SDC

Evaluating the Quality and Usability of Open Data for Public Health Research: A Systematic Review of Data Offerings on 3 Open Data Platforms

Erika G. Martin, PhD, MPH; Jennie Law, MPA; Weijia Ran, MPhil; Natalie Helbig, PhD, MPA; Guthrie S. Birkhead, MD, MPH

ABSTRACT

Context: Government datasets are newly available on open data platforms that are publicly accessible, available in nonproprietary formats, free of charge, and with unlimited use and distribution rights. They provide opportunities for health research, but their quality and usability are unknown.

Objective: To describe available open health data, identify whether data are presented in a way that is aligned with best practices and usable for researchers, and examine differences across platforms.

Design: Two reviewers systematically reviewed a random sample of data offerings on NYC OpenData (New York City, all offerings, n = 37), Health Data NY (New York State, 25% sample, n = 71), and HealthData.gov (US Department of Health and Human Services, 5% sample, n = 75), using a standard coding guide.

Setting: Three open health data platforms at the federal, New York State, and New York City levels.

Main Outcome Measures: Data characteristics from the coding guide were aggregated into summary indices for intrinsic data quality, contextual data quality, adherence to the Dublin Core metadata standards, and the 5-star open data deployment scheme.

Results: One quarter of the offerings were structured datasets; other presentation styles included charts (14.7%), documents describing data (12.0%), maps (10.9%), and query tools (7.7%). Health Data NY had higher intrinsic data quality ($P < .001$), contextual data quality ($P < .001$), and Dublin Core metadata standards adherence ($P < .001$). All met basic "web availability" open data standards; fewer met higher standards of "hyperlinked to other data."

Conclusions: Although all platforms need improvement, they already provide readily available data for health research. Sustained effort on improving open data websites and metadata is necessary for ensuring researchers use these data, thereby increasing their research value.

KEY WORDS: data quality, data reporting, data sources, government, open data

Introduction

Motivated by President Obama's 2009 Memorandum on Transparency and Open Government,[1] government agencies are rapidly developing open data platforms containing thousands of datasets meeting several "openness" criteria: publicly accessibility, availability in nonproprietary formats, free of charge, and with unlimited use and distribution rights. Although public health agencies have historically posted datasets online, the open data movement has

Author Affiliations: *Nelson A. Rockefeller Institute of Government, Albany, New York (Dr Martin); Rockefeller College of Public Affairs & Policy (Dr Martin and Ms Law), College of Computing and Information (Ms Ran), and School of Public Health (Dr Birkhead), University at Albany, Albany, New York; and New York State Department of Health, Albany, New York, NY (Drs Helbig and Birkhead).*

The authors are grateful to Courtney Burke, Patricia Lynch, Theresa Pardo, and Ozlem Uzuner for providing comments on an early draft; Christopher Kotfila for providing JSON technical support to assist with the metadata scrape; and Oscar Alleyne, Erich Bremmer, Sharon Dawes, Janine Jurkowski, Jacqueline Lawler, Kimberly Libman, Rachel Manes, Erin Pascaretti, Giri Tayi, Johnson Qian, and Mike Zdeb for providing feedback on how health data are used, characteristics of data and metadata with high quality and usability, and the conceptual model.

This work was 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.). G.S.B. and N.H. are employees of the New York State Department of Health, which maintains the Health Data NY open data platform reviewed in this study.

Supplemental digital content is available for this article. Direct URL citations appear in the printed text and are provided in the HTML and PDF versions of this article on the journal's Web site (http://www.JPHMP.com).

The authors declare no conflicts of interest.

Correspondence: *Erika G. Martin, PhD, MPH, Rockefeller College of Public Affairs & Policy, University at Albany, 1400 Washington Ave, Milne 300E, Albany, NY 12222 (emartin@albany.edu).*

DOI: 10.1097/PHH.0000000000000388

promoted interest in proactively publishing many types of government data including administrative data that were not always considered valuable. Open data platforms contain capabilities for users to browse datasets and related materials, generate maps and summary statistics, and export results to other sites. Standardized metadata providing information about how the data were collected, coded, etc, and different data formats allow users to readily locate, understand, and use the data. Platforms also offer data in open formats such as comma-separated values or JavaScript Object Notation or EXtensible Markup Language that do not require proprietary software such as SAS or Excel.[2-4]

Open data platforms provide new opportunities for health research and practice. For example, open datasets in Health Data NY, New York State (NYS) platform devoted to health data, were used in the emergency response to Hurricane Irene, to highlight childhood obesity, to advance medical curricula, and to document disparities in health care costs, quality, and outcomes.[5] Although open data platforms cannot contain data with protected identifiable health information, early feedback from health researchers suggests they could improve their capacity to conduct interdisciplinary research and reduce administrative burdens to obtain data.[3] As technical capabilities to link and process large datasets improves, these platforms may become increasingly important resources for health researchers and practitioners.

