3 R’s of Basic Epidemiology: Rates, Ratios, and Risks

Robert Ferrer MD MPH
Dept. of Family & Community Medicine
University of Texas Health Science Center at San Antonio
Dec 3, 2005
Agenda

- Measures of frequency
  - How common is a health event relative to size of population?

- Measures of association
  - How strong is the relationship between studied exposure and disease?

- Measures of potential impact
  - What is the expected contribution of an exposure to the frequency of a disease in a population?
Measuring stuff

- **Count:** no denominator
  - There are 8 people in this room

- **Proportion:** numerator and denominator drawn from same population
  - Of the 300 people now in this building, 8 are in this room = 8/300 = 2.7%
Measuring stuff

- **Rate**: change in one quantity per unit change in another quantity
  - e.g. 78 heart beats per minute
- **Ratio**: a fraction in which numerator is not included in denominator
  - e.g. hospital beds/100,000 persons; neonatal deaths/1000 live births
Measures of frequency

- Prevalence - existing cases
  - Bias
- Incidence - new cases
Frequency measures

*Prevalence*

- Number of *existing cases* of disease in a specific population at a particular point in time / numerator + number of persons in population who are at risk but do not have the disease at the specified time
- e.g. 17% of San Antonians age 18-64 have hypertension as of December 31, 2001 (*point prevalence*)
- e.g. 7% of the US population suffered from major depression between 1995 and 2000 (*period prevalence*)
- ranges from 0 to 1
Frequency measures

Cumulative incidence

- Number of persons with *new occurrences* of a disease during a specified time period / number of persons initially at risk
- A proportion. Has no units.
- e.g. 4% of non-diabetic adults 18-64 years old in our practice became diabetic over a five-year follow-up.
- Ranges from 0 to 1
Confidence interval for a proportion

- Using normal approximation
  - Holds when \( n \) is large (\( np \geq 5 \) and \( nq \geq 5 \))
- \( SE (p) = \sqrt{pq/n} \)
- 95% C.I. = \( p \pm 1.96 \times se(p) \)
- Example: A new surgical procedure has a 3% mortality rate in the first 200 patients - what is the 95% C.I. for the surgical mortality?
  - \( SE (p) = \sqrt{(0.03 \times 0.97/200)} = 0.01206 \)
  - 95% C.I. = \( 0.03 \pm 1.96(0.01206) = (0.0064, 0.054) \)
Rule of thumb for a zero numerator

If, in a series of $N$ patients (or $N$ procedures), no instances of $X$ are observed, how confident can we be that $X$ will never occur?

- Rule of thumb (accurate for $N > 30$): We can be 95% sure that the “truth” is that $X$ would not occur more than $3/N$ times in $N$ patients.

- Example: Using the Ottawa ankle rules, $0/74$ fractures were missed. How confident can we be that we will never miss a fracture by applying the rules?
  
  $3/N = 3/74 = 0.04 = 4%$

So... based on this series, we can be 95% certain that we will misclassify no more than 4% of patients with ankle fractures as not having a fracture.
Testing $H_0$ that a proportion = prespecified proportion

- Normal approximation, two-sided
- $z = (|p - P_0| - 1/2n) / (\sqrt{P_0Q_0/n})$
  - Where $P_0$ is the prespecified proportion and $Q_0 = 1 - P_0$
  - $(-1/2n)$ is a continuity correction applied only when it is smaller than $(p - P_0)$
  - Example: 25 patients known to have a certain disorder undergo a new diagnostic test. 23/25 have a + test. Is the new test’s sensitivity greater than 75%?
    - $z = (|0.92 - 0.75| - 1/(2\times25)) / (\sqrt{0.75\times0.25/25}) = 1.73$
    - For $z = 1.73$, $p = 0.08$, so fail to reject null hypothesis that $p = P_0$
Frequency measures

- Incidence density
  - Number of new occurrences of disease in a population observed over a specified amount of time / total amount of person-time at risk experienced by the population during the time period. Units of time\(^{-1}\).
  - e.g. incidence of first major coronary event in male smokers age 40-64 is 13 events per 1000 person-years (0.013/year)
  - ranges from 0 to 1
Relationships among frequency measures

Cumhulative incidence = 1 - e^{-\text{ID}(\Delta t)} where ID is the incidence density and \Delta t is the elapsed time.

