Evaluating the Quality, Usability, and Fitness of Open Health Data for Public Health Research

Erika G. Martin PhD MPH1,2, Jennie Law MPA2, Weijia Ran MPhil3, Natalie Helbig PhD MPA4, Guthrie S. Birkhead MD MPH4,5

1 Nelson A. Rockefeller Institute of Government, 2 Rockefeller College of Public Affairs & Policy, University at Albany, 3 College of Computing & Informatics, University at Albany, 4 New York State Department of Health, 5 School of Public Health, University at Albany-SUNY

Contact: emartin@albany.edu

Supported by a grant from the Robert Wood Johnson Foundation’s Public Health Services & Systems Research Program [Grant ID#71597 to E.G.M. and G.S.B]

OPEN DATA FOR PUBLIC HEALTH SERVICES AND SYSTEMS RESEARCH (PHSSR)

- Open government data is a new source of information for public health research
  - Motivated by President Obama’s Memorandum on Transparency and Open Government (2009)

- Thousands of government datasets released on open data platforms meeting several “openness” criteria
  - Publicly accessible
  - Available in non-proprietary format
  - Free of charge
  - Unlimited use and distribution rights

- Features of open data platforms
  - Search engines locate data objects
  - Capabilities to interact directly with data in the platform
  - Offer incentives and resources for developers
  - Opportunities to submit ideas for new datasets
  - Provide user feedback

RESEARCH QUESTIONS

- To what extent are open health data usable and fit for public health research?
- How could government agencies improve the quality of the data and corresponding metadata, to make these data more usable and fit for public health researchers and practitioners?

RESEARCH DESIGN: SYSTEMATIC REVIEW OF DATA OBJECTS

- Systematic review of open health data offerings on federal, state, and local platforms
- Adapted from Institute of Medicine and Patient Centered Outcomes Research Institute guidelines for systematic review
- Interdisciplinary research team with collective expertise in epidemiology, health services research, informatics, ontology development, digital government, database management, and the production of public health datasets.
- Two reviewers ensured consistency in scoring
- All data offerings examined with a structured coding guide to evaluate
  - Data quality (etrimis, contextual)
  - Consistency with five-star deployment
  - Adherence to Dublin Core metadata standards
  - Descriptive information
  - Platform usability

RESEARCH DESIGN: SAMPLING PROCEDURES

- Final selection (N=183)
  - 44 New York City open data offerings related to health (N=44) https://www.health.data.ny.gov
  - 35 random sample of New York State open data offerings (N=35) http://www.health.ny.gov
  - 35% random sample of federal open data offerings (N=75) https://www.healthdata.gov

- Systematic random sampling of data offerings
- Metadata from platforms scraped into three Excel spreadsheets
- Random number generator used to select objects

PLATFORM USABILITY

- Common Features
- Areas for Improvement
  - Hosting data on platforms, with external links
    (Health Data NY, NYC Open Data)
  - Handbooks to standardize metadata
    (Health Data NY, NYC Open Data)
  - Multiple functions
    - Search and download
    - Post comments and ideas
    - Engage developers and the public with pictures, story boards, social media
  - Help functions: tutorials, help email
  - Visualizations in external pages
    (Health Data NY, NYC Open Data)

FINDINGS: DESCRIPTIVE INFORMATION ABOUT DATA OFFERINGS

- Only 25% of open data offerings are structured datasets
- Most offerings do not contain demographic variables commonly used in public health research

FINDINGS: INTRINSIC AND CONTEXTUAL DATA QUALITY AND ADHERENCE TO DUBLIN CORE METADATA STANDARDS

- Health Data NY scores highest on Intrinsic Data Quality measures
- Health Data NY scores highest on Contextual Data Quality measures
- Health Data NY scores highest on Adherence to Dublin Core Metadata Standards

FINDINGS: GAPS IN MEETING FIVE-STAR OPEN DATA CRITERIA

- Fewer met higher standards of being hyperlinkable to other data (fifth star criteria)
- 35% of offerings met all five criteria

IMPLICATIONS FOR POLICY AND PRACTICE: IMPROVING FITNESS AND USABILITY OF OPEN DATA

- Government agencies have little guidance on how to release open data for different user communities
- All three platforms have areas needing improvement, but Health Data NY scored the highest on all indices
- Sustained effort on improving the usability and quality of open data is necessary for improving their value for public health
- Future work is needed to develop standards of quality and usability
- Additional research on the factors that make some open data sites more successful
- Development of checklists of “best practices” for open data managers

STUDY LIMITATIONS

- New York platforms are not nationally representative
- Coding guides limited to simple fact-based questions
  - Coding guides do not capture:
    - Subjective nature of data quality, which depends on intended use
    - Time constraints
    - Unanticipated finding that most data objects are not tabular datasets

DISCLOSURES: Guthrie Birkhead and Natalie Helbig are employees of the New York State Department of Health, which maintains the Health Data NY open data platform reviewed in this study.

Contact: emartin@albany.edu