Confounding and Bias in Case-Control Studies

Ching-Lan Cheng (鄭靜蘭), Ph.D.
Assistant Professor
Institute of Clinical Pharmacy and Pharmaceutical Sciences,
National Cheng Kung University
Disclosures

• There is no potential conflict of interest relevant to this presentation

• Materials in this presentation are adopted from the lectures in this year provided by Dr. Tobias Gerhard!
Outline

• Bias that might occur in case-control studies
  – Selection Bias
  – Information Bias

• Summary
SELECTION BIAS
Selection Bias

• Selection bias occurs when a systemic error in the ascertainment of cases or controls in case-control studies.

• If exposure status is differentially distributed between cases and controls, leading to a distortion of the exposure-disease association.
Population base

Study population

Should include equal proportions from each category
Selection Bias

- **Population base**
- **Study population**
- **Distorted picture of the population base**
Example I: Selection Bias in Case Control Studies

- Imagine a cumulative case-control study conducted in one large hospital. The study aims to explore whether smoking increases the risk of experiencing a stroke. Cases are patients admitted for stroke, controls are patients admitted for everything else. In order to have an unbiased result, the controls need to be representative of the non-cases in the source population, particularly in regards to the exposure of interest (smoking). However, because smokers are also at higher risk for other diseases that lead to hospitalizations than non-smokers (lung cancer, COPD, etc), smoking is more common among hospitalized non-cases than among non-cases in the source population. This will result in an underestimation of the effect of smoking on stroke risk.
Unbiased Control Selection

Source Population (Exposure odds in non-cases = 0.5)

<table>
<thead>
<tr>
<th></th>
<th>Stroke (Cases)</th>
<th>No Stroke (Controls)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Smoker</td>
<td>60</td>
<td>30000</td>
</tr>
<tr>
<td>Non-Smoker</td>
<td>40</td>
<td>60000</td>
</tr>
</tbody>
</table>

True OR = 3.0

Random Sample

Cumulative Case-Control Study (4:1); (Exposure odds in non-cases = 0.48)

<table>
<thead>
<tr>
<th></th>
<th>Stroke (Cases)</th>
<th>No Stroke (Controls)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Smoker</td>
<td>60</td>
<td>130</td>
</tr>
<tr>
<td>Non-Smoker</td>
<td>40</td>
<td>270</td>
</tr>
</tbody>
</table>

Estimated OR = 3.1
## Biased Control Selection

Source Population  (Exposure odds in non-cases = 0.5)

<table>
<thead>
<tr>
<th></th>
<th>Stroke (Cases)</th>
<th>No Stroke (Controls)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Smoker</td>
<td>60</td>
<td>30000</td>
</tr>
<tr>
<td>Non-Smoker</td>
<td>40</td>
<td>60000</td>
</tr>
</tbody>
</table>

**True OR = 3.0**

Hospitalization

Hospitalized Population

<table>
<thead>
<tr>
<th></th>
<th>Stroke (Cases)</th>
<th>No Stroke (Controls)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Smoker</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Non-Smoker</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
### Biased Control Selection

Source Population  (Exposure odds in non-cases = 0.5)

<table>
<thead>
<tr>
<th></th>
<th>Stroke (Cases)</th>
<th>No Stroke (Controls)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Smoker</td>
<td>60</td>
<td>30000</td>
</tr>
<tr>
<td>Non-Smoker</td>
<td>40</td>
<td>60000</td>
</tr>
</tbody>
</table>

True OR = 3.0

Hospitalization

All cases are hospitalized

Hospitalized Population

<table>
<thead>
<tr>
<th></th>
<th>Stroke (Cases)</th>
<th>No Stroke (Controls)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Smoker</td>
<td>60</td>
<td></td>
</tr>
<tr>
<td>Non-Smoker</td>
<td>40</td>
<td></td>
</tr>
</tbody>
</table>
Biased Control Selection

Source Population  (Exposure odds in non-cases = 0.5)

<table>
<thead>
<tr>
<th></th>
<th>Stroke (Cases)</th>
<th>No Stroke (Controls)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Smoker</td>
<td>60</td>
<td>30000</td>
</tr>
<tr>
<td>Non-Smoker</td>
<td>40</td>
<td>60000</td>
</tr>
</tbody>
</table>

True OR = 3.0

Among the possible controls (i.e. the source population) smokers are more likely to be hospitalized than non smokers (1.8% vs. 0.6%).

Hospitalized Population

<table>
<thead>
<tr>
<th></th>
<th>Stroke (Cases)</th>
<th>No Stroke (Controls)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Smoker</td>
<td>60</td>
<td>540</td>
</tr>
<tr>
<td>Non-Smoker</td>
<td>40</td>
<td>360</td>
</tr>
</tbody>
</table>
Biased Control Selection

Source Population  (Exposure odds in non-cases = 0.5)

<table>
<thead>
<tr>
<th></th>
<th>Stroke (Cases)</th>
<th>No Stroke (Controls)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Smoker</td>
<td>60</td>
<td>30000</td>
</tr>
<tr>
<td>Non-Smoker</td>
<td>40</td>
<td>60000</td>
</tr>
</tbody>
</table>

True OR = 3.0

Hospitalization

Among the possible controls (i.e. the source population) smokers are more likely to be hospitalized than non smokers (1.8% vs. 0.6%).

