5. Multivariate analysis: Analysis
- Stratified analysis does not allow for simultaneous adjustment of many variables because cell sizes get too small and are unstable
- Most often, we adjust for many confounders at once
- Variables such as age, gender, race, and education are often incorporated into a model with the exposure to establish the independent effect of the exposure after adjusting for other variables
- Multivariate analysis is a tool to adjust for multiple variables simultaneously.

Multivariate analysis, con’d
- Involves modeling the relation between E and D by means of a mathematical equation
- In the equation, the outcome (D) is considered the dependent variable
- The exposure(s) of interest, other strong risk factors for the D and one or more PCF are considered independent (predictor, explanatory) variables.

Before including a variable in a model,
- We conduct stratified PCF-E and PCF-D and also stratified M-H crude and adjusted analysis to discern if each variable should be introduced into the mathematical model.

Multivariate analysis equation
(general form)

\[ Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n + \ldots + \gamma X_1 X_2 \]

\( \beta \) = coefficients from mathematical model
X = independent variable
Y = outcome (state you wish to predict)

Outcome = intercept + exposure \((X_1)\) + confounder \((X_2)\) + \ldots + interaction term (effect-modification) for \(X_1 \times \) confounder \(X_2\)

Multiple regression model
\[ Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n \]
- Outcome = continuous (BP, cholesterol level, Ab titers)
- The coefficients of each independent (X) variable are directly interpretable in relation to the outcome
- The presence of each variable in the model means that we know their relation to \( Y \) independent of the other variables
- Commonly used for cross-sectional data, cohort data looking at baseline or longitudinal change (also incorporates random effects)
- Many variations of the multiple regression model
Logistic regression model

- Outcome = natural log odds of disease
- \( \log(\text{odds of disease}) = \alpha + \beta_1X_1 + \beta_2X_2 + \ldots + \beta_nX_n \)
- Coefficients (\( \beta \)) can be directly converted to the odds ratio, which describes the independent association of that parameter with the outcome.
- Commonly used for case-control (frequency or unmatched) and cross-sectional data. Can be used for cohort data if time to event is not important and can adjust for follow-up time.
- A special type of logistic regression (conditional logistic regression) is used for 1:1 matched case-control data.

Proportional hazards model (Cox regression)

- Dependent variable is natural logarithm of the incidence of disease:
  \[ \log[\text{incidence rate (t)}] = \alpha + \beta_1X_1 + \beta_2X_2 + \ldots + \beta_nX_n \]
- Commonly used for incidence density data in cohort studies when the "time to event" is important to quantify (e.g., do people with the E get the disease sooner than people without the E?).

6. Rate adjustment

- Idea:
  - Statistical technique to compare rates among populations with differing underlying structures (age, gender, education, etc).
- Two ways to standardize:
  - Direct adjustment
  - Indirect adjustment (SMR)

Two ways to adjust (correct, standardize) rates:

- Direct adjustment: Use specific rates from study sample and distribution by third variable (e.g., age) from standard population.
- Indirect adjustment: Use distribution by third variable from study sample and rates from a referent (outside population), such as U.S.
  - Used commonly in occupational studies to compare observed numbers of deaths/disease among workers exposed to some agent to what would be expected in general population.

Why do we adjust?

Crude death rates between cities/counties/states/countries may look different but are different only because of confounding by a third variable.

- Consider:
  - Crude death rate in State A
  - Crude death rate in State B

Example of why we might want to adjust

- Crude death rate in Florida = 921/100,000/yr
- Crude death rate in Alaska = 506.7/100,000/yr

- Question:
  - Should we all move to Alaska?
  - Does living in Florida really mean we are almost twice as likely to die than if we lived in Alaska?
Purpose of adjusting rates

When we wish to compare mortality/morbidity rates:

- Between populations
- At different periods of time in a single population

We need to take into account possible differing structures of the underlying populations by age, sex, race, SES, etc.

