Secondary Data Analysis

Villanova College of Nursing Colloquium
Michael A. Posner, Ph.D.
Department of Mathematical Sciences
Villanova University
April 27, 2007

1

Who am I?

- Assistant Professor, Department of Mathematical Sciences, since 2005
- Public Health Researcher, 1996-2005
  - New England Research Institutes
  - Boston University / Boston Medical Center
    - Data Coordinating Center (BUSPH)
    - Research Data Assistance Center (CMS)
    - Health Care Research Unit
    - National Center of Excellence in Women's Health
- Ph.D. in Biostatistics, Boston University

2

Goals

1. Secondary data are valuable resources for (public health / nursing) researchers

2. Understand issues in use of secondary data

3

Outline

- What is Secondary Data?
- How do I Use Secondary Data?
- Secondary Data Examples
- Causal Inference for Intervention Studies

4

Outline

- What is Secondary Data?
- How do I Use Secondary Data?
- Secondary Data Examples
- Causal Inference for Intervention Studies

5

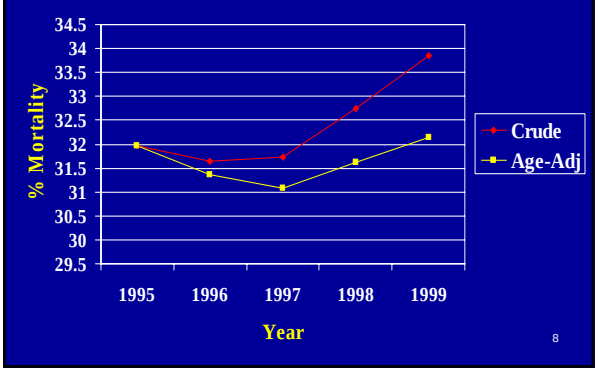
6

Consider a Research Question

- Mortality following acute myocardial infarction (AMI) was falling through mid 1990s but has begun increasing
- Why is this happening?
  - Older population?
  - Patients are sicker?
  - Revascularization techniques?
  - More patients with second AMI?

7

Post-AMI Mortality Trend in Medicare



8

Designing a Randomized Controlled Trial (RCT)

- Study Approval (IRB)
- Identify patients with AMI hospitalization
- Solicit them to join the study
  - Bias due to refusal?
- Inquire as to their medical history
  - Recall bias? Information bias?
- Track them over time (1 year?)
  - Computer systems, loss to follow-up?, time!!!
- Collect and clean the data
- Etc…
- Millions of $$$ and many years

9

Secondary Data Analysis

- Evaluation of hypotheses using data where the original purpose of data collection was to answer a different (primary) question
- Also includes
  - Data with no primary hypothesis
  - Analysis of publicly-available or population-based data

10

Benefits of Secondary Data

- Less resources (time, money)
- Expertise used
  - Data gathering (sampling)
  - Survey design
  - Instrument development
- Clean data
- Easier IRB approval
- Less burden to society

11

Benefits of Secondary Data II

- Larger sample sizes
- Often longitudinal
  - Gathered over time
  - Using same instruments and formats
- Population-based (vs. individual-based)
  - Some people don't want to participate in research studies
- Breadth of availability of secondary data

12

Drawbacks of Secondary Data

- May lack information / variables
  - Didn't collect time in U.S. for immigrants
  - Collapsed race into White, Black, Other
- Might be limited to a population different than the one about which you wish to infer
  - Different time
  - Different geographic region
  - Different sampling frame

13

Drawbacks of Secondary Data II

- Some data withheld for confidentiality
  - HIPAA, unique identifiers, etc.
- Details of data gathering not available
  - Trust the integrity of the data collectors, database managers, and analysts

14

Drawbacks of Secondary Data III

- CORRELATION DOES NOT IMPLY CAUSATION!
  - *All those who drink of this remedy recover in a short time, except those whom it does not help, who die. Therefore, it is obvious that it fails only in incurable cases. --Galen (circa 100 A.D.)*

15

Which Should I Use? Primary or Secondary Data

- It depends on your study and resources
- Secondary and primary data both have benefits and limitations
- Some analysts propose mixed designs that involve both experimental (primary) and observational (secondary) data

