous re-assessment is necessary as information accumulates, therapeutic alternatives changes, medical advances evolve, resistance evolves, etc. |

<table>
<thead>
<tr>
<th>Concluding Remarks</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Prospective planning of the collection and analysis of safety data is important</td>
</tr>
</tbody>
</table>
| • Enforce the ITT principle  
  - Follow all randomized subjects until study completion |