Hospitalized Population (Exposure odds in non-cases = 1.5)

<table>
<thead>
<tr>
<th></th>
<th>Stroke (Cases)</th>
<th>No Stroke (Controls)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Smoker</td>
<td>60</td>
<td>540</td>
</tr>
<tr>
<td>Non-Smoker</td>
<td>40</td>
<td>360</td>
</tr>
</tbody>
</table>
Biased Control Selection

Source Population (Exposure odds in non-cases = 0.5)

<table>
<thead>
<tr>
<th></th>
<th>Stroke (Cases)</th>
<th>No Stroke (Controls)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Smoker</td>
<td>60</td>
<td>30000</td>
</tr>
<tr>
<td>Non-Smoker</td>
<td>40</td>
<td>60000</td>
</tr>
</tbody>
</table>

True OR = 3.0

Among the possible controls (i.e. the source population) smokers are more likely to be hospitalized than non smokers (1.8% vs. 0.6%).

Hospitalization

Hospitalized Population → sample controls for study

<table>
<thead>
<tr>
<th></th>
<th>Stroke (Cases)</th>
<th>No Stroke (Controls)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Smoker</td>
<td>60</td>
<td>540 (\rightarrow) 240</td>
</tr>
<tr>
<td>Non-Smoker</td>
<td>40</td>
<td>360 (\rightarrow) 160</td>
</tr>
</tbody>
</table>
### Biased Control Selection

Source Population  (Exposure odds in non-cases = 0.5)

<table>
<thead>
<tr>
<th></th>
<th>Stroke (Cases)</th>
<th>No Stroke (Controls)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Smoker</td>
<td>60</td>
<td>30000</td>
</tr>
<tr>
<td>Non-Smoker</td>
<td>40</td>
<td>60000</td>
</tr>
</tbody>
</table>

**True OR = 3.0**

Hospitalization

Among the possible controls (i.e. the source population) smokers are more likely to be hospitalized than non smokers (1.8% vs. 0.6%).

Study Population  (Exposure odds in non-cases = 1.5)

<table>
<thead>
<tr>
<th></th>
<th>Stroke (Cases)</th>
<th>No Stroke (Controls)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Smoker</td>
<td>60</td>
<td>240</td>
</tr>
<tr>
<td>Non-Smoker</td>
<td>40</td>
<td>160</td>
</tr>
</tbody>
</table>
Biased Control Selection

Source Population (Exposure odds in non-cases = 0.5)

<table>
<thead>
<tr>
<th></th>
<th>Stroke (Cases)</th>
<th>No Stroke (Controls)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Smoker</td>
<td>60</td>
<td>30000</td>
</tr>
<tr>
<td>Non-Smoker</td>
<td>40</td>
<td>60000</td>
</tr>
</tbody>
</table>

True OR = 3.0

Hospitalization

Among the possible controls (i.e. the source population) smokers are more likely to be hospitalized than non smokers (1.8% vs. 0.6%).

Study Population (Exposure odds in non-cases = 1.5)

<table>
<thead>
<tr>
<th></th>
<th>Stroke (Cases)</th>
<th>No Stroke (Controls)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Smoker</td>
<td>60</td>
<td>240</td>
</tr>
<tr>
<td>Non-Smoker</td>
<td>40</td>
<td>160</td>
</tr>
</tbody>
</table>

Study OR = 1.0
## Biased Control Selection

### Source Population (Exposure odds in non-cases = 0.5)

<table>
<thead>
<tr>
<th></th>
<th>Stroke (Cases)</th>
<th>No Stroke (Controls)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Smoker</td>
<td>60</td>
<td>30000</td>
</tr>
<tr>
<td>Non-Smoker</td>
<td>40</td>
<td>60000</td>
</tr>
</tbody>
</table>

True OR = 3.0

### Study Population (Exposure odds in non-cases = 1.5)

<table>
<thead>
<tr>
<th></th>
<th>Stroke (Cases)</th>
<th>No Stroke (Controls)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Smoker</td>
<td>60</td>
<td>240</td>
</tr>
<tr>
<td>Non-Smoker</td>
<td>40</td>
<td>160</td>
</tr>
</tbody>
</table>

Study OR = 1.0

Hospitalization

Exposure distribution in study controls ≠ exposure distribution in source population controls
Selection Bias in Case Control Studies

• In the example, the selection process for the controls — sampled from hospitalized patients instead of randomly sampled from the non-cases in the source population — changed the distribution of the exposure of interest (smoking) in the control patients of the study from the true distribution in the source population.