Despite directives to release data, government agencies have received little guidance on how to release data that are usable for researchers.[6] Although all datasets should report standardized metadata, various users such as software application developers or researchers may have different data and metadata requirements. In the absence of health-specific metadata standards, the Dublin Core metadata element set is a widely-used international standard with specifications to describe 15 desired elements of metadata including contributor (people, organizations, or services contributing to the data resource), description (an abstract, table of contents, geographical representation, or text narrative describing the resource), data rights (a statement about intellectual property or other rights), and subject (often described using keywords, key phrases, or classification codes).[7] Even with a common metadata standard, government agencies may inconsistently apply these definitions. For example, Open NY, the NYS platform serving all state agencies, uses the Dublin Core elements but also adds metadata that may not appear on other government platforms such as a supplemental document explaining the data collection methodology.[8] Operational definitions for Open NY's required metadata elements may also vary. For example, "data provided by," defined as "the agency that provided the data," will likely generate a standard outcome; but the "define any limitations" metadata element, defined as "description of any limitations of the dataset or exclusions," may yield different detail.

Although open data platforms provide new opportunities for researchers and practitioners, the extent to which these sites and their content are currently usable for health research is unclear. A systematic review of a sample of federal, state, and local open data offerings was therefore conducted to describe available content, identify whether data presentations are aligned with best practices and usable for researchers, and examine differences across platforms. The results inform researchers, practitioners, and policymakers about what data and other content are available on these platforms, and inform government agencies on how to improve the quality of open datasets to make them more usable by the public health and health care communities.

Methods

A systematic review was conducted on a random sample of health-related data offerings on 3 open data platforms: HealthData.gov (the federal platform hosted by the US Department of Health and Human Services), Health Data NY (NYS health data platform), and NYC OpenData (New York City [NYC]'s open data platform). These "data offerings" are unique searchable data items that can be discovered through the platform search tools, which can take the form of datasets, charts, maps, application programming interfaces, or other types. The Institute of Medicine's and the Patient-Centered Outcomes Research Institute's procedures for systematic literature reviews were adapted.[9,10] The project team included a state health department-based public health physician/epidemiologist, an open data director, and collective expertise in health services research, epidemiology, informatics, ontology development, digital government, database management, and the production of public health datasets. Academic literature in public health, health informatics, and information science and stakeholder feedback were used to develop a conceptual framework of the important characteristics of data quality and usability, which were used as outcomes for this analysis. These characteristics were defined in a coding guide and subsequently abstracted from a sample of open data offerings and placed in a database for study.

Development of coding instrument

A conceptual framework (Figure 1) of the factors influencing the use of governmental data and

