Overview of Epidemiologic Study Design

Daniel Hryhorczuk, MD, MPH
Goal of Epidemiologic Research

• Determine if there is an association between an exposure (E) and a disease (D)
Risk Factors for Disease

The causes of disease are often multifactorial
Measures of Disease Frequency

- **Incidence rate** =
  \[
  \text{Number of } \textit{new} \text{ cases during a period of time} \quad \text{Total population (or person-time) at risk}
  \]

- **Prevalence rate** =
  \[
  \text{Number of } \textit{existing} \text{ cases} \quad \text{at a designated time or period of time}
  \text{Total population}
  \]
Measures of Disease Frequency

"... just what we needed -- a puddle too big to walk around!"

**Incidence:** the rain arriving

**Prevalence:** the water in the puddle, new and old

**Period Prevalence:** during a period

**Point Prevalence:** at one point in time

The water draining away into the soil or into drains reduces the puddle (i.e. the prevalence), just as recovery or death reduces the number of patients with a problem.

Measures of Disease Frequency

• Two types of incidence rates:
  – Cumulative incidence
  – Person-time incidence rate
Cumulative Incidence

Used for fixed cohorts

\[ \hat{P} = \frac{d}{N} \]

\[ \hat{P} = \frac{3}{10} \]

\[ \hat{P} = 30\% \text{ over 5 years} \]
Person Time Incidence Rate

Year
1 2 3 4 5
Person Time Incidence Rate

Year

<table>
<thead>
<tr>
<th>Year</th>
<th>Person Time</th>
<th>Incidence Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>X</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>X</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>X</td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>X</td>
<td>4</td>
</tr>
<tr>
<td>5</td>
<td>X</td>
<td>5</td>
</tr>
<tr>
<td>6</td>
<td>X</td>
<td>1.5</td>
</tr>
<tr>
<td>7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

$t_i$
Person Time Incidence Rate

Year

\[\sum t_i = T\]

\[T = 30.5\]
Person Time Incidence Rate

*Used for dynamic cohorts*

\[ \tilde{p} = \frac{d}{T} \]

\[ \tilde{p} = \frac{3 \text{ persons}}{30.5 \text{ person-years}} \]

\[ \tilde{p} = 0.098/\text{year} \]
At time $N_t$, $\hat{p} = 1 - N_t / N_0$

$\hat{P} = -\frac{dN_t}{dt} \frac{1}{N_t}$

$\hat{P} = 1 - e^{\tilde{p}t}$
Point Prevalence

Prevalence = 2/10 = 20%

Year

Person
Associations

• The association between an exposure (E) and a disease (D) has two properties:
  – Strength
  – Stability
2 x 2 Table

<table>
<thead>
<tr>
<th></th>
<th>Yes</th>
<th>No</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>A</td>
<td>C</td>
</tr>
<tr>
<td>No</td>
<td>B</td>
<td>D</td>
</tr>
</tbody>
</table>

$E$  \quad  $N_1$  \quad  $N_2$  \quad  $N$

$M_1$  \quad  $M_2$
Measures of Association

• Measure the strength of the association
• For several (relative risk, rate ratio, prevalence ratio, odds ratio) if the value is one, then there is no association
• If it is less than one (between 0 and 1), it is protective
• The choice of the measure of association depends on the type of study
Relative Risk

\[ RR = \frac{A}{M_1} \div \frac{C}{M_2} \]

\[ RR = \frac{200}{500} \div \frac{50}{950} = 7.55 \]
Odds Ratio

\[
\text{OR} = \frac{AD}{BC}
\]

Lung Cancer

\[
\begin{array}{c|c|c|c}
& Y & N \\
Y & 200 & 300 \\
N & 50 & 950 \\
\end{array}
\]

Smoking

\[
\begin{array}{c|c|c|c}
& Y & N \\
Y & 200 & 300 \\
N & 50 & 950 \\
\end{array}
\]

\[
\text{OR} = \frac{200 \times 950}{50/300} = 12.6
\]
Rate ratio = $\frac{A/T_{exp}}{C/T_{non}}$

