Evaluating the Quality, Usability, and Fitness of Open Health Data:

A Systematic Review of Open Data Objects on Federal, State, and Local Platforms

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

Prepared for Thursday Policy Lunch, University at Albany
Albany, NY
September 18, 2014

Acknowledgements

- Funding from the Robert Wood Johnson Foundation's Public Health Services & Systems Research Program (grant ID #71597 to Martin and Birkhead)

- Early feedback from: Courtney Burke, Patricia Lynch, Theresa Pardo, Ozlem Uzuner

- JSON technical support from: Chris Kotfila

Agenda

- Open data background
- Research questions
- Research methods
  - Overview
  - Sampling
  - Coding instrument
- Preliminary findings
- Next steps

Open data background

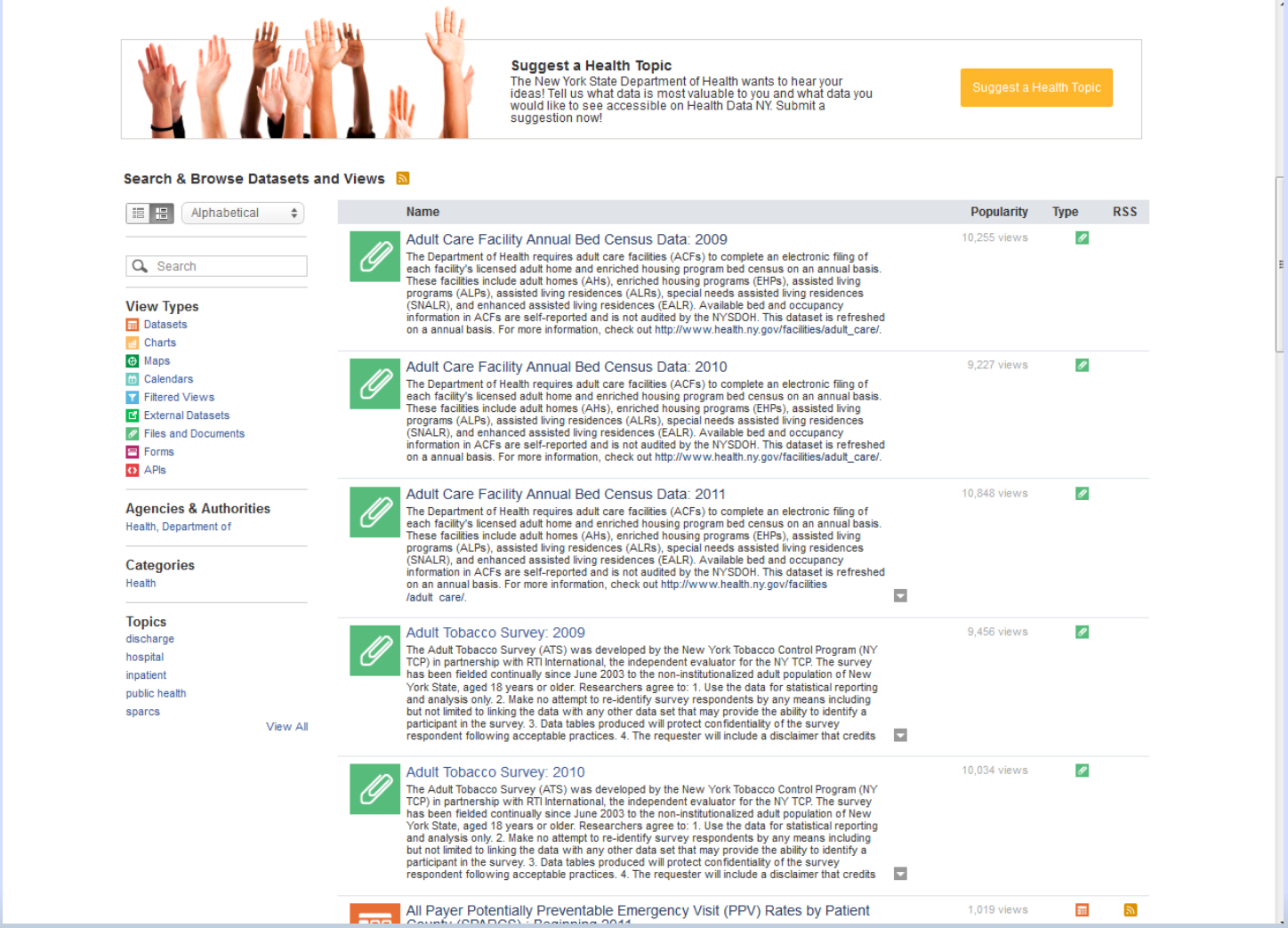
- New source of information for public health research
  - Martin, Helbig, Birkhead *J Public Health Manag Pract* 2014
- Motivated by government transparency movement, including President Obama's memorandum on open government
- Thousands of government datasets released on open data platforms at federal, state, and local levels meeting several "openness" criteria
  - Publicly accessible, available in non-proprietary formats, free of charge, unlimited use and distribution rights
- New opportunities for public health research and practice
  - New York State examples in Martin, Helbig, Shah *JAMA* 2014