Example: probability that male smoker will have first coronary event between ages of 40 and 64 is 1 - e^{-0.013\text{ per year} \times 25\text{ years}} = 0.277
Relationships among frequency measures

- Cannot just multiply incidence rate by length of observation period
  - Example: 0.013 x 25 years = 0.325
- … because person with an event is removed from population at risk and contributes shorter period of follow-up
Example: Cohort study over 10 years

<table>
<thead>
<tr>
<th></th>
<th>Exposed</th>
<th>Non-exposed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cases</td>
<td>1,813 (a)</td>
<td>952 (b)</td>
</tr>
<tr>
<td>Non-cases</td>
<td>8,187 (c)</td>
<td>9,048 (d)</td>
</tr>
<tr>
<td>Initial population</td>
<td>10,000 (N₁)</td>
<td>10,000 (N₀)</td>
</tr>
<tr>
<td>Person-years</td>
<td>90,635 (Y₁)</td>
<td>95,163 (Y₀)</td>
</tr>
<tr>
<td>Incidence rate</td>
<td>0.0200/yr</td>
<td>0.0100/yr</td>
</tr>
<tr>
<td>Cumulative incidence</td>
<td>0.1813</td>
<td>0.0952</td>
</tr>
</tbody>
</table>

Wrong: 0.02/year x 10 years = 0.20 cumulative incidence
Right: \(1 - e^{-0.02 \times 10} = 0.1813\) cumulative incidence
Measures of association

- **Difference measures**
  - Useful for assessing public health impact of a certain exposure
  - Exposed group minus reference group

- **Ratio measures**
  - Useful for assessing causal relationships
  - Exposed group divided by reference group
Relative Risk - 3 Methods

- **Rate Ratio**
  - Incidence rate in exposed divided by incidence rate in unexposed (cohort study)

- **Risk Ratio**
  - Cumulative incidence in exposed divided by incidence in unexposed (cohort study)

- **Odds ratio**
  - Ratio of exposure odds among cases to exposure odds among controls (case-control study)
    - Valid as estimate of RR only when disease outcome occurs < 10%
Risk measures (measures of association)

Risk ratio (relative risk)

- Risk of disease among exposed / risk of disease among unexposed. \( \frac{l_e}{l_0} = \frac{a}{a+b}/\frac{c}{c+d} \)

- e.g. risk of lung ca among smokers 2/1000 per year; among nonsmokers 0.19/1000

\[ RR = 10.5 \]

\[
\begin{array}{ccc}
\text{Disease +} & \text{Disease -} \\
\text{Exposure +} & a & b \\
\text{Exposure -} & c & d \\
\end{array}
\]
Risk measures

*The odds ratio (OR)*

Used in cross-sectional or case-control designs

*Ratio of exposure odds among cases to exposure odds among controls*

<table>
<thead>
<tr>
<th></th>
<th>Disease +</th>
<th>Disease -</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exposure +</td>
<td>a</td>
<td>b</td>
</tr>
<tr>
<td>Exposure -</td>
<td>c</td>
<td>d</td>
</tr>
</tbody>
</table>
More math

- **Odds**
  - Probability of an event / probability that it doesn’t occur \( (p / 1-p) = \text{odds} \)
  - Odds / Odds + 1 = probability
  - e.g. Probability of 0.25 = odds of 0.25/0.75 = odds of 1:3
  - Odds of 1/3 = probability of 1/3 / 4/3 = 0.25
### Cohort study

<table>
<thead>
<tr>
<th></th>
<th>Disease</th>
<th>No disease</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exposed</td>
<td>$n_1p_1$</td>
<td>$n_1q_1$</td>
<td>$n_1$ fixed</td>
</tr>
<tr>
<td>Nonexp.</td>
<td>$n_2p_2$</td>
<td>$n_2q_2$</td>
<td>$n_2$ fixed</td>
</tr>
<tr>
<td>Total</td>
<td>$n_1p_1 + n_2p_2$</td>
<td>$n_1q_1 + n_2q_2$</td>
<td>$N$</td>
</tr>
</tbody>
</table>

