DISCUSSION OF

Statistical Analysis Using Combined Data Sources
By Ray Chambers

Nathaniel Schenker*
Associate Director for Research and Methodology
National Center for Health Statistics
Centers for Disease Control and Prevention
nschenker@cdc.gov

JPSM Distinguished Lecture
University of Maryland
April 7, 2011

* Thanks to Van Parsons of NCHS for helpful discussions.

CONTENTS

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 data sources**
- **Use one data source to supply information lacking in another**
- **Handle various non-sampling errors; e.g.,**
  - **Coverage error**
  - **Errors due to missing data**
  - **Measurement or response error**
- **Lower sampling error, i.e., improve precision**

2. EXAMPLES OF WORK AT NCHS ON COMBINING INFORMATION, PLACED IN CONTEXT OF RAY'S LECTURE

A. Using information from the National Health and Nutrition Examination Survey (NHANES) to improve on analyses of self-reported data from the larger National Health Interview Survey (NHIS) (Schenker *et al*. 2010, *Stat Med*)

- **Motivation: Some self-reported data on health conditions from large, interview-based surveys might not accurately reflect prevalences of conditions**

- **NHANES asks self-report questions during an interview; obtains clinical measures for many interviewees based on a physical examination**

- **Fitted "measurement error" models to NHANES data predicting clinical outcome from self-reported answer and covariates**

- **Applied the fitted models to the NHIS; used multiple imputation to account for variability**

- **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.

- **In the context of Ray's lecture**
  - **A version of Ray's Example 5 ("Values of 'accurate' zero-one variable *Y* from a small survey *A*. Values of 'rough' zero-one approximation *X* from much larger survey *B*.")**
    - ♦ **But not restricted to a zero-one variable**
  - **Used a "relative" of MIP**
    - ♦ **Multiple-imputation analysis averages the complete-data inference over the predictive distribution of the missing data given the observed data**
  - **Research aimed at producing public-use file with built-in adjustment for self-reporting**
    - $\Rightarrow$ **Imputation problem rather than MLE problem**
  - **Comparability a key issue; e.g., sample designs, contexts**

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

- **Motivation: Interest in local (e.g., county-level) prevalences of cancer risk factors and screening**
- **Surveys used**
  - ♦ **Behavioral Risk Factor Surveillance System (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, which can be identified; higher response rates**
    - **- Smaller; only about 25% of counties in sample**

- **Used Bayesian methods to combine information from the two surveys**
  - **Trivariate Fay-Herriot (1979, J Amer Statist Assoc) type of model**
  - **Approximate posterior distributions obtained via Gibbs sampling**
- **National Cancer Institute released small-area estimates on-line for 1997-9 and 2000-3 (http://sae.cancer.gov/)**
- **Current work involves including component for cell-phone-only households**

- **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**

Range of Telephone Non-Coverage Rates (%)	Mean (Standard Deviation) of County-Level Estimates (%)	
	BRFSS-Alone	BRFSS/NHIS
< 2	20.6 (3.7)	20.4 (4.4)
2 – 3	21.1 (3.8)	23.0 (3.7)
3 – 5	21.9 (4.0)	24.3 (3.9)
5 – 8	23.0 (4.4)	25.7 (3.9)
8 – 10	24.1 (4.7)	26.6 (3.8)
10 – 15	24.4 (4.7)	27.7 (4.1)
15 – 20	25.4 (4.1)	29.8 (3.8)
≥ 20	24.1 (5.0)	30.8 (5.8)

- **In the context of Ray's lecture**
  - **Ray mentioned this as a version of Example 5 ("Values of 'accurate' zero-one variable *Y* from a small survey *A*. Values of 'rough' zero-one approximation *X* from much larger survey *B*.")**
  - **More information available in survey *A* (NHIS) than was available in work of Elliott and Davis (2005, *Applied Stat*)**
    - **⇒ Used "relative" of MIP, i.e., Gibbs sampling**
  - **Use of Bayesian iterative sampling helped to make the problem tractable**
    - **♦ Directly maximizing likelihood would have been difficult**
  - **Again, comparability a key issue; e.g., questions, modes**

