JOHNS HOPKINS
BLOOMBERG
SCHOOL of PUBLIC HEALTH

Analysis Issues II

Mary Foulkes, PhD
Johns Hopkins University

JOHNS HOPKINS
BLOOMBERG
SCHOOL of PUBLIC HEALTH

Section A

Missing Data

Causes of Missing Data

- Dropout
- Noncompliance with measurement procedure
  - Missed visit
  - Refused procedure
- Error
  - Test not done
  - Test done improperly
  - Results lost

Ideal Study

- No dropouts
- No noncompliance
- No errors
- No loss to follow-up

Big Worry

- Missing data may be associated with outcome
- Form of this association is unknown
- Failing to account for the (true) association, may **bias** results

Bias

- Null hypothesis: no difference
- If null hypothesis is true, and it systematically eliminates patients with poorer prognosis on one treatment arm, results will be biased towards the alternative hypothesis
- If results are biased in this way, the type one error (false positive rate) is inflated
- Presence of bias means significance tests are unreliable

Missing Data and the Intention-to-Treat (ITT) Principle

- ITT: analyze all randomized patients in groups
- What to do when data are unavailable?
- Implication of ITT principle for design: collect all required data on all patients, regardless of compliance with treatment
  - Thus, you avoid missing data

Dilemma

- Excluding patients with missing values can bias results and increase type one error (false positives)
- Collecting and analyzing outcome data on non-compliant patients may dilute results and increase type two error (false negatives)
  - But we can compensate for this dilution with sample size increases

Different Questions

- Is this treatment safe and effective when used as directed in a well-defined population?
- Is this drug safe and effective when used for its intended purpose by practicing physicians?

Relevance of Methodology

- Quadratic function of am't missing
- Methodology matters little when . . .
  - No data are missing: analysis is straightforward
  - Lots of data are missing: no analysis is credible
- Methodology is more important when modest/moderate amounts of data are missing
  - Not enough to demolish credibility
  - Enough to potentially shift conclusions

Handling Missing Data: Design and Conduct

- Select primary endpoints that are more “robust” to missing data
  - Change from first to final visit
  - Slope
- Expend effort on collecting outcome data on all patients entered
  - Visit reminders
  - General patient encouragement
  - Collection of data regardless of compliance with treatment

Handling Missing Data: Analysis

- Exclude subjects with missing values
- Last Observation Carried Forward (LOCF)
- Group means
- Multiple imputation
- Best/worst case scenarios

Exclusions of Subjects

- **Excluding subjects with missing values is the simplest approach**
  - Requires the assumption that excluded subjects are a random subset of all randomized subjects
  - Analysis is straightforward
- **Assumption noted above is difficult to justify in most instances**

Exclusion: Source of Bias

- Randomization ensures that there are no systematic differences between treatment groups
- Excluding subjects from analysis on the basis of post-randomization events, e.g., (non-compliance), may introduce systematic differences between treatment groups, thereby biasing the comparison

Imputation

- Determine a value that is “best guess” of true value of missing data point
- Several approaches proposed and/or in use
- Simplest approaches are statistically most problematic
- Any approach involving “made-up” data is problematic to some degree

Imputation

- Last Observation Carried Forward (LOCF)
  - Use last measurement available in patients with missing data after a certain point
- Group means
  - Assign average value of outcome variable among those in that treatment group with complete data
- Multiple imputation*
  - Predict missing outcome on the basis of outcomes for other patients with similar characteristics

*Source: Rubin, D.B. (1987). *Multiple Imputation for Nonresponse in Surveys*. Wiley. New York.

Assumptions Required

- LOCF: last available measure is a good estimate of final (missing) measure
- Mean substitution: assigning group average to missing values will result in a good estimate of treatment effect
- Multiple imputation: one can create a model using data on covariates to accurately estimate missing value

Problems with Assumptions

- LOCF: data may be missing because of some aspect of treatment
  - Patient not responding
  - Treatment not tolerable so that early data may not reflect true treatment effect
- Mean substitution is like excluding patients with missing data, except that power is (artificially) maintained
- Multiple imputation approach accounts for variability in estimating missing values, but assumption that we can predict outcome accurately based on measured covariates remains problematic

Bottom Line

- Assessments based on imputation models are reliable only insofar as one has confidence in the assumptions

Worst/Best Case Scenarios (Sensitivity Analysis)

- Neither exclusions nor imputation
- Provides bounds for the "true" results if all planned data points had been observed
- Provides a sense of how far off any of the other analyses could be
- Does not provide a single "answer" but aids in the interpretation of other "answers"

An Ounce of Prevention . . .

