Combining Information from Multiple Data Systems to Enhance Analyses Related to Health: Examples and Lessons Learned*

Nathaniel Schenker
National Center for Health Statistics
Centers for Disease Control and Prevention
(nschenker@cdc.gov)

Presented at the
Symposium of the U.S. Statistical Agencies
Washington, DC
November 13, 2013

*** The findings and conclusions in this presentation are those of the author and do not necessarily represent the views of the National Center for Health Statistics, Centers for Disease Control and Prevention.**

CDC
CENTERS FOR DISEASE CONTROL AND PREVENTION

CONTENTS

1. REASONS FOR COMBINING INFORMATION

2. THREE PROJECTS THAT INVOLVED COMBINING INFORMATION (PLUS A BIT ABOUT SOME OTHERS)

3. LESSONS LEARNED ABOUT COMBINING INFORMATION

4. POTENTIAL APPLICATIONS TO “BIG DATA”?

1. REASONS FOR COMBINING INFORMATION (Schenker and Raghunathan 2007, *Stat Med*)

- **Want more information in the face of limited resources**
  - **Cannot conduct a new study for every new problem of interest**
- **Take advantage of different strengths of different surveys**
- **Use one survey to supply information lacking in another**
- **Lower bias and improve precision**

2. THREE PROJECTS THAT INVOLVED COMBINING INFORMATION (PLUS A BIT ABOUT SOME OTHERS)

A. Combining data from a survey of households and a survey of nursing homes to extend coverage (Schenker *et al*. 2002, *Public Health Rep*)

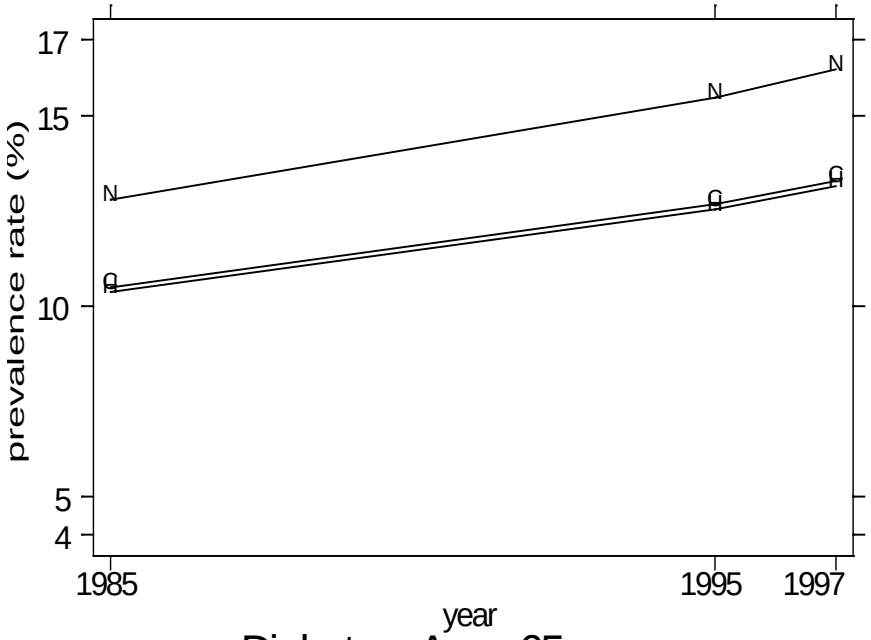
- **Motivation**
  - **More comprehensive estimates of prevalences of chronic conditions for the elderly**
  - **Avoid misleading results due to concentrating on a subset of the population**

- **Surveys used**
  - **National Health Interview Survey (NHIS) for 1985, 1995, 1997 (http://www.cdc.gov/nchs/nhis.htm)**
    - **Principal source of information on the health of the civilian non-institutionalized population of U.S.**
    - **Household members asked about specific conditions**
  - **National Nursing Home Survey (NNHS) for 1985, 1995, 1997 (http://www.cdc.gov/nchs/nnhs.htm)**
    - **Continuing series of nationally representative sample surveys of U.S. nursing homes, their services, their staff, and their residents**
    - **Staff asked to list primary and limited number of other diagnoses for selected residents, with the aid of medical records**