Health Data NY
Open.ny.gov
Health Data NY
Health.ny.gov
Developers
Help
Check out Prevention Agenda 2013-2017
The Prevention Agenda 2013-17 is the blueprint for state and local action to improve the health of New Yorkers in five priority areas and to reduce differences among racial, ethnic, disability, socioeconomic and other groups with health disparities.
Recently Added
Featured Datasets
Most Viewed
View Full Data Catalog
Hospital Inpatient Prevention Quality Indicators for Pediatric Discharges by Patient County
Hospital Inpatient Prevention Quality Indicators for Pediatric Discharges by Patient Zip Code
All Payer Potentially Preventable Emergency Visit Rates by County
All Payer Potentially Preventable Emergency Visit Rates by Zip Code
prevention
Suggest a Health Topic

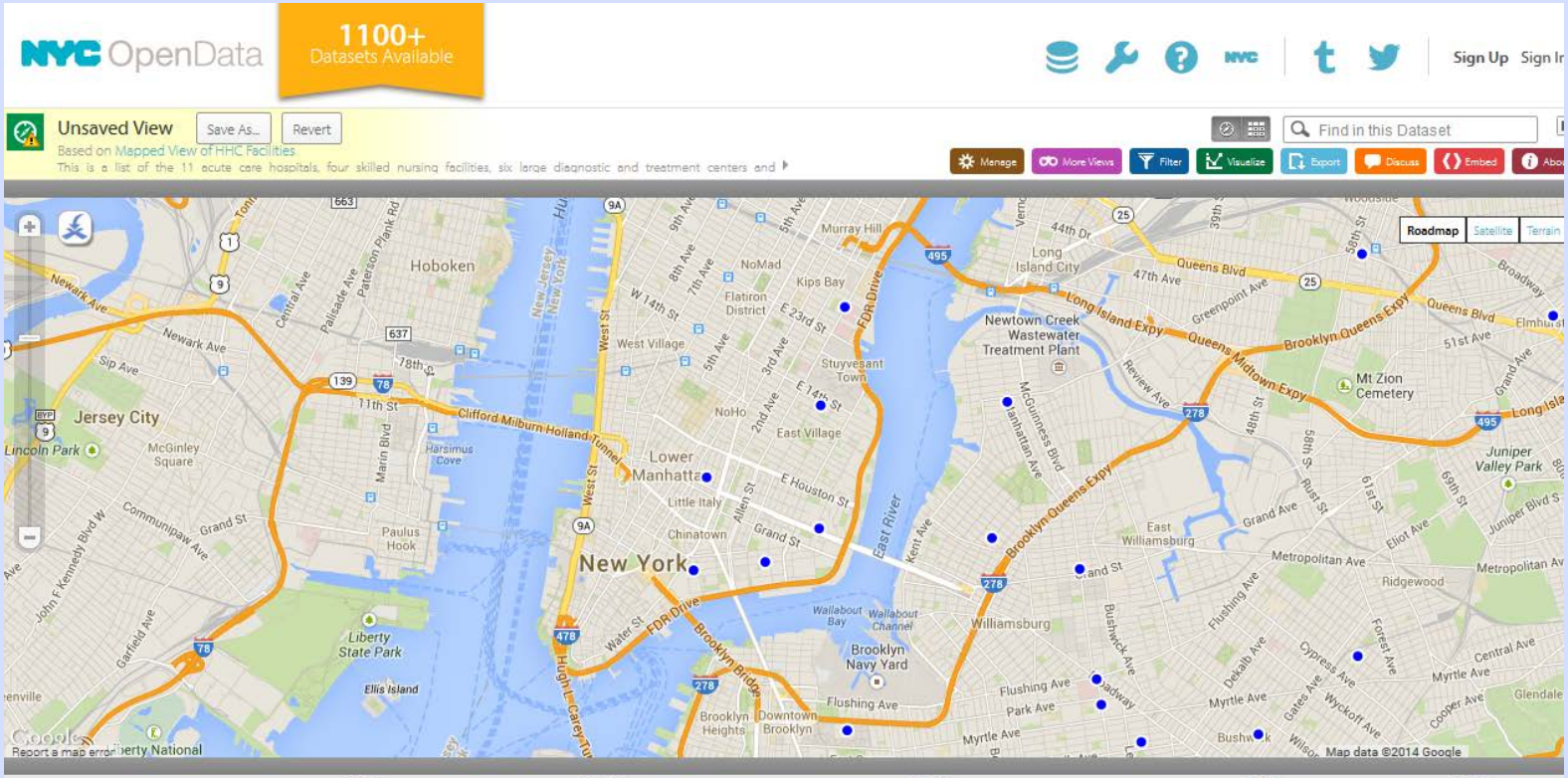
HealthData.gov
Home
Data
Blog
Q & A
Ideas
Developers
Only 1 week left to apply to HHS Entrepreneurs!
Last week to apply for HHS Entrepreneurs!
Six all new projects and we're looking for the best talent to come into government and solve critical problems in health care and government. Apply today! Read more »
HDI Starter Kit - Learn about all of the HHS data available to you.
Get the Kit!
Found a great health-related dataset on another site? Tell us about it!
Suggest a Dataset
Search the Data
Search for
Sub-Agency
Subject Area
Search
Recent Datasets
NNDSS - Table II. Babesiosis to...
NNDSS - Table I. Infrequently reported...
Medicare Hospital Spending Per Patient -...
Hospital General Information
Timely and Effective Care - Hospital
View more »
Recent Blog Entries
Open Data for Transparent and Effective...
HHS Open Government Plan 3.0 is Now Posted...
NYS Health Challenge Needs Your Ideas to...
Last week to apply for HHS Entrepreneurs!
Using Data to Advance Health Equity for Men...
View more »
Medicare
Medicaid
Epidemiology
Treatments
Population Statistics

NYC OpenData
1100+ Datasets Available
Sign Up
Sign In
Featured Datasets for NYC BigApps 2014!
NYC BigApps is a competition that empowers the sharpest minds to solve New York City's toughest challenges through technology, data, and collaboration.
View
More Stories
Search
Click here for the official list of NYC datasets
Business
City Government
Education
Environment
Health
Housing & Development
Public Safety
Recreation

