NATIONAL ACADEMIES Sciences Engineering Medicine

Toward a 21st Century National Data Infrastructure: Enhancing Survey Programs by Using Multiple Data Sources

Using Multiple Data Sources to Enhance Data Equity

Kimberlyn Leary

Harvard Medical School

COMMITTEE ON NATIONAL STATISTICS

Data Equity

- Panel charge included:
  - Implications of new data sources for population subgroup coverage
  - Approaches for linking sources to assess and enhance representativeness
- Data equity involves
  - Ability to produce disaggregated statistics (includes measuring group membership)
  - Accuracy of data for population subgroups
  - Assessment of representation

Executive Order 13985, January 20, 2021

Advancing Racial Equity and Support for Underserved Communities Through the Federal Government

- Charges agencies to understand disparities in programs they administer and identify roadblocks
- "Many Federal datasets are not disaggregated by race, ethnicity, gender, disability, income, veteran status, or other key demographic variables. This lack of data has cascading effects and impedes efforts to measure and advance equity. A first step to promoting equity in Government action is to gather the data necessary to inform that effort" (p. 7011).

Equitable Data Working Group

"Equitable data are those that allow for rigorous assessment of the extent to which government programs and policies yield consistently fair, just, and impartial treatment of all individuals. Equitable data illuminate opportunities for targeted actions that will result in demonstrably improved outcomes for underserved communities."

Measuring Subpopulation Membership

- Which subpopulations should be measured?
  - Depends on purpose and context of analysis; may change over time. For example,
    - May want crime statistics by victim/offender race, ethnicity, gender, sexual orientation, …
    - May want agriculture statistics by size of farm and demographic characteristics of operator/owner
  - Stakeholder and community engagement
- Importance of standards and definitions
  - OMB Statistical Policy Directive 15 on measuring race and ethnicity (currently under revision)
  - Best methods for measuring sexual orientation, gender identity, ability/disability, partnership status, …
  - Shared definitions are particularly important when data sources are integrated

Producing Disaggregated Statistics

- EO 13895: statistics for race and ethnicity subgroups
- Some data sources do not collect race/ethnicity information, e.g.
  - Tax records
  - Satellite data used to measure crop cover
- Or race/ethnicity information may be inaccurate, e.g.
  - Race on death certificates less accurate for Native Americans (Arias et al., 2016, 2021)
- Or sample size too small for desired disaggregation, e.g.
  - Probability surveys such as Current Population Survey

Producing Disaggregated Statistics

- Add race/ethnicity information to data records lacking it
  - Link with records from census or survey that measures race/ethnicity
  - Impute race/ethnicity using statistical model
  - Use area-level information from American Community Survey (e.g. Urban Institute Spatial Data Equity Tool)
- Use models to produce statistics for small population groups
- Combine datasets to obtain better population coverage

CONCLUSION 3-1: Many data sources include or represent only part of the population of interest. Multiple data sources can be used to assess and improve the coverage of underrepresented groups, and to enable the production of disaggregated statistics. It is important to examine the representativeness and coverage of combined data sources to ensure data equity.

Equitable Data Linkage

- Linkage quality depends on quality of identifying information
- Linkage errors and uncertainty
  - Missed link: Two records belong to the same person, but the match is not found
  - False link: Two records are linked but actually belong to different persons
- Numerous studies have found that linkage errors can be more common for Black, Hispanic, Asian Americans

CONCLUSION 3-2: Record linkage can merge information from separate data sources and add variables that are needed to produce disaggregated statistics. But linkage procedures may also introduce biases because linkage errors can disproportionately affect members of some population subgroups. It is important to assess data-equity implications of record-linkage methods.

Promoting Equitable Data Linkage

- Improve quality of identifying information in datasets to be linked
- Study linkage quality for subpopulations, not just for population as a whole
  - Report linkage consent and eligibility rates for subpopulations
  - Provide disaggregated estimates of linkage error rates
  - Analyze differences between linked, unlinked records
- Consider possible unintended consequences of data linkage
  - Could linking data result in harm to a community?
  - Transparency and communication with community members

Data Equity as a Core Value

- The panel views data equity as a core value to be considered when
  - Designing data-collection or data-integration systems
  - Evaluating quality of data products
- The future is equitable data science
- Research needed on equity aspects of data collection and integration

CONCLUSION 3-3: Data equity is an essential aspect of any data system. Documentation of equity aspects, including a discussion of the decisions to include or exclude population subgroup information and an evaluation of data quality for subpopulations of interest, will promote transparency. Development of standards for data equity, and procedures for regularly reviewing equity implications of statistical programs, would enhance efforts to improve data equity across the federal statistical system.