Lecture 10
Missing data management and sensitivity analysis

Goals:
• To examine the different kinds of missing data
• To understand the potential threats to validity when non-trivial amounts of data are unavailable
• To review analytic remedies for missing data
• To examine sensitivity analysis in the context of missing data

Consequences of missing data

At the least, missing data have a negative effect on the precision of estimates

When there is much missing data there is also a concern about representation of the target population.

An even more serious concern is the risk of associated bias in estimates.
Avoiding different kinds of missing data

The single best way to manage missing data is to avoid it. This involves understanding why data may be missing and taking steps to minimize the impact of these reasons. Nevertheless, in most circumstances at least some data will be missing and needs to be considered with regard to potential bias and analytic remedies for this bias.

Data that are entirely missing

Failure to capture all segments of a cohort or control population; participation rates are linked to exposures of interest. Coping mechanisms include:

1. Minimize such risk in the design of the study.
2. Include variables related to missingness in the analysis.
3. Examine the potential magnitude of the bias attributable to these factors (sensitivity analysis).

Study design to minimize missing data

Recruitment and completion of information is generally highest when:

1. Carried out through or with the support of existing organizations
2. Collect data in settings convenient to the participant.
3. Collect data in person or telephone interview rather than by mail
4. Include motivating communications to participants.
5. Provide motivating factors commensurate with participant effort.
Study drop out

Incomplete data collection is so common in epidemiological studies as to almost be considered universal.
Study drop out is more likely at times of stress and distress.
Factors related to study refusal are also related to study drop out.
Partial data provided on those who fail to complete the study protocol can be helpful in identifying bias.

Missing cases and controls in case-control studies

Differential participation requires examination of the potential confounding or mediating effects of variables correlated to such participation differential

See, for example, Glaser, SL & Stearns CB, 2002, Reliability of random digit dialing calls to enumerate an adult female population, American Journal of Epidemiology, 155, 972 -975.

Found that random digit dialing tends to under-enumerate women.
Researchers also find that younger persons are less likely to respond to random digit dialing.

Minimize impact on study estimates

Participation biasing factors may be confounders or effect modifiers.

   e.g., age, sex, and social class

Missing study components.

In this study the researchers investigated whether the 52% participation rate in the clinical assessment portion of the study may have accounted for an unexpected finding.

Missing data of the “inevitable” type: individual variables

- R & G suggest that usually selecting for complete data in the analysis is not a problem (based on the relatively large N in many epidemiological studies).
- Once data involve a number of variables, such selection can cumulatively lose enough participants too affect power, as well as threaten bias in estimates.
- When a study is relatively small such risks are especially great.

Study design to minimize missing data for individual items

The following protocol problems can increase the number of respondents who do not answer:

- In language not understood by respondent
- In a frame or context unfamiliar to the respondent
- One of many questions that respondent cannot answer adequately
What can the data analysts do about missing data?

- Eliminate subjects from the analyses
- Statistically estimate the missing values
- Add the fact of “missingness” as study variables
- Use multiple imputation procedures to produce estimates to be employed

Choice of analytic strategy depends upon the pattern of missingness

- Missing at random (MAR)
- Missing completely at random (MCAR)
- Missing for reasons not associated with measured variables but relevant to the estimated effects (non-ignorable missingness)

MCAR data

- Unbiased
- Potentially inefficient

Strategy:
If sample is large enough dropping subjects will not bias.
Multiple imputation will produce appropriate and efficient estimates.
MAR data

- Missingness is associated with other measured variables.
- When these associations are taken into account, it is expected that estimates will be unbiased.

Strategic alternatives:
- Estimate missing values by OLS
- Estimate at the mean
  - May also add a variable reflecting “missingness”
- Use multiple imputation (or a single imputation that includes random variance)

Estimate missing values by OLS

- Use when there are powerful predictors of the missing values.
- Particularly useful when variables in question will be included in many analyses/ by many analysts
- Particularly useful when proportion of missing data is small
  - Uses an empirical prediction model for all participants who have data.
  - Applies this model to estimate values for participants for whom this data is missing.
- Problem: An absence of individual level “error” or random fluctuation -> underestimates in standard errors.