Direct adjustment of rates, method 1

Need:

- (Internal) Age-specific mortality/morbidity rates for each population (study) to compare
- (External) Number of people in each age stratum in one, both populations and/or another standard population, such as U.S. population

Age-specific death/morbidity rates must match population strata (e.g., 1-4, 5-9, 10-14, 15-19 etc, or 1-9, 10-19, 20-29 etc)

Direct adjustment of rates, method 2

For each population to be compared:

1. Multiply each (internal) age-specific rate by the (external) standard population (each age stratum);
2. This generates the number of deaths/cases you would expect to see in each (age) stratum in the "normal" or standard population
3. This answers the question: how many deaths/cases would I expect to see in a "normal" population if they had the same rates as I see in my study?
4. Sum the expected number of deaths/cases to obtain the total number of deaths;
5. Divide the total number of expected deaths/cases by the total standard population to obtain the (age)-adjusted death/morbidity rate;
6. Generate an age-adjusted death rate (using the same standard population) for each population you are comparing.

Direct adjustment of rates, method 3

Now compare the age-adjusted death/morbidity rates between the populations:

Differences in mortality/morbidity experience can now be compared by getting rid of the confounding effect of age. Differences now are INDEPENDENT of age. We have ADJUSTED for the effects of age.

To summarize direct adjustment in statistical notation:

\[
\sum_{i=1}^{n} \frac{(\text{internal raw study rate}) \times \text{external standard population}}{\text{total standard population } N}
\]

\(i = 1\) (first age stratum)
\(2\) (second age stratum)
\(\ldots\) \(n\)

= Total expected # deaths/cases in standard population
= Total # people in standard population

Direct adjustment, caveats

- Adjusted (standardized) rates are fictitious numbers
- Their magnitude will reflect the size of the standard population used
- Standard population can be one of the two pops being compared, both pops, combined or a third pop.
- It doesn’t really matter which one is chosen, since it is used as the “great equalizer” in all cases.
- They are used only for comparison purposes
Direct Age Standardization for AL and FL, using 2009 US population as standard

<table>
<thead>
<tr>
<th>Age stratum</th>
<th>ALASKA</th>
<th>FLORIDA</th>
<th>US Population, 2009 (STANDARD)</th>
<th>Expected deaths (AL)</th>
<th>Expected deaths (FL)</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;1</td>
<td>643.5</td>
<td>705.8</td>
<td>4,261,000</td>
<td>27,420</td>
<td>30,074</td>
</tr>
<tr>
<td>1-3</td>
<td>34.8</td>
<td>34.6</td>
<td>17,039,000</td>
<td>5,947</td>
<td>5,861</td>
</tr>
<tr>
<td>4-14</td>
<td>28.9</td>
<td>16.0</td>
<td>40,584,000</td>
<td>11,729</td>
<td>6,493</td>
</tr>
<tr>
<td>15-24</td>
<td>120.4</td>
<td>100.6</td>
<td>43,078,000</td>
<td>51,866</td>
<td>43,337</td>
</tr>
<tr>
<td>25-34</td>
<td>119.1</td>
<td>127.8</td>
<td>41,567,000</td>
<td>49,506</td>
<td>53,123</td>
</tr>
<tr>
<td>35-44</td>
<td>226.4</td>
<td>208.0</td>
<td>41,531,000</td>
<td>9,4026</td>
<td>8,665</td>
</tr>
<tr>
<td>45-54</td>
<td>423.0</td>
<td>459.2</td>
<td>44,592,000</td>
<td>51,866</td>
<td>43,337</td>
</tr>
<tr>
<td>55-64</td>
<td>796.1</td>
<td>888.2</td>
<td>34,787,000</td>
<td>276,939</td>
<td>308,978</td>
</tr>
<tr>
<td>65-74</td>
<td>2008.6</td>
<td>1808.8</td>
<td>20,792,000</td>
<td>417,628</td>
<td>376,086</td>
</tr>
<tr>
<td>75-84</td>
<td>432.5</td>
<td>478.1</td>
<td>13,148,000</td>
<td>671,468</td>
<td>546,378</td>
</tr>
<tr>
<td>85+</td>
<td>11754.3</td>
<td>10258.1</td>
<td>5,631,000</td>
<td>661,885</td>
<td>577,378</td>
</tr>
<tr>
<td>ALL</td>
<td>506.7</td>
<td>921.0</td>
<td>307,000,000</td>
<td>2,457,308</td>
<td>2,247,543</td>
</tr>
</tbody>
</table>