Relative risk = \( \frac{a}{a+b} / \frac{c}{c+d} \)

\[ = \frac{n_1p_1/n_1}{n_2p_2/n_2} \]

\[ = p_1/p_2 \]
# Case-control study

<table>
<thead>
<tr>
<th></th>
<th>Case</th>
<th>Control</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exposed</td>
<td>$n_1p_1$</td>
<td>$n_1q_1$</td>
<td>$n_1$</td>
</tr>
<tr>
<td>Nonexp.</td>
<td>$n_2p_2$</td>
<td>$n_2q_2$</td>
<td>$n_2$</td>
</tr>
<tr>
<td>Total</td>
<td>$n_1p_1 + n_2p_2$ fixed</td>
<td>$n_1q_1 + n_2q_2$ fixed</td>
<td>$N$</td>
</tr>
</tbody>
</table>
## Calculating the OR

<table>
<thead>
<tr>
<th>Exposure</th>
<th>Case</th>
<th>Control</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>11 a</td>
<td>38 b</td>
<td>49</td>
</tr>
<tr>
<td>No</td>
<td>7 c</td>
<td>43 d</td>
<td>50</td>
</tr>
<tr>
<td>Total</td>
<td>18</td>
<td>81</td>
<td>99</td>
</tr>
</tbody>
</table>

Odds of exposure among cases is $11/7$. Odds of exposure among controls is $38/43$. Ratio of these odds is $(11/7) / (38/43) = (11 \times 43) / (7 \times 38) = \frac{ad}{bc}$
The OR in a case-control study - a useful property:

<table>
<thead>
<tr>
<th></th>
<th>Case</th>
<th>Control</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exposed</td>
<td>$n_1 p_1$</td>
<td>$n_1 q_1$</td>
<td>$n_1$</td>
</tr>
<tr>
<td>Nonexp.</td>
<td>$n_2 p_2$</td>
<td>$n_2 q_2$</td>
<td>$n_2$</td>
</tr>
<tr>
<td>Total</td>
<td>$n_1 p_1 +$</td>
<td>$n_1 q_1 +$</td>
<td>$N$</td>
</tr>
</tbody>
</table>

$$ad/bc = \frac{(n_1 p_1)(n_2 q_2)}{(n_1 q_1)(n_2 p_2)} = \frac{p_1 q_2}{p_2 q_1} = \frac{p_1}{p_2}$$ for $q_1$ and $q_2$ near 1.

$p_1/p_2 = $ risk ratio from cohort study!
### 2 x 2 funny business with relative risk

<table>
<thead>
<tr>
<th>Rx</th>
<th>Alive</th>
<th>Dead</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>11</td>
<td>38</td>
<td>49</td>
</tr>
<tr>
<td>B</td>
<td>7</td>
<td>43</td>
<td>50</td>
</tr>
<tr>
<td>Total</td>
<td>18</td>
<td>81</td>
<td>99</td>
</tr>
</tbody>
</table>

Survival rate for Rx A = $\frac{11}{49} = 22\%$
Survival rate for Rx B = $\frac{7}{50} = 14\%$
Death rate for Rx A = $\frac{38}{49} = 78\%$
Death rate for Rx B = $\frac{43}{50} = 86\%$

RR = 1.57
RR = 0.91
No 2 x 2 funny business with odds ratio

<table>
<thead>
<tr>
<th>Rx</th>
<th>Alive</th>
<th>Dead</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>11</td>
<td>38</td>
<td>49</td>
</tr>
<tr>
<td>B</td>
<td>7</td>
<td>43</td>
<td>50</td>
</tr>
<tr>
<td>Total</td>
<td>18</td>
<td>81</td>
<td>99</td>
</tr>
</tbody>
</table>

Survival odds for Rx A = \( \frac{11}{38} = 0.29 \)
Survival odds for Rx B = \( \frac{7}{43} = 0.16 \)  
Death odds for Rx A = \( \frac{38}{11} = 3.45 \)  
Death odds for Rx B = \( \frac{43}{7} = 6.16 \)

OR = 1.81
OR = 0.56
Measures of association and “true” relationships

- The relationship of exposure x and disease y can vary according to the presence or absence of another factor z
  ▲ = effect modification

- The relationship of exposure x and disease y is actually due to the association of a third variable with both x and y
  ▲ = confounding
Effect modification - interaction

