What to do today (Feb 16, 2023)?

Part I. Introduction
Part II. Epidemiologic Concepts and Designs
Part III. Clinical Trials

Part IV. Modern Biostatistical Approaches

Part IV.1 Incomplete Data Analysis

Part IV.2 Some Other Important Topics (Chp 8 - 18, Koepsell and Weiss, 2003)

IV.2.1 Measures of Risks
IV.2.2 Measurement Error Revisit
IV.2.3 Confounding and Its Control
IV.2.4 Causation vs Association

Part IV.3 Selected Widely-Used Algorithms

Part IV.2.2 Measurement Error Revisit

Mismeasurement of exposure status or level is "present to at least some degree in nearly every epidemiologic study, since nearly every means of ascertaining the presence or level of exposure is imperfect" – Koepsell and Weiss (2003)

- *Measure* refers broadly to any way of capturing data on a certain characteristic of study subjects.
- *Measurement error* is the discrepancy between the true value and the measured value.
- *The scale of measurement* is usually categorized into
  - continuous: e.g. body weight; any positive real number
  - categorical: ordinal vs nominal; e.g. disease serverity – mild, moderate, severe vs gender – male, female

  Misclassification; *Fine to Coarse Measurement Scales*

Part IV.2.2 Measurement Error Revisit

Assessing Measurement Error

- **Reliability.** A good measurement should yield the same value if applied repeatdly under circumstances in which the underlying characteristic is believed to remain the same.
  - e.g. for binary measures and 2 × 2 table of outcomes, concordance [percent agreement]: $p_O = \frac{n_{11}}{n_{++}} + \frac{n_{22}}{n_{++}}$
  - e.g. for binary measures, Kappa: $\kappa = \frac{p_O - p_e}{1 - p_e}$ with $p_e = \left(\frac{n_{1+}}{n_{++}}\right)\left(\frac{n_{+1}}{n_{++}}\right) + \left(\frac{n_{2+}}{n_{++}}\right)\left(\frac{n_{+2}}{n_{++}}\right)$, expected overlap by chance
  - e.g. for continuous measures, intraclass correlation coefficient (reliability ratio): $\lambda = \frac{\sigma_x^2}{\sigma_x^2 + \sigma_u^2} \leq 1$

Part IV.2.2 Measurement Error Revisit

- **Validity** A good measurement method should yield the correct value. [Being consistent is not good enough if the results are consistently wrong.]

  A gold standard (a criterion measure) is required to evaluate the validity of a measure.

  - sensitivity and specificity: 2×2 table of outcomes with a diagnosis test and the condition presence
    Sensitivity$=P(T_+|C_+)$, estimated by n_{11}/n_{+1}
    Specificity$=P(T_-|C_-)$, estimated by n_{22}/n_{+2}
  - when a test yields an ordinal or continuous scale, often is to select a cutoff value $\Longrightarrow$ receiver operating characteristic (ROC) curve: (1-specificity, sensitivity) at different cutoff values
    uninformative test; good test; perfect test

Part IV.2.2 Measurement Error Revisit

Consequences of Measurement Error

- with Continuous Variables
  - when the variable is the response: if the errors sum up to zero? if the errors don't sum up to zero?
  - when the variable is explanatory: if the errors sum up to zero? [Part IV.1.4] if the errors don't sum up to zero?
- with Categorical Variables (misclassification)
  - non-differential (non-selective) – a form of random measurement errors?
  - differential – bias to a particular direction?

Part IV.2.2 Measurement Error Revisit

"*Nondifferential misclassification of exposure is ubiquitous in epidemiology, and usually leads to an attenuation of the estimated size of a true association between exposure and disease.*" (Thomas, 1995)

Example. In a case-control study

A. When the exposure was measured perfectly

Exposure	case	control	Odds Ratio
yes	150	75	$\frac{150}{150} \div \frac{75}{225}$
no	150	225	$= 3.0$
Total	300	300	

Part IV.2.2 Measurement Error Revisit

Example. (cont'd)

B. When 1/3 of exposured subjects were misclassified

Exposure	case	control	Odds Ratio
yes	150-50	75-25	$\frac{100}{200} \div \frac{50}{250}$
no	150+50	225+25	$= 2.5$
Total	300	300	

C. In addition to B., 20% of non-exposured subjects were misclassified

Exposure	case	control	Odds Ratio
yes	150-50+30	75-25+45	$\frac{130}{170} \div \frac{95}{205}$
no	150+50-30	225+25-45	$= 1.65$
Total	300	300	

Part IV.2.3 Confounding and Its Control

What is confounding?

“Confounding occurs in epidemiologic research when the measured association between an exposure and disease occurrence is distorted by an imbalance between exposed and non-exposed persons with regard to one or more other risk factors for the disease.”