$RR = \frac{100/8000}{20/10000} = 6.25$
Stability

• Two Schools of Thought
  – Epidemiologic studies are experiments
    • Apply statistical tests such as the Chi Square Test
  – We are observing natural phenomena
    • Measure the stability of the association by applying a confidence interval around the estimate of the strength of the association
Confounding

1. Associations of a hypothetical exposure, disease, and confounding variable
Stratified Analysis

Stratify on the confounder
Calculate stratum specific odds ratios
Look for confounding or interaction
Use Mantel Haenszel methods
Example of Confounding

**Univariate Analysis**

<table>
<thead>
<tr>
<th></th>
<th>+ Lung Cancer</th>
<th>- Lung Cancer</th>
</tr>
</thead>
<tbody>
<tr>
<td>+ EtOH</td>
<td>36</td>
<td>24</td>
</tr>
<tr>
<td>- EtOH</td>
<td>24</td>
<td>36</td>
</tr>
</tbody>
</table>

Odds ratio = 2.25

**Multivariate Analysis**

**True Smokers**

<table>
<thead>
<tr>
<th></th>
<th>+ Lung Cancer</th>
<th>- Lung Cancer</th>
</tr>
</thead>
<tbody>
<tr>
<td>+ EtOH</td>
<td>34</td>
<td>6</td>
</tr>
<tr>
<td>- EtOH</td>
<td>18</td>
<td>2</td>
</tr>
</tbody>
</table>

Odds ratio = 0.63

**Non-smokers**

<table>
<thead>
<tr>
<th></th>
<th>+ Lung Cancer</th>
<th>- Lung Cancer</th>
</tr>
</thead>
<tbody>
<tr>
<td>+ EtOH</td>
<td>2</td>
<td>18</td>
</tr>
<tr>
<td>- EtOH</td>
<td>6</td>
<td>34</td>
</tr>
</tbody>
</table>

Odds ratio = 0.63

*Figure 7a.* Effect of alcohol abuse on the development of lung cancer.

*Figure 7b.* Effect of alcohol abuse on the development of lung cancer when patients are stratified by tobacco abuse.
Interaction

**Effect Modification in a Cohort Study**

<table>
<thead>
<tr>
<th>Heart Disease</th>
<th>yes</th>
<th>no</th>
</tr>
</thead>
<tbody>
<tr>
<td>Active  yes</td>
<td>51</td>
<td>76</td>
</tr>
<tr>
<td>no</td>
<td>800</td>
<td>600</td>
</tr>
</tbody>
</table>

Crude RR = 0.57

**Stratified Analysis**

**Men**

<table>
<thead>
<tr>
<th>Heart Disease</th>
<th>yes</th>
<th>no</th>
</tr>
</thead>
<tbody>
<tr>
<td>Active  yes</td>
<td>35</td>
<td>315</td>
</tr>
<tr>
<td>no</td>
<td>60</td>
<td>290</td>
</tr>
</tbody>
</table>

RR = 0.58 (0.40 - 0.86)

**Women**

<table>
<thead>
<tr>
<th>Heart Disease</th>
<th>yes</th>
<th>no</th>
</tr>
</thead>
<tbody>
<tr>
<td>Active  yes</td>
<td>11</td>
<td>339</td>
</tr>
<tr>
<td>no</td>
<td>9</td>
<td>341</td>
</tr>
</tbody>
</table>

RR = 1.22 (0.51 - 2.91)

Effect modification is present if the stratum-specific estimates of association differ from each other!
Multivariable Analysis

- For dichotomous outcome variables, use logistic regression
- For continuous outcome variables, use linear regression
Logistic Regression

\[ y = b_0 + b_1 x \]  \hspace{1cm} \text{Linear Model}

\[ p = \frac{1}{1 + e^{-(b_0 + b_1 x)}} \]  \hspace{1cm} \text{Logistic Model}
Types of Designs

- Observational
- Experimental
Types of Designs

Observational

Experimental

Clinical Trial

Community Intervention Trial
Types of Designs

- Observational
  - Descriptive
    - Case Report
  - Analytic
    - Case Series
- Experimental
  - Incomplete
    - Outbreak Investigation
Types of Designs