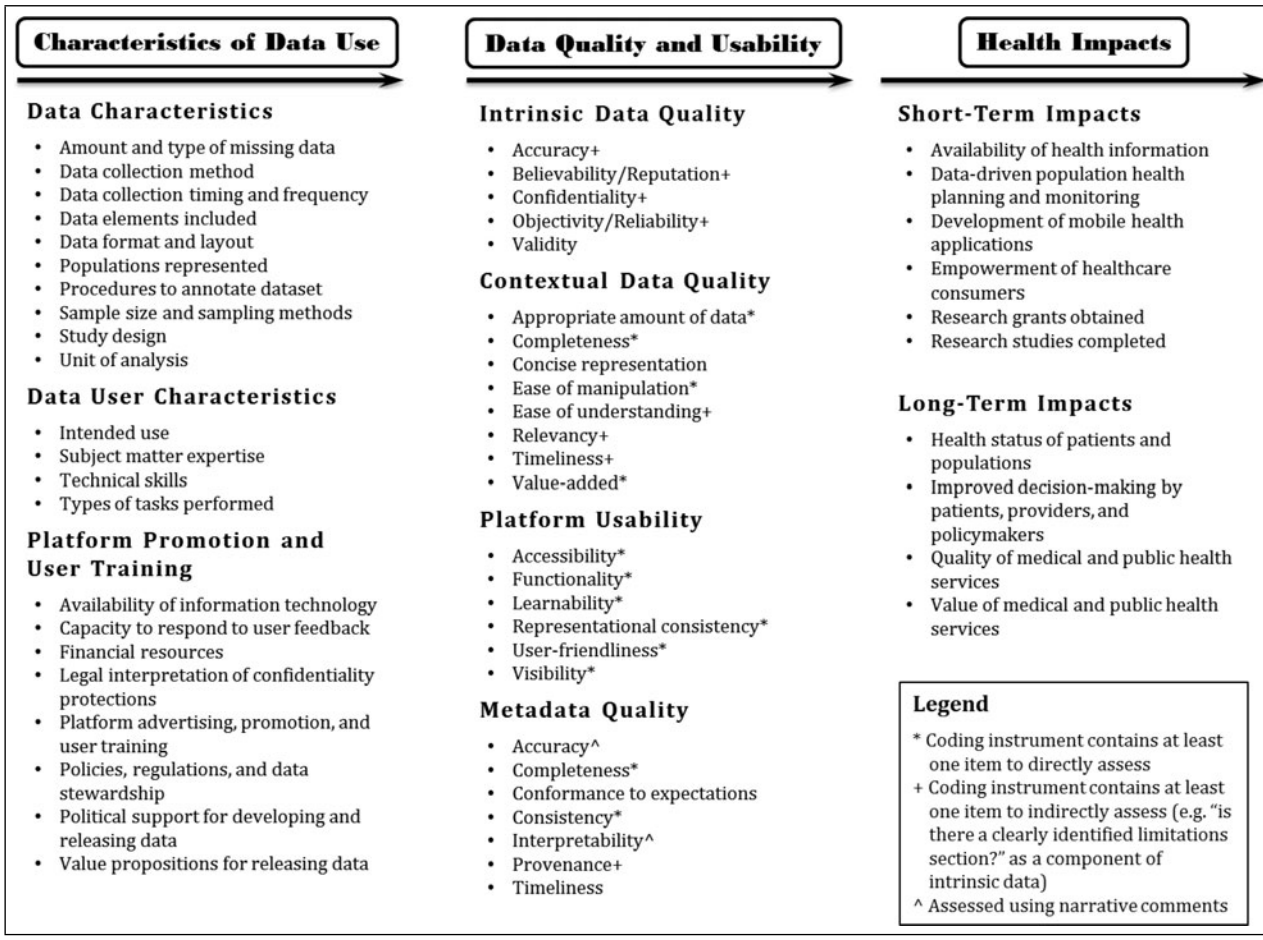

FIGURE 1 Factors That Influence the Use of Governmental Data From Open Data Websites and the Possible Impacts on Subsequent Health Outcomes The extent to which data characteristics, data user characteristics, and platform promotion and user training align influences the amount and types of meaningful use. Intended use, which differs across data users, influences the manner in which quality and usability are defined and their degree of importance. Leveraging data that are high quality, usable, and fit for intended use has the potential to improve short- and long-term outcomes for individuals, populations, and health care systems. The dimensions of intrinsic and contextual data quality, platform usability, and metadata quality were operationalized on the coding instrument.

subsequent health outcomes was developed from the literature review and refined through the following: conversations with 11 key informants with collective expertise in the semantic web, data quality, academic and applied policy research, and local health department practice about how health data is used, characteristics of datasets with high quality and usability, metadata needed to evaluate datasets, and comments on the draft conceptual framework; transcripts from focus groups of public health researchers and practitioners at a recent open health data workshop;[3] a blog post to the NYS department of health's SAS user group; and a review of stakeholder feedback comments on the NYS department of health's Prevention Agenda website.[11] Inputs into the conceptual framework (Figure 1, first column) are characteristics of the data, intended data users, and open data platforms that influence data usability, such as factors related to data production and regulatory restrictions limiting the data available.[8,12-28] These predata use considerations influence the data's quality and usability (Figure 1, the second column) for the intended use. Measurable outputs include intrinsic and contextual data quality, such as accuracy, validity, timeliness, and ease of understanding,[12,17,22,25,29-37] and platform usability, such as accessibility and visibility.[17,22,25,29,33-38] The conceptual framework also includes a dimension for metadata quality, to highlight the importance of this factor in determining data quality and relevance for the intended use.[15,16,30,39-43] Using data that are of high quality, usable, and relevant for intended use will enable the data to be leveraged to influence short- and long-term outcomes for individuals, populations, and health care systems (Figure 1, the third column).[5,13,17,21,23,44]

This framework was operationalized into a 99-item coding guide to document attributes of data offerings related to intrinsic data quality, contextual data

quality, and metadata quality (see the supplement digital content appendix available at http://links.lww.com/JPHMP/A204). For example, items related to "ease of manipulation" include the data offering's primary presentation style, such as dataset, map, or application programming interface, whether visualizations contain statistics, the availability of additional presentation styles, whether the data can be downloaded from the platform, whether the data are available as structured data with rows and columns or in nonproprietary formats, and whether the data offering is viewable in a web browser. The coding guides also included items measuring adherence to Dublin Core metadata standards.[7] Platform usability was assessed with a separate 29-item coding guide with questions corresponding to the conceptual framework (see the supplement digital content appendix available at http://links.lww.com/JPHMP/A204). As data quality is dependent on context and user, the coding guide is from the perspective of public health researchers and practitioners.

The 99-item coding guide also assessed data offerings' adherence to the 5-star open data deployment scheme.[45,46] This concept describes 5 technical standards for open data: online availability (first star), availability as structured data such as an Excel spreadsheet for reuse (second star), availability in nonproprietary formats such as comma-separated values instead of Excel (third star), using uniform resource identifiers such as a web address so users can locate the data online (fourth star), and hyperlinking the data to other relevant data to provide context (fifth star).

Sampling protocol

A random sampling protocol was used to select 5% of HealthData.gov offerings (n = 75) and 25% of Health Data NY offerings (n = 71). All NYC OpenData offerings in the health category were selected (n = 37), for a total of 183 data offerings. During the review, it became apparent that most offerings were not structured datasets, but the original sampling protocol was retained as the study purpose was to describe and evaluate the content of the platforms. As content is continuously updated, the final sampling frame for the NYC and federal sites were determined on June 27, 2014, and for the NYS site on June 30, 2014. The complete data catalogues for Health Data NY and NYC OpenData were available online in Excel spreadsheets and used as sampling frames. The HealthData.gov sampling frame was constructed using the JavaScript Object Notation listing on all data offerings' metadata and converting to Excel, with a random number generator from Kutools for Excel[46] to randomly sample offerings.