C. National Center for Health Statistics 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

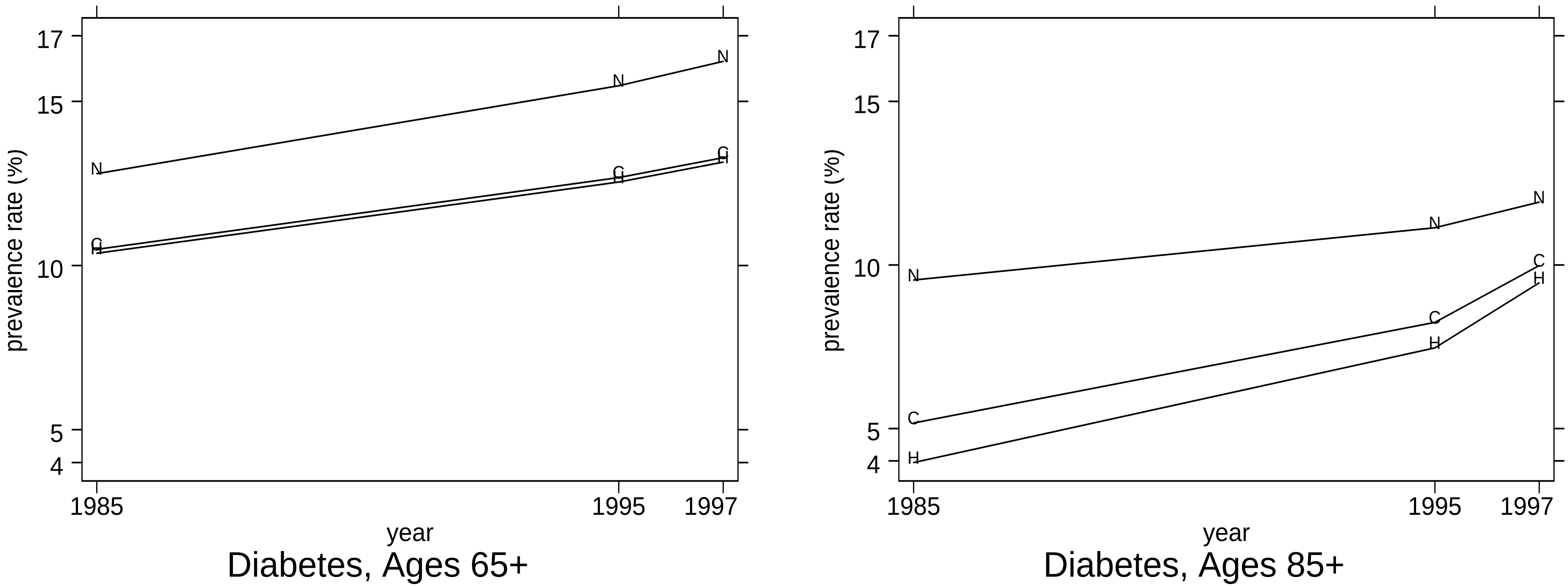
- Data being linked to various NCHS surveys
  - Air quality data from the Environmental Protection Agency
  - Death certificate records from the National Death Index
  - Medicare enrollment and claims data from the Centers for Medicare and Medicaid Services
  - Benefit history data from the Social Security Administration

- **In the context of Ray's lecture**
  - **A version of Ray's Example 3 ("Values of *Y* from register *A*. Values of *X* from register *B*. Registers probabilistically linked. Interest in modeling *Y* – *X* relationship.")**
  - **Data can have complex longitudinal structure**
  - **Records with insufficient information are ineligible for linking**
    - ♦ **Post-stratification weighting adjustments sometimes used; perhaps calibration extensions could be useful**
  - **Often interested in both links and non-links (e.g., as deaths and censored cases, respectively)**
  - **How to estimate probabilities of linkage/non-linkage errors and incorporate into analyses?**
  - **How to incorporate adjustments into public-use data?**

D. Combining information from the NHIS and the National Nursing Home Survey (Schenker *et al*. 2002, *Public Health Rep*)