- Exists when the strength of the association between variables is influenced by one or more other factors

<table>
<thead>
<tr>
<th></th>
<th>Circ. dz. mortality per 100,000 woman-years</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Smoker</strong></td>
<td></td>
</tr>
<tr>
<td>OC user</td>
<td>51.90</td>
</tr>
<tr>
<td>OC nonuser</td>
<td>11.76</td>
</tr>
<tr>
<td><strong>Nonsmoker</strong></td>
<td></td>
</tr>
<tr>
<td>OC user</td>
<td>15.12</td>
</tr>
<tr>
<td>OC nonuser</td>
<td>5.19</td>
</tr>
</tbody>
</table>
Effect modification

- When present, it is inappropriate to calculate summary rates across groups without adjustment for modifying factors.
- Can see effect modification in rate difference or rate ratio, or both.
## Effect modification - example

<table>
<thead>
<tr>
<th>Inactive</th>
<th>CD deaths</th>
<th>Woman-years x100,000</th>
<th>CD rate difference</th>
<th>Rate difference</th>
<th>Rate ratio difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>OC user</td>
<td>6</td>
<td>0.29</td>
<td>20.69</td>
<td></td>
<td></td>
</tr>
<tr>
<td>OC nonuser</td>
<td>3</td>
<td>0.44</td>
<td>6.82</td>
<td>13.87</td>
<td>3.03</td>
</tr>
<tr>
<td>Active</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>OC user</td>
<td>7</td>
<td>0.57</td>
<td>12.28</td>
<td></td>
<td></td>
</tr>
<tr>
<td>OC nonuser</td>
<td>1</td>
<td>0.33</td>
<td>3.03</td>
<td>9.83</td>
<td>4.05</td>
</tr>
<tr>
<td>All</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>OC user</td>
<td>13</td>
<td>0.86</td>
<td>15.12</td>
<td></td>
<td></td>
</tr>
<tr>
<td>OC nonuser</td>
<td>4</td>
<td>0.77</td>
<td>5.19</td>
<td>9.93</td>
<td>2.91</td>
</tr>
</tbody>
</table>
# Calculation of adjusted rate ratio in cohort studies

**General:**

<table>
<thead>
<tr>
<th>Ill</th>
<th>Exposed</th>
<th>Nonexposed</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$a_i$</td>
<td>$b_i$</td>
<td>$T_i$</td>
</tr>
<tr>
<td>Person-years</td>
<td>$N_{ei}$</td>
<td>$N_{0i}$</td>
<td></td>
</tr>
</tbody>
</table>

Adjusted rate ratio = \[
\frac{\sum a_i N_{0i} / T_i}{\sum b_i N_{ei} / T_i}
\]

From OC example: \[
= \frac{(6(.44)/.73) + (7(.33)/.90)}{(3(.29)/.73) + (1(.57)/.90)}
\]

\[= 3.39\]
Calculation of adjusted odds ratio in x-sec studies

General:

<table>
<thead>
<tr>
<th></th>
<th>Disease</th>
<th>Well</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exposed</td>
<td>a_i</td>
<td>b_i</td>
<td></td>
</tr>
<tr>
<td>Nonexposed</td>
<td>c_i</td>
<td>d_i</td>
<td>T_i</td>
</tr>
</tbody>
</table>

Adjusted odds ratio = \( \frac{\sum a_i d_i / T_i}{\sum b_i c_i / T_i} \)

where a_i, b_i, c_i, and d_i are the cell contents within each stratum and T_i is the subtotal n of subjects in each stratum
Confounding

A confounding factor must

- Be a “risk factor” for outcome among the nonexposed (a cause, a correlate, or influential in the recognition of the outcome)
- Be associated with the exposure variable in the population (but not be a result of exposure)
- Not be an intermediate step in the causal pathway from exposure to outcome
Confounding

Yes

Factor → Exposure → Disease

No

Exposure → Factor → Disease

No

Factor → Expose → Disease

No

Exposure → Factor

Disease
Managing confounding

*Study design*
- Matching
- Restriction
- Randomization

*Analysis*
- Stratification
- Multivariate modeling
Standardized rates

- Comparison of crude rates can be misleading if populations differ on key characteristics, such as age, sex, race/ethnicity.
- Standardization procedure recalculates *adjusted* rates as if populations had identical demographics.
# Hypothetical example