Data collection procedures

Two reviewers (J.L. and W.R.) pilot-tested the data collection instrument on 16 data offerings from the 3 platforms with different features, such as administrative versus survey data, format type, and size. They coded all offerings, continuously comparing their responses, discussing discrepancies with E.G.M., and refining coding instrument with extensive decision rules to increase reliability. Pilot-testing ended when there was uniform agreement between reviewers.

After the pilot phase, 1 reviewer was assigned to each data offering, with frequent team meetings to discuss coding. Responses were inputted into an Access database[47] with form views to minimize data entry errors. All datasets and static artifacts, such as codebooks and data access pages, were archived. Platform usability was evaluated after all offerings were coded.

Data analysis

The Access database was exported to Stata[48] for analysis. Summary statistics described data offerings' characteristics. Items from the coding guide were aggregated into indices for (1) intrinsic data quality, (2) contextual data quality, and (3) adherence to the Dublin Core metadata standards (specific outcomes derived from the conceptual framework), and the percentage meeting the 5-star criteria. Indices were constructed using the mean value of all items related to each index. The supplement digital content appendix, available at http://links.lww.com/JPHMP/A204, contains the full list of variables included in each summary measure, their response categories, and how they were recoded for inclusion in the indices. Analysis of variance tests compared differences in means across platforms. Features related to platform usability (another outcome derived from the conceptual framework) were summarized in narrative form, as the small sample size (n = 3 platforms) did not allow for quantitative comparison.

Compliance with protection of human participants

The study analyzed publicly available data offerings, not human subjects, and therefore the study protocol did not need to be reviewed by an institutional review board.

Results

Characteristics of open data offerings

One quarter of the offerings had, as their primary presentation format, structured datasets that could be opened in spreadsheet programs or statistical

software; other primary presentation styles included charts (14.7%), documents about the data (12.0%), maps (10.9%), query tools (7.7%), and application programming interfaces (1.6%). Offerings on Health Data NY were more likely to have different presentation formats (NYC = 29.0%, NYS = 59.2%, federal = 13.5%). Three quarters of the offerings on the NYC (73.7%) and NYS (79.9%) platforms were viewable in a browser, but only 36.5% of federal offerings could be viewed. Common reasons for not being viewable were external files requiring users to open in a separate software program (19.7%) or technical problems with the data access page (6.0%). NYC and NYS commonly had data available for download from the platform or an external data access page (NYC = 84.2%, NYS = 93.0%), but 43.2% of federal offerings were not available for download and among those offerings that were downloadable (n = 42), 45.2% were only accessible through external data access pages. One fifth (21.6%) of the federal data offerings were federated from other platforms, with an original version available elsewhere such as other state open data sites.

Data offerings varied in size and scope. Among those with a structured data presentation style, offerings on the NYS and federal sites were larger with respect to number of rows and columns. Demographic variables were included in less than a quarter of all data offerings, with NYC offerings including these variables with the lowest frequency. Where age groups were included, they were general in broad categories such as 10- or 25-year age groups. Geographic identifiers (NYC = 44.7%, NYS = 63.4%, federal = 37.8%) and variables related to providers or health facilities (NYC = 47.4%, NYS = 50.7%, federal = 32.4%) were included more frequently. Additional descriptive statistics about the data offerings and differences across platforms are in the supplement digital content appendix available at http://links.lww.com/JPHMP/A204.

Data quality, metadata quality, and 5-star open data deployment

The NYS offerings had the highest mean values for the intrinsic data quality index (NYC = 0.22, NYS = 0.67, federal = 0.47, $P < .001$), contextual data quality index (NYC = 0.44, NYS = 0.84, federal = 0.42, $P < .001$), and index of adherence to the Dublin Core metadata standards (NYC = 0.65, NYS = 0.86, federal = 0.55, $P < .001$). Figure 2 displays these results graphically, and the supplement digital content appendix available at http://links.lww.com/JPHMP/A204 contains mean values for the items included in each index.

