Study Design and Analysis in Epidemiology:
Where does modeling fit?

Meaningful Modeling of Epidemiologic Data, 2011
AIMS, Muizenberg, South Africa

Steve Bellan
MPH Epidemiology
PhD Candidate
Department of Environmental Science, Policy & Management
University of California at Berkeley
Defining Epidemiology

“The study of the distribution and determinants of health related states and events in populations, and the application of this study to control health problems.”

John M Last
Dictionary of Epidemiology
Varieties of Infectious Disease Epidemiology

• Risk Factors & Intervention Epidemiology

Risk Factor: A characteristic that is correlated with a measure of disease.

• Often used synonymously with covariate.

• Protective factors: Risk factors that are negatively associated with disease
Varieties of Infectious Disease Epidemiology

- Risk Factors & Intervention
- Outbreak
  - Clinical
  - Molecular & Genetic
  - Surveillance
How does mathematical modeling fit?

• A subfield of epidemiology:
  Linking pattern with process across scales

BUT ALSO

• A set of methodologies to be used in any field of epidemiology

*Importance of knowledge breadth*
What do *Introductory Epidemiology* courses teach?

- Measures of Disease
- Measures of Effect (of a risk factor)
- Study Designs for Measuring Effects
  - Dealing with random error
  - Dealing with confounding
  - Dealing with bias
- Biostatistical analyses for analyzing data
Measures of Disease

• Incidence
  – Cumulative Incidence
  – Incidence Density

• Prevalence
  – Point Prevalence
  – Period Prevalence

• Survivorship
  (time to event, such as mortality)
Measures of Covariates (risk factors)

- **Binary**: gender, smoker, circumcised
- **Nominal/Categorical**: geographic region
- **Continuous**: birth weight, T-cell count
- **Ordinal**: education, socioeconomic status (SES)
Measures of Effect

• How do you measure the effect of a risk factor on a disease?

Example

*How could you measure whether circumcision reduces the risk of HIV infection?*
Measures of Effect

- Compare measure of disease across levels/values of risk factors

- Relative Risk
  Ratio of rates or proportions
  - Prevalence Ratio
  - Cum. Incidence Ratio
  - Incidence Density Ratio
  - Odds Ratio

- Attributable Risk
  Subtract rates or proportions

Relative Risk (RR) & Attributable Risk (AR)

- Attributable Risk = 1000 per 100,000 people
- Relative Risk = 11000/10000 = 1.1

Example:

- Brush Incidence per 100,000 people
  - > 8 hours of sleep: 8000
  - < 8 hours of sleep: 12000

Graph showing the comparison between > 8 hours of sleep and < 8 hours of sleep in terms of flu incidence per 100,000 people.
### Contingency Tables: Relative Risk (RR)

<table>
<thead>
<tr>
<th></th>
<th>Disease</th>
<th>No Disease</th>
<th>Total (Margins)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exposed</td>
<td>a</td>
<td>b</td>
<td>a+b</td>
</tr>
<tr>
<td>Not exposed</td>
<td>c</td>
<td>d</td>
<td>c+d</td>
</tr>
<tr>
<td>Total (Margins)</td>
<td>a+c</td>
<td>b+d</td>
<td>a+b+c+d</td>
</tr>
</tbody>
</table>

Cumulative Incidence Ratio (CIR) is the ratio of cumulative incidence in the exposed population divided by the cumulative incidence in the unexposed population.

\[
CIR = \frac{\frac{a}{a+b}}{\frac{c}{c+d}}
\]

- CIR < 1 means exposure correlates with reduced risk of disease
- CIR > 1 means exposure correlates with increased risk of disease
Epidemiologic Studies

• Descriptive Epidemiology
  – Baseline data on distribution of disease
  – Surveillance

• Analytic Epidemiology – Measure Effect
  – Prospective Cohort Studies
  – Cross-sectional Studies
  – Retrospective Case-Control Studies
  – Ecologic Studies
  – Randomized Controlled Trials

Observational

Experimental
Cohort Studies

• Follow a selected population through time
  – Establishes temporal relationships
  – Can measure incidence

• Takes lots of resources, money, & time!

• Poor design for rare diseases.
## Relative Risk: Incidence Density Ratios

<table>
<thead>
<tr>
<th>Disease</th>
<th>No Disease</th>
<th>Total (Margins)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exposed</td>
<td>a</td>
<td>PY&lt;sub&gt;e&lt;/sub&gt;</td>
</tr>
<tr>
<td>Not exposed</td>
<td>c</td>
<td>PY&lt;sub&gt;0&lt;/sub&gt;</td>
</tr>
<tr>
<td>Total (Margins)</td>
<td>a+c</td>
<td>PY&lt;sub&gt;e&lt;/sub&gt; + PY&lt;sub&gt;0&lt;/sub&gt;</td>
</tr>
</tbody>
</table>

Incidence Density Ratio (IDR) is the ratio of incidence density of the exposed population to that of the unexposed population.

\[
IDR = \frac{\frac{a}{PY_e}}{\frac{c}{PY_0}}
\]

- IDR < 1 means exposure correlates with reduced risk of disease
- IDR > 1 means exposure correlates with increased risk of disease
Cross-Sectional Studies

- Snapshot of diseases & risk factors.
- Cannot establish temporal relationship.
- Relatively cheap & easy.
- Population must be large to study rare disease.
- Not great for diseases of short duration. Why?
Case-Control Studies

• Compare diseased individuals to chosen controls.
  – Quality of study depends entirely on how controls are chosen.