<table>
<thead>
<tr>
<th>Age</th>
<th># deaths</th>
<th>Pop.</th>
<th>Rate per 1000</th>
<th>#deaths</th>
<th>Pop.</th>
<th>Rate per 1000</th>
</tr>
</thead>
<tbody>
<tr>
<td>Young</td>
<td>1</td>
<td>100</td>
<td>10</td>
<td>20</td>
<td>1000</td>
<td>20</td>
</tr>
<tr>
<td>Middle</td>
<td>25</td>
<td>500</td>
<td>50</td>
<td>50</td>
<td>500</td>
<td>100</td>
</tr>
<tr>
<td>Old</td>
<td>100</td>
<td>1000</td>
<td>100</td>
<td>20</td>
<td>100</td>
<td>200</td>
</tr>
<tr>
<td>Total</td>
<td>126</td>
<td>1600</td>
<td>78.8</td>
<td>90</td>
<td>1600</td>
<td>56.3</td>
</tr>
</tbody>
</table>
Direct Standardization

<table>
<thead>
<tr>
<th>Age</th>
<th>Rate per 1000</th>
<th>Ref. Pop.</th>
<th>Expected deaths</th>
<th>Rate per 1000</th>
<th>Ref. Pop.</th>
<th>Expected Deaths</th>
</tr>
</thead>
<tbody>
<tr>
<td>Young</td>
<td>10</td>
<td>1100</td>
<td>11</td>
<td>20</td>
<td>1100</td>
<td>22</td>
</tr>
<tr>
<td>Middle</td>
<td>50</td>
<td>1000</td>
<td>50</td>
<td>100</td>
<td>1000</td>
<td>100</td>
</tr>
<tr>
<td>Old</td>
<td>100</td>
<td>1100</td>
<td>110</td>
<td>200</td>
<td>1100</td>
<td>220</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td></td>
<td>171</td>
<td></td>
<td></td>
<td>342</td>
</tr>
</tbody>
</table>

Developed: \( \frac{171}{3200} = 53.4 \text{ per 1000} \)
Underdeveloped: \( \frac{342}{3200} = 106.9 \text{ per 1000} \)
Indirect standardization

- Direct standardization: weigh a set of observed category-specific rates according to a standardized distribution
- Indirect standardization: weigh a set of observed population distributions according to a standardized set of category-specific rates

(used when category-specific rates in study population are unknown)
Standardizing rates

- Check to be sure no important information is lost when rates are standardized

![Graph showing age vs. rate for Country 1 and Country 2]
Risk measures

- Risk difference (AKA attributable risk, absolute risk difference)
  - Incidence in exposed - incidence in unexposed. \( I_e - I_0 \)
  - E.g. smokers 2/1000 - nonsmokers 0.19/1000 = risk difference of 1.81/1000
Risk measures (impact)

- **Attributable risk percent**

  - percentage of disease in exposed individuals due to exposure: \( \frac{I_e - I_0}{I_e} \)
  
  - e.g. \( \frac{2/1000 - 0.19/1000}{2/1000} = 90\% \)
  
  - that is, 90% of lung ca in smokers is attributable to smoking
  
  - also calculable as \( \frac{(RR-1)}{RR} \), where \( RR = \frac{I_e}{I_0} \)
Risk measures

- Population attributable risk
  \[ I - I_0 \]
  \[ \text{Risk of disease incidence in population as a whole due to exposure} \]
  \[ \text{e.g. if incidence of lung ca in population is} \]
  \[ 1.1/1000 \text{ then PAR is} \]
  \[ (1.1/1000 - 0.19/1000) = 0.91/1000 \]
  \[ \text{PAR of lung ca is} \]
  \[ 0.91/1000 \text{ for smoking} \]
Risk measures (impact)

Population attributable risk percent

\[
(I - I_0 / I) \times 100
\]

Percent of all cases occurring in a mixed population of exposed and unexposed individuals that is attributable to exposure.

\[
(1.1/1000 - 0.19/1000)/(1.1/1000) \times 100 = 83\%
\]

of all lung cancer cases in the population are attributable to smoking.