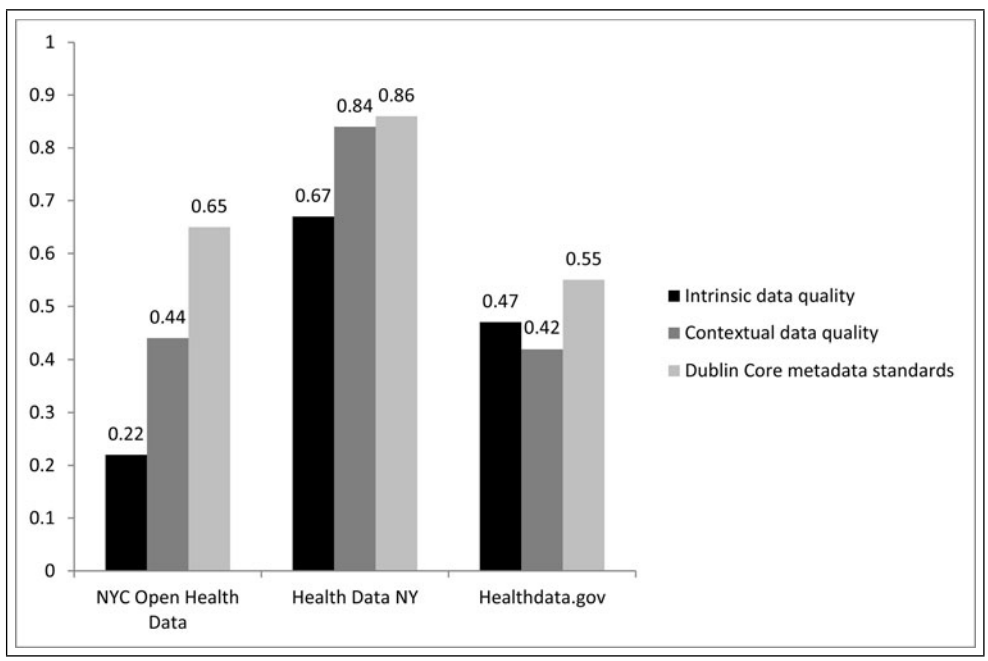

FIGURE 2 Summary Indices of Intrinsic Data Quality, Contextual Data Quality, and Adherence to the Dublin Core Metadata Standards for Health Data Offerings Reviewed From Federal, State, and Local Open Data Platforms

Bars represent the mean value of data offerings' scores on the 3 indices, by platform. Indices were generated using items from the systematic review coding guide, with specific questions listed in the appendix. Mean differences were statistically significant across platforms for the intrinsic data quality index ($F = 180.6$, $P < .001$), contextual data quality index ($F = 281.4$, $P < .001$), and Dublin Core metadata Standards index ($F = 263.8$, $P < .001$).

All offerings met "1-star" criteria (web availability), and most met "2-star" (available as structured data, rather than a Portable Document Format; NYC = 81.6%, NYS = 93.0%, federal = 64.9%, $P < .001$), "3-star" (nonproprietary formats; NYC = 79.0%, NYS = 90.1%, federal = 47.3%, $P < .001$), and "4-star" criteria (uniform resource identifiers; NYC = 100%, NYS = 98.6%, federal = 31.1%, $P < .001$), but fewer met "5-star" criteria (hyperlinked to other data sources for context) (NYC = 0%, NYS = 77.5%, federal = 12.1%, $P < .001$). Overall, only 35.0% of offerings met all 5-star criteria.

Platform usability

Platforms performed well on some dimensions of platform usability from the perspective of public health researchers and practitioners, but there were limitations, described in Table 1. All platforms include search engines to discover and download data offerings in multiple formats, post comments and ideas, develop application programming interfaces, and announce innovation challenges to engage developers. The Health Data NY metadata pages frequently included hyperlinks to other websites or datasets with additional contextual information. The NYS and NYC platforms contained downloadable open data handbooks, with procedures to standardize metadata and vocabulary, capabilities to analyze data in the platform such as generating descriptive statistics, and an ability to embed visualizations into external websites. To improve its visibility and encourage use, the link to the NYS site appears in Google searches for "New York State Department of Health" and a large Health Data NY icon appears at the top of the Department of Health's homepage.

There are several areas where platform usability could be improved. First, the federal website primarily functions as a search engine, with data offerings hosted on external webpages. Compared with the NYC and NYS platforms, it is more difficult to locate these offerings because users are redirected to external sites. Second, there are technical problems limiting functionality, such as frequent broken links on the federal site and problems loading map visualizations on the NYC site. Third, although all platforms provide a help function, the reviewers did not receive responses after sending test questions to each site.

Discussion

Health-related open data offerings on 3 government platforms (NYC, NYS, and US Department of Health and Human Services) were systematically reviewed, with a detailed coding guide to evaluate their quality and usability for public health research. This review showed that these sites publish a wide range of data offerings, although only one quarter are structured datasets typically used in public health research. Data offerings varied in size and scope, with most data offerings not including demographic variables desired for many public health studies or else using broad demographic categories that may limit their usefulness. Although these platforms provide valuable data, there is room for improvement. Most offerings did not contain the complete set of items listed in the Dublin Core metadata element set, and only a minority met the 5-star criteria for open data deployment. Each platform also has areas where its usability could be improved. Despite its limitations, the NYS site scored the highest on all dimensions of data and metadata quality and could serve as a good example of open data deployment. Differences in functionalities between the federal platform and other sites are likely due to evolving technologies since HealthData.gov's initial development. For example, vendors such as Socrata now offer commercial products that government agencies can readily customize for their jurisdictions. The large volume of data assets from 11 federal agencies and inclusion of federated offerings from state and local platforms (where data and metadata quality depends on the source) on HealthData.gov create challenges in achieving consistently high dataset quality. Finally, HealthData.gov may have a deliberate focus on users that typically submit requests for protected data that will never be available as open data.

These findings suggest multiple benefits of open data to researchers. This is a rich new source of information, as these data are publicly available without requiring data requests or data use agreements. Potential use cases include ecological studies, combining data across sectors including education or transportation, and conducting exploratory analyses and pilot studies before requesting more detailed data. However, open data platforms are unlikely to post data that can be linked at the individual level because of regulations designed to maintain confidentiality of personal health information. There will be a continued role for researchers requesting individual-level data from medical records, administrative claims databases, and surveillance systems.