Search engine to locate data objects

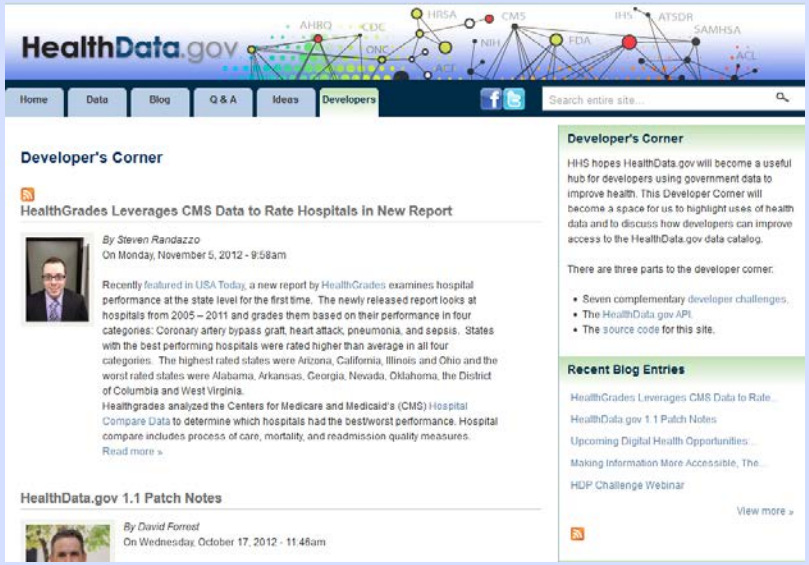


Capabilities to interact directly with data in the platform



	Facility Type	Borough	Facility Name	Cross Streets	Phone	Location
9	Child Health Center	Manhattan	Baruch Houses Family Health Center	corner of Columbia St.	212-673-5990	280 Delanc
10	Child Health Center	Manhattan	Judson Health Center		212-925-5000	34 Spring S
11	Child Health Center	Manhattan	Smith Communicare Health Center	corner of Catherine St.	212-346-0500	60 Madisor
12	Child Health Center	Manhattan	Roberto Clemente Health Center		212-387-7400	540 13th St
13	Child Health Center	Queens	Elmhurst Hospital Center		718-334-4000	79 01
14	Child Health Center	Queens	Ridgewood Communicare Clinic	between Woodbine St. & Madison St.	718-334-6190	769 Onder
15	Child Health Center	Queens	Woodside Houses Child Health Clinic	between Northern Blvd. & 50th St.	718-334-6140	50 53 Newt

Challenges and resources for developers



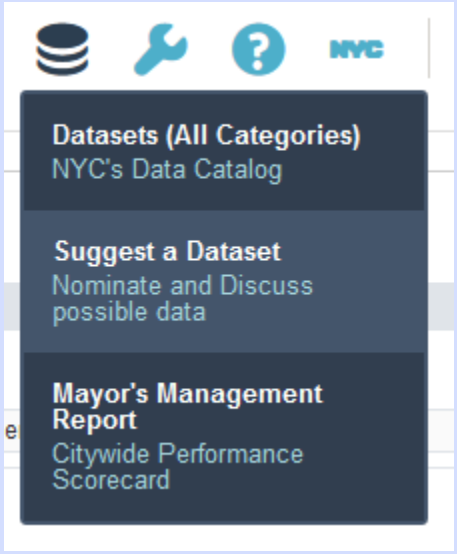
Opportunities to submit ideas for new dataset, and user feedback

payers by the patient's county.

Suggest a Health Topic

The New York State Department of Health wants to hear your ideas! Tell us what data is most valuable to you and what data you would like to see accessible on Health Data NY. Submit a suggestion now!

Suggest a Health Topic



Research questions

- Open data are promising but…
- 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 overview

- Systematic review of open health data objects on federal, state, and local platforms
  - Adapted from Institute of Medicine and Patient-Centered Outcomes Research Institute guidelines for systematic literature reviews
- Health-related data objects randomly sampled from three platforms
  - Healthdata.gov (federal)
  - Health Data NY (state)
  - NYC Open Data (city)
- All data objects examined using a coding guide to evaluate:
  - Data quality (intrinsic, contextual)
  - Metadata quality
  - Platform usability

Sampling design

- Final selection
  - All NYC Open Data objects related to health (N=37)
  - 25% simple random sample of Health Data NY data objects (N=71, of 308 available)
  - 5% simple random sample of Healthdata.gov data objects (N=75, of 1,526 available)
  - Total of 183 data objects