– Koepsell and Weiss (2003)

Part IV.2.3A Confounding and Its Control

Example. Crude Death Rate (per 100,000 person-years):

$$\frac{\text{total deaths in a year}}{\text{average popluation in the year}} \times 10^5$$

- U.S. Global Health Policy: (http://www.globalhealthfacts.org/data/topic/map.aspx?ind=90) Crude Death Rate (per 100,000 people) in 2012:

  Canada 8.09; Mexico 4.90

- Mexican age specific moratlity rates are greater: The World Bank (http://data.worldbank.org/indicator/SH.DYN.MORT)

  age 5 or under group: Canada 6; Mexico 16

  Why?

Part IV.2.3A Confounding and Its Control

Example. Mortality Rates in Two Hypothetical Communities

	Community A			Community B		
Age	No. of Deaths	Mid-Year Population	Rate[a]	No. of Deaths	Mid-Year Population	Rate[a]
young	1	1000	1	10	5000	2
middle	15	3000	5	40	4000	10
old	50	5000	10	20	1000	20
Total	66	9000	7.3	70	10,000	7.0

[a] Deaths per 1000 person-year

- Crude Death Rates (1000 per-year): A, 7.3; B, 7.0
- Mortality Rates in A and B both sharply increase with increasing age
- Difference in the age distributions on average: people in A older

A has higher proportion of older people and is "penalized" in comparison to B: the Simpson's Paradox

Part IV.2.3B Confounding and Its Control

Methods of Accounting for Confounding Variables:

- **in the Study Design:**
  - random assignment
  - matching - select matched pairs (sets) from each age group in Mexico and Canada
  - restriction - compare death rate within a specific age group
- **as Part of Data analysis:**
  - stratification - obtain separate comparisons of death in each selected age groups using age-specific mortality rates
  - covariate adjustment

Advantage vs disadvantage for each?

Part IV.2.3C Confounding and Its Control

Standardization: to calculate what would have been the overall mortality rates in A and B if they had the same age composition (i.e. by using a common set of weights).

- Step 1. Pick a reference population to construct weights

  Choice of a Standard Population:

  - regional comparisons may use the combined population of a specified date as the standard
  - the non-exposed group

- Step 2. Calculate weighted average using age-specific rates in each population and the selected weights.

  The common confounding factor distn is taken from the standard population; hence, the term of "standardization".

Part IV.2.3C Confounding and Its Control

Example. Mortality Rates in Two Hypothetical Communities (cont'd)

- Step 1. Select the combined mid-year population of Community A and B to construct the reference population:

Age	Standard Weights	
young	(1000+5000)/19,000	= 0.316
middle	(3000+4000)/19,000	=0.368
old	(5000+1000)/19,000	=0.316
Total		1.000

- Step 2. Calculate weighted average

	Community A			Community B		
Age	rate	weight		rate	weight	
young	1 ×	.316	=.316	2 ×	.316	=.632
middle	5 ×	.368	=1.84	10 ×	.368	=3.68
old	10 ×	.316	=3.16	20 ×	.316	=6.32
		5.3[a]			10.6[a]	

[a] Age standardized mortality rates in Community A and B

Part IV.2.3C Confounding and Its Control

Direct vs Indirect Standardization

- Direct Standardization: all disease rates from strata are (weighted) averaged using the distribution of the standard population for the weights
    - It gives the crude rate would have been if the study population(s) had the same distribution as the standard population.
      Other adjusted measures ... ...
      e.g. $\hat{\theta}_{XY,MH} = \frac{\sum_k N_{11k}N_{22k}/N_{++k}}{\sum_k N_{12k}N_{21k}/N_{++k}}$ [adjusted OR]
    - It may be inefficient when there are few events per stratum

Part IV.2.3C Confounding and Its Control

- Indirect Standardization:
  - "Multivariate regression analysis" (Multiple regression?)
    e.g. multiple logistic regression analysis [adjusted log-OR]: an additional covariate to adjust for the effect of a confounder
  - Propensity scores
    to control multiple potential confounders simultaneously by using a propensity score:
- *First modeling the exposure variable as a function of the potential confounders by logistic regression or a related method:*

  to calculate an expected probability ("propensity") of exposure for each study subject
- *Then examining the exposure-outcome association while controlling for the propensity score by stratification, matching, or covariate adjustment*

Part IV.2.3C Confounding and Its Control

Stratification:

to separate data into several subgroups (e.g. by age and sex)

- 1st step in standardization
- stratified analysis: rationale for reporting it vs a combined result?

"conditioning"

Part IV.2.3C Confounding and Its Control

Conditional vs Marginal Associations

- **X-Y conditional odds ratios**: [describe conditional X-Y association] For $Z = k$, $k = 1, \ldots, K$,

$$\theta_{XY(k)} = \frac{\pi_{11k}\pi_{22k}}{\pi_{12k}\pi_{21k}} = \frac{\mu_{11k}\mu_{22k}}{\mu_{12k}\mu_{21k}}$$

  If $\theta_{XY(k)} \equiv$ *constant,* $\Longrightarrow$ *"homogeneous" conditional X-Y association*

- **X-Y marginal odds ratios**: [describe marginal X-Y association]

$$\theta_{XY} = \frac{\pi_{11+}\pi_{22+}}{\pi_{12+}\pi_{21+}} = \frac{\mu_{11+}\mu_{22+}}{\mu_{12+}\mu_{21+}}$$