This study provides several contributions to health informatics and public health. First, it offers a new focus on the health research user community. The Sunlight Foundation produces state report cards assigning letter grades on the public availability of state legislative data,[49] but these scores do not capture relevant attributes for researchers. Similarly, the open data movement generally focuses on application developers through innovation challenges like "hackathons." Second, although there is extensive dialogue on aspirational goals such as the 5-star open data deployment scheme, there is less work on how to

TABLE 1
Usability of Federal, New York State, and New York City Open Data Platforms[a]

Criteria	Measures	Findings and Differences Across Platforms
Accessibility	Open data offerings located in the platform rather than on external websites Data offerings can be exported in standard formats Hyperlinks to external websites function	HealthData.gov has frequent broken links to external websites and data downloads; all data offerings are hosted on external webpages, with 4 standard download formats Health Data NY hosts all data offerings on platform, with extensive download format options; frequent links on metadata pages to direct users to additional contextual information NYC OpenData hosts all data offerings on the platform, with extensive download format options; few additional links on metadata or data access pages for contextual information; frequent problems with loading map visualizations
Representational Consistency	Presence of an open data handbook Use of controlled vocabulary or ontology Standard metadata format	HealthData.gov has no handbook; inconsistent number of metadata elements to describe data offerings Health Data NY has a handbook that describes the standardization of metadata elements and domain categories for tags NYC OpenData has a handbook that describes the standardization of metadata and vocabulary for some metadata elements
Functionality	Availability of multiple ways in which users can interact with data on the platform	All platforms have functions to: search for and download data offerings, post comments and ideas, develop application programming interfaces, and announce innovation challenges to engage developers and the public HealthData.gov provides limited interaction with the data; some external sites allow users to run queries to filter data, use interactive maps, and run simulations although some tools are not functional or else require registration Health Data NY and NYC OpenData allow users to interact directly with data in the platform and embed data into other websites such as news articles NYC OpenData allows users to use Twitter to highlight data and provide comments
User-friendliness	Presence of "help" functions Presence of "help" e-mail address Ease of navigation	All platforms provide a help function and e-mail address, although it is unclear who responds to the e-mail HealthData.gov help focused on user-posted questions; Health Data NY help includes structured information about browsers, navigation, and videos of basic functions; NYC OpenData explains how to navigate the platform All platforms are intuitive to navigate after an initial learning curve or completing a tutorial, but have specific limitations: difficult to locate data offerings from HealthData.gov because of redirection to external sites, difficult to differentiate health-specific content on state and city platforms (eg, Health Data NY vs Open NY), NYC OpenData contains multiple data offerings with similar titles and instances where maps do not load correctly
Learnability	Presence of tutorials Presence of frequently asked questions	HealthData.gov's Health Data Initiative starter kit provides an introduction to the site; external sites hosting data offerings may contain offering-specific tutorials; user questions posted for 1-month period (June 2012) Health Data NY provides videos to describe basic functions and additional sources for programmers; no frequently asked questions section NYC OpenData provides tutorials for developers only; no frequently asked questions section
Visibility	Open data site hyperlink visible in a Google search for the agency's name Open data site hyperlink available on the agency's homepage Specific open data offerings hyperlinked from the agency's homepage	HealthData.gov not viewable from a search for "Department of Health and Human Services"; a small link to HHS/Open appears at the bottom of the agency's homepage Health Data NY is a visible link with a search for "New York State Department of Health," and a large-font icon appears on the top of the agency's homepage NYC OpenData not viewable from a search or agency website, although specific open data offerings are available

[a] *Data compiled from reviewers' experiences with the 3 open data platforms under review.*

operationalize these outcomes. This coding guide is a first step toward providing practical tools for open data managers to improve the quality and usability of their health data as they are collected and published as open data. This could improve their value to researchers wanting to use the data for broader purposes such as advancing the evidence base and communicating about public health issues.

The findings inform several suggestions for improvement as open data platforms develop. First, government agencies releasing data should proactively engage multiple stakeholders to promote awareness and encourage data use, improve data release strategies and platform functionality, and cultivate the emerging open data ecosystem. Open data platforms target multiple users, and the federal platform has historically been oriented toward developers. Although both the federal and state sites have been updated because our review to have a new look-and-feel and be more visually appealing, their underlying content and technical functionalities have not changed substantially. Our findings offer areas where open data platforms could be enhanced to meet the needs of public health researchers and practitioners. As government agencies strive to engage additional users that are less technically savvy, stakeholders could be consulted on interface design, desired data, optimal data formats and presentation styles, required metadata, and controlled vocabulary ontologies for keywords describing the datasets and terms in the metadata. When discussing our work with practitioners, they have been particularly interested in ideas for improving their metadata and moving toward 5-star data including hyperlinks to provide context. Finally, sustained leadership and financial resources are necessary to ensure that agencies have adequate capacity to release data and continuously update platform content.