- Sampling methods
  - Scraped metadata from three platforms into three Excel spreadsheets
  - Used Excel-based random number generator to assign random integer values from 1 to N, then selected every dataset assigned a 1

Development of coding guide

- Cross-disciplinary literature review to develop a preliminary conceptual framework of data quality, usability, and fitness

- Stakeholder conversations to refine conceptual framework
  - Respondents: experts in computer science/semantic web (1) and data quality (2); academic health researchers (3); local epidemiologists (3); analysts at health policy and advocacy center (2)
  - Topics covered: how health data are used; which health datasets are useful; how respondents decide whether a dataset is of high quality, usable, and fit; metadata needed to evaluate datasets; comments on conceptual framework
  - Internal vetting with interdisciplinary research team

Development of coding guide, cont.

- Additional stakeholder input on the quality, usability, and fitness of data for health research obtained from:
  - Focus groups of public health researchers and practitioners, conducted at November 2013 open data workshop in Albany, NY (Martin, Helbig, Birkhead *J Public Health Manag Pract* 2014)
  - Blog post to NYSDOH SAS user group to solicit comments
  - Review of stakeholder feedback comments on the Prevention Agenda dashboard
  - Review of a sample of data-based County Health Assessments
  - Grant reviewers' feedback
- Extensive pilot-testing and refinement

Characteristics of Data Use

Data Characteristics

- Populations represented
- Sample size and sampling methods
- Unit of analysis
- Data elements included
- Data collection method
- Study design
- Data collection timing and frequency
- Data format and layout
- Amount and type of missing data
- Procedures to annotate dataset

Data User Characteristics

- Subject matter expertise
- Technical skills
- Types of tasks performed
- Intended use

Platform Promotion and User Training

- Policies, regulations, and data stewardship
- Legal interpretation of confidentiality protections
- Political support for developing and releasing data
- Capacity to respond to user feedback
- Financial resources
- Value propositions for releasing data
- Availability of information technology
- Platform advertising, promotion, and user training

Data Quality and Usability

Intrinsic Data Quality

- Accuracy+
- Believability/Reputation+
- Objectivity/Reliability+
- Confidentiality+
- Validity

Contextual Data Quality

- Relevancy+
- Value-added*
- Timeliness+
- Completeness*
- Appropriate amount of data*
- Ease of understanding+
- Ease of manipulation*
- Concise representation

Platform Usability

- Accessibility*
- Representational consistency*
- Functionality*
- User-friendliness*
- Learnability*
- Visibility*

Metadata Quality

- Completeness*
- Interpretability^
- Accuracy^
- Provenance+
- Consistency*
- Timeliness
- Conformance to expectations

Health Impacts

Short-Term Impacts

- Research studies completed
- Research grants obtained
- Development of mobile health applications
- Data-driven population health planning and monitoring
- Availability of health information
- Empowerment of healthcare consumers

Long-Term Impacts

- Quality of medical and public health services
- Value of medical and public health services
- Health status of patients and populations
- Improved decisionmaking by patients, providers, and policymakers

Legend

* Coding instrument contains at least one item to directly assess
+ Coding instrument contains at least one item to indirectly assess (e.g. "is there a clearly identified limitations section?" as a component of intrinsic data)
^ Assessed using narrative comments

Example of coding guide questions

- Contextual data quality – ease of manipulation
  - What is the data object's primary presentation format (table, chart, map, external file, API, filter, other)?
  - If primary format is a visualization, are simple statistics available?
  - Are there different presentation formats for the data object (if so, list available formats)?
  - Can the data be downloaded from the platform (if so, what download options are available)?
  - Can the data be downloaded from the data access page (if so, what download options are available)?
  - Are the data available as structured data?
  - Are the data available in non-proprietary formats?
  - Is the selection a data artifact?
  - Is the data object viewable in a browser (if no, why not)?