- **Motivation**
  - **More comprehensive estimates of the prevalences of chronic conditions for the elderly**
  - **Avoid misleading results due to concentrating on a subset of the population**
- **Estimated distribution into households and nursing homes (from data for 1985, 1995, 1997)**
  - **Ages 65+: 95% in households, 5% in nursing homes**
  - **Ages 85+: 79% in households, 21% in nursing homes**

- **Estimated prevalence rates for diabetes, by age group, 1985, 1995, and 1997 (H = households; N = nursing homes; C = combined)**

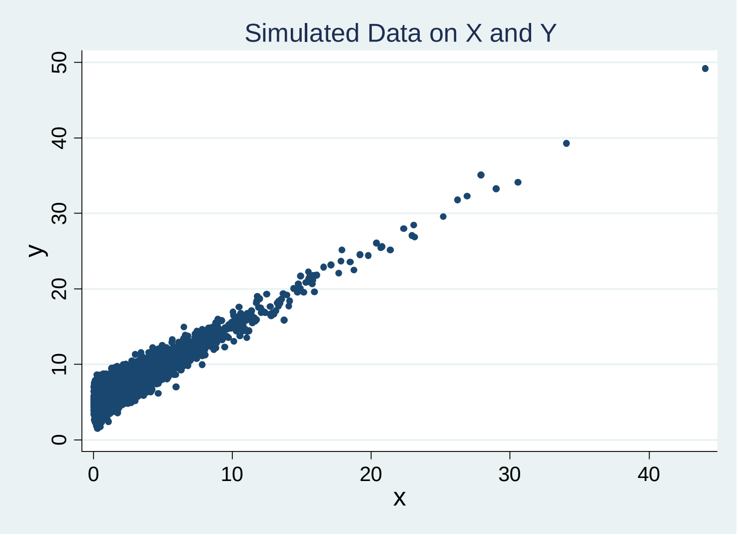


- **In the context of Ray's lecture**
  - **Relatively simple problem**
    - ♦ **Target populations for the two surveys (nearly) disjoint**
    - ♦ **Standard design-based methods applicable (with households and nursing homes treated as separate strata)**
  - **Again, comparability a key issue; e.g., sources of information**

3. RANDOM BUT HOPEFULLY INFORMATIVE COMMENTS ON RAY'S EXAMPLES

- **Calibrating to Out of Date Constraints (Example 1)**
  - **Simulation results intuitively reasonable**
  - **Multivariate extensions desirable**
  - **Often *X* (variable for which population mean is known) is categorical**
  - **How important is it to have a good model?**
  - **Variance estimation difficult, especially with use of outlier robust method**
    - ♦ **Use replication methods?**

- **Combining Survey Data and Marginal Population Information – Comparing MIP with Calibrated Weighting (Example 2)**
  - **Often only have marginal population information on auxiliary variables (*X*) and not on outcome variable (*Y*)**
  - **Why does MIP method achieve substantial gains even under PPY sampling?**
    - ♦ **Perhaps in part because of strong relationship between *X* and *Y* (correlation ≅ 91%)**
      - · **PPY sampling not very informative, given *X*?**



4. CONCLUDING GENERAL POINTS

- **Combining data sources can yield substantial gains, especially when:**
  - **- Data sources have complementary strengths**
  - **- Strong predictors are available**

- **Analyses with combined data sources can often be viewed as a missing-data problem**

  $\Rightarrow$ Models and "relatives" of MIP can be very helpful

  - **♦ Model checking is important**

- **Comparability is crucial**

- **Combining data sources across organizations can require a great deal of care and cooperation**
  - **- Confidentiality concerns**
  - **- Differing policies and priorities across organizations**

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.

Elliott, M.R. and Davis, W.W. (2005), "Obtaining Cancer Risk Factor Prevalence Estimates in Small Areas: Combining Data from Two Surveys, *Applied Statistics*, 54, 595-609.

Fay, R.E., and Herriot, R.A. (1979), "Estimates of Income for Small Places: An Application of James–Stein Procedures to Census Data," *Journal of the American Statistical Association*, 74, 269–277.

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 Estimat2e 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.