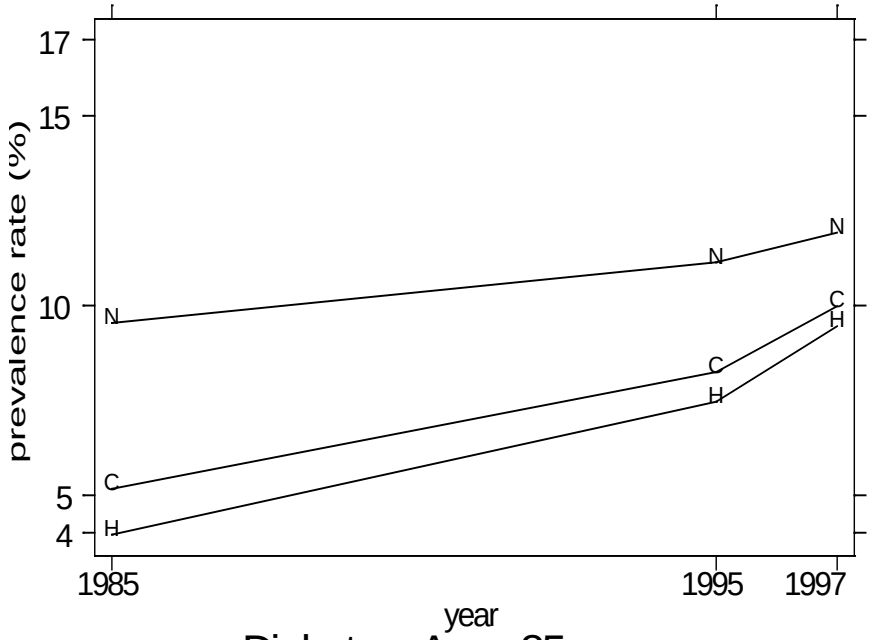
- **Estimated distribution into households and nursing homes during time of study**
  - **Ages 65+: 95% in households, 5% in nursing homes**
  - **Ages 85+: 79% in households, 21% in nursing homes**
- **Calculated combined, design-based prevalence estimates for chronic conditions**
- **Relatively simple problem**
  - **Target populations for the two surveys (nearly) disjoint**

- **Separate and combined estimated prevalence rates for diabetes, by age group, 1985, 1995, and 1997**

(H = households; N = nursing homes; C = combined)



Diabetes, Ages 65+



Diabetes, Ages 85+

B. Using information from an examination-based health survey to improve on analyses of self-reported data in a larger interview-based survey (Schenker *et al*. 2010, *Stat Med*)

- **Motivation**
  - Data on health conditions often from large surveys (e.g., NHIS) using questions such as:

    "Has a doctor or other health professional ever told you that you have <condition of interest>?"

    OR

    "What is your <height/weight>?"
  - Such self-reported data might not accurately reflect prevalences of health conditions

- **Method for improving on analyses of self-reported data**
  - **National Health and Nutrition Examination Survey (NHANES) (http://www.cdc.gov/nchs/nhanes.htm)**
    - ♦ **Program of studies designed to assess health and nutritional status of adults and children in U.S.**
    - ♦ **Asks self-report questions during an interview**
    - ♦ **Also obtains clinical measures for many interviewees based on a physical examination**
  - **Fitted "measurement error" models to NHANES data predicting clinical outcome from self-report answer and covariates**
  - **Used the fitted models to multiply impute clinical outcomes for persons in the NHIS**

- **Comparison of 1999-2002 NHIS Estimated Prevalence Rates for Persons of Ages 20 Years and Above: Self-Reported (SR) Data versus Multiply Imputed Clinical (MICL) Data**