16

Sources of Secondary Data

- Clinical setting
  - Patient charts
  - Satisfaction surveys
- State level
  - Labor bureau, vital stats, health records
- Federal Agencies
  - Census Bureau
  - National Council on Health Statistics (CDC)
  - National Institutes of Health
  - Center for Medicare and Medicaid Services (CMS)
  - See resources (last slide)

17

Secondary Data Sources

- Secondary Data Sources for Public Health (2007). Sarah Boslaugh. Cambridge University Press.
- *Existing Population-Based Health Databases: Useful Resources for Nursing Research* (2007). Zeni and Kogan. Nurs Outlook 2007;55:20-30.
- *Secondary Data Analysis: Using Existing Data to Answer Clinical Questions*. (2005) Rosenberg, Greenfield, Dimick. Prepared for chapter of Clinical Research Methods for the Surgical Disciplines.

18

AMI Mortality Trend Secondary Data Analysis

- Identify secondary data sources
  - Medicare data (MedPAR, Inpatient, Outpatient, Carrier (part B), Denominator)
- Get approval to use data
  - IRB, Original Source (CMS)
- Limitations
  - Only generalize to Medicare beneficiaries
  - Some desired variables not available

19

Is This a Good Data Set To Use?

- Purpose of study/data
- Sponsor
- Data collector and manager
- Mode of data collection
- Quality of data
- When were data collected
- Sampling procedures
- Consistency with other sources
  - Sample size, demographic information, disease rates, etc.

20

Outline

- What is Secondary Data?
- How do I Use Secondary Data?
- Secondary Data Examples
- Causal Inference for Intervention Studies

21

Statistical Inference

- Use sample data to infer about population
- Identify sampling frame / generalizability
  - From what population are you gathering data?
  - Who are you excluding?
  - Non-response bias / refusal to participate
- AMI Example: Is it appropriate to generalize from Medicare beneficiaries to U.S. population?

22

Biases

- Non-response bias
  - The people who don't answer are different than those that do answer
  - Repeated request for information (3-tier)
- Response, recall, or information bias
  - The data you get are not good data due to people lying, not remembering, misestimating, or misinterpreting questions

23

Sampling

- What is sampling?
  - Subset of population used for inference
- Types of Sampling
  - Simple random sampling
  - Complex sample designs
    - Stratified
    - Cluster
    - Mixed mode

24

Sampling Weights

- Sampling techniques
  - Allow valid inferences to population of interest
  - Increase efficiency of study
- Secondary data analysis has to reverse effects of sampling techniques
  - Using weights or other design adjustments

25

Inferences Using Weights

- Consider a study comparing STD rates by race
- Sample 1000 White and 1000 Blacks
- 20% of Whites and 30% of Blacks have STDs
- Conclusion: Blacks have higher rates (10% higher or 1.5 times higher) of STD that Whites

26

Inference Using Weights II

- Secondary question: What percent of people in U.S. have STDs?
  - Naïve answer
    - (20% of 1000) + (30% of 1000) = 500 / 2000
    - Rate of STD in population = 25%
  - Weighted answer
    - 85% of population is White, 15% is Black
    - 20% of 85 + 30% of 15 = 21.5%
    - Rate of STD in population = 21.5%

27

Inference Using Weights III

- Sampling weight = $p_i * N / n_i$
  - p_i = proportion of population for group i
  - N = total sample size
  - n_i = sample size for group i
  - Whites = 0.85 * 2000 / 1000 = 1.7
  - Blacks = 0.15 * 2000 / 1000 = 0.3
  - These weights can be used in other types of analyses (contingency tables, regression, etc.)