Example of coding guide questions, cont.

- Intrinsic data quality – accuracy/objectivity/reliability
  - Is a limitations section clearly and explicitly identified?*
  - Is there a codebook or data dictionary?
  - Is any information about the purpose of the data collection listed?*
  - Is there a description of the sample design?*
  - Is there a description of how the data were collected?*
  - Is the data collection instrument available?*
  - Is there any notation about random checks for data accuracy, auditing procedures, validity checks, etc.?*
  - Is there any notation about the data preparation/processing steps that happened as the data were transformed into open data?*

** if yes, coders copy and paste relevant text*

Example of coding guide questions, cont.

- Contextual data quality – relevancy/value-added
  - Is there a data object description?*
  - Is the granularity clearly and specifically identified?*
  - Is the unit of analysis clearly and specifically identified?*
  - Is the data object available via a URI on the metadata page?*
  - Are there examples of how data have been used in research/practice?*
  - Does the platform list any ideas for how data could be used?*
  - Is there mention of other data objects that would be of interest?*
  - Are the data available in RDF format?
  - Do variable names hyperlink to contextual information?
  - Series of questions on presence of demographic, provider, and health facility variables, and their response categories
    - Demographics: age, gender, race/ethnicity, insurance status, income, education

** if yes, coders copy and paste relevant text*

Additional coding guide considerations

- Includes questions to address adherence to international Dublin Core Metadata Standards

- Static documents archived on hard drive
  - Codebooks, data dictionaries, dataset downloads, other available materials online
  - Metadata and data access pages saved as complete webpages

- Questions very specific and direct, to improve inter-rater reliability

Data collection procedures

- Extensive pilot-testing of coding guide
  - Purposive selection of 16 data objects from the three platforms which varied widely (e.g. administrative vs survey, simple tabular format vs large SAS-file download, small vs large size)
  - J.L. and W.R. double-coded and compared responses, discussing discrepancies with E.M.
  - Interim feedback from N.H. and G.B.
  - Coding guide continuously updated until uniform agreement
- Coding guide transformed into Access database for data entry
  - Form view and fixed response categories to minimize data entry errors
  - Flags for queries to discuss with the team
- Separate coding guide for platform usability
  - Assessed after object-by-object coding

Limitations

- Smaller N than anticipated
- Limited to fact-based questions (e.g. “is there a clearly identified limitations section?”)
  - Subjective nature of data quality, which depends on intended use
  - Time constraints – limited to a cursory examination of each object
  - Unanticipated finding that many data objects are not tabular datasets
  - (Somewhat anticipated) finding that the three platforms present information in inconsistent formats and locations
- Coding guide does not capture:
  - Representational consistency (platform usability)
  - Metadata consistency (metadata quality)

Characteristics of data objects

Characteristic	NYC Open Data (city, N=38)[1]	Health Data NY (state, N=71)	Healthdata.gov (federal, N=74)
Primary presentation format, % (N)			
Table			16.2 (12)
Chart			--
Map			1.4 (1)
External file	26.3 (10)	12.7 (9)	36.5 (27)
API	--	2.8 (2)	--
Filter	10.5 (4)	2.8 (2)	10.8 (8)
Other	--	--	1.4 (1)
Data artifact	7.9 (3)	--	1.4 (1)
Not applicable[2]	10.5 (4)	5.6 (4)	32.4 (24)
Availability of additional presentation formats, % (N)	29 (11)	59.2 (42)	13.5 (10)
Availability of data related to visualizations[3], % (N)			
Raw data available	13.1 (5)	47.9 (34)	1.4 (1)
Simple statistics available	--	70.7 (29)	--