Categories		Hypertension		Diabetes		Obesity	
		SR	MICL	SR	MICL	SR	MICL
Education	< HS Grad.	30.9	39.5	11.1	14.2	25.7	30.1
	HS Grad.	22.9	30.1	6.6	8.8	23.5	28.1
	> HS Grad.	16.5	22.8	4.2	6.5	18.7	23.1
Race/ Ethnicity	Hispanic	14.1	20.8	6.9	9.7	23.2	28.2
	N.H. Black	26.7	35.1	8.8	11.3	29.9	34.8
	N.H. White	20.8	27.6	5.6	7.9	19.8	23.1

Note: Certain records were excluded from the data for this study due to missing covariate values. NHANES sample size = 6,110. NHIS sample size = 105,252.

- **Ratios of Estimated Standard Errors of Estimated Prevalence Rates for Persons of Ages 20 Years and Above Based on Data from Survey Years 1999-2002: (NHANES Clinical) ÷ (NHIS Multiply Imputed Clinical)**

Categories		Hypertension	Diabetes	Obesity
Education	< HS Grad.	2.2	3.2	2.4
	HS Grad.	1.9	2.1	3.5
	> HS Grad.	2.7	1.9	4.1
Race/ Ethnicity	Hispanic	3.3	3.9	2.8
	N.H. Black	2.1	3.3	2.9
	N.H. White	2.5	2.2	3.7

Note: Certain records were excluded from the data for this study due to missing covariate values. NHANES sample size = 6,110. NHIS sample size = 105,252.

C. Combining information from two health surveys to enhance small-area estimation (Raghunathan *et al*. 2007, *J Amer Statist Assoc*; Davis *et al*. 2010, *Public Health Rep*)

- **Project led by National Cancer Institute, with collaboration by:**
  - **National Center for Health Statistics**
  - **National Center for Chronic Disease Prevention and Health Promotion**
  - **University of Michigan**
  - **University of Pennsylvania**
  - **Information Management Services**

- **Motivation**
  - **Interest in local (e.g., county-level, state-level) prevalences of cancer risk factors and screening**
  - **Different surveys have different strengths**
  - **Combining information from surveys could improve small-area estimates**

- **Surveys used**
  - ♦ **Behavioral Risk Factor Surveillance System (BRFSS) (http://www.cdc.gov/brfss/)**
    - **+ Large; almost all counties in sample**
    - **- Telephone survey**
      - **⇒ Non-coverage of non-telephone households; high nonresponse rates**
  - ♦ **NHIS**
    - **+ Face-to-face survey**
      - **⇒ Includes non-telephone households (and a question identifying type of household); higher response rates**
    - **- Smaller; only about 25% of counties in sample**
- **Problem can be viewed as one of coverage error and missing data**

- **Project developed Bayesian methods to combine information from the two surveys; also incorporated telephone coverage rates from the census**

