Cohort Studies – Part 2

Dr. Borenstein

NO AUDIO ON THIS SLIDE. CLICK NEXT TO CONTINUE >>

Back to Cohorts

- People accrue time in the study. We need to document each person’s time accurately.
- Always use time unit when expressing either CI or ID rates.
- ID rate is a true “rate” – implies “speed” of developing disease.
- CI is a rate because we use a time unit (e.g., 10 per 1,000 people per year) but does not imply “speed” the way ID does.
- Prevalence “rate” is not really a rate, just the proportion of people who have the disease in a population at a given (specified) time.
- We can estimate person-time (N x t) but this number (denominator) will always be larger than actual person-time: incidence density is (a) more accurate and (b) will reflect a rate that is a bit larger than a rate where T is assumed to be complete.

- Ex: 10/1000 py= 0.01 (CI=Nxt) 10/867=0.015 (ID)

Cohort studies

- Selection of the comparison (unexposed) group
  - The unexposed group should be similar to the exposed group except for the exposure
    - In RCT, Rx and non-Rx groups will be exactly similar if (a) N is large enough and (b) randomization is accomplished
    - Can match exposed to unexposed on important characteristics on which they might differ and which are related to risk of disease, but very difficult to do and not commonly done
    - If unexposed group comes from same base population as exposed group, the two groups will be similar in that they arose from the same population
Design issues in cohort studies: sources of exposure data

- Pre-existing records (medical records, employment records)
  - Advantages:
    - Cheap
    - Relatively easy to work with
    - Usually unbiased (exposures documented before outcomes)
    - Good info on meds, procedures, etc.
  - Disadvantages:
    - May be inaccurate
    - May have missing data
    - May not contain data needed, esp. re: potential confounders/effect-modifiers

- Self-report (in-person, telephone, mail)
  - Advantages:
    - Can design sophisticated, structured questionnaires and questions on as many factors as necessary
  - Disadvantages:
    - Subject to response (recall) bias (but not selectively by case status)
    - Subject to interviewer/observer bias (but not selectively by case status)
    - Expensive

- Physical examinations, psychological testing, biologic specimen collection, laboratory testing, environmental monitoring of home/person/ workplace
  - Advantages:
    - Sophisticated measures, direct measures of exposures (e.g., stress test vs. question on CVD)
    - As detailed as you wish, can minimize missing data
    - Unbiased
  - Disadvantages:
    - Poor participation rates
    - Limited to time of data collection, may need to repeat over follow-up period
    - Very expensive
Exposure data, con’d

- Always try to conduct a pilot study and/or validation study to make sure your data will measure what you are actually trying to measure (validity and reliability): what is the gold standard for this measure?
- May need to build in repeated measurements of exposure over time (e.g., blood pressure in HHP; biennial mailed interviews in NHS)
  - Misclassification of exposure (Unit 7)
    - A person may have been a smoker at baseline but quit – should you code him as a current smoker or non-smoker? Hmmm...
    - Take into account in analysis (time-dependent exposures)

Design issues in cohort studies: sources of outcome data

- Some of the same sources as exposure data: records, self-report, measurements on subjects, including medical evaluations (can include biologic specimens, laboratory, imaging)
- Death and birth certificates: beware! Death certificates are not a good source for some diseases...
- Registry data (SEER, birth defects, trauma)
- HMO medical records (Kaiser, Mayo)

Follow-up in cohort studies

- Principal source of bias: Attrition (loss to follow-up)
  - Particularly problematic if losses (to outcome) differ by exposure status (if there is an association between E and D). Ex: individuals at high risk for AD who have poor sense of olfaction at baseline drop out disproportionately: RR will be under-estimated (more in Unit 7)
Other important concerns in cohort studies

- Non-participation at baseline (can be internally valid but not generalizable – more in Unit 6)
- Confounding (measurement of all possible factors that could affect E-D association)
- Length of follow-up (must be long enough to allow outcome to develop in large enough number of individuals)
- Sample size (N)
- Different sources of information on study subjects (e.g., proxy vs. self-report)
- Exposure misclassification
- Outcome misclassification

Three choices for unexposed group:

- **Internal comparison**
- External comparison (separate, unexposed cohort)
- Comparison with available rates from the general population

Internal comparison group:

- A single sample (representative of some population) is selected
- Exposure status is ascertained
- Unexposed becomes the comparison group
  - Okay if exposed and unexposed groups are large enough
  - Okay if outcome is common enough for a sufficient number of cases to develop over the follow-up time (design features)

*Example:*

- N = 1,836 Japanese-Americans in Seattle, WA
- Drink fruit and vegetable juices 
  - 3+ times/week
  - < 1/week
- Incidence (density) AD in E
- Incidence (density) AD in U

E.g.,
Three choices

- Internal comparison
- External comparison (separate, unexposed cohort)
- Comparison with available rates from the general population

Separate unexposed populations: special exposures

- Most often used in occupational epi studies
  - Select a group of individuals who are in the same industry but unexposed
    - Problem: These people tend to be quite different from E cohort – office workers, higher income
  - Select a group from individuals who work in a different industry where exposure is uncommon
  - Careful! As in case-control studies, where the choice of controls can generate or mask a finding, so can the inappropriate selection of unexposed individuals in a cohort study.

Selection of a comparison group: the general population

- The general population will include some exposed subjects (okay if exposure is rare)
- The general population experiences a higher mortality rate than an occupational population (workers are healthier on average than the general population, which contains both working and non-working individuals)
  - Note: may need more than one comparison group to make sure associations are consistent between groups
The Healthy Worker Effect

- People who work are healthier than people who do not work.
- Any excess risk associated with a particular occupation will tend to be underestimated when compared with the general population.

\[
\begin{array}{c|c|c}
\text{Exposed} & \text{YES} & \text{NO} \\
\text{Disease/Outcome} & a & b \\
\text{General population} & c & d
\end{array}
\]

\[RR = \frac{a}{a+b} \]

Measures of association from cohort studies

- Prospective and retrospective (historical) cohort studies:
  - Using one sample: risk ratio, rate ratio
    - Both are known as the relative risk because they compare the risk among exposed individuals relative to that in the unexposed.
    - If baseline risk is 10/100 (0.10) then a RR of 2 = 20/100 (0.20)
    - If baseline risk is 10/100,000 (0.0001) a RR of 2 = 20/100,000 (0.0002)
  - Using two samples: Standardized mortality/morbidity ratio (SMR) and Proportional Mortality ratio (PMR).

Standardized mortality/morbidity ratio (SMR)

- Compares the observed amount of D among exposed cohort to what would be expected in an unexposed cohort.

\[
SMR = \frac{\text{# Observed cases (O)}}{\text{# Expected cases (E)}} \times 100
\]

Or can be interpreted as a relative risk (1.0 is null association).
Example: Underground potash workers vs. general population: comparison of mortality rates due to cardiovascular disease

Underground workers exposed to heat

<table>
<thead>
<tr>
<th>Age</th>
<th>P-Y (Observed)</th>
<th>Deaths</th>
<th>Gen. Pop Death Rate (per 100,000)</th>
<th>Expected Deaths</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt; 45</td>
<td>12140</td>
<td>1</td>
<td>50 (.0005)</td>
<td>50</td>
</tr>
<tr>
<td>45-49</td>
<td>8863</td>
<td>4</td>
<td>79 (.00079)</td>
<td>79</td>
</tr>
<tr>
<td>50-54</td>
<td>10095</td>
<td>11</td>
<td>1005 (.00105)</td>
<td>105</td>
</tr>
<tr>
<td>55-59</td>
<td>5924</td>
<td>14</td>
<td>120 (.00120)</td>
<td>120</td>
</tr>
<tr>
<td>&gt; 60</td>
<td>1280</td>
<td>6</td>
<td>250 (.00250)</td>
<td>250</td>
</tr>
<tr>
<td>TOTAL</td>
<td>38303</td>
<td>36</td>
<td>34</td>
<td>34</td>
</tr>
</tbody>
</table>

SMR = O / E (x 100) = 36 / 34 = 1.059 (x 100) = 105.9

Interpretation

- SMR = 1.059
- We estimate that underground potash miners have a 5.9% (RR/OR/SMR-1) excess risk of mortality compared with the general population for exposure to heat.
- Caution! Remember Healthy Worker Effect: we expect workers to have a lower mortality than the general population.
- In general, when we compare an occupational group to the general population, we will underestimate the true effect since the general population has more disease than workers.