- Observational
  - Descriptive
    - Cross Sectional
  - Analytic
    - Case Control
- Experimental
  - Incomplete
    - Cohort
Types of Designs

Observational
- Descriptive
  - Cluster Analysis

Experimental
- Analytic
  - Ecologic
- Incomplete
  - PMR
Observational Analytic Designs

• Differ from each other
  – Time reference
  – Method of sampling from the target population
Time Reference

Cross Sectional

Cohort

Case Control

E

D

nonE

nonD
Validity

- **External**: how well does the sample represent the target population?
- **Internal**: is the study free of other types of bias?
Bias

• Bias: any systematic error in the design or conduct of a study

• Common types of bias
  – Sampling bias
  – Measurement bias
Sampling from Target Population

<table>
<thead>
<tr>
<th></th>
<th>D</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>Y</td>
<td>800</td>
<td>9,200</td>
</tr>
<tr>
<td>N</td>
<td>200</td>
<td>89,800</td>
</tr>
</tbody>
</table>

Target Population

<table>
<thead>
<tr>
<th></th>
<th>D</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>Y</td>
<td>10,000</td>
<td>100,000</td>
</tr>
<tr>
<td>N</td>
<td>90,000</td>
<td>99,000</td>
</tr>
</tbody>
</table>

Sample
Cross-Sectional Sampling

[Diagram showing a 2x2 contingency table with labels A, B, C, D, a, b, c, d, Y, N, E, n.]
Cross-Sectional Sampling

100,000 1,000
Case Control Sampling

- Category A: 1000
- Category B: 99,000
- Category C: 100
- Category D: 200

Diagram:

- Two tables with categories A, B, C, D and a, b, c, d
- Y (Yes) and N (No) for each category
- 1000 and 99,000
- 100 and 200
Cohort Sampling

\[
\begin{array}{cc}
Y & D \\
N & \end{array}
\]

\[
\begin{array}{cc}
A & B \\
C & D \\
\end{array}
\]

\[
\begin{array}{cc}
a & b \\
\ \ & \end{array}
\]

\[
\begin{array}{cc}
c & d \\
\ \ & \end{array}
\]
Cohort Sampling

\[ \begin{array}{c|c|c|c}
Y & D & N \\
\hline
90,000 & Y & E \quad 10,000 & Y \\
\hline
1000 & N & 1000 & N
\end{array} \]
Cross Sectional Study
Types of Designs

Observational
  - Descriptive
    - Cross Sectional
  - Analytic
    - Case Control

Experimental
  - Incomplete
    - Cohort
Cross-Sectional Study

- Evaluation of a sample or entire population for the simultaneous presence or absence of disease and exposure
Time Reference

Cross Sectional

E

D

Time
Cross-Sectional Study

- Common study design
- Inexpensive
- Cross-sectional studies are typically morbidity (measuring current health status) studies
Cross-Sectional Sampling

• Random sample from N in the target population
  – Example: a survey of a representative sample of a population

• Simultaneous determining the prevalence of disease and presence of a risk factor
  – Example:
    • Do you currently smoke?
    • Do you currently have chronic bronchitis?
Cross-Sectional Sampling

\[
\begin{array}{c|c|}
A & B \\
\hline
C & D \\
\end{array}
\]

\[
\begin{array}{c|c|}
a & b \\
\hline
c & d \\
\end{array}
\]
Hybrid: Cross-sectional/Cohort

• We are more often interested in sampling on exposed ($M_1$) vs unexposed ($M_2$) than we are in doing a survey of a general population

• We can do a hybrid design that combines elements of both cohort and cross-sectional studies
Hybrid: Cross-sectional/Cohort

- The cohort component is sampling on exposed ($M_1$) vs unexposed ($M_2$)
- The cross-section component is using prevalence as the measure of disease frequency rather than incidence
Hybrid: Cross-sectional/Cohort

<p>| | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>E</td>
<td>Y</td>
<td>A</td>
<td>B</td>
</tr>
<tr>
<td>N</td>
<td>Y</td>
<td>C</td>
<td>D</td>
</tr>
</tbody>
</table>

<p>| | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>E</td>
<td>Y</td>
<td>a</td>
<td>b</td>
</tr>
<tr>
<td>N</td>
<td>Y</td>
<td>c</td>
<td>d</td>
</tr>
</tbody>
</table>

M₁ → M₂ → M₁

m₁ → m₂ → m₁
Defining the Research Question

- What is the research question?
- What are the exposures of interest?
- What is the disease of interest?
Who is the Target Population?