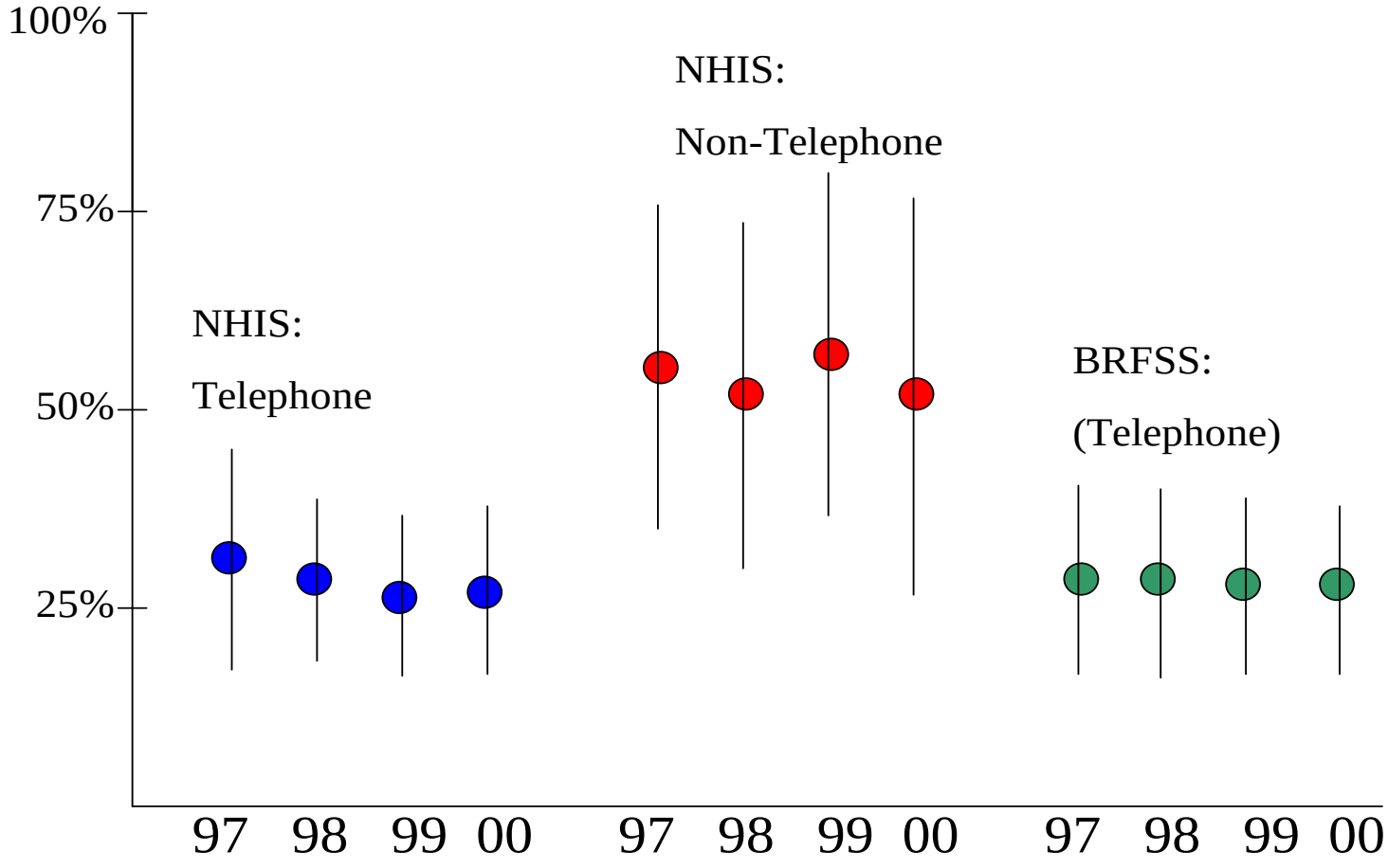
- **National Cancer Institute released estimates for two time periods: 1997-99 and 2000-03 (http://sae.cancer.gov/)**

  - Smoking, mammography, and pap smear

  - Counties, health service areas, and states

- **Current work involves more recent years and including component for cell-phone-only households**

- **Means and standard deviations of county-level direct estimates of current-smoking rates for men, 1997 – 2000**



- **Summaries of Bayesian BRFSS-alone and BRFSS/NHIS county-level estimates of prevalence rates for current smoking among adult males in 2000, by range of telephone non-coverage rates (based on work described in Raghunathan et al. 2007)**

Range of Telephone Non-Coverage Rates (%)	Mean of County-Level Estimates (%)	
	BRFSS-Alone	BRFSS/NHIS
< 2	20.6	20.4
2 – 3	21.1	23.0
3 – 5	21.9	24.3
5 – 8	23.0	25.7
8 – 10	24.1	26.6
10 – 15	24.4	27.7
15 – 20	25.4	29.8
≥ 20	24.1	30.8