28

Outline

- What is Secondary Data?
- How do I Use Secondary Data?
- Secondary Data Examples
- Causal Inference for Intervention Studies

29

Examples of Secondary Data

- Framingham Heart Study (1948)
  - Goal: Identify factors associated with CVD
  - n=5209, 5124 offspring, Recruiting Gen III
  - 1288 publications through 2004
    - Genetics, diet vs. bone density, back symptoms, …
- Nurses Health Study (1976, 1989)
  - Goal: Examine long-term effect of oral contraceptives
  - 122,000 nurses
  - 700 publications (estimated) through 2005
    - Carbohydrate intake vs. stroke, diet vs. pancreatic cancer, …

30

AMI Mortality Trends Using Secondary Data

- Generalizability
  - Can you use Medicare beneficiaries to infer about the entire U.S. population?
  - Medicare patients with a principal inpatient diagnosis code of AMI
  - Include only if eligible for Medicare for entire prior year
  - Transfers (re-admission w/in 1 day of discharge) rolled up into 1 record (11.3%)

31

AMI Mortality Trends References and Additional Studies

- *Final Report (to CMS): Risk Adjustment Models to Examine AMI Mortality Trend*
- *Using Claims Data to Examine Mortality Trends Following Hospitalizations for Heart Attack in Medicare*
  - *Ash, Posner, Speckman, Franco, Yacht, Bramwell. Health Services Research 38:5 (Oct 2003)*
- *Additional studies using AMI data*
  - *Missed opportunities, previous MI affecting survival*

32

ADD Health Another Example of Secondary Data

- Identified adolescent health data set
  - "I feel like a kid in a candy store"
- Issues with secondary data
  - Sampling weights
  - Inability to gather more data/refine questions

33

ADD Health - Background

- Goal: Explore causes of health-based behavior
- School-based, nationally representative, probability-based sample of grade 7-12 adolescents
- Wave I: 1994-1995
  - School, home, school admin questionnaires
  - Genetic sample (twins, full sibs, half sibs, etc.)
- Wave II: 1996
- Wave III: 2001-2002
- http://www.cpc.unc.edu/addhealth

34

ADD Health – Sampling

- Cluster sampling within schools
  - Probability of selection proportional to school size
- Samples of ~200 students selected within each school
- Saturation sample from 16 schools – <u>all</u> students selected into sample
- Alternate school samples included as well
- Oversampled certain demographic groups
  - Cubans, Chinese, Disabled, Puerto Ricans, High SES African-Americans

35

ADD Health – Public Use Data

- Public use data includes
  - Half of core sample, chosen at random
  - Half of over-sampled high SES African-Americans
- Data confidentiality issues
  - Only subset of cases available
  - Geocodes are not available
- Must incorporate design into analyses to produce unbiased results

36

ADD Health - Cancer

- Dataset obtained (and converted)
- Determined correct design adjustments
  - Approximation used for public use data
- Resolved data issues
  - 25% of data were missing (intentional skips)
  - Sample size didn't match published articles
    - Public use vs. restricted data
- Performed analysis

37

ADD Health – Cancer Results

- Cancer patients similar to healthy adolescents in terms of
  - Self Esteem
    - Have good qualities, proud, like self, give in to peer pressure
  - Future Orientation/Outlook
    - Marry, live to 35, earn middle class income
- …different in terms of
  - General health
    - Cancer patients felt in poorer general health
  - Affect
    - More often felt depressed or sad

38

ADD Health – Cancer Results II

		Raw Numbers		Design Adjusted	
		Cancer	Healthy	Cancer	Healthy
How Often Bothered By Things	Never/Rarely	44%	57%	49%	57%
	Sometimes	44%	35%	29%	35%
	A Lot of Time	10%	6%	13%	6%
	Most/All of Time	3%	2%	9%	2%

P-value = 0.39 with raw data
P-value = 0.047 with design adjustment

39

ADD Health - Asthma

- Same design adjustment / data issues
- New variables/outcome examined
  - Include changes from Wave I to Wave III
  - Drinking, smoking, etc.
- Results TBA (Dowdell, Posner)

40

Outline

- What is Secondary Data?
- How do I Use Secondary Data?
- Secondary Data Examples
- Causal Inference for Intervention Studies

41

What is an Intervention Study?

- Goal: Estimate the effect of an intervention (treatment) on outcome of interest
  - Reduced cholesterol using Lipitor vs. Placebo
  - Increased rate of early detection of cancer for mammography use (vs. no mammography)
  - Lower hospital readmission rate after being sent to respite unit vs. no respite intervention
- Consider situation without randomization
  - Observational studies
  - Secondary data analysis

42

Randomized Controlled Trials (RCT)

- Gold standard
- Randomization washes out the effect of other covariates (in theory)
- Have good internal validity
- …but lack external validity (generalizability)
  - Ex: Hormone Replacement Therapy, Vioxx
- Expensive and time consuming
- Randomization may "fail", protocols get violated
- Not feasible or ethical in some cases

43

Go Jump Out of a Plane!