Divide expected deaths/standard population x 10^5
- ALASKA: [2,457,308/307,000,000] x 10^5 = 800.43 per 100,000
- FLORIDA: [2,247,543/307,000,000] x 10^5 = 732.10 per 100,000

Adjusted death rates show that death rate is a bit higher in Alaska than Florida (in agreement with trends seen in age-specific data)

Indirect adjustment of rates, method 1
- Need:
  - Number of people in each (age) stratum for each population to compare
  - Rates from outside source (NCHS mortality data for U.S., SEER incidence rates for site-specific cancer, etc)

Indirect adjustment of rates, method 2
- Rates from a standard population are used to calculate the expected number of cases/events using the (age) distribution of the study population.
- Note:
  - In indirect adjustment, we use the study’s population numbers as weights and rates from outside (standard) to calculate expected # outcomes for each (age) stratum.
  - In direct adjustment, we use the numbers from outside as weights (standard) and rates from inside the study to calculate the expected # outcomes for each stratum.

Standardized morbidity/mortality ratio (SMR)
- The results of an indirect standardization are presented as a “standardized morbidity/mortality ratio” (SMR)

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

Example: Indirect adjustment
Szklo & Nieto, 2000 Table 7-8

<table>
<thead>
<tr>
<th>Age (yrs)</th>
<th>Study Group A</th>
<th>Study Group B</th>
<th>External reference rates</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt; 40</td>
<td>N</td>
<td>Deaths</td>
<td>Rate</td>
</tr>
<tr>
<td>≥40</td>
<td>500</td>
<td>100</td>
<td>10%</td>
</tr>
<tr>
<td>Total</td>
<td>600</td>
<td>110</td>
<td>18.3%</td>
</tr>
</tbody>
</table>

Answers the question: If my study population had the same rates as (external reference rates), how many deaths/cases would I expect to see in my study?
Example: Indirect adjustment, con’d

<table>
<thead>
<tr>
<th>Age (yrs)</th>
<th>Study Group A</th>
<th>Study Group B</th>
<th>External reference rates</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>N</td>
<td>Deaths</td>
<td>Rate</td>
</tr>
<tr>
<td>&lt; 40</td>
<td>100</td>
<td>10</td>
<td>10%</td>
</tr>
<tr>
<td>≥40</td>
<td>500</td>
<td>100</td>
<td>20%</td>
</tr>
<tr>
<td>Total</td>
<td>600</td>
<td>110</td>
<td>18.3%</td>
</tr>
</tbody>
</table>

Study Group A
Study Group B
SMR\textsubscript{A} = \frac{12\% \times 100}{12\% \times 500} = \frac{12000}{60000} = 0.2 \text{ or } \frac{120}{600} = 0.20 \text{ or } 20\%<br>
SMR\textsubscript{B} = \frac{50\% \times 500}{50\% \times 100} = \frac{25000}{50000} = 0.5 \text{ or } \frac{250}{500} = 0.50 \text{ or } 50\%<br>

Total expected
\sum {262} \times 110 = 42\% \text{ or } \frac{28220}{66000} = 0.42 \text{ or } 42\%<br>

Under what circumstances to adjust and not to adjust?

<table>
<thead>
<tr>
<th>Distribution of characteristic</th>
<th>Rates</th>
<th>Adjust</th>
<th>Do not adjust</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Same</td>
<td>Different</td>
<td>Adjust</td>
<td>X</td>
</tr>
<tr>
<td>2. Different</td>
<td>Same</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>3. Different</td>
<td>Different</td>
<td>X</td>
<td></td>
</tr>
</tbody>
</table>

Notes – when to use stratum-specific rates:
In situation 3: Use caution – adjusted rate may mask important stratum-specific differences in rates. Best to present stratum-specific rates and briefly mention results of adjustment.

If several stratum-specific rates differ in opposite directions (e.g., if in two age strata, Pop I has higher death rates, while in 2 other age strata, Pop II has higher death rates), adjustment may not be indicated; stratum-specific rates should be emphasized.

