CASE-CONTROL STUDIES

Daniel E. Ford, MD, MPH
Vice Dean of Clinical Investigation
Johns Hopkins School of Medicine
Introduction to Clinical Research
July 17, 2013

STUDY DESIGNS AND CORRESPONDING QUESTIONS

- Ecologic
  - What explains differences between groups?

- Case Series
  - How common is this finding in a disease?

- Cross-sectional
  - How common is this disease or condition?

- Case-control
  - What factors are associated with having a disease?

- Prospective
  - How many people will get the disease?
  - What factors predict development?

CASE-CONTROL STUDY DESIGN

- Select participants based on presence or absence of disease or condition
- Compare frequency of risk factor (exposure) between persons with disease or condition and those without
- Differs from cross sectional studies in that the investigator determines the numbers of cases and controls rather than being determined by prevalence
CASE-CONTROL STUDIES

ASSUMPTIONS

• Cases are representative of everyone with the disease

• Controls are representative of everyone without the disease

• Information on risk factors is collected in the same way for cases and controls

RISK OF MORTALITY ASSOCIATED WITH LOW Kt/V IN 200 HEMODIALYSIS CASES AND CONTROLS

<table>
<thead>
<tr>
<th>Death</th>
<th>Survivors</th>
</tr>
</thead>
<tbody>
<tr>
<td>N=100</td>
<td>N=100</td>
</tr>
<tr>
<td>Low Kt/V</td>
<td>70</td>
</tr>
<tr>
<td>High Kt/V</td>
<td>30</td>
</tr>
</tbody>
</table>

Odds Ratio = \( \frac{(70)(70)}{(30)(30)} = 5.4 \)
Locating Cases

- Cases may be
  - Undiagnosed
  - Misdiagnosed
  - Unavailable for study
  - Dead (from the disease)
- Need to ask if cases are representative of all cases
- Examples, patients with myocardial infarctions or patients with diabetes

CASE-CONTROL STUDIES

SOURCES OF PARTICIPANTS

<table>
<thead>
<tr>
<th>Cases</th>
<th>Controls</th>
</tr>
</thead>
<tbody>
<tr>
<td>All cases in a community</td>
<td>Random sample of community</td>
</tr>
<tr>
<td>All hospitalized cases</td>
<td>Sample of hospitalized non-cases</td>
</tr>
<tr>
<td>All cases in a single hospital</td>
<td>Non-cases from the same neighborhood</td>
</tr>
<tr>
<td>Any of the above</td>
<td>Neighborhood, spouses, sibs, co-workers</td>
</tr>
</tbody>
</table>

SELECTION OF CONTROLS

- Hardest part of doing a case-control study
- No one perfect control group
- Run the risk of selection bias or overmatching
  - Selection bias: control group not representative of all noncases, e.g., coffee and cancer of the pancreas
  - Overmatching: creates similar exposure in cases and controls
SELECTION OF CONTROLS

• Community-based controls are preferable but often no single best control group
• Often there may be some selection of cases, i.e., single hospital, illness status, that can be offset by similar selection of controls
  – “compensatory bias”
• Use multiple control groups

CASE-CONTROL STUDIES
CONFOUNDING

• Factor associated with both exposure and outcome creates a noncausal association

Alcohol Abuse  ➔  CAD

Tobacco Use  ➔  CAD
CASE-CONTROL STUDIES
CONFOUNDING

• Factor associated with both exposure and outcome creates a noncausal association

Male

Gender

Smoking

Lung Cancer

CASE-CONTROL STUDIES
MATCHING

• Makes case and control groups more comparable
• Easy to explain and understand
• Time-consuming, logistically difficult
• Frequency or group matching may be easier
• Cannot study association of matched factor with disease

CASE-CONTROL STUDIES
ADJUSTMENT

• Accomplished during analysis using stratified analysis (young and old, diabetic and nondiabetic) or multivariate models
• Easier than matching but not as intuitive to some readers
• Need large subgroups (age, race, etc.)
CASE-CONTROL STUDIES

ADVANTAGES

• May be only way to study etiology of rare disease
• Study multiple etiologic factors simultaneously, generate hypotheses
• Less time-consuming and cheaper than prospective studies
• Require smaller sample size than prospective studies

CASE-CONTROL STUDIES

DISADVANTAGES

• Temporal relationship between exposure and disease unclear
• Subject to bias because disease has already occurred
• Cannot estimate incidence or prevalence
• Cannot study rare exposures
• Limited to one outcome

NESTED CASE-CONTROL STUDIES

• Uses incident cases from a prospective study
• Controls drawn from non-cases in the same study
• Data collected at baseline prior to development of disease
• Efficient and avoids usual limitations of case-control studies
NESTED CASE-CONTROL DESIGN

1. identify cases
2. draw representative sample of controls
3. measure risk factors

NESTED CASE-CONTROL STUDIES WHEN SHOULD THEY BE DONE?