D. Other Projects

- **Combining multiple years of survey data (e.g., NCHS 2012, Appendix VI)**
  - **Increases sample size $\Rightarrow$ can increase precision**
  - **Need to check comparability of variables**
  - **Need to be careful about weights, strata, PSUs**

- **NCHS record linkage program (http://www.cdc.gov/nchs/data_access/data_linkage_activities.htm)**
  - **Enables researchers to examine factors that influence disability, chronic disease, health care utilization, morbidity, and mortality**
  - **Data being linked to various NCHS surveys**
    - **Environmental Protection Agency air quality data**
    - **National Death Index death certificate records**
    - **Medicare enrollment and claims data**
    - **Social Security benefit data**
  - **Analytic issues:**
    - **Accounting for sample design**
      - **e.g., Schenker et al. (2011, *Stat Med*)**
    - **Adjusting for cases not linked**

3. LESSONS LEARNED ABOUT COMBINING INFORMATION

- **Technical lessons**
  - Can yield gains, especially when the data systems have complementary strengths
  - Comparability is crucial
  - Use care in dealing with different sample designs
  - Try to find good predictors
  - Might need to deal with small samples, sparse data

- **Administrative lessons**
  - Sharing data and estimates among multiple organizations can require a lot of work
  - Important to educate secondary users on methods used and limitations of results

4. POTENTIAL APPLICATIONS TO "BIG DATA"?

- **Advantages and disadvantages of Big Data**

  + **Big**
  + **Timely**
  + **Predictive (sometimes)**
  + **Cheap (?)**

  - **Unknown population representation**
  - **Issues of data quality**
  - **Typically not very multivariate (at the person level)**
  - **Privacy and confidentiality issues**
  - **Difficult to assess accuracy and uncertainty**

- **Combine Big Data with carefully collected small data to take advantage of strengths of both?**

REFERENCES

Davis, W.W., Parsons, V.L., Xie, D., Schenker, N., Town, M., Raghunathan, T.E., and Feuer, E.J. (2010), "State-Based Estimates of Mammography Screening Rates Based on Information from Two Health Surveys," *Public Health Reports*, 125, 567-578.

National Center for Health Statistics (2013). *2012 National Health Interview Survey (NHIS) Public Use Data Release: NHIS Survey* Description, Division of Health Interview Statistics, National Center for Health Statistics, Hyattsville, MD (ftp://ftp.cdc.gov/pub/Health_Statistics/NCHS/Dataset_Documentation/NHIS/2012/srvydesc.pdf).

Raghunathan, T.E., Xie, D., Schenker, N., Parsons, V.L., Davis, W.W., Dodd, K.W., and Feuer, E.J. (2007), "Combining Information From Two Surveys to Estimate County-Level Prevalence Rates of Cancer Risk Factors and Screening," *Journal of the American Statistical Association*, 102, 474-486.

Schenker, N., Gentleman, J.F., Rose, D., Hing, E., and Shimizu, I.M. (2002), "Combining Estimates from Complementary Surveys: A Case Study Using Prevalence Estimates from National Health Surveys of Households and Nursing Homes," *Public Health Reports*, 117, 393-407.

Schenker, N., and Raghunathan, T.E. (2007), "Combining Information from Multiple Surveys to Enhance Estimation of Measures of Health," *Statistics in Medicine*, 26, 1802-1811.

Schenker, N., Raghunathan, T.E., and Bondarenko, I. (2010), "Improving on Analyses of Self-Reported Data in a Large-Scale Health Survey by Using Information from an Examination-Based Survey," *Statistics in Medicine*, 9, 533-545.

Schenker, N., Parsons, V.L., Lochner, K.A., Wheatcroft, G., and Pamuk, E.R. (2011), "Estimating Standard Errors for Life Expectancies Based on Complex Survey Data with Mortality Follow-Up: A Case Study Using the National Health Interview Survey Linked Mortality Files," *Statistics in Medicine*, 30, 1302-1311.