- Specify your target population
- It could be as large or as small as you like
How Will You Sample?

- Decide how you will select your sample
- Will you study the entire population?
- Will you sample on exposure?
- How well your sample reflects your target population determines the external validity of your study
Ascertaining Exposure

• How will exposure be measured?
  – Questionnaire?

• How will exposure be classified?
  – Yes/No
  – Categories?
  – Quantitative estimate?
Ascertaining Health Status

• Questionnaire
  – Exposure questionnaire
  – Health questionnaire
  – Quality control
    • Use validated questionnaires, if available
    • Pilot test your questionnaire
    • The exposure and health questionnaires should be administered by different interviewers
    • Check completed questionnaires on site for completeness, consistency
Data Management

• Need protocols and procedures for data management
  – Data editing
  – Data reduction
  – Data entry
Data Analysis

• Crude
• Stratified
• Multivariable
Point Prevalence

Prevalence

= 2/10
= 20%
Prevalence Ratio

\[
\begin{array}{cc}
\text{E} & \text{D} \\
\text{Yes} & A & C \\
\text{No} & B & D \\
\end{array}
\]

\begin{align*}
M_1 &= A \\
M_2 &= B \\
N_1 &= \text{Yes} \\
N_2 &= \text{No} \\
N &= \text{Numerator} \\
M_1 + M_2 &= \text{Numerator} \\
B + M_2 &= \text{Numerator} \\
\end{align*}

\[PR = \frac{A}{M_1} \quad \frac{B}{M_2} \]

Chi Square Test
Limitations

• The major limitation of the cross-sectional study is using prevalence rather than incidence as the measure of disease frequency
• Results in incidence-prevalence bias
When will the Prevalence Ratio Approximate the Relative Risk?

Exposed

Unexposed

Time

One case

Two cases
When will the Prevalence Ratio Approximate the Relative Risk?

• When the disease status does not affect the exposure status

• Example
  – If an exposure is resulting in asthma, and sick workers leave the job, then the prevalence of asthma in that department will be erroneously thought to be low
Case Control Study
Types of Designs

- Observational
  - Descriptive
    - Cross Sectional
  - Analytic
    - Case Control
- Experimental
  - Incomplete
    - Cohort
Time Reference

Case Control

D

nonD

Time
Case Control Studies

- Sample on disease yes/no
- Useful for rare diseases
- Select incident cases of disease
- Need strict case definition
- Single disease – can ask multiple exposure questions
Case Control Studies

• Population-based
• Hospital based
Case Control Studies

• Selection of the control group is the Achilles heel of the case control study
Case Control Studies

• Ask about past exposures
• Typically with questionnaire
Case Control Sampling

A  B  C  D
Y   D  N
Y   A  B
E   C  D
N   N_1  N_2

a  b
Y   D  N
Y   a  b
E   c  d
N   n_1  n_2
Odds Ratio

\[ \text{OR} = \frac{AD}{BC} \]

Chi Square Test
Odds Ratio

\[
\text{OR} = \frac{AD}{BC}
\]

\[
\text{OR} = \frac{200 \times 950}{50/300} = 12.6
\]
Cohort Study
Types of Designs

- Observational
  - Descriptive
    - Cross Sectional
  - Analytic
    - Case Control
- Experimental
  - Incomplete
    - Cohort
Time Reference

Cohort

E →
nonE →

Time
Cohort Studies

- Exposed yes/no
- Single exposure – multiple disease outcomes
- Prospective or historical
- Not good for rare diseases
Cohort Sampling
Cohort Sampling