• Limited number of cases in prospective study, e.g., rare outcomes like sudden death
• Efficient use of biological specimens and resources, e.g., serum bank
• Not enough resources to study all cases
• Avoids usual limitations of case-control studies, i.e., temporality relationship clear

ODDS RATIOS

• Measure of association in case-control studies
• Odds of exposure among cases compared with controls
• Approximates relative risk if disease is rare
**ODDS**

**DEFINITION**

Odds = probability / 1 - probability

Odds of randomly picking Sunday as a day of the week is 1/6 not 1/7.

---

**ODDS RATIO OF EXPOSURE**

<table>
<thead>
<tr>
<th></th>
<th>Cases</th>
<th>Controls</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exposed +</td>
<td>a</td>
<td>b</td>
</tr>
<tr>
<td>Exposed -</td>
<td>c</td>
<td>d</td>
</tr>
</tbody>
</table>

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

---

**CALCULATION OF ODDS RATIOS**

<table>
<thead>
<tr>
<th></th>
<th>Yes</th>
<th>No</th>
</tr>
</thead>
<tbody>
<tr>
<td>N=100</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>30</td>
<td>30</td>
</tr>
<tr>
<td>No</td>
<td>70</td>
<td>70</td>
</tr>
</tbody>
</table>

Odds ratio = \frac{ad}{bc} = \frac{(30)(70)}{(70)(30)} = 1
**CALCULATION OF ODDS RATIOS**

<table>
<thead>
<tr>
<th></th>
<th>Yes</th>
<th>No</th>
</tr>
</thead>
<tbody>
<tr>
<td>N=100</td>
<td>40</td>
<td>25</td>
</tr>
<tr>
<td>No</td>
<td>60</td>
<td>75</td>
</tr>
</tbody>
</table>

Odds ratio = \( \frac{(40)(75)}{(60)(25)} \)

Odds ratio = 2

---

**PREVALENCE OF LOW Kt/V AND MORTALITY JANUARY TO DECEMBER, 1996**

<table>
<thead>
<tr>
<th></th>
<th>Dead</th>
<th>Alive</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low Kt/V</td>
<td>400</td>
<td>1,000</td>
<td>1,400</td>
</tr>
<tr>
<td>High Kt/V</td>
<td>350</td>
<td>1,250</td>
<td>1,600</td>
</tr>
<tr>
<td></td>
<td>750</td>
<td>2,250</td>
<td>3,000</td>
</tr>
</tbody>
</table>

Odds Ratio = \( \frac{(400)(1,250)}{(350)(1,000)} = 1.4 \)

---

**NICE THINGS ABOUT THE ODDS RATIO**

- When an event is common, upper range of relative risk limits are constrained—less so for odds ratio
- Very sensitive to small changes in exposure
- Estimated by logistic regression analysis, a commonly used form of multivariate analysis
CAUTIONS ABOUT ODDS RATIOS

• Overestimates the relative risk
• Must be careful in interpretation
• Classic example: NEJM article on using videotaped vignettes to diagnose coronary heart disease
ODDS OF REFERRAL IN WOMEN COMPARED TO MEN

<table>
<thead>
<tr>
<th>Referral</th>
<th>+</th>
<th>-</th>
</tr>
</thead>
<tbody>
<tr>
<td>Women</td>
<td>85</td>
<td>15</td>
</tr>
<tr>
<td>Men</td>
<td>90</td>
<td>10</td>
</tr>
</tbody>
</table>

| OR        | $\frac{(85)(10)}{(90)(15)} = 0.63$ |

Percent of men not referred for cath: 10%
Percent of women not referred for cath: 15%

Why not use relative risk as a measure of association in case-control studies?
### CASE-CONTROL STUDY WITH 1:1 MATCHING