"Parachute use to prevent death and major trauma related to gravitational challenge: a systematic review of RCTs"
BMJ, December 2003

Conclusion: "The effectiveness of parachutes has not been subjected to rigorous evaluation using RCTs. [Some researchers] criticize the adoption of interventions evaluated using only observational data. Everyone might benefit if the radical protagonists…organized and participated in a double blind, randomized, placebo-controlled, crossover trial of the parachute."

44

Mammography in Older Women

- Feasibility/Ethical issues with RCT
- Linked Medicare-Tumor registry (SEER)
  - NCI cancer registry
  - CMS's Medicare utilization
  - Added income (median income by zip)

45

Mammography Data

- Outcome - Stage of Diagnosis
  - Early vs. Late (lymph node involvement)
- Exposure: User of Mammography
  - Not explicitly recorded in the data
  - Our def: 2 mammos in the last 2 years vs none
- Other variables gathered
  - Age, Race, # of Primary Care Visits, Income (by zip code), Comorbidities, Region

46

Standard Regression

- Logistic regression (stage is dichotomous)
- Control for all covariates by using them as variables in the model
- Effect of user status on stage of diagnosis determined by coefficient in the analysis (odds ratio)

47

Standard Regression - Results

- OR = 2.97 (95% CI: 2.56, 3.45)
  - Users of mammography have 2.97 times the odds of non-users for detecting at early stage
- Results estimate the average effect for the entire population

48

When Standard Analysis Fails Two Necessary Conditions

- Model misspecification
- Uneven covariate distribution by experimental group
  - Often called "Selection Bias"
  - Example: Income
    - Rich women get mammography
    - Poor women don't get mammography
    - Income and mammography and confounded
  - Stratification may solve this problem

49

Propensity Score Analysis

- Situation
  - Uneven baseline groups -> Potential Bias
- Goal
  - Even out groups in baseline characteristics
  - Facilitates matching on lots of variables
- Rosenbaum & Rubin, 1983

50

Propensity Score Matching II

- Three stages
  - Logistic model to determine the probability of being a user of mammography (propensity score)
  - Select sub-samples of the data (with deciles or quintiles, nearest neighbor, etc.)
  - Standard analysis on reduced data
- The cases included in the propensity score analysis are now comparable across covariates

51

Propensity Score Matching Pre- and Post-Matched Samples

	Pre-Matching		Post-Matching	
	Non-User	User	Non-User	User
Total Sample	2140	2516	1274	1274
Decile 1	416	**57**	57	**57**
Decile 2	339	**89**	89	**89**
Decile 3	359	**136**	136	**136**
Decile 4	239	**205**	205	**205**
Decile 5	**204**	305	**204**	204
Decile 6	**158**	287	**158**	158
Decile 7	**135**	321	**135**	135
Decile 8	**112**	366	**112**	112
Decile 9	**100**	379	**100**	100
Decile 10	**78**	371	**78**	78

52

Propensity Score Matching Age, Pre- and Post-Matching

	Pre-Matching			Post-Matching		
	Non-User	User	p-value	Non-User	User	p-value
Age at Dx						
67-69	37.8%	62.2%		48.9%	51.1%	
70-74	39.1%	60.9%		51.4%	48.6%	
75-79	44.1%	55.9%		49.7%	50.3%	
80-84	51.6%	48.4%		49.1%	50.9%	
85+	76.5%	23.5%	0.001	50.0%	50.0%	0.919

53

Propensity Score Matching - Results

- Results consistent with standard analysis
  - OR = 3.27 (95% CI: 2.72, 3.93) vs. 2.97
- Results apply to population with characteristics similar to matched sample

54

Summary of Causal Inference

- Need to modify standard analytic techniques to make valid inferences in intervention studies

55

Summary

- Secondary data are
  - Available
  - Invaluable
    - Save time, money, make analyses feasible
- Pay attention to
  - Differences between intended and current use
  - Generalizability, bias, and sampling
  - Causal Inference for Intervention Studies

56