<table>
<thead>
<tr>
<th>Y</th>
<th>D</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>E</td>
<td>90,000</td>
<td>10,000</td>
</tr>
<tr>
<td>N</td>
<td>1000</td>
<td>1000</td>
</tr>
</tbody>
</table>

1000
Relative Risk

RR = \frac{A/M_1}{C/M_2}

RR = \frac{200/500}{50/950} = 7.55
Rate ratio = \frac{A}{T_{\text{exp}}} = \frac{C}{T_{\text{non}}} \quad \text{RR} = \frac{100}{8000} = \frac{20}{10000} = 6.25
Ecologic Studies
Types of Designs

Observational
- Descriptive
  - Cluster Analysis

Experimental
- Analytic
- Incomplete
  - Ecologic
  - PMR
Definition

- An ecologic study is an epidemiologic study in which the *group* rather than the *individual*, is the unit of observation.
- The groups chosen for study are typically geographically-defined areas, e.g. city, state, country.
Aims

• The major aims of ecologic studies are to:
  – Generate etiologic hypotheses
  – Evaluate the effectiveness of a population intervention

• Ecologic studies are “incomplete” study designs

• Because of their inherent flaws, they are generally not used to test etiologic hypotheses
Example of Ecologic Study

- Aim: explore association between air pollution and infant mortality
- Select several different cities as your groups for study
- Measure air pollution in those cities
- Measure infant mortality in those cities
- Analyze whether increased air pollution is associated with increased infant mortality between these cities
### Group % Exposed % Diseased

<table>
<thead>
<tr>
<th>Group</th>
<th>% Exposed</th>
<th>% Diseased</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>( M_{11} / N_1 )</td>
<td>( N_{11} / N_1 )</td>
</tr>
<tr>
<td>2</td>
<td>( M_2 / N_2 )</td>
<td>( N_2 / N_2 )</td>
</tr>
<tr>
<td>3</td>
<td>( M_3 / N_3 )</td>
<td>( N_3 / N_3 )</td>
</tr>
</tbody>
</table>
Advantages of Ecologic Studies

• Typically done by combining existing data files
  – Exposure data file
  – Disease data file
  – E and D linked by geographic unit (group)
• Less expensive than other designs
• Can be done more quickly than other types of studies
Disadvantages

• Ecologic Fallacy
  – Making a causal inference about individual phenomena on the basis of observations on groups
Analysis of Data

• Two approaches
  – Correlation analysis
  – Regression analysis
Regression Analysis

\[ y = B_0 + B_1 x \]

Disease (y)

Exposure (x)

Group data

\((x_1, y_1)\)
Individual Data

Disease \((y)\)

Exposure \((x)\)
Individual Data

Relationship holds but more spread out due to random error or effects of other variables.
Individual Data

- No relationship
- Inverse relationship
- No relationship

Disease (y)

Exposure (x)
Four data sets
Still relationship, but now
Variability between groups
Individual Data

Two data sets. Relationship holds. Variability at the individual level.
Ecologic Level

We assume this is the nature of the Relationship.
Ecologic Level

But it could just as easily be this!
Ecologic Fallacy

\[ r_T = r_w \sqrt{1 - C_x} \sqrt{1 - C_Y} + r_B \sqrt{C_X C_Y} \]

Where
- \( r_T \) = total correlation
- \( r_w \) = within groups correlation
- \( r_B \) = between groups correlation
- \( C_x \) is a measure of homogeneity of exposure; when everyone in a Group has the same level of exposure, \( C_x = 1 \)
- Similarly, \( C_Y \) is a measure of homogeneity of disease
Individual Data

Homogeneity of disease
\( C_y = 1 \)
Individual Data

Disease \((y)\)

Exposure \((x)\)

Homogeneity of exposure

\(C_x = 1\)
Individual Data

Homogeneity on exposure and disease; When all individual variability is eliminated, then there is no ecologic fallacy.
Minimizing Problems with Ecologic Studies

• Look for homogeneity on exposure within the groups you select; i.e. everyone in the geographic unit has a similar level of exposure

• Use multiple regression rather than correlation; allows you to control for confounding