<table>
<thead>
<tr>
<th>MI</th>
<th>+</th>
<th>-</th>
</tr>
</thead>
<tbody>
<tr>
<td>CIG +</td>
<td>50</td>
<td>25</td>
</tr>
<tr>
<td>-</td>
<td>25</td>
<td>50</td>
</tr>
<tr>
<td>75</td>
<td>75</td>
<td>150</td>
</tr>
</tbody>
</table>

RR = \( \frac{50}{75} \div \frac{25}{75} = 2.0 \)

OR = \( \frac{(50)(50)}{(25)(25)} = 4.0 \)

### CASE-CONTROL STUDY WITH 10:1 MATCHING

<table>
<thead>
<tr>
<th>MI</th>
<th>+</th>
<th>-</th>
</tr>
</thead>
<tbody>
<tr>
<td>CIG +</td>
<td>50</td>
<td>250</td>
</tr>
<tr>
<td>-</td>
<td>25</td>
<td>500</td>
</tr>
<tr>
<td>75</td>
<td>750</td>
<td>825</td>
</tr>
</tbody>
</table>

RR = \( \frac{50}{300} \div \frac{25}{525} = 3.5 \)

OR = \( \frac{(50)(500)}{(25)(250)} = 4.0 \)

### CASE-CONTROL STUDIES BIAS

- Systematic difference from the truth
- Can occur at any phase of the study: design, data collection, analysis
- Main types are selection, information, analysis
SELECTION BIAS  
GENERAL POPULATION

<table>
<thead>
<tr>
<th>Disease Status</th>
<th>Yes</th>
<th>No</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Risk Factor</td>
<td>A</td>
<td>B</td>
<td>A+B</td>
</tr>
<tr>
<td>Yes</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No</td>
<td>C</td>
<td>D</td>
<td>C+D</td>
</tr>
<tr>
<td>Total</td>
<td>A+C</td>
<td>B+D</td>
<td></td>
</tr>
</tbody>
</table>

OR = AD/BC

STUDY SAMPLE

<table>
<thead>
<tr>
<th>Disease Status</th>
<th>Yes</th>
<th>No</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Risk Factor</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>S1A</td>
<td>S3C</td>
<td>S1A+S3C</td>
</tr>
<tr>
<td>No</td>
<td>S2B</td>
<td>S4D</td>
<td>S2B+S4D</td>
</tr>
<tr>
<td>Total</td>
<td>S1A+S2B</td>
<td>S3C+S4D</td>
<td></td>
</tr>
</tbody>
</table>

OR = ad/bc=(S1A S4D / S2B S3C)

Selection bias occurs when (S1) (S4)/(S2 S3) not equal to 1.0

PREVALENCE/INCIDENCE BIAS

Case Control Study

Prevalent Cases at Exam 6

<table>
<thead>
<tr>
<th>Serum Cholesterol at Exam 1</th>
<th>Yes</th>
<th>No</th>
</tr>
</thead>
<tbody>
<tr>
<td>High Quartile</td>
<td>39</td>
<td>34</td>
</tr>
<tr>
<td>Lower Quartiles</td>
<td>113</td>
<td>117</td>
</tr>
<tr>
<td>151</td>
<td>151</td>
<td>302</td>
</tr>
</tbody>
</table>

OR = (39) (117)/ (113) (34)=1.16
PREVALENCE/INCIDENCE BIAS

**Cohort Study**

Incident Cases by Exam

<table>
<thead>
<tr>
<th>Serum Cholesterol at Exam 1</th>
<th>High Quartile Yes</th>
<th>High Quartile No</th>
<th>Lower 3 Quartiles Yes</th>
<th>Lower 3 Quartiles No</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>85</td>
<td>462</td>
<td>116</td>
<td>1511</td>
</tr>
<tr>
<td></td>
<td>547</td>
<td></td>
<td>1627</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>201</td>
<td>1973</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>2174</td>
<td></td>
</tr>
</tbody>
</table>

RR = (85/547) / (116/1627) = 2.18
OR = (85) (1511) / (116) (462) = 2.40

PREVALENCE/INCIDENCE BIAS

- Late look at an association may underestimate risk because most severe cases die first
- Can even get paradoxical results where a harmful exposure appears beneficial
- Also called survival bias
- Avoid by using incident cases