Chance, bias and confounding

- Once we have eliminated chance, bias and confounding as potential explanations for E-D association, we can pronounce the relation to be “valid”.
- Our job as epidemiologists is to “kill” associations. If they can not be killed, maybe they are real.

Introduction to effect-modification

- Does the magnitude of the association between E and D vary by the presence/level of a third variable?
- When we limit the analysis to a level of a third variable (e.g., gender), is the E-D association different for the two groups?
- Ex: smoking-LC is much stronger in individuals who have been occupationally exposed to asbestos than in those who have never been exposed to asbestos.

Introduction to effect-modification

- Unlike bias and confounding, which need to be investigated and eliminated in order to declare a valid association, effect-modification is a biologic phenomenon that needs to be investigated, described and reported.

Introduction to effect-modification

- PCF is now considered a potential “effect-modifier”
The association between smoking and lung cancer depends on asbestos exposure

- Smoke + Asbestos + → Lung Cancer (RR=18)
- Smoke + Asbestos - → RR=5.0
- Smoke - Asbestos + → RR=3.0
- Smoke - Asbestos - → RR=1.0

Facts about effect-modification
Oleckno, WA (2002), p.221

- Effect-modification is distinct from confounding
- Confounding is an annoyance that needs to be controlled so we can get an accurate (valid) look at the relation between an E and a D.
- Effect-modification is a real effect that helps elucidate the relation between an E and a D in the presence of other factors
- Therefore, it should always be described in a study.
- We can see effect-modification most clearly in a stratified analysis.

E-D can have opposing effects by the third variable

- Among women < age 50:
  - High BMI at age 18 in the NHS decreases risk for breast cancer (0.57; 95% CI, 0.41 - 0.81)
- Among women > age 50:
  - High BMI (obesity) increases risk for breast cancer (OR=1.6 - 2.1)
  - PMID: 9020494 [PubMed]
- "Body size during the early phases of adult life seems to be particularly important in the development of premenopausal breast cancer," the authors of the NHS write.

The third variable can be:

- An effect-modifier and a confounder
  - Report as an effect-modifier
- An effect-modifier but not a confounder
  - Report as an effect-modifier
- A confounder but not an effect-modifier
  - Adjust and eliminate confounding
- Neither a confounder nor an effect-modifier
  - Report crude
- While confounding and effect-modification are completely different concepts, we can use stratified analyses as one way to evaluate simultaneously.

Example of effect-modification

<table>
<thead>
<tr>
<th>OC user</th>
<th>Stroke</th>
<th>No stroke</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>200</td>
<td>200</td>
</tr>
<tr>
<td>Y</td>
<td>150</td>
<td>500</td>
</tr>
<tr>
<td>OR = 3.33</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

```
Stroke  No stroke
-150    100
50      250
Smokers OR = 7.5
```

```
OC user | Stroke | No stroke |
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>350</td>
<td>700</td>
</tr>
<tr>
<td>Y</td>
<td>50</td>
<td>100</td>
</tr>
</tbody>
</table>

Non-smokers OR=1.25
```

Questions to ask yourself about confounding and effect-modification:

Confounding:

- Is the crude measure between E-D distorted?
- Is the crude RR/OR outside of the stratum-specific interval?
- Is the adjusted RR/OR different from the crude?

Effect modification:

- Does the E-D association vary according to a level of a third variable?
- Are the stratum-specific measures different from one another?
Interpretation

- The OC-stroke association is stronger among women who also smoke
- The OC-stroke association is (almost) absent among women who do not smoke
- Smoking modifies the effect of OCs on the risk of stroke
- Smoking is not a confounder (e.g., crude OR lies within the stratum-specific ORs) – therefore we should not adjust for smoking but show the effect-modification, i.e., women who smoke should not use OCs.

Confounding

- Statistically:
  - (RR crude – RR adjusted)/RR adjusted
    - If > 10%, there is confounding
  - If crude RR/OR lies outside of stratum-specific RR/OR interval
    - In multivariate modeling, if adding the PCF to the model changes the model by >10%
- Biologically:
  - If the adjusted RR/OR “looks” (visually) different from the crude (given N) and it makes biologic sense (plausibility).