So, make a more fair comparison: Underground potash workers vs. copper miners

<table>
<thead>
<tr>
<th>Age</th>
<th>P-Y (Observed)</th>
<th>Deaths (Observed)</th>
<th>Copper Miners Death Rate (per 100,000)</th>
<th>Expected Deaths</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt; 45</td>
<td>12140</td>
<td>1</td>
<td>35</td>
<td>4.25</td>
</tr>
<tr>
<td>45-49</td>
<td>8863</td>
<td>4</td>
<td>64</td>
<td>5.67</td>
</tr>
<tr>
<td>50-54</td>
<td>10095</td>
<td>11</td>
<td>89</td>
<td>8.98</td>
</tr>
<tr>
<td>55-59</td>
<td>5924</td>
<td>14</td>
<td>87</td>
<td>5.15</td>
</tr>
<tr>
<td>&gt; 60</td>
<td>1280</td>
<td>6</td>
<td>170</td>
<td>2.18</td>
</tr>
<tr>
<td>TOTAL</td>
<td>38303</td>
<td>36</td>
<td>26.23</td>
<td>26.23</td>
</tr>
</tbody>
</table>

SMR = O / E (x 100) = 36/26.23 = 1.3725 (x 100) = 137.25
Interpretation

- SMR=1.3725
- We estimate that underground potash miners who are exposed to heat have a 37.25% (RR=1) increased risk of mortality compared with copper miners.
- This interpretation is probably closer to the most "valid" value because healthy worker effect is present in both groups. First comparison (gen. pop) was underestimated (SMR closer to 1.0 or 100% = no excess mortality/morbidity).

<table>
<thead>
<tr>
<th>SMR</th>
<th>1.0</th>
<th>1.059</th>
<th>1.37</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(100%)</td>
<td>(105.9)</td>
<td>(137.25)</td>
</tr>
</tbody>
</table>

Standardized mortality/morbidity ratio (SMR)

- Caution:
  - It is usually inappropriate to compare a number of SMRs to one another that have been calculated from different study populations (e.g., potash workers from different studies)
  - This is because the category-specific weights (person-years or number of people in each age stratum) will usually be different for each study population.

SMR – Unit 8: Confounding

- SMR also is a method to adjust for underlying characteristics of a population when you are trying to compare frequency of disease in two separate populations.
Proportional Mortality Ratio (PMR)

- Proportion of deaths from a specific cause relative to all deaths in the exposed cohort is compared with the proportion of deaths from a specific cause relative to all deaths in the unexposed cohort.
- Used when you have the numbers and causes of deaths in the exposed group but NOT the structure of the population from which these persons arose.
- Only numerator data are used (from exposed cohort) (i.e., number of deaths grouped by cause).

Proportional Mortality Ratio (PMR), con’d

\[
PMR = \frac{\text{proportion of deaths from specified cause (exposed)}}{\text{proportion of deaths from specified cause (comparison population)}} \times 100
\]

= ratio of two proportions

Example: PMR calculation

<table>
<thead>
<tr>
<th></th>
<th>Deaths among Exposed Population</th>
<th>Deaths among Comparison Population</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>n</td>
<td>proportion</td>
</tr>
<tr>
<td>Cancer</td>
<td>24</td>
<td>0.32</td>
</tr>
<tr>
<td>Heart Disease</td>
<td>27</td>
<td>0.36</td>
</tr>
<tr>
<td>AIDS</td>
<td>1</td>
<td>0.01</td>
</tr>
<tr>
<td>Other</td>
<td>23</td>
<td>0.31</td>
</tr>
<tr>
<td>Total</td>
<td>75</td>
<td>1.0</td>
</tr>
</tbody>
</table>

PMR<sub>Ca</sub> = \( \frac{0.32}{0.30} \times 100 = 106.7 \)
PMR<sub>HD</sub> = \( \frac{0.36}{0.42} \times 100 = 85.7 \)
PMR<sub>AIDS</sub> = \( \frac{0.0133}{0.019} \times 100 = 70.0 \)
PMR<sub>Other</sub> = \( \frac{0.3067}{0.2619} \times 100 = 117.1 \)
**Interpretation**