Part IV.2.3C Confounding and Its Control

- Homogeneous conditional association
  If $\theta_{XY(k)} = c$ for all k, not necessarily $\theta_{XY} = c$
  e.g. the Simpson's Paradox

- Marginal vs conditional independence
  - $X \perp Y | Z \leftrightarrow$ (iff) $\theta_{XY(k)} = 1$ for all k
  - $X \perp Y \leftrightarrow$ (iff) $\theta_{XY} = 1$
  - $X \perp Y | Z \not\leftrightarrow X \perp Y$

Part IV.2.3C Confounding and Its Control

- **Cohran-Mantel-Haenszel Test.** with a $2 \times 2 \times K$ table, to test $X \perp Y | Z$ – H_0 : "$\theta_{XY(k)} = 1$ for all $k = 1, \ldots, K$" vs H_1 : otherwise
  - CMH-test works well if conditional X-Y associations are similar
- **Mantel-Haenszel Estimator.** with a $2 \times 2 \times K$ table, when $\theta_{XY(1)} = \ldots = \theta_{XY(K)}$, to estimate the common conditional odds ratio: $\hat{\theta}_{XY,MH} = \frac{\sum_k N_{11k}N_{22k}/N_{++k}}{\sum_k N_{12k}N_{21k}/N_{++k}} \neq \frac{N_{11+}N_{22+}}{N_{12+}N_{21+}}$
- **Breslow-Day Test.** with a $2 \times 2 \times K$ table, to test for homogeneity of conditional odds ratios – $H_0 : \theta_{XY(1)} = \ldots = \theta_{XY(K)}$ vs H_1 : otherwise

Part IV.2.3D Confounding and Its Control

Confounding vs Mediating Variables

- Mediators are also known as intervening or intermediate variables.
- Confounders are associated with but not caused by exposure; adjusting for variables on the causal pathway biases estimated odds ratios towards one (Leon, 1993).

e.g. Birth weight is on the causal pathway between maternal smoking and infant mortality:

$$\begin{array}{ccccc} \textit{Maternal} & & \textit{Birth} & & \textit{Infant} \\ \textit{Smoking} & \Rightarrow & \textit{Weight} & \Rightarrow & \textit{Mortality} \\ (\textit{Exposure}) & & (\textit{Mediator}) & & (\textit{Outcome}) \end{array}$$

The odds ratio for infant mortality comparing smokers to non-smokers was:

- 1.3 (95% CI (1.2,1.4)), after adjusting for marital status, education, maternal age and parity;
- 1.0 (95% CI (0.9,1.1)), after further adjustment for infant birth weight!

Part IV.2.3E Confounding and Its Control

Residual Confounding

Our ability to obtain unconfounded estimates for the effect of exposure in observational studies is limited by residual confounding due to:

- unknown confounding variables,
- known confounders are not measured,
- random measurement error (non-differential misclassification) of confounders biasing adjusted estimates of the exposure-disease association towards estimates of the unadjusted association.

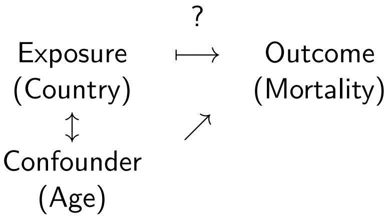
For example,

- Mothers who smoke while pregnant tend to have smaller babies.
- Male babies tend to be bigger than female babies.
- To what extent could the observed association between maternal smoking and infant birth weight be confounded by infant gender?

Part IV.2.3F Confounding and Its Control

When is confounding present?

- **classical criteria**
  A variable is a confounder if it is associated with exposure and causally related to the outcome:



the question mark ? about the association of Country and Mortality

Part IV.2.3F Confounding and Its Control

- **collapsibility criterion**

  Confounding is present when there is a substantive difference between the crude and adjusted odds ratios.

  - A common application of the collapsibility criterion concern for the effects of confounding occur when the crude and adjusted estimates of excess risk differ by at least 10%.

Part IV.2.3F Confounding and Its Control

How to Use the Criteria for Confounding?

- The classical criteria may be used when designing a study to: (i) develop a conceptual framework and (ii) identify potential confounding variables.

  The classical criteria may also prove useful in identifying the source of confounding.

- The collapsibility criteria is most useful when deciding how best to describe study results.

What to study next?

Part I. Introduction
Part II. Epidemiologic Concepts and Designs
Part III. Clinical Trials

Part IV. Analytic Epidemiology

Part IV.1 Incomplete Data Analysis (*supplementary*)

Part IV.2 Some Other Important Topics (Chp 8 - 18, Koepsell and Weiss, 2003)

IV.2.1 Measures of Risks
IV.2.2 Measurement Error Revisit
IV.2.3 Confunding and Its Control
IV.2.4 Causation vs Association

Part IV.3 Selected Widely-Used Algorithms

IV.3.1 Bootstrap and Related
IV.3.2 EM Algorithm and Related