This analysis has several limitations suggesting avenues of future research. The small sample size makes it difficult to conduct extensive comparisons across platforms, and some important datasets may have been omitted from the final sample. Future research could use the coding instrument to document systematic differences across state and local platforms. The coding guide was limited to objective questions such as, "is there a clearly identified limitations section?" as a proxy indicator for accuracy, because of the following: the subjective nature of data and metadata quality, which depend on intended use; the finding that many data offerings are not structured datasets, which required a more flexible coding guide that was applicable to other offering types; and inconsistencies in the presentation of information on the 3 platforms. The coding guide does not capture representational consistency or metadata consistency; future work could systematically assess these outcomes. This review was a broad inventory of the overall quality and usability of offerings from multiple open data sites rather than the quality of individual data elements in specific datasets. Future research should focus on documenting data quality on a dataset-by-dataset basis with techniques such as benchmarking to other datasets, evaluating the amount and type of missing data, or examining the internal consistency of similar data elements. Two other areas for future study are whether usability failures such as broken links are more common for some data categories, or whether common data elements such as geographic identifiers are sufficiently consistent across sources to support linkage studies. Only having one reviewer code each data offering limited the ability to assess interrater reliability, although reliability was increased through extensive pilot-testing and ongoing discussions among reviewers. Finally, open data platforms are dynamic, with content changing hourly. This sample represents a snapshot in time and does not document recent improvements. As these platforms continue to evolve on the basis of user feedback and the availability of improved technical capabilities, it would be valuable to track how sites are adapting.

Open data is a great advance making data widely available to researchers, practitioners and the public, but needs to strive toward high quality and usable data. Standardizing open data offerings and metadata can increase the usefulness of open data, allowing users to more efficiently determine which data will fit their purposes. Standardizing metadata may also decrease the chances that invalid conclusions will be drawn because of misunderstanding of data limitations and better understanding of data collection. Improving the value of open data requires a sustained look at the data, theory, methods, analysis, and visualization of data.

Implications for Policy & Practice

- This review showed that the NYC OpenData, Health Data NY, and HealthData.gov sites publish a wide range of data offerings, although only one quarter are structured datasets typically used in public health research.
- These findings suggest multiple benefits of open data to researchers, as these data are publicly available without requiring data requests or data use agreements.
- Open data is a great advance making data widely available to researchers, practitioners and the public, but needs to strive toward high quality and usable data through standardizing open data offerings and metadata.