Estimate at the mean or median

- Use when proportion of data is small and no good predictors are available.
- Use with or without a “missing data” categorical variable.
- If the variable with missing data is categorical, you will simply add another category.
- Inclusion of a binary missing data variable produces effect estimates based only on those with actual data.
- Inclusion of a binary missing data variable makes obvious what other variables are related to missingness.
- Like other analytic solutions, it presumes the model is equivalent, on average, for those with and without data.
Multiple imputation of missing data

- Used for predictor or outcome variables.
- Produces estimated values with enough random variance to produce appropriate standard errors in subsequent analyses.
  - Multiple imputation programs create new data sets with imputed values and appropriate random variance.
  - In general, five imputations are sufficient to produce stable estimates.
  - A data combining program produces a "best" estimate with greater efficiency and appropriate standard errors.

Example: Taylor et al study of urologic symptoms

- As noted, only 52% of the entire sample agreed to participate in the clinical assessment.
- The study findings indicated a (biologically implausible) decline of symptoms and a lack of test specificity among the oldest study participants.
- The oldest participants also agreed to the clinical assessment at the highest rate.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Age 40-49</th>
<th>Age 50-59</th>
<th>Age 60-69</th>
<th>Age 70-79</th>
</tr>
</thead>
<tbody>
<tr>
<td>Observed mean symptoms</td>
<td>6.33 (0.58)</td>
<td>7.41 (0.52)</td>
<td>8.86 (0.72)</td>
<td>7.86 (0.97)</td>
</tr>
<tr>
<td>Imputation estimate mean symptoms</td>
<td>6.85 (0.42)</td>
<td>7.86 (0.48)</td>
<td>8.77 (0.56)</td>
<td>9.63 (0.77)</td>
</tr>
<tr>
<td>Observed extreme antigen in biopsy neg men</td>
<td>2.36</td>
<td>3.30</td>
<td>5.38</td>
<td>8.20</td>
</tr>
<tr>
<td>Estimated</td>
<td>2.64</td>
<td>3.41</td>
<td>5.35</td>
<td>6.54</td>
</tr>
</tbody>
</table>
Taylor et al study of urologic symptoms

- Conclusion: The peculiar findings were probably due to biased participation in the clinical assessment.
- The relationship with age in the analyses including imputed data was linear.
- The sensitivity changed much less with age in the analyses including the imputed data.

Planned missing data

Large cohort studies may administer portions of lengthy protocols to randomly selected individuals to save money and participant burden.

Studies with expensive assays or clinical assessments may do so on a random basis to save money.

Because missing data are MCAR, estimated values can be used to achieve the greatest possible efficiency (statistical power).

- As noted, may be done because some parts of the study are too expensive to apply to all participants.
- Alternatively, the burden to participants may be too great.
- Access may be limited in time.

- This is a likely “wave of the future” in epidemiology, especially if large national cohorts are initiated.
Planned missing data

- Example: The New York City public school study (Dr. Christina Hoven and colleagues)
  - Time limited access to children in classrooms
  - More data required in order to answer study questions than could be answered in that time period.
  - The protocol was divided into 4 sections: a Core section given all students and 3 non-Core sections.
  - Every study responded to Core plus two sections.

Planned missing data: NYC public school study

- The three versions of the protocol (CoreAB, CoreAC, CoreBC) were distributed randomly
- Multiple imputation methods are employed whenever study questions “cross-over” non-Core sections.
- Efficiency is maximized by the fact that every relationship is estimated by a minimum of 2/3 of the entire sample.
- Bias is not a problem because of the randomness of the design.

Non-ignorable missingness

If there are strong concerns that the data model for missing observations is not the same as for the completed observations,

And

The variables that might produce valid estimations are unavailable,

Sensitivity analysis can explore the possible biasing impact of these problems.
Non-ignorable missingness

- Such a problem may occur when those who are reluctant to provide data are more likely to have a particular response.
  - The question answer itself influences the response probability.
  - Such problems are more likely if the answer is embarrassing or suggests illegal or socially undesirable behavior.

