STUDY DESIGN
CASE-CONTROL

Daniel E. Ford, MD, MPH
Vice Dean of Clinical Investigation
Johns Hopkins School of Medicine
Introduction to Clinical Research
July 11, 2012

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
### 2 x 2 TABLE

**CASE-CONTROL STUDY**

<table>
<thead>
<tr>
<th></th>
<th>Disease</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Yes</td>
<td>No</td>
<td>Total</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Exposure</td>
<td>Yes</td>
<td>a</td>
<td>b</td>
<td>a + b</td>
<td></td>
</tr>
<tr>
<td></td>
<td>No</td>
<td>c</td>
<td>d</td>
<td>c + d</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>a + c</td>
<td>b + d</td>
<td>N</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- Select 100 patients who died in first year on dialysis (cases)
- Select 100 patients who did not die (controls)
- Compare Kt/V between the groups

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

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

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

### 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
Locating Cases

- Cases may be
  - Undiagnosed
  - Misdiagnosed
  - Unavailable for study
  - Dead
- Need to ask if cases are representative of all cases
- Ex. can only include patients with myocardial infarctions who survive the initial event

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

CONFOUNDING

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

CONFOUNDING IN CASE CONTROL STUDIES

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

![Diagram of X, Y, Z, Ear Crease, CAD, Age relationships]
CONFOUNDING

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

Male Gender → Lung Cancer
Smoking

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

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.)
ADVANTAGES FOR CASE CONTROL STUDIES

- 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

DISADVANTAGES FOR CASE CONTROL STUDIES

- Temporal relationship between exposure and disease unclear
- Subject to bias because disease has already occurred
- Do not 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 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 fulfilled

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>
<tr>
<td>OR _{exp} = \frac{a}{c} = \frac{ad}{bd} = \frac{bc}</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**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 = \( \frac{ad}{bc} = \frac{(30)(70)}{(70)(30)} = \frac{1}{1} \)
**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>Yes</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 2010**

<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

The New England Journal of Medicine

THE EFFECT OF RACE AND SEX ON PHYSICIANS' RECOMMENDATIONS FOR CARDIAC CATHETERIZATION

Kevin A. Sacksman, M.D., Jesse A. Brown, Sc.D., William H. Grumbach, M.D., John F. Kressel, Ph.D., Samuel S. Sisson, M.D., Bernard J. Goldberg, M.B., Ch.B., D.R.C.P., Ross D. Davis, Christopher K. Talwalker, M.D., Amanda E. Brown, M.A., M.S., Susan Williams, M.D., Sven M. Eichhorn, M.D., and John E. Ingors, M.D., Ph.D.

Volume 340 Number 8 618 - February 25, 1999

percent for those with definite angina; P < 0.001). Logistic-regression analysis indicated that women (odds ratio, 0.60; 95 percent confidence interval, 0.4 to 0.9; P = 0.02) and blacks (odds ratio, 0.60; 95 percent confidence interval, 0.4 to 0.9; P = 0.02) were less likely to be referred for cardiac catheterization than men and whites, respectively. Analysis of race-sex interactions showed that black women were significantly less likely to be referred for catheterization than white men (odds ratio, 0.4; 95 percent confidence interval, 0.2 to 0.7; P = 0.004).

Conclusions Our findings suggest that the race and sex of a patient independently influence how physicians manage chest pain. (N Engl J Med 1999; 340:618-25.)

©1999, Massachusetts Medical Society.
ODDS OF REFERRAL IN WOMEN COMPARED TO MEN

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

175 25 200

\[ \text{OR} = \frac{(85)(10)}{(90)(15)} = 0.63 \]

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>-</td>
<td>75</td>
<td>75</td>
</tr>
</tbody>
</table>

$RR = \frac{50/75}{25/75} = 2.0$

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

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>-</td>
<td>75</td>
<td>750</td>
</tr>
</tbody>
</table>

$RR = \frac{50/300}{25/525} = 3.5$

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

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><strong>Risk Factor</strong></td>
<td><strong>Yes</strong></td>
<td><strong>No</strong></td>
<td><strong>Total</strong></td>
</tr>
<tr>
<td>Yes</td>
<td>A</td>
<td>B</td>
<td>A+B</td>
</tr>
<tr>
<td>No</td>
<td>C</td>
<td>D</td>
<td>C+D</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>A+C</td>
<td>B+D</td>
<td></td>
</tr>
</tbody>
</table>

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

### STUDY SAMPLE

<table>
<thead>
<tr>
<th>Disease Status</th>
<th>Yes</th>
<th>No</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Risk Factor</strong></td>
<td><strong>Yes</strong></td>
<td><strong>No</strong></td>
<td><strong>Total</strong></td>
</tr>
<tr>
<td>Yes</td>
<td>S1A</td>
<td>S1B</td>
<td>S1A+S1B</td>
</tr>
<tr>
<td>No</td>
<td>S2C</td>
<td>S2D</td>
<td>S2C+S2D</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>S1A+S2C</td>
<td>S1B+S2D</td>
<td></td>
</tr>
</tbody>
</table>

\[ \text{OR} = \frac{ad}{bc} = \frac{(S1A \cdot S2D)}{(S1B \cdot S2C)} \]

Selection bias occurs when \((S1) \cdot (S4)/(S2S3)\) not equal to 1.0

### PREVALENCE/INCIDENCE BIAS

<table>
<thead>
<tr>
<th>Case Control Study</th>
<th>Prevalent Cases at Exam 6</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Serum Cholesterol at Exam 1</strong></td>
<td><strong>High</strong></td>
</tr>
<tr>
<td><strong>Quartile</strong></td>
<td><strong>Yes</strong></td>
</tr>
<tr>
<td>High</td>
<td>39</td>
</tr>
<tr>
<td>Lower</td>
<td>113</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>151</td>
</tr>
</tbody>
</table>

\[ \text{OR} = \frac{(39 \cdot 117)}{(113 \cdot 34)} = 1.16 \]
**PREVALENCE/INCIDENCE BIAS**

<table>
<thead>
<tr>
<th>Serum Cholesterol at Exam 1</th>
<th>Incident Cases by Exam</th>
<th>Cohort Study</th>
</tr>
</thead>
<tbody>
<tr>
<td>High Quartile</td>
<td>Yes 6 No</td>
<td></td>
</tr>
<tr>
<td>Lower 3 Quartiles</td>
<td>85 462 547</td>
<td></td>
</tr>
<tr>
<td></td>
<td>116 1511 1627</td>
<td></td>
</tr>
<tr>
<td></td>
<td>201 1973 2174</td>
<td></td>
</tr>
</tbody>
</table>

\[
RR = \frac{85/547}{116/1627} = 2.18
\]

\[
OR = \frac{85 \times 1511}{116 \times 462} = 2.40
\]

**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**

- Type of selection bias
- Occurs when using hospitalized cases and controls; not representative of all cases and controls
- Two diseases appear to be associated because more likely to be admitted if have comorbidity
VOLUNTEER BIAS

• Willingness to volunteer or to follow healthy lifestyle may be associated with confounding variables or systematic differences compared to the general population that are not known.

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
• 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 \textit{a priori}
• Be honest in reporting results