• Good for rare diseases.

• Relatively cheap & quick.
### Case Control Studies: Odds Ratios

Odds ratio is the ratio of odds in the diseased population divided by the odds in the non-diseased population.

\[
OR = \frac{a/d}{b/c} = \frac{ad}{bc}
\]

- OR < 1 means exposure correlates with reduced risk of disease
- OR > 1 means exposure correlates with increased risk of disease
Randomized Controlled Trials

- Experimental or Intervention Studies
- Establishes temporal relationships
- Addresses confounding (more to come)
Ecologic Studies

• Measurements made at population rather than individual level.

• Weaker inference, but easier to gather data.
Measures of Covariates (risk factors)

- **Binary**: gender, smoker, circumcised
- **Nominal/Categorical**: geographic region
- **Continuous**: birth weight, T-cell count
- **Ordinal**: education, socioeconomic status (SES)
What do *Introductory Epidemiology* courses teach?

- Measures of Disease
- Measures of Effect (of a risk factor)
- Study Designs for Measuring Effects
  - Dealing with random error
  - Dealing with confounding
  - Dealing with bias
- Biostatistical analyses for analyzing data
Random Error

• How many people must be in a study for the measure of effect to be believable?

• Statistical Approach:
  Assign probabilities to our findings being a product of random error rather than a real phenomenon.
Bias

*Difference between observed value and true value due to all causes other than random error.*

Bias does not go away with greater sample size!

Bias must be dealt with during study design!
Selection Bias

*Error due to systematic differences between those who take part in the study and those who do not.*

John Last, Dictionary of Epidemiology

Information Bias

*A flaw in measuring exposure or outcome data that results in different quality (accuracy) of information between comparison groups.*

John Last, Dictionary of Epidemiology
### Confounding

<table>
<thead>
<tr>
<th>Literacy</th>
<th>HIV Status</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>HIV+</td>
</tr>
<tr>
<td>Literate</td>
<td>660</td>
</tr>
<tr>
<td>Illiterate</td>
<td>180</td>
</tr>
</tbody>
</table>

\[
PR = \frac{660/1000}{180/1000} = 3.67
\]

What if some of the study population were much younger than others?
## Confounding

<table>
<thead>
<tr>
<th></th>
<th>HIV+</th>
<th>HIV-</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Pooled</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Literate</td>
<td>660</td>
<td>340</td>
</tr>
<tr>
<td>Illiterate</td>
<td>180</td>
<td>820</td>
</tr>
<tr>
<td><strong>6-15 years old</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Literate</td>
<td>30</td>
<td>270</td>
</tr>
<tr>
<td>Illiterate</td>
<td>90</td>
<td>810</td>
</tr>
<tr>
<td><strong>16-24 years old</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Literate</td>
<td>630</td>
<td>70</td>
</tr>
<tr>
<td>Illiterate</td>
<td>90</td>
<td>10</td>
</tr>
</tbody>
</table>

6-15 year olds: Literacy = \(\frac{300}{1200} = 25\%\)

16-24 year olds: Literacy = \(\frac{700}{800} = 87.5\%\)

\[
PR_{\text{all}} = \frac{660/1000}{180/1000} = 3.67
\]

\[
PR_{6-15\text{yrs}} = \frac{30/300}{90/900} = 1
\]

\[
PR_{16-24\text{yrs}} = \frac{630/700}{90/100} = 1
\]
Confounding

<table>
<thead>
<tr>
<th></th>
<th>HIV+</th>
<th>HIV-</th>
</tr>
</thead>
<tbody>
<tr>
<td>Literate</td>
<td>660</td>
<td>340</td>
</tr>
<tr>
<td>Illiterate</td>
<td>180</td>
<td>820</td>
</tr>
</tbody>
</table>

\[
PR = \frac{660/1000}{180/1000} = 3.67 \\
PR = \frac{30/300}{90/900} = 1 \\
PR = \frac{630/700}{90/100} = 1
\]
Biostatistical Analyses

• Permutation Tests

• Chi Squared Test

• Generalized Linear (Mixed) Models
  – Normal Regression
  – Logistic Regression
  – Poisson Regression
  – Negative Binomial Regression

• Survival Analysis
<table>
<thead>
<tr>
<th>Statistical Models</th>
<th>Dynamic Models</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Account for bias and random error to find correlations that may imply causality.</td>
<td>• Systems Approach: Explicitly model multiple mechanisms to understand their interactions.</td>
</tr>
<tr>
<td>• Often the first step to assessing relationships.</td>
<td>• Links observed relationships at different scales.</td>
</tr>
<tr>
<td>• Assume independence of individuals (at some scale, <em>i.e.</em> clusters).</td>
<td>• Explicitly focuses on dependence of individuals</td>
</tr>
</tbody>
</table>

By developing dynamic models in a probabilistic framework we can account for dependence, random error, and bias while linking patterns at multiple scales.
Questions in Epidemiology

Statistical Models

• Is HIV status positively associated with the risk of TB infection?

Dynamic Models

• Based on increased TB risk due to HIV, how much should we expect TB notification rate to increase for a given HIV prevalence?
Questions in Epidemiology

Statistical Models

• Are Insecticide Treated Bednets (ITNs) or Indoor Residual Spraying (IRS) more effective for controlling malaria?

Dynamic Models

• How do we expect the age-distribution of malaria incidence to change after implementing ITNs or IRS?