Biostatistics for Health Care Researchers: A Short Course

Study Design

Presented by:
Patrick O. Monahan, PhD
Division of Biostatistics
Indiana University School of Medicine
Objectives

• Describe the elements of a research study
• Describe the main observational study designs
  – Epidemiologic
  – Outcomes
• Describe estimates of exposure-disease association, including relative, attributable, and population attributable risk and odds ratios.
How Does Medicine Advance

• Observation
• Informal Experimentation
• Formal Experimentation
  – Evidence based medicine (RCT)
  – Causal Inference
Elements of a Study

- QUESTION and TESTABLE HYPOTHESIS
- Outcome variable and its measurement
- Experimental Study Design
- Population
- Protocol
- Analyses
- Conclusion should answer question
Question and Hypothesis

• The **QUESTION** is the original problem that prompted the study

• The **HYPOTHESIS** is a rephrasing of the question in a statistically testable form

• Example
  – Question - does smoking cause lung cancer?
  – Hypothesis – Do people who smoke have a greater incidence of lung cancer than non-smokers?
Outcome Measure

• Reflects the hypothesis to be tested
• Clearly defined
• Appropriate for hypothesis (and question)
Epidemiologic Study Designs

- Case Series
- Assess association between risk factors and disease
- Ecologic
- Cross-Sectional
- Case-Control
- Prospective and Historical Prospective
Measures of Disease Frequency

- **Prevalence (P)** - proportion of individuals who have the disease at a specific time
- **Cumulative Incidence (CI)** - proportion of individuals who become diseased during a specified time interval
- **Incidence Rate (I)** - number of individuals who become diseased divided by person-time observed
Risk vs. Rate

• Prevalence - risk that an individual will be ill at a given point in time
• Cumulative Incidence - risk that an individual will develop the disease in a specific time interval
• Incidence Rate - instantaneous rate of development of disease in a population
Ecologic Studies

• Purpose:
  – Describe patterns or trends on a geographic level
  – explore potential associations between community-level risk factors and disease

• Design:
  – Collect group level data (Country, state, city, etc.)

• Data:
  – Disease incidence or prevalence in each population
  – Risks and confounding factors of populations
Ecological Study: Breast Cancer

- Data collected in the early 1970’s from 39 different countries
- Relationship between per capita fat consumption and death rate from breast cancer
- Examined animal and vegetable fat separately
Fat Intake and Breast Cancer

[Graph showing correlation between fat intake and breast cancer death rates across various countries.]

- Netherlands
- Canada
- Denmark
- U.K.
- New Zealand
- Belgium
- U.S.A.
- Ireland
- Sweden
- Austria
- Norway
- Germany
- Australia
- France
- Finland
- Hungary
- Poland
- Portugal
- Bulgaria
- Hong Kong
- Venezuela
- China
- Panama
- Greece
- Romania
- Spain
- Yugoslavia
- Puerto Rico
- Philippines
- Columbia
- Mexico
- Taiwan
- Ceylon
- El Salvador

Age adjusted death rate/100,000 pop (y)

Animal fat intake ($x_a$), gm/day
Fat Intake and Breast Cancer
Ecologic Studies

• Advantages:
  – relatively quick and inexpensive
  – Allows estimation of effects not easily measurable for individuals
  – Permits exploratory analyses of potential factors in disease etiology

• Disadvantages:
  – Risk factors and disease endpoints are not measured on the same individuals
  – Difficult to control for confounders
Cross-Sectional Studies

• **Purpose:**
  - explore potential associations between individual-level characteristics and disease endpoints

• **Design:**
  - A single examination of a group of individuals
  - Physical Examination, Questionnaires

• **Data:**
  - Presence of disease (Prevalence)
  - Measurement of risk and confounding factors
Cross-Sectional Study: NHANES

• NHANES (National Health and Nutrition Examination Survey) Study
  – Random sample of the entire country
  – Physical and laboratory examinations
  – Dietary questionnaires
  – Looks at relationships between dietary intake and disease presence
  – Dietary data is self-report
Cross-Sectional Studies

• Advantages:
  – Disease and risk factor data are collected from the same individuals
  – Complete data collection

• Disadvantages:
  – Higher proportion of long term survivors
  – No data on time relationship between risk factors and disease development
Case-Control Studies

• Purpose:
  – Determine whether those with a disease, i.e., prevalent cases, differ from those without the disease

• Design:
  – Define and identify ‘cases’
  – Identify a comparison group - ‘controls’
  – Measure risk factors

• Data:
  – exposure to risk factors in case and controls
# Measures of Association - Odds Ratios