Synergistic vs. antagonistic E-M

- Synergy (positive interaction): two factors work together to produce more disease than would be expected based on the action of either one working alone: risk is greater than additive or multiplicative
  - Smoking + asbestos = much larger risk of LC
- Antagonism (negative interaction): two factors working together produce less disease than would be expected based on the action of either one working alone
  - third factor mitigates (annuls) E-D association
  - Ex: contaminated shellfish – gastroenteritis
    - If eaten in the presence of alcohol, no association

Summary on what to do with confounding and effect-modification

<table>
<thead>
<tr>
<th>OR1=6.13</th>
<th>OR2=1.85</th>
<th>ORcrude=2.85</th>
<th>E-M</th>
<th>No confounding Report and describe</th>
</tr>
</thead>
<tbody>
<tr>
<td>p=0.0001</td>
<td>p=0.009</td>
<td>p=0.003</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>OR1=6.13</th>
<th>OR2=1.85</th>
<th>ORcrude=1.56</th>
<th>E-M</th>
<th>Confounding Report and describe</th>
</tr>
</thead>
<tbody>
<tr>
<td>p=0.0001</td>
<td>p=0.009</td>
<td>p=0.003</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>OR1=2.01</th>
<th>OR2=1.85</th>
<th>ORcrude=1.56</th>
<th>E-M</th>
<th>No E-M Confounding Adjust/summary</th>
</tr>
</thead>
<tbody>
<tr>
<td>p=0.0001</td>
<td>p=0.009</td>
<td>p=0.003</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>OR1=2.01</th>
<th>OR2=1.85</th>
<th>ORcrude=1.90</th>
<th>E-M</th>
<th>No E-M No confounding Report crude</th>
</tr>
</thead>
<tbody>
<tr>
<td>p=0.0001</td>
<td>p=0.009</td>
<td>p=0.003</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Also see p.222 Oleckno

How do we decide whether there is an ‘important’ amount of confounding or effect-modification?

- Two schools of thought:
  - Evaluate statistically
  - Evaluate biologically
  - Usually, some balance of the two.

Steps to evaluate confounding and effect-modification - summary

- Look at crude E-D association
- Stratify E-D association by levels of the third variable
- Calculate stratum-specific RR/ORs
- Evaluate whether crude RR/OR is outside stratum-specific interval (confounding)
- Evaluate RR/OR for similarity (effect-modification) – “eyeballing” (clinical) vs. “statistical” (test)
  - If effects are thought to be uniform, no confounding: use crude OR
  - If effect are thought to be uniform, confounding, use ORadj or modeling
  - If effects are not uniform, effect-modification may be present: report stratum-specific estimates and describe (biologic plausibility). There is usually a biologic reason that the effect is limited to a subgroup of people!

Summary on what to do with confounding and effect-modification

<table>
<thead>
<tr>
<th>OR1=6.13</th>
<th>OR2=1.85</th>
<th>ORcrude=2.85</th>
<th>E-M</th>
<th>No confounding Report and describe</th>
</tr>
</thead>
<tbody>
<tr>
<td>p=0.0001</td>
<td>p=0.009</td>
<td>p=0.003</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>OR1=6.13</th>
<th>OR2=1.85</th>
<th>ORcrude=1.56</th>
<th>E-M</th>
<th>Confounding Report and describe</th>
</tr>
</thead>
<tbody>
<tr>
<td>p=0.0001</td>
<td>p=0.009</td>
<td>p=0.003</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>OR1=2.01</th>
<th>OR2=1.85</th>
<th>ORcrude=1.56</th>
<th>E-M</th>
<th>No E-M Confounding Adjust/summary</th>
</tr>
</thead>
<tbody>
<tr>
<td>p=0.0001</td>
<td>p=0.009</td>
<td>p=0.003</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>OR1=2.01</th>
<th>OR2=1.85</th>
<th>ORcrude=1.90</th>
<th>E-M</th>
<th>No E-M No confounding Report crude</th>
</tr>
</thead>
<tbody>
<tr>
<td>p=0.0001</td>
<td>p=0.009</td>
<td>p=0.003</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Also see p.222 Oleckno