Workshop on Sensitivity Analysis with Integrated Data hosted by the Federal Committee on Statistical Methodology and the Washington Statistical Society

June 10, 2019

Bureau of Labor Statistics Conference Center
Washington, DC

This presentation is released to inform interested parties of ongoing research and to encourage discussion. The views expressed are those of the presenter and not necessarily those of the U.S. Census Bureau.

United States™ Census Bureau | U.S. Department of Commerce
Economics and Statistics Administration
U.S. CENSUS BUREAU
census.gov

- Federal statistical system of the United States is **decentralized**, with 13 principal statistical agencies
- **FCSM**, with representatives from various agencies, pools expertise and advises OMB and agency heads on methodological and statistical issues
- Given the increasing use of **non-survey and integrated data** to create statistical products, an FCSM working group was established in 2017 to begin work on developing **quality standards** for programs and products involving integrated data
- Reports available at the FCSM website `https://nces.ed.gov/fcsm/`

Taxonomy (dim 1): Sources of Integrated Data

(courtesy of Mike Elliott)

1. Combining data from probability samples (e.g., multiple frames)
2. Combining data from probability and non-probability samples (e.g., generalizing results from a clinical trial to a population)
3. Combining data from probability samples with administrative data*
4. Combining data from non-probability samples with administrative data*

* loosely defines as data collected for non-research purposes, covering a substantial part of the target population

Taxonomy (dim 2): Structure of Variables

1. Key variables are available from multiple sources, but do not always agree
2. Some key variables available from multiple sources, others are not
3. Key variables are from different sources
4. Key variables from different sources are similar but with important differences (requires harmonization)

Methods for Combining Across Sources

(depends on dim1 and dim2, may use more than one)

1. Quasi-randomization (modeling the selection mechanism)
2. Probabilistic matching, record linkage, entitity resolution
3. Statistical matching, superpopulation modeling, imputation (modeling the data-generating mechanism)
4. Latent-variable modeling (harmonization without a gold standard)

Sensitivity Analysis

- Each combining method makes assumptions that may be **difficult to test** or **completely untestable** from observed data
- Assumptions are not always explicit
- Scientific integrity and transparency requires that we **report assumptions** and investigate what happens **if assumptions are violated**
- Biostatisticians and epidemiologists have been developing frameworks for sensitivity analysis in various settings, some with integrated data
  - estimating causal effects
  - dropout in longitudinal studies
  - addressing measurement error
  - generalizing results from experiments to populations

Workshop Schedule

- Session 1: "Understanding Sources of Uncertainty with Integrated Data" (Sharon Lohr, with discussion by John Eltinge)
- Session 2: "Perspectives on Sensitivity Analysis from Health-Outcomes Research" (Juned Siddique, Benjamin Ackerman)
- Audience participation, discussion, brainstorming is strongly encouraged