<table>
<thead>
<tr>
<th>Exposure</th>
<th>Disease</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>Yes</td>
<td>$P_1$</td>
<td>A</td>
</tr>
<tr>
<td></td>
<td>No</td>
<td>$1 - P_1$</td>
<td>B</td>
</tr>
<tr>
<td>No</td>
<td>Yes</td>
<td>$P_2$</td>
<td>C</td>
</tr>
<tr>
<td></td>
<td>No</td>
<td>$1 - P_2$</td>
<td>D</td>
</tr>
</tbody>
</table>

- $P_1 = \text{prop. of exposed with disease} = A/(A + B)$
- $P_2 = \text{prop. of non-exposed with disease} = C/(C + D)$
- Odds ratio $\text{OR} = \frac{P_1}{P_2} = \frac{P_1(1-P_2)}{(1-P_1)P_2} = \frac{AD}{BC}$
## Case-Control Study: Smoking

<table>
<thead>
<tr>
<th>Cigarette Smoking</th>
<th>Lung Cancer</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Cases</td>
<td>Controls</td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>1350</td>
<td>1296</td>
<td></td>
</tr>
<tr>
<td>No</td>
<td>7</td>
<td>61</td>
<td></td>
</tr>
</tbody>
</table>

OR = $\frac{(1350)(61)}{(1296)(7)} = 9.1$

Case-Control Study: Febrile Seizures

<table>
<thead>
<tr>
<th></th>
<th>Crude OR</th>
<th>P-value for trend</th>
<th>OR adjusted for maternal age, education, and race</th>
<th>P-value for trend</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non-smoker</td>
<td>1.0</td>
<td></td>
<td>1.0</td>
<td></td>
</tr>
<tr>
<td>Quit during pregnancy</td>
<td>1.6 (1.0-2.7)</td>
<td>1.2 (0.7-2.1)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Smoked throughout pregnancy</td>
<td>2.1 (1.5-2.9)</td>
<td>&lt;0.001</td>
<td>2.0 (1.3-2.8)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>1-10 cigs per day</td>
<td>2.1 (1.2-3.5)</td>
<td></td>
<td>1.6 (0.9-2.9)</td>
<td></td>
</tr>
<tr>
<td>11-20 cigs per day</td>
<td>2.0 (1.2-3.5)</td>
<td></td>
<td>2.0 (1.2-3.4)</td>
<td></td>
</tr>
<tr>
<td>=&gt; 21 cigs per day</td>
<td>2.7 (1.2-6.1)</td>
<td>&lt;0.001</td>
<td>2.6 (1.0-6.6)</td>
<td>&lt;0.001</td>
</tr>
</tbody>
</table>

Numbers in parentheses are 95% confidence intervals

Case-Control Studies

• Advantages:
  – Quick ascertainment of cases
  – Requires administration of questionnaires to a relatively small study population
  – useful for rare disease

• Disadvantages:
  – Potential for recall bias
  – Potential bias due to participation of non-representative group of controls (or cases)
  – generally can’t assess incidence or prevalence
Prospective Studies

• Purpose:
  – Establish incidence rates for disease
  – Estimate associations between risk factors and incidence of disease or survival

• Design:
  – Identify disease free cohort, follow over time

• Data:
  – collected repeatedly over time on presence of risk and confounding factors and disease development
  – Use physical exams and questionnaires
Prospective Study: British Male Doctors

  - \( N = 34,440 \)
  - Smokers: 140 lung cancer deaths per 100,000 subjects per year
  - Non-smokers: 10 lung cancer deaths per 100,000 subjects per year
Measures of Association

- Relative Risk or Rate (RR) = $\frac{CI_e}{CI_u}$ or $\frac{I_e}{I_u}$

- Attributable Risk or Rate (AR) = $CI_e - CI_u$ or $I_e - I_u$

- Population Attributable Risk (PAR) = AR $\times$ proportion exposed
British Male Doctors

- RR = 140/10 = 14. The risk of lung cancer death is 14 times higher in smokers than non-smokers.
- AR = 140 - 10 = 130. The excess occurrence of lung cancer in smokers due to smoking is 130 per 100,000 subjects per year.
- PAR = 130(.1)=13
  - if proportion exposed=.1
Prospective Studies

- **Advantages:**
  - Measures risk factors prior to the development of disease (i.e. time relationship)
  - Permits ascertainment of true incidence rates
  - Useful for rare exposures

- **Disadvantages:**
  - Time
  - Expensive
  - Loss to follow-up bias
  - Difficult to study rare diseases
Historical Prospective Studies