References

1. Obama B. *Transparency and Open Government*. Washington, DC: White House; 2009.
2. Sunlight Foundation. *Guidelines for Open Data Policies*. Sunlight Foundation; 2014. http://assets.sunlightfoundation.com/policy/Open%20Data%20Policy%20Guidelines/OpenDataGuidelines_v3.pdf. Accessed April 18, 2014.
3. Martin EG, Helbig N, Birkhead GS. Opening health data: what do researchers want? Early experiences with New York's open health data platform. *J Public Health Manag Pract*. 2015;21(5):E1-E7.
4. New York State Open Data Initiative. Open data handbook. http://ny.github.io/open-data-handbook/index.html?1427817087569. Accessed March 31, 2015.
5. Martin EG, Helbig N, Shah NR. Liberating data to transform healthcare: New York's open data experience. *J Am Med Assoc*. 2014;311(24):2481-2482.
6. Harrison TM, Pardo TA, Cook M. Creating open government ecosystems: a research and development agenda. *Future Internet*. 2012;4(4):900-928.
7. Initiative DCM. Dublin Core Metadata Initiative wiki. 2014; http://wiki.dublincore.org/index.php/Main_Page.
8. New York State Office of Information Technology Services. Open Data Handbook: New York State Open Data Initiative. 2013; http://nys-its.github.io/open-data-handbook/.
9. Institute of Medicine. *Finding What Works in Health Care: Standards for Systematic Reviews*. Washington, DC: National Academies Press; 2011.
10. Patient-Centered Outcomes Research Institute. The PCORI Methodology Report. 2013; http://www.pcori.org/assets/PCORI-Methodology-Standards1.pdf. Accessed September 10, 2014.
11. New York State Department of Health. Prevention agenda 2013-2017. New York State's health improvement plan. https://www.health.ny.gov/prevention/prevention_agenda/2013-2017/. Accessed March 31, 2015.
12. Ballou DP, Tayi GK. Enhancing data quality in data warehouse environments. *Commun ACM*. 1999;42(1):73-78.
13. Bloomrosen M, Detmer D. Advancing the framework: use of health data—a report of a Working Conference of the American Medical Informatics Association. *J Am Med Inform Assoc*. 2008;15(6):715-722.
14. Cook ME, Dawes SS, Helbig NC, Lishnoff RJ. *Use of Parcel Data in New York State: A Reconnaissance Study*. Center for Technology in Government; 2005.
15. Dawes SS, Helbig N. The Value and Limits of Government Information Resources for Policy Informatics. In: Desouza KC, Johnston EW, eds. *Policy Informatics Handbook*. Routledge; 2014.
16. Dawes SS, Pardo TA, Cresswell AM. Designing electronic government information access programs: a holistic approach. *Gov Inform Q*. 2004;21(1):3-23.
17. Friedman DJ, Parrish RG. *Issues in Evaluating Health Department—Web-Based Data Query Systems: Working Papers*. Princeton, NJ: The Robert Wood Johnson Foundation; 2008.
18. Harrison TM, Guerrero S, Burke GB, et al. Open government and e-government: democratic challenges from a public value perspective. *Inform Polity*. 2012;17(2):83-97.
19. Janssen M, Charalabidis Y, Zuiderwijk A. Benefits, adoption barriers and myths of open data and open government. *Inform Syst Manage*. 2012;29(4):258-268.
20. Neely MP, Cook JS. Fifteen years of data and information quality literature: developing a research agenda for accounting. *J Inform Syst*. 2011;25(1):79-108.
21. Safran C, Bloomrosen M, Hammond WE, et al. Toward a national framework for the secondary use of health data: an American Medical Informatics Association White Paper. *J Am Med Inform Assoc*. 2007;14(1):1-9.
22. Strong DM, Lee YW, Wang RY. Data quality in context. *Commun ACM*. 1997;40(5):103-110.
23. Verhulst S, Noveck BS, Caplan R, Brown K, Paz C. *The Open Data Era In Health and Social Care*. National Health Service (NHS England); 2014.
24. Wang RY, Lee YW, Pipino LL, Strong DM. Manage your information as a product. *Sloan Manage Rev*. 1998;39(4):95-105.
25. Wang RW, Strong DM. Beyond accuracy: what data quality means to data consumers. *J Manag Inform Syst*. 1996;12(4):5-34.
26. Wang RY, Storey VC, Firth CP. A framework for analysis of data quality research. *IEEE Trans Knowledge Data Eng*. 1995;7(4):623-640.
27. Woodall P, Borek A, Parlikad AK. Data quality assessment: the hybrid approach. *Inform Manage*. 2013;50:369-382.
28. Department of Health & Human Services. Health Data Initiative: Strategy and Execution Plan. 2013; http://www.healthdata.gov/blog/health-data-initiative-strategy-execution-plan-released-and-ready-feedback#alWwVLmXhtFPKSpq.99. Accessed September 15, 2014.
29. Batini C, Cappiello C, Francalanci C, Maurino A. Methodologies for data quality assessment and improvement. *ACM Comput Surv*. 2009;41(3):16:11-16:52.
30. Hovenga EJ, Grain H. Health data and data governance. *Stud Health Technol Inform*. 2013;193:67-92.
31. Pipino LL, Lee YY, Wang RY. Data quality assessment. *Commun ACM*. 2002;45(4):211-218.
32. Redman TC. The impact of poor data quality on the typical enterprise. *Commun ACM*. 1998;41(2):79-82.
33. Xiao Y, Lu LYY, Liu JS, Zhou Z. Knowledge diffusion path analysis of data quality literature: a main path analysis. *J Informetrics*. 2014;8:594-605.
34. Yeganeh NK, Sadiq S, Sharaf MA. A framework for data quality aware query systems. *Inform Syst*. 2014;46:24-44.
35. Wand Y, Wang RY. Anchoring data quality dimensions in ontological foundations. *Commun ACM*. 1996;39(11):86-95.
36. Caro A, Calero C, Caballero I, Piattini M. A proposal for a set of attributes relevant for Web portal data quality. *Softw Qual J*. 2008;16(4):513-542.
37. Chen CC, Tseng Y-D. Quality evaluation of product reviews using an information quality framework. *Decis Support Syst*. 2011;50(4):755-768.
38. FlowingData. How to Make Government Data Sites Better. 2014; http://flowingdata.com/2014/06/10/how-to-make-government-data-sites-better/. Accessed June, 2014.
39. Bruce TR, Hillmann DI. The continuum of metadata quality: Defining, expressing, exploiting. In: Hillmann DI, Westbrooks EL, eds. *Metadata in Practice*. Chicago, IL: American Library Association; 2004:238-256.
40. Edvardsen LFH, Sølvberg IT, Aalberg T, Trætteberg H. Automatically generating high quality metadata by analyzing the document code of common file types. In: Proceedings of the 9th ACM/IEEE-CS joint conference on Digital libraries; 2009; Austin, TX.
41. Greenberg J. Metadata and the World Wide Web. In: Drake M, ed. *Encyclopedia of Library and Information Science*. Vol 3. 2nd ed. Boca Raton, FL: Taylor & Francis; 2005:1876-1888.
42. Ochoa X, Duval E. Automatic evaluation of metadata quality in digital repositories. *Int J Digit Libr*. 2009;10(2-3):10:67-91.
43. Tani A, Candela L, Castelli D. Dealing with metadata quality: the legacy of digital library efforts. *Inform Process Manage*. 2013;49(6):1194-1205.
44. Manyika J, Chui M, Groves P, Farrell D, Van Kuiken S, Doshi Almasi E. *Open Data: Unlocking Innovation and Performance with Liquid Information*. McKinsey Global Institute; McKinsey Center for Government; McKinsey Business Technology Office; 2013. http://www.mckinsey.com/insights/business_technology/open_data_unlocking_innovation_and_performance_with_liquid_information.
45. Wong KK, Izaguirre DI, Kwan SY, et al. Poor survival with wild-type TP53 ovarian cancer? *Gynecol Oncol*. 2013;130(3):565-569.
46. Detong Technology Ltd Hainan. ExtendOffice professional add-ins and tools for Microsoft Office. http://www.extendoffice.com/product/kutools-for-excel.html. Accessed April 15, 2015.
47. Microsoft. *Access Software*. Redmond, WA: Microsoft; 2014.
48. StataCorp. *Stata Statistical Software: Release 13*. College Station, TX: StataCorp LP; 2013.
49. Sunlight Foundation. Open data legislative report card. http://openstates.org/reportcard/. Accessed April 13, 2015.