<table>
<thead>
<tr>
<th>Cause</th>
<th>PMR Calculation</th>
<th>PMR Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ca</td>
<td>0.32 / 0.30</td>
<td>= 1.067 (x 100) = 106.7</td>
</tr>
<tr>
<td>HD</td>
<td>0.36 / 0.42</td>
<td>= 0.857 (x 100) = 85.7</td>
</tr>
<tr>
<td>AIDS</td>
<td>0.0133 / 0.019</td>
<td>= 0.70 (x 100) = 70.0</td>
</tr>
<tr>
<td>Other</td>
<td>0.3067 / 0.2819</td>
<td>= 1.171 (x 100) = 117.1</td>
</tr>
</tbody>
</table>

- Ca: The proportion of deaths attributable to cancer was approximately (RR/SMR - 1) 6.7% higher in the exposed population compared to the proportion of deaths from cancer in the comparison population.
- HD: The proportion of deaths due to heart disease was 14.3% (1 - 0.857) lower in the exposed cohort compared to that in the unexposed population.
- AIDS: The proportion of deaths due to AIDS was 30% lower in the exposed pop. compared with the unexposed population.
- Other: The proportion of deaths due to other causes was 17.1% higher...

**SMR vs. PMR: Advantages of PMR**

- SMR requires knowledge of structure of exposed population in each stratum; PMR requires only number who died, what they died of.
- PMR is quick and inexpensive. Data usually gleaned from death certificates.

**SMR vs. PMR: Disadvantages of PMR**

- Competing causes of death/competing risk
  - The relative frequency of other causes of death can affect the proportional mortality for the cause of interest (i.e., an observed excess in one cause of death may reflect a deficit of deaths due to other causes. E.g., more cancer but less CVD).
  - PMRs can suggest that an excess (or decreased) risk exists, but need to evaluate hypothesis taking into account the population at risk and compare cause-specific mortality rates across populations (PMR can provide a ‘first peek’ but is more Ho generating than Ho testing).
  - PMR may yield unstable estimates for rare causes of death (due to very small n’s in strata).
Overview of Unit 5: Cohort studies

- Overview of study designs
- Overview of associations: non-directional and uni-directional
- Think about causal pathways: would like to infer temporal association between E and D
- Viewing all of epidemiologic methods in context of the prospective cohort study
- Experimental (RCT) studies vs. the observational prospective cohort study
- Prospective, retrospective (historical) cohort and bi-directional study designs
- When to choose a cohort design
- Refresher: Calculation of the relative risk and person-time
- Scale of the RR/OR
- Selection of the unexposed cohort in cohort studies: internal, external (separate unexposed cohort vs. general population)
- Healthy worker effect
- Sources of exposure data
- Sources of outcome data
- Sources of concern in cohort studies
- Measures of association in cohort studies: risk ratio, rate ratio, SMR, PMR
- Advantages and disadvantages of the PMR vs. SMR

SMR Question Rationale

<table>
<thead>
<tr>
<th>Age Group</th>
<th>Canada Lyme Disease Rate per 100,000</th>
<th>Population Podunk</th>
<th># Cases of Lyme in Podunk</th>
<th>Expected # Cases in Podunk</th>
</tr>
</thead>
<tbody>
<tr>
<td>0-9</td>
<td>600</td>
<td>37,030</td>
<td>1.12</td>
<td>222.1800</td>
</tr>
<tr>
<td>10-19</td>
<td>4.60</td>
<td>16,980</td>
<td>2.85</td>
<td>20.8300</td>
</tr>
<tr>
<td>20-24</td>
<td>40.40</td>
<td>23,050</td>
<td>3.19</td>
<td>442.0460</td>
</tr>
<tr>
<td>25-44</td>
<td>25.90</td>
<td>23,450</td>
<td>3.33</td>
<td>385.5200</td>
</tr>
<tr>
<td>45-64</td>
<td>12.60</td>
<td>26,600</td>
<td>3.87</td>
<td>973.8760</td>
</tr>
<tr>
<td>Total</td>
<td>122,600</td>
<td>726</td>
<td>973.8760</td>
<td></td>
</tr>
</tbody>
</table>

To calculate the expected number of Lyme disease cases in Podunk, multiply Canada’s age-specific rates by Podunk’s population.

For example, the expected number of cases in Podunk for age group (0-9) is:

\[(600/100,000) \times 37,030 = 222.1800\]

The End