Quality of Data Processing

Lisa B. Mirel
Office of Analysis and Epidemiology,
National Center for Health Statistics

FedCASIC Conference
April 18, 2018

Disclaimer and Acknowledgements

- The views expressed in this presentation are those of the author and no official endorsement by the Department of Health and Human Services, the Centers for Disease Control and Prevention, or National Center for Health Statistics is intended or should be inferred
- I would like to thank FCSM Interagency working group and, specifically, Joe Schafer and Wendy Martinez for help in developing these slides

Three Workshops

Input data

Process

Output data

Workshop 1: Quality of Input Data

December 1, 2017

Workshop 2: Quality of Data Processing

January 25, 2018

Workshop 3: Quality of Output Data / Synthesis

February 26, 2018

Key questions for data processing: blended data

Input data
Process
Output data

- Fitness for use
  - What are key quality features that need to be considered when deciding to use a data source?
  - What are key quality features that should be identified to understand the strengths and weaknesses of a final product?
- Communication
  - What is the best way to communicate quality features to a diverse audience?

Defining Data Processing

Input data
Process
Output data

- All the steps taken between ingesting inputs from multiple sources
  - surveys, administrative lists, commercially purchased data, scraped data, sensors
- Releasing the final products
  - estimates, analytic reports, actual or synthetic microdata…
- Diverse methodologies; many are new/cutting edge/experimental
- None of us has experience or expertise in all of these areas
  - Understanding how data processing happens and its impacts on quality must be a team effort

Diverse Methods for Data Processing

Record Linkage	Using Multiple Frames	Statistical Matching
Models for Combining Statistics	Dimension Reduction	Harmonization
Editing and Imputation	Adjusting for Representativeness	Estimation
Disclosure Avoidance	Origin / curation of metadata	

Data Processing: Blended Data

- **Record linkage**: exact or probabilistic match, privacy-preserving
- **Statistical matching**: Joining two or more non-overlapping samples by variables shared in common, then applying modeling or imputation techniques to handle missing values
- **Harmonization**: Standardizing information across data sets
- **Disclosure Avoidance**: Techniques for preventing re-identification of de-anonymization of individual records

Record Linkage Speakers: Rebecca Steorts; William Winkler

- Defined as practice of joining multiple data sets by removing duplicate entries, often in the absence of a unique identifier
- Issues:
  - Entity resolution: is the entity the same across data sets?
  - Matching in a quick and automated way
  - Metrics to evaluate quality of the match
- Take Away Messages
  - Need for high quality sets where true matches are known
  - Transparency – statistical agencies showing what they are producing and how they do it
  - Additive error (Winkler) – 5% error in each of two linked data sets and a 5% matching error, resulting data set has 15% error

Statistical Matching Speakers Jerry Reiter; Ed Mulrow

- Defined as blending data sets without unique identifiers; may be used to match data sets without overlapping observations
- Issues:
  - Joint distribution cannot be estimated from data alone
  - Some form of external information is needed
  - Potential for selection bias
- Take Away Messages:
  - Quality measures need to be reported
  - Description of models used and quality for model fit
  - Important to present results of sensitivity analyses, edits performed and steps taken to harmonize variables

Harmonization Speakers: Ben Reist; Don Jang; Scott Holan

- Defined as “the process of mapping and synchronizing data derived from multiple sources into a coherent data file for analysis.” (Jang)
- Issues:
  - Data sources are hard to link
  - Data can vary in who/what they represent
  - No universal data quality measures to evaluate harmonized data
  - Integration and harmonization requires significant resources
- Take Away Messages:
  - Survey estimates can be used to assess the quality of administrative record data and possibly adjust/improve estimates from administrative records (Reist)
  - Harmonization is implemented at the question level – naming, formats, coding and editing rules are standardized across surveys

Disclosure Avoidance Speakers: Latanya Sweeney; John Abowd

- Defined as protecting privacy while preserving data utility
- Issues:
  - How to prevent re-identification of individuals in surveys and administrative records
  - Improvements may be needed based on current approaches (example: governor in MA)
- Take Away Messages:
  - Should report what disclosure methods were used
  - Introduce random noise that is statistically independent of any other distribution (Abowd)

Summary

- Data harmonization is a fundamental first step in blending multiple data sources
- Transparency is necessary for each step
  - Original need to collect data
  - Harmonization steps
  - Matching procedures
  - Models used and assumptions
  - Evaluation techniques used
  - How privacy was maintained
- With transparency reporting users can assess the utility of the data

Potential Future Research

- Communication strategies about linkage methods
- Truth decks for validation
- Linkage consent implications