Maximizing Internal and External Validity in Epidemiology Studies

American Society of Human Genetics
October 24, 2007

Robert N. Hoover, M.D., Sc.D.
Director
Epidemiology and Biostatistics Program
October 24, 2007
Validity (Internal Validity)

- Both patients and comparison groups are representative of the same “study-base” (population).

- If selective factors enter into the choice of patients, the same factors should enter into the selection of the comparison group.
External Validity
(Representativeness)

- The study-base for your cases and controls is similar in all relevant aspects to the population to which you wish to extrapolate the results.
Study Base (Internal Validity) Questions

- If one of your controls had contracted the disease, would he/she be in your case group?
- If one of your cases had not been a case would he/she have been equally likely to be chosen as a control as any of the controls in your study?
- How did you end up with the cases you have?
Common Starting Places for A Study Base

- General Population
- Special Population
  - HMO
  - Cohort
- Hospital/Clinic
- Neighborhood
- Families (Siblings, Spouses)
Hospital/Clinic Study - Base

Controls

Cases
General Population-Based Study of Lymphoma

<table>
<thead>
<tr>
<th>Interview + DNA</th>
<th>Cases</th>
<th>Controls</th>
</tr>
</thead>
<tbody>
<tr>
<td>Participation Rate</td>
<td>68%</td>
<td>47%</td>
</tr>
<tr>
<td>Response Rate</td>
<td>53%</td>
<td>40%</td>
</tr>
</tbody>
</table>
Population Study-Base With Response Bias

Cases

Controls
Maternal Urinary Mercury & Neural-Tube Defects in Mexican-Americans

<table>
<thead>
<tr>
<th></th>
<th>Cases</th>
<th>Study Controls</th>
<th>Natl Sample</th>
</tr>
</thead>
<tbody>
<tr>
<td>% ≥ 5.62mg/L</td>
<td>28%</td>
<td>17%</td>
<td>5%</td>
</tr>
<tr>
<td>Odds Ratio</td>
<td>1.8</td>
<td>7.4</td>
<td></td>
</tr>
</tbody>
</table>
### Endometrial Cancer & Estrogen Use

<table>
<thead>
<tr>
<th>Other Cancer Controls</th>
<th>D &amp; C Controls</th>
</tr>
</thead>
<tbody>
<tr>
<td>Relative Risk</td>
<td>10.8</td>
</tr>
<tr>
<td>P-value</td>
<td>1.7</td>
</tr>
<tr>
<td></td>
<td>0.005</td>
</tr>
<tr>
<td></td>
<td>0.12</td>
</tr>
</tbody>
</table>
The study-base for your cases and controls is similar in all relevant aspects to the population to which you wish to extrapolate the results.
## “Relevant Aspects”

<table>
<thead>
<tr>
<th>Factor</th>
<th>Measure Affected</th>
</tr>
</thead>
<tbody>
<tr>
<td>Underlying Biology</td>
<td>Relative Risk</td>
</tr>
<tr>
<td></td>
<td>Absolute Risk</td>
</tr>
<tr>
<td></td>
<td>Attributable Risk</td>
</tr>
<tr>
<td>Prevalence of Risk Factors That Modify The Effect</td>
<td>Same as above</td>
</tr>
<tr>
<td>Prevalence of The Study Risk Factor</td>
<td>Absolute Risk</td>
</tr>
<tr>
<td></td>
<td>Attributable Risk</td>
</tr>
</tbody>
</table>
Underlying Biology/Presence of Modifiers

BrCa1 in high-risk families → 85% lifetime risk
BrCa1 in population sample → 50% lifetime risk

Heavy Smoking/Heavy Alcohol Use and Esophageal Cancer

<table>
<thead>
<tr>
<th>Socioeconomic Status</th>
<th>Relative Risk</th>
</tr>
</thead>
<tbody>
<tr>
<td>High</td>
<td>34</td>
</tr>
<tr>
<td>Medium</td>
<td>95</td>
</tr>
<tr>
<td>Low</td>
<td>421</td>
</tr>
</tbody>
</table>
## Prevalence of Risk Factors

### Estrogen Therapy and Uterine Cancer

<table>
<thead>
<tr>
<th></th>
<th>California</th>
<th>Minnesota</th>
</tr>
</thead>
<tbody>
<tr>
<td>Relative Risk</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>% Exposed</td>
<td>50%</td>
<td>3%</td>
</tr>
<tr>
<td>Etiologic Fraction</td>
<td>67% (Attributable Risk)</td>
<td>11%</td>
</tr>
</tbody>
</table>
Selection of Cases
(Issues of External Validity)

- Self Report – validated or not
- Clinically Defined (Hospital, Registry, Physician)
- All
- Advanced
- Incident
- Prevalent
Prevalent Conditions

Prevalence = Incidence x Duration

Prevalent samples → weighted with long-term survivors
Selection of Controls

Cautionary Note Re: Extreme Phenotype
And Violation of Study-Base Principle
Caspase 8 (CASP8) D302H Variant Decreases Breast Cancer Risk

Overall OR (95% CI) 0.88 (0.84, 0.92) P=1x10^-7

Studies (sorted by size)
- Kuopio
- Helsinki
- CNIO
- LSHTM
- USRTS
- ABCFS
- GENICA
- HBCS
- Mayo_Clinic
- Sheffield
- CAHRES
- PBCS
- SEARCH

Total Sample Size
- 16,423 cases
- 17,109 controls

Histidine (H) allele in 13% of controls

Nat Genet 2007;39:352-8
# Lung Cancer Risk and CYP2D6*

<table>
<thead>
<tr>
<th>Relative Risk</th>
<th>Study 1</th>
<th>Study 2</th>
<th>Study 3</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>15.6 (4.8 – 55.9)</td>
<td>6.1 (2.2 – 17.1)</td>
<td>0.6 (0.3 – 1.2)</td>
</tr>
<tr>
<td>Epidemiologic Quality</td>
<td>Low</td>
<td>Intermediate</td>
<td>High</td>
</tr>
<tr>
<td>(% participation)</td>
<td>(?)</td>
<td>(26%)</td>
<td>(80%)</td>
</tr>
</tbody>
</table>

* Risk of homozygous extensive metabolizers compared to homozygous poor metabolizers.
Study-Base Flaws in GWAS

**False Positives**
- Better Designs in Replications

**False Negatives**
- No GXE
  - Likely not great
- GXE
  - Could be Substantial
Summary

- Internal Validity is paramount to achieve case and control series that are comparable in order to confidently uncover causal risk factors.
- The key to achieving this comparability is to define the “study-base” for which the cases are all, or a representative sample of all, of those affected in the study-base, and to draw a representative sample of unaffected from this same study-base as controls.
External Validity relates not to the validity of the association noted, but to how it may be extrapolated to other circumstances.

The greater the difference between the study-base for the study and the populations to which the results are extrapolated, the more problematic the extrapolation.
Conclusion

- Genetic risk factors are likely to be less affected by potential selection biases introduced by study-base flaws than environmental and life-style risk factor.

- There will, however, be some effect, and for genetic risk factors operating through GXE interaction, the effects could be substantial.
Selection Bias in Genetic Studies

- Genetic variants associated with selection factors
- Genes associated with risk factors for the disease that are biased by sampling (e.g. obesity, smoking, alcohol)
- Genetic effect from altered GXE for Biased Exposures