Big Data: Uses and Limitations

Nathaniel Schenker
Associate Director for Research and Methodology
National Center for Health Statistics
Centers for Disease Control and Prevention

Presentation for discussion at the meeting of the
NCHS Board of Scientific Counselors

September 19, 2013

CDC
Centers for Disease Control and Prevention

CONTENTS

Definitions of Big Data (or lack thereof)

- Wikipedia: "**Big data** is the term for a collection of data sets so large and complex that it becomes difficult to process using on-hand database management tools or traditional data processing applications."
- Horrigan (2013): "I view Big Data as nonsampled data, characterized by the creation of databases from electronic sources whose primary purpose is something other than statistical inference."
- Rodriguez (2012): "For years, statisticians have been working with large volumes of data in fields as diverse as astronomy, bioinformatics, and data mining. Big Data is different because it is generated on a massive scale by countless online interactions among people, transactions between people and systems, and sensor-enabled machinery."

- Arbesman (2013, “Five myths about big data”)
  - Myth 1: “‘Big data’ has a clear definition.”

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

Skills needed with Big Data (Rodriguez 2012)

- Management and processing of distributed data
- New tools for data analysis and visualization
  - E.g., unstructured text data

Current and potential uses of Big Data (not including administrative data) in the Federal Statistical System

- Current
  - Bureau of Labor Statistics (Horrigan 2013)
    - Web scraping to obtain prices for various goods and services
    - Use of retail scanner data in research on distributions of items within expenditure classes

- Potential
  - NCHS:
    - EHRs; pilot tests in National Health Care Surveys (http://www.cdc.gov/nchs/dhcs.htm)
  - Bureau of Labor Statistics (Horrigan 2013)
    - Replacement of traditional data collection from establishments by corporate data from parent company
  - Bureau of Economic Analysis (Lohr 2013)
    - Use of data from Intuit on small businesses for national income accounting
  - Census Bureau (Capps and Wright 2013)
    - Auxiliary data for stratification, improving survey estimates, compensating for nonresponse, small-area estimation, …
    - Helping to check estimates
    - More timely, preliminary estimates (to be revised using survey data)

Robert Groves's COPAFS presentation (COPAFS 2013)

- Two extreme approaches one could take with respect to Big Data
  1. Replace existing measures with big data indicators
     - Tools becoming available, but there are issues of:
       - Quality; e.g., coverage error
       - Inability to examine subgroups due to lack of multivariate nature
  2. Assume that the present system will endure and win out over Big Data
     - Perhaps optimal now, but what happens when big data become so prevalent and are used so widely in business that they cannot be ignored?
- Groves's view: Traditional survey data (although challenged) are not going away, and Big Data are too powerful to ignore
  - Only choice: pursue path of blending Big Data and survey data

Some recent work at NCHS on blending data

- Combining information from complementary surveys (the National Health Interview Survey and the National Nursing Home Survey) to extend coverage (Schenker, Gentleman, Rose, Hing, and Shimizu 2002)
  - Big Data analogue: Combining Big Data with data from a smaller survey to adjust for non-coverage in Big Data
- Combining information from the Behavioral Risk Factor Surveillance System and the National Health Interview Survey via Bayesian modeling for small-area estimation (http://sae.cancer.gov)
  - Big Data analogue: Combining Big Data with data from a smaller survey to adjust for nonresponse and non-coverage in Big Data via modeling

- Bridging the transition from single-race reporting to multiple-race reporting in the census using information from the National Health Interview Survey (http://www.cdc.gov/nchs/nvss/bridged_race.htm)
  - Big data analogue: Adjusting for a change in reporting systems for Big Data using information from a smaller survey with data collected under both reporting systems
- Enhancing the scientific value of surveys by linking their data with administrative and other data (http://www.cdc.gov/nchs/data_access/data_linkage_activities.htm)
  - Big Data analogue: Linking survey data with Big Data
    - Probably more feasible at area level than at person level
- New project, joint with the Census Bureau: Using information from the American Community Survey to create predictors in small-area estimation for outcomes measured in NCHS surveys
  - Big Data analogue: Using local summaries of Big Data as predictors in small-area estimation

Lessons learned from work at NCHS on blending data

- Can yield gains, especially when data systems being blended have complementary strengths
- Comparability is key
- Methods can become “obsolete” quickly
- Need to use care in dealing with different sample designs
- Try to find good predictors
- Sharing information among multiple organizations can require a lot of work (and cost?)
- Important to safeguard privacy and confidentiality
- Important to educate secondary users on methods used and limitations of results

Cukier and Mayer-Schoenberger (2013)

- With Big Data, "three profound changes in how we approach data":
  - "… collect and use a lot of data rather than settle for small amounts or samples …"
  - "… shed our preference for highly curated and pristine data and instead accept messiness: in an increasing number of situations, a bit of inaccuracy can be tolerated, because the benefits of using vastly more data of variable quality outweigh the costs of using smaller amounts of very exact data."
  - "… in many instances, we will need to give up our quest to discover the cause of things, in return for accepting correlations. … "Big data helps answer what, not why, and often that's good enough."

[My view: Need to take the above statements with a big grain of salt. But Big Data could indeed be very useful in combination with survey data, e.g., as predictors for small-area estimation.]

Some Questions for Discussion

1. What Big Data sources could provide information useful to NCHS?
   a. For enhancing the information we provide
   b. For decreasing our costs

2. How could NCHS use the sources identified in Question 1 to improve its work?

3. How could we assess the quality of potential sources of Big Data?

4. In what situations would NCHS be willing to sacrifice accuracy and data quality to obtain much more data?

5. Should we form a Big Data working group?
   a. Within NCHS? Interagency? Elsewhere?

References

Arbesman, S. (2013), "Five Myths about Big Data," Washington Post, August 16, 2013. (http://articles.washingtonpost.com/2013-08-16/opinions/41416362_1_big-data-data-crunching-marketing-analytics)

Capps, C., and Wright, T. (2013), "Toward a Vision: Official Statistics and Big Data," *Amstat News*, August 2013, 9-13.

COPAFS (2013), Minutes of the March 1, 2013 COPAFS quarterly meeting. (http://www.copafs.org/minutes/march2013.aspx)

Cukier, K., and Mayer-Schoenberger, V. (2013), "The Rise of Big Data," *Foreign Affairs*, May/Juen 2013, 28-40.

Horrigan, M.W. (2013), "Big Data: A Perspective from the BLS," *Amstat News*, January 2013, 25-27.

Lohr, S. (2013), "More Data Can Mean Less Guessing About the Economy," New York Times, September 7, 2013.

Rodriguez, R.N. (2012), "Big Data and Better Data," *Amstat News*, June 2012, 3-4.

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.