• Purpose:
  – same as usual prospective study

• Design:
  – track disease occurrence between baseline and the present

• Data:
  – records for a cohort established some time in the past
Historical Prospective Study: Allegheny Co. Steelworkers

• All steelworkers employed for at least one month between 1/1/52 and 12/31/52
• Estimate exposure. Time employed in specific occupations (e.g., coke oven)
• Identify comparison populations, e.g. county workers, other mill workers
### Observed and Expected Mortality, 1953-1975, for Allegheny County Steelworkers

<table>
<thead>
<tr>
<th>Cause</th>
<th>Obs.</th>
<th>Exp.</th>
<th>RR</th>
<th>Obs.</th>
<th>Exp.</th>
<th>RR</th>
<th>Obs.</th>
<th>Exp.</th>
<th>RR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Respiratory Organs</td>
<td>121</td>
<td>83.4</td>
<td>*1.55</td>
<td>90</td>
<td>50.1</td>
<td>*2.03</td>
<td>31</td>
<td>32.4</td>
<td>0.95</td>
</tr>
<tr>
<td>Digestive Organs and Peritoneum</td>
<td>84</td>
<td>79.6</td>
<td>1.06</td>
<td>34</td>
<td>44.4</td>
<td>.74</td>
<td>50</td>
<td>34.6</td>
<td>*1.48</td>
</tr>
<tr>
<td>Genito-Urinary Organs</td>
<td>41</td>
<td>32.2</td>
<td>1.32</td>
<td>28</td>
<td>18.9</td>
<td>*1.56</td>
<td>13</td>
<td>12.8</td>
<td>1.02</td>
</tr>
</tbody>
</table>

*significant at .05 level
Association of Hepatitis C and Chronic Kidney Disease

• Using RMRS, identify subjects with Hep C test between 1994 and 2004 and no previous or concurrent CKD (N=8,224)

• CKD defined by elevated serum creatinine

• Use database to follow forward for CKD
Results

<table>
<thead>
<tr>
<th></th>
<th>CKD</th>
<th>No CKD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hep C positive</td>
<td>428 (17.5)</td>
<td>2,049</td>
</tr>
<tr>
<td>Hep C negative</td>
<td>821 (14.3%)</td>
<td>4,926</td>
</tr>
</tbody>
</table>

OR = \frac{428 \times 4926}{821 \times 2049} = 1.25
95% CI = (1.10, 1.42)
RR = 1.28
Historical Prospective Studies

• Advantages:
  – Same as usual prospective studies, plus …
  – Data already exist, so study can be done almost as quickly as case-control studies

• Disadvantages:
  – limited data available
  – may be unrecorded changes in risk factors
  – few suitable cohorts exist, or have data available for study
Case Control vs. Cohort

• CASE CONTROL
  – Collect data on exposure presence in the past
  – Look back in time
  – Good for rare diseases
  – Small Samples
  – Inexpensive
  – Factors related to presence of disease

• COHORT
  – Collect data on disease development in future
  – Follow over time
  – Good for rare exposures
  – Large samples
  – Expensive
  – Factors related to development of disease
Outcomes Studies

• Purpose:
  – relate health care delivery to outcomes
• Design:
  – develop instrument to measure outcome if necessary
  – relate outcome to health care delivery (study design can be observational or experimental)
Common Outcomes Measures

• Mortality
• Morbidity
• Disease Severity - PASI, NYHA
• Health Status - SF-36, SIP
• Quality of Life - QWB

Instruments are often specific to disease: e.g. AIMS
Health Care Delivery Studies

• Often Cross-sectional in design, using existing data bases
• Advantages:
  – Inexpensive
  – Questions not amenable to formal experiment
• Disadvantages:
  – Sparse and incomplete data
  – Population limitations
  – Treatment selection bias
Effect of Hospital Volume on Angioplasty Complications

![Bar chart showing the effect of hospital volume on mortality and bypass procedures each year. The chart displays the percentage of mortality and bypass procedures for hospitals with volumes <50, 50-100, and >100 procedures each year.]
Outcome Studies of Treatment

• BE VERY CAREFUL!
• Biased assignment to treatment
  – Treatments given to optimize outcomes
• Some statistical techniques available (e.g. propensity scores) but can only adjust for observed differences
• Cannot adjust for gross imbalance
Review

• Main observational study designs
  – Ecologic
  – Cross-sectional
  – Case-control
  – Cohort (prospective and historical)

• Estimates of exposure-disease association,
  – relative, attributable, and population attributable risk and odds-ratios