Sensitivity Analysis

Explore biases associated with:
- Study design
- Participant non-response
- Measurement error
- Non-random missingness on study effects

Model data estimates that might characterize the complete, accurate data compared to the incomplete or inaccurate data actually collected.

Sensitivity Analysis

- Important to distinguish this set of analytic techniques from the other methods we have discussed.
- Sensitivity analyses are "what if" analyses: What if our assumption that some aspect of the data was not problematic (e.g., was random or unbiased) is wrong? How far off can we be?
- Therefore, they inevitably involve "making up" data, or guessing at parameters we can't actually measure.
Sensitivity analysis and missing subjects

• In our study of the relationship of congenital rubella to later psychotic disorder we used a simple sensitivity analysis to determine whether attrition of % of the rubella cohort could have biased results.

• The analysis showed that even if none of the missing participants with congenital rubella had such a disorder, the rate of psychotic disorder in this group as a whole would still be significantly greater than in the reference group.

Nonaffective psychosis after prenatal exposure to rubella


<table>
<thead>
<tr>
<th></th>
<th>N</th>
<th>Number with schizophrenia symptoms</th>
<th>Relative risk (CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unexposed youth</td>
<td>164</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>Exposed youth</td>
<td>70</td>
<td>11</td>
<td>5.2 (1.9 – 14.3)</td>
</tr>
<tr>
<td>Lost to follow-up</td>
<td>67</td>
<td>unknown</td>
<td></td>
</tr>
<tr>
<td>Combining exposed youth &amp; lost to follow-up</td>
<td>137</td>
<td>11</td>
<td>2.6 (0.94 – 7.40)</td>
</tr>
</tbody>
</table>

Another application of sensitivity analysis

Another application of sensitivity analysis

Megan Murray and David Alland, Methodological Problems in the Molecular Epidemiology of Tuberculosis. American Journal of Epidemiology, 155, 565 – 571.

The presence of different genetic polymorphisms (the genetic “fingerprint”) of tuberculosis bacteria is used to distinguish reactivation of the disease from new cases. Since the bacteria mutate over time, multiple cases with identical bacteria are probably recently transmitted cases from a single source of contagion. Cases with unique genetic fingerprints are most likely old cases.

These two kinds of cases have different relationships to risks and different implications for public health policy.
Methodological Problems in the Molecular Epidemiology of Tuberculosis, continued

Potentially biasing problems:
The source of patients too limited in some studies. If one fails to find genetic “matches” for some cases that actually exist in the population there will be an overestimation of old cases.

If so, there may also be an underestimation of the effects of some exposures on new cases.

These investigators used data from one such study to determine the sensitivity of study estimates to the sampling fraction of total TB cases included in the study.

In that study of 104 cases 40% were in clusters (thus designated new cases) 60% were unique (reactivated cases). They estimated 1076 cases occurring in the total study area.

Logically, smaller sampling fractions will mean that more cases are considered unique because other cases in their “cluster” were not sampled.

Proportion of unique isolates and OR for several hypothetical cluster distributions

<table>
<thead>
<tr>
<th>N</th>
<th>SF</th>
<th>Proportion of cases due to reactivation</th>
<th>Odds ratio for HIV</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>n</td>
<td>n minus 1</td>
</tr>
<tr>
<td>104</td>
<td>.1</td>
<td>.6</td>
<td>.7</td>
</tr>
<tr>
<td>300</td>
<td>.35</td>
<td>.43</td>
<td>.58</td>
</tr>
<tr>
<td>1076</td>
<td>.1</td>
<td>.35</td>
<td>.52</td>
</tr>
<tr>
<td>2000</td>
<td>.05</td>
<td>.21</td>
<td>.47</td>
</tr>
</tbody>
</table>

N = number of cases in the hypothetical complete cluster distribution
SF = sampling fraction based on (104/N)
Conclusions from sensitivity analysis

- The proportion of cases that were new may have been grossly underestimated.

- The risk of new cases in comparison with reactivated cases associated with HIV may have been grossly underestimated.