- Discourage dropout
- Encourage compliance
- Institute QC checks to minimize errors

Summary

- The less there is missing data, the there is less concern about . . .
  - Bias
  - Reliable conclusions
  - Appropriate methods of analysis
- Methods to replace or otherwise account for missing data are all flawed in important ways
- Sensitivity analyses are essential to evaluating reliability of conclusions

In the Next Section We'll Look at . . .

- Multiplicity
  - Multiple treatment arms
  - Multiple controls
  - Multiple subgroups

JOHNS HOPKINS
BLOOMBERG
SCHOOL of PUBLIC HEALTH

Section B

Multiplicity

Multiplicity in Trials

- Refers to the multiple judgments and inferences
  - Hypothesis tests
  - Confidence intervals
  - Graphical analyses
- Leads to a concern about false positives or an inflation of type one error

Example

- Chance (probability) of drawing the ace of clubs by randomly selecting a card from a complete, shuffled deck is 1/52
- Chance of drawing the ace of clubs at least once by randomly selecting a card from a complete, shuffled deck 100 times is . . . ?

Sources of Multiplicity

- Multiple treatment arms
  - Doses, regimens
  - Treatments
- Multiple controls
  - Internal
  - External (historical)
- Multiple evaluation groups
  - Evaluable subgroup
  - Per protocol subgroup
  - Demographic and disease-based subgroups

Sources of Multiplicity

- Multiple endpoints
  - Efficacy variables
  - Evaluation time-points
- Multiple analyses
  - Statistical tests
  - Dichotomization and cut-points
  - Approaches to covariates or missing data
- Multiple interim analyses of accumulating data
- Multiple studies, sets of studies

Young's Rules

- With enough testing, false positives will occur
- Internal evidence will not contradict a false positive results (i.e., don't expect to figure out which are the false positives)
- Good investigators will come up with possible explanations
- "It only happens to the other guy"

Approaches to the Problem

- Do only one test
- Adjust the p-values
- Perform the desired tests at the nominal level and warn reader that no account has been taken for multiple testing
- Ignore the problem

Bonferroni Adjustment

- Most common approach
- Can severely reduce power when many comparisons are made
- Conservative when comparisons are not independent
- Methods exist to account for inter-correlation of results in adjusting significance thresholds

Multivariate Testing

- Mutually exclusive groups
- Multivariate (“global”) tests assess whether all groups are similar
- If conclusion of global test is no, then do pair-wise tests
- If conclusion of global test is yes, stop
- Problem: if all are similar but one, the power for a global test may be low
- Note: if multiple arms represent different doses, you should test for dose-response rather than heterogeneity

Limitation of Comparisons

- Select a single primary hypothesis; treat others as exploratory
- Pre-specify comparisons of interest
- Create composite variables
- Refrain from data-dredging

Multiplicity when "Lumping" or "Splitting"

- Splitting a trial with focus on the positive subgroup generally leads to a misleading result
  - ISIS-2 trial [*Lancet.* (1988). 349-360]
  - The subgroup result by astrological signs at birth of Gemini or Libra came out to be slightly unfavorable for aspirin
  - However, the subgroup result by the other astrological sign gave a result in favor of aspirin with p-value < 0.00001
- Lumping trials (say, given two trials) only when at least one of them does not give statistically significant result inflates the type I error rate

Source: Lu and Huque. (2001). Proper planning for multiplicity, replicability and clinical validity at the design stage is necessary.

Some Final Comments

- Address multiplicity issues at the design stage
- Different questions require different approaches
- Clinical subset decision rules, involving multiple endpoints, inflate type one error rate and require adjustments
- With multiple step procedures, interpretation of conditional results and computation of confidence intervals can be problematic
- Testing for large families of hypotheses using methods such as FDR (False Discovery Rate) should be considered exploratory strategies

In the Next Lecture We'll Look at . . .

- Non-inferiority
  - Active control trials
  - Specifying "delta"
  - Assay sensitivity
  - Potential problems