Statistical and Computational Challenges in Combining Information from Multiple data Sources

T. E. Raghunathan
University of Michigan

Opportunities

- Computational ability and cheap storage has made digitally stored data an important source for policy research
- Social media, credit card transactions, purchasing, electronic health records, banking data, real estate, etc. are becoming accessible non-survey data sources
- Survey data based on probability samples for policy research is facing challenges
  - Declining response rates
  - Increasing costs
  - Leading to question the need for survey data

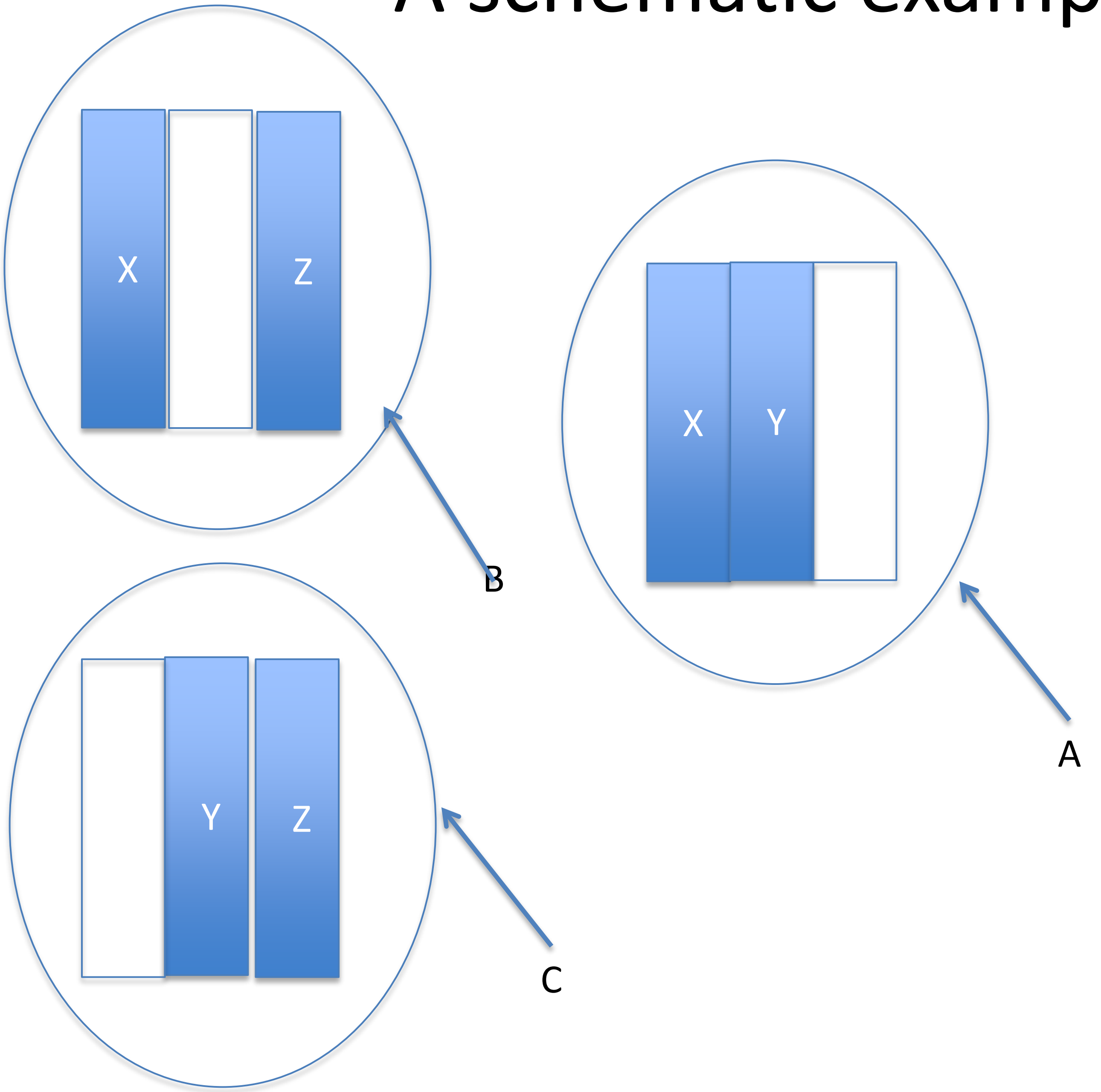
Big Data versus Survey Data

- It is tempting to substitute the “big data” from administrative sources for survey data
- Problem: Inclusion into big data may be related to outcome variable of interest
- Compare the MSE of a proportion from the big data subject to selection bias versus survey data
  - For the population prevalence rate of an attribute: 0.1
  - 20% differential inclusion in the big data with and without the attribute
  - A random sample of size about 290 has a smaller MSE than infinitely large big data
  - For 5% differential rate of inclusion the random sample size needed is around 4,900
- Without knowing the selection issues in the big data abandoning survey data can lead to severely based inferences
- Survey data with nonresponse can be corrected through auxiliary variables, frame variables and post-stratification etc.

It is not Big data vs. Survey Data
But
Big Data +Survey Data (small)

- Survey data carefully crafted can leverage information from Big data
- Survey data can be used to “correct” for selection into big data
- Modeling framework is available to address this issue
- Missing data framework can provide useful methods for combining information from survey and non-survey data sources

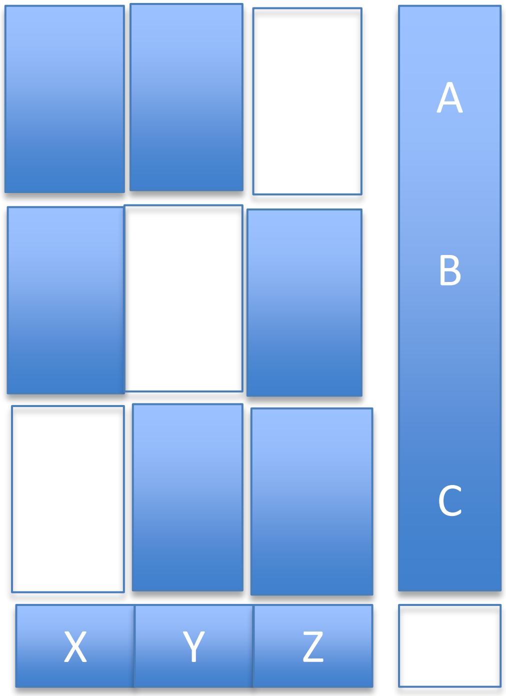
A schematic example



All three data sources are subject to unknown selection bias