BERKSON’S BIAS

HOSPITALIZED CASES

<table>
<thead>
<tr>
<th>Hypertension</th>
<th>Skin Cancer</th>
<th>Bone Fracture</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>193</td>
<td>320</td>
</tr>
<tr>
<td>No</td>
<td>200</td>
<td>800</td>
</tr>
</tbody>
</table>

OR = (193)(800)/(200)(320) = 2.41
**COMMUNITY POPULATION**

<table>
<thead>
<tr>
<th>Hypertension</th>
<th>Skin Cancer (Yes)</th>
<th>Bone Fracture (Yes)</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>1000</td>
<td>1000</td>
<td>2000</td>
</tr>
<tr>
<td>No</td>
<td>4000</td>
<td>4000</td>
<td>8000</td>
</tr>
</tbody>
</table>

\[ \text{OR} = 1 \]

**BERKSON’S BIAS**

- Type of selection bias
- Occurs when using hospitalized cases and controls; not representative of all cases and controls
- One of Raymond Pearl’s contributions

**MEMBERSHIP OR VOLUNTEER BIAS**

<table>
<thead>
<tr>
<th>Exercise</th>
<th>Recurrent MI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>Yes: 7</td>
</tr>
<tr>
<td></td>
<td>No: 59</td>
</tr>
<tr>
<td>Total</td>
<td>66</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Exercise</th>
<th>Recurrent MI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>Yes: 18</td>
</tr>
<tr>
<td></td>
<td>No: 46</td>
</tr>
<tr>
<td>Total</td>
<td>64</td>
</tr>
</tbody>
</table>

\[ \text{OR} = \frac{(7)(46)}{(18)(59)} = 0.30 \]
MEMBERSHIP OR VOLUNTEER BIAS

<table>
<thead>
<tr>
<th>Recurrent MI</th>
<th>Randomized to Exercise</th>
<th>Yes</th>
<th>No</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td></td>
<td>28</td>
<td>359</td>
<td>387</td>
</tr>
<tr>
<td>No</td>
<td></td>
<td>21</td>
<td>345</td>
<td>366</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>49</td>
<td>704</td>
<td>753</td>
</tr>
</tbody>
</table>

\[ OR = \frac{(28)(345)}{(21)(359)} = 1.28 \]

\[ RR = 1.26 \]

VOLUNTEER BIAS

- Membership in a group or willingness to volunteer may be associated with confounding variables or systematic differences compared to the general population

CASE-CONTROL STUDIES MINIMIZING SELECTION BIAS

- Take all cases
- Standardize eligibility criteria
- Match
- Sample cases and controls in similar way
- Use multiple control groups
- Whenever possible, use population cases and controls
EXPOSURE SUSPICION BIAS

- Knowledge of patient’s disease status may influence intensity and outcome of search for exposure
- May be subtle -- interviewer may ask same questions differently
- Occurs when interviewer not masked or assessment methods differ for cases and controls

RECALL BIAS

- Repeated questions about specific exposures may be asked several times of cases but only once of controls
- May inflate exposures in case groups
- Examples -- MI cases versus normal controls
- Congenital malformations may cause parents to “ruminate”

FAMILY INFORMATION BIAS

- Flow of family information about exposures is stimulated by a case of disease
- May especially be a problem in studying family history as an exposure
- Example -- congenital malformations
CASE-CONTROL STUDIES MINIMIZING INFORMATION BIAS

• Standardize questionnaire and exposure assessment
• Always treat cases and controls in the same way
• Train observers
• Mask observers and participants
• Use information collected in past
• Independently assess exposure
• Use “hard” endpoints

“HARD” OUTCOMES

• Less subject to misclassification
• Angina versus transmural MI
• TIA versus completed stroke
• Serum creatinine of 1.2 versus ESRD
• “Hardest” outcome is death (but some misclassification there as well!)

ANALYTIC BIAS

• Change criteria for outcomes or exposures during analysis
• Repeated analysis, data dredging
• Excluding participants in analysis phase
• Reporting subgroup analysis as primary analysis
SUBGROUP ANALYSIS

• Distribution of p values of all possible subgroup analyses follows a normal distribution with a mode at the overall p value
• Demonstrates that, if you do enough subgroup analyses, you will find a significant result

AVOIDING ANALYTIC BIAS

• Specify hypotheses before conducting the study or, in a secondary data analysis, prior to analyzing the data
• Specify subgroup analyses and hypotheses a priori
• Be honest in reporting results