21.6% of federal objects federated from a different platform

Availability of data objects

Characteristic	NYC Open Data (city, N=38)[1]	Health Data NY (state, N=71)	Healthdata.gov (federal, N=74)
Ability to view data object in browser, % (N)			
Object is viewable in a browser	73.7 (28)	78.9 (56)	36.5 (27)
Problem with the data access page	13.2 (5)	1.4 (1)	6.8 (5)
Data object is an external file	5.3 (2)	18.3 (13)	28.4 (21)
Data object requires subscription or registration	2.6 (1)	--	8.1 (6)
Data object is only viewable in a proprietary format	2.6 (1)	--	--
Not applicable	2.6 (1)	1.4 (1)	20.3 (15)
Ability to download data, % (N)			
Available via platform	--	14.1 (10)	--
Available via data access page	--	--	25.7 (19)
Available from both sources	84.2 (32)	78.9 (56)	31.1 (23)
Not available for download	15.8 (6)	7 (5)	43.2 (32)

Accessibility of data objects relevant for health research

Characteristic	NYC Open Data (city, N=38)1	Health Data NY (state, N=71)	Healthdata.gov (federal, N=74)
Inclusion of demographic variables, % (N)			
Age	5.3 (2)	29.6 (21)	24.3 (18)
Gender	5.3 (2)	18.3 (13)	18.9 (14)
Race/ethnicity	5.3 (2)	11.3 (8)	13.5 (10)
Insurance status	5.3 (2)	28.1 (20)	24.3 (18)
Education	5.3 (2)	14 (10)	2.7 (2)
Income	18.4 (7)	7 (5)	10.8 (8)
Geographic identifier	44.7 (17)	63.4 (45)	37.8 (28)
Provider and/or health facilities	47.4 (18)	50.7 (36)	32.4 (24)

Platform usability

- Common features
  - Hosting data on platforms, with links to external pages where relevant (Health Data NY, NYC Open Data)
  - Open data handbooks to guide standardization of metadata (Health Data NY, NYC Open Data)
  - Multiple functions to search for and download data objects, post comments and ideas, develop APIs, and announce innovation challenges to engage developers and the public
  - Help functions such as tutorials, help email address
  - Designed to engage the public, with pictures, story boards, ways for users to provide comments

Platform usability

- Interesting features
  - Ability to embed visualizations into external webpages (Health Data NY, NYC Open Data)
  - Allowing users to Tweet posts highlighting data (NYC Open Data)
  - Health Data NY has a visible presence in Google (appears in general search for “New York State Department of Health,” and on the NYSDOH webpage)
  - Rotating story board with public health messaging, and list of recently added/featured/most viewed datasets (Health Data NY)

Platform usability

- Healthdata.gov scores low on many platform usability criteria
  - Accessibility-frequent broken links; all objects hosted on external webpages
  - Representational consistency-no handbook, inconsistent metadata elements
  - Functionality- limited interaction with data on the platform
  - User-friendliness- difficult to locate objects due to redirection to other sites
  - Visibility- Healthdata.gov not viewable from a search of “Department of Health & Human Services”; small link to HHS/Open at the bottom of agency webpage

Next steps

- Develop scales to combine items for:
  - Intrinsic data quality
  - Contextual data quality
  - Metadata quality
  - Adherence to the Dublin Core metadata standards

- Run additional tests to see how items cluster, to make recommendations on how to improve the instrument for use in future comparative research

Planned research products

- Presentation to NYSDOH, Socrata
- Primary manuscript of main findings
- Data collection tools, to post to project webpage
- Potential commentaries
  - Dimensions of data quality, and how to evaluate whether a dataset is usable and fit for public health research
  - Evolution of the open data movement
  - Ideas for improving the design of open data platforms and presentation of data and metadata

Other project activities

- Key informant interviews with public health practitioners to understand the value propositions of integrating researchers into the open data ecosystem, and barriers to releasing data

- Pilot geospatial analysis of the relationship between childhood obesity and the built environment in NYS, using open data resources
  - Collaboration with Health Data NY team and Socrata
  - Comparison of results from "gold standard data ecosystem" data analysis model to: 1) no interaction with practitioners, and 2) automated platform-based findings

Questions?

For additional information on the project, fall 2013 workshop, and links to open data resources:

www.rockinst.org/ohdoo