Subject to item missing data

Small Survey carefully crafted to be representative



Impute the missing values simultaneously X, Y, Z and match data source

Restrict imputation of Z to observed and imputed A

Restrict imputation of Y to observed and imputed B

Restrict imputation of X to observed and imputed C

This can be implemented, for example, using IVEWARE

- Additional Adjustments
  - Post-stratification of imputed data to population characteristics
- Multiple imputation to incorporate imputation uncertainty
- Extends the purpose beyond each data set and more powerful than using just the survey data alone
- Information from Data sources A, B and C can be used to design the small survey
- Extendable to multiple data sources and complicated missing data patterns
- Any takers?

Another Possible Project

- Use the 1940 Census data to develop a cohort
- Link probabilistically and/or deterministically to later period digitized data
- Take a careful sample and digitize data not yet available to supplement
- Create a “record based” cohort-longitudinal data with missing values
- Multiply impute the missing values

Current Project

- Goal: Investigate relationship between covariates, disease and expenditure
- NIH Funded Study (PI: David Cutler, Harvard University)
- No one data source measuring all these variables
- Piecing together MCBS, NHANES, NHIS, HRS, MEPS, PSID and NCS
- Period 1999- 2012
- So far completed for age 65 and older, 1999-2011
- Current work: 45-64, 18-44 and <18.

Final Thoughts

- Combining information from survey and non-survey data sources provides great opportunities for policy research
- Designing surveys to supplement rich non-survey data can provide the leverage needed to harness information
- Modeling and missing data framework can be used to develop methods needed to construct inferences
- Private-public partnership is needed for this line of research to work
- Secured environment will be needed for sharing data to maintain confidentiality