Data sources for a proposed course on secondary data analysis

Stephen D. Simon
Part-time consultant
P.Mean Consulting

Part-time faculty
Department of
Biomedical and
Health Informatics

Full-time father

Slides available at:
http://www.pmean.com/13/secondary.html

1

Outline of this talk

- Overview
- Proposed data sources
  - Nationally representative samples
  - Data registries
  - Genetics data
  - Insurance claims data
- Conclusion

2

Definition: Secondary data analysis is analyzing "someone else's data set"

- Re-purposing
  - Data set collected for one reason
  - Analyzed for a different reason
- Using a secondary data set requires compromises
  - It may not have all the variables you want.
  - It may not have all the subjects you want.
  - It may not be in the format you want.

4

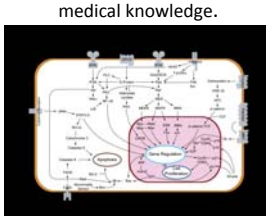
Secondary data analysis has many advantages.

- It is an ideal entry point for a new researcher.
- NIH mandated data sharing plans will increase the availability of secondary data sets.
- Secondary data analysis is cheap.
- Secondary data analysis is fast.
- Secondary data analysis is FUN!

5

Some secondary data set sources WILL NOT work for a course.

- Too expensive
  - A cost that would be fine for research is often unacceptable for a classroom setting.
- Too slow
  - The need for registration and signed agreement to access may not fit in the tight teaching time frame.
- Too complex
  - Often require specialized medical knowledge.



Source: en.m.wikipedia.org/wiki/Akt/PKB_signaling_pathway

6

Some data sources WILL work for a course on secondary data analysis.

- Many data sets are free and easy to obtain.
- Many do not require specialized expertise.
- I will focus on health care issues.
  - My apologies to researchers in education, sociology, etc.
- I want to highlight "one of each type."
- I would love your feedback on these choices and additional data sets to consider.

7

There are four broad classes of secondary data in health care.

1. Nationally representative data sets
2. Data registries
3. Genetics data sets
4. Insurance claims

Feedback: Are you aware of a fifth broad class of data that I am missing?

8

Nationally representative data sets use a carefully crafted sample.

- This allows you to safely extrapolate your results nationwide.
- Nationally representative data sets have some special issues
  - Sample weights
  - Clustering, stratification
  - Oversampling of important demographic subgroups

10

NHAMCS is an excellent teaching example.

http://www.cdc.gov/nchs/ahcd.htm

- "The National Hospital Ambulatory Care Survey (NHAMCS) is designed to collect data on the utilization and provision of ambulatory care services in hospital emergency and outpatient departments and in ambulatory surgery centers. "

11

NHAMCS has many advantages.

- You can download the data without any registration process.
- You can find code for SAS, Stata, and SPSS and detailed guidance on data analysis issues.
- This data set is concrete and easy to visualize.

12

A data registry represents a census or partial census of patients.

- Doctors forward information on anyone with a specific medical condition (e.g., tumor, lead poisoning).
- Data registries have special epidemiological issues:
  - The registry may produce prevalence data or it may produce incidence data.
  - Comparison group may be internal or external.

14

SEER is the granddaddy of all data registries

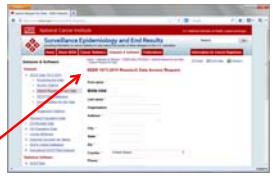
http://seer.cancer.gov

"The Surveillance, Epidemiology, and End Results (SEER) Program of the National Cancer Institute (NCI) is an authoritative source of information on **cancer incidence and survival** in the United States. SEER currently collects and publishes cancer incidence and survival data from population-based cancer registries covering approximately 28 percent of the US population. "

15

Advantages of SEER

- SEER has broad coverage across many years.
- Concrete and easy to visualize.
- BUT…SEER requires registration and signed agreement prior to use.



16

In the broad class of genetics data sets, microarray data sets are ideal for teaching.

- Microarrays measure the expression of genes in human or animal tissue.
- Microarray data sets offer several advantages
  - Fewer privacy concerns
  - Simpler than most other genetics databases
- There are special issues for microarray data sets.
  - You have more variables than observations
  - You need to pay attention to the normalization methods.

18

Son et al article in Genome Research provides an excellent microarray data set.

- This paper provides "a high-density gene expression database of 18,927 unique genes for 158 normal human samples from19 different organs."

http://genome.cshlp.org/content/15/3/443.long

19

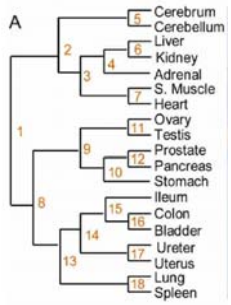
The Son et al data set has several advantages.

- This is one of my favorite data sets of all time.
- The data using in this paper is freely downloadable.
- It does not require highly specialized knowledge to understand and use.

20

The Son et al database has several advantages.

- The Son et al database allows a wide array of hypotheses can be explored/tested
- The 19 organs are easy to visualize: Adrenal, bladder, cerebellum, cerebrum, colon, heart, ileum, kidney, liver, lung, ovary, pancreas, prostate, s. muscle, spleen, stomach, testicle, ureter, uterus.



21

Insurance claims data offer a very rich source of information.

- Very broad coverage and often quite detailed.
- With care, it can provide information on economics of health.
- Special issues
  - Private insurers have no economic incentive to share their data broadly.
  - Insurance claims data have many potential privacy concerns.

23

A good source for insurance claims comes from Medicare.

- "CMS is committed to increasing access to its Medicare claims data through the release of de-identified data files available for public use. These files are available to researchers as free downloads in CSV format. They contain non-identifiable claim-specific information and are within the public domain."

http://www.cms.gov/bsapufs/

24

The Medicare Claims Data has several advantages.

- Free downloads
- Large sample size/broad coverage
- Well defined data dictionary
- SAS code for inputting and labeling variables

25

Conclusion

- There are four data sets across four broad classes that could be used in a course on secondary data analysis.
  - NHAMCS (Nationally representative samples)
  - SEER (Registries)
  - Son et al (Genetics data)
  - Medicare (Insurance claims data)
- These data sets are easy to get, easy to understand, and fun to analyze.
- Slides available at: http://www.pmean.com/13/secondary.html

27