Solution → Population-based sampling of controls
Selection Bias in Case Control Studies

Population base

Study population (sampled from hospitalized patients)

Exposed

With disease

Smokers w/o stroke \(\rightarrow\) overrepresented in the hospital

Not exposed

Without disease

Nonsmokers w/o stroke \(\rightarrow\) underrepresented in the hospital
Example II – Prevalent User Bias

- Those who develop outcomes stop taking the drug (depletion of susceptibles, sick stoppers)
- Prevalent users tend to be healthy adherers and those that benefit from treatment (healthy users)
- In sum, inclusion of prevalent users will distort the study population (oversampling of subjects / person time at low risk) and result in underestimation of harms and overestimation of benefits

Solution → New user design
Selection Bias – Prevalent User Bias

- Population base
- Study population
- With disease
- Without disease
- Exposed
- Not exposed
- “Healthy Users”
- “Sick Stoppers”
INFORMATION BIAS
Information Bias

• Often referred to as measurement bias
• Occurs due to poor measurement (classification) of study variables (exposure)
• Distinguish two basic types of information bias
  – Non-differential
    - Misclassification between groups is approximately equal
  – Differential
    - Amount of misclassification differs between groups
Misclassification of Exposure

• **Binary, non-differential** → **Bias towards the null**
  – 20% of exposed subjects classified as unexposed (used OTC version of the drug)
  – 10% of unexposed subjects classified as exposed (non-compliers)

<table>
<thead>
<tr>
<th>Truth</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>AE+</td>
<td>AE-</td>
<td></td>
</tr>
<tr>
<td>Exp+</td>
<td>20</td>
<td>10</td>
<td></td>
</tr>
<tr>
<td>Exp-</td>
<td>80</td>
<td>90</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Non-differential misclassification of exposure</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
<tr>
<td>Exp+</td>
</tr>
<tr>
<td>Exp-</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Observation</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>AE+</td>
<td>AE-</td>
</tr>
<tr>
<td>Exp+</td>
<td>24</td>
<td>17</td>
</tr>
<tr>
<td>Exp-</td>
<td>76</td>
<td>83</td>
</tr>
</tbody>
</table>

True OR = 2.25  
\[(20 \times 90) / (80 \times 10)\]

Estimated OR = 1.54  
\[(24 \times 83) / (76 \times 17)\]
Misclassification of Exposure

- **Binary, differential** → Direction of bias is unpredictable

<table>
<thead>
<tr>
<th>Truth</th>
<th>AE+</th>
<th>AE-</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exp+</td>
<td>20</td>
<td>10</td>
</tr>
<tr>
<td>Exp-</td>
<td>80</td>
<td>90</td>
</tr>
</tbody>
</table>

True OR = 2.25  
\[
\frac{(20\times90)}{(80\times10)}
\]

- **Observation I**

<table>
<thead>
<tr>
<th>AE+</th>
<th>AE-</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exp+</td>
<td>20</td>
</tr>
<tr>
<td>Exp-</td>
<td>80</td>
</tr>
</tbody>
</table>

Estimated OR = 3.32  
\[
\frac{(20\times93)}{(80\times7)}
\]

Bias away from null

- **Observation II**

<table>
<thead>
<tr>
<th>AE+</th>
<th>AE-</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exp+</td>
<td>20</td>
</tr>
<tr>
<td>Exp-</td>
<td>80</td>
</tr>
</tbody>
</table>

Estimated OR = 1.07  
\[
\frac{(20\times81)}{(80\times19)}
\]

Bias towards null

- **Exposure not binary** → Direction of bias is unpredictable
Misclassification of Confounders

- Adjustment with a binary non-differentially misclassified confounder reduces bias and produces a partially adjusted effect estimate that falls between the crude and true effect – residual confounding
  
  Greenland and Robins, AJE 1985

  - Residual confounding decreases with increasing sensitivity and specificity of the misclassified confounder
    
    Savitz and Baron, AJE 1989

  - Necessary assumption (likely to hold in most applications in epidemiology) – Effect of the confounder on the outcome is in the same direction among the treated and the untreated (i.e., there is no qualitative interaction between the treatment and the confounder)
    
    Ogburn and VanderWeele, Epidemiology 2012
Addressing Misclassification

• Prospective studies with primary data collection
  – Ensure accurate measurement (instruments, procedures, quality control, etc)

• Studies that rely on secondary data
  – Use validated measures for exposure, outcome, and confounding factors
  – Rule out recall and detection biases
In summary…

• Best remedy for bias is prevention!

• RCTs
  – Randomization
  – Blinding
  – Primary data collection

• Observational Studies
  – Sample selection
  – Choice of comparator
  – Use validated measures
  – Statistical analysis
Thank you

Ching-Lan (Rebecca) Cheng

clcheng@mail.ncku.edu.tw