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

Erika Martin, PhD MPH
Rockefeller Institute of Government & University at Albany

PHSSR Research-in-Progress Webinar
National Coordinating Center for PHSSR
Lexington, KY
March 11, 2015

Acknowledgements & Disclosures

- Funding from the Robert Wood Johnson Foundation's Public Health Services & Systems Research Program (grant ID #71597 to Martin and Birkhead)
- Coauthors: Gus Birkhead, Natalie Helbig, Jennie Law, Weijia Ran
- Early feedback: Courtney Burke, Patricia Lynch, Theresa Pardo, Ozlem Uzuner
- JSON technical support: Chris Kotfila
- Gus 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

Agenda

- Promises of open data
- Research and practice gaps
  - Making open data usable and high quality for public health research
- Research methods to document characteristics of open data offerings and differences across platforms
  - Sampling design
  - Coding instrument
  - Statistical analysis
- Findings and implications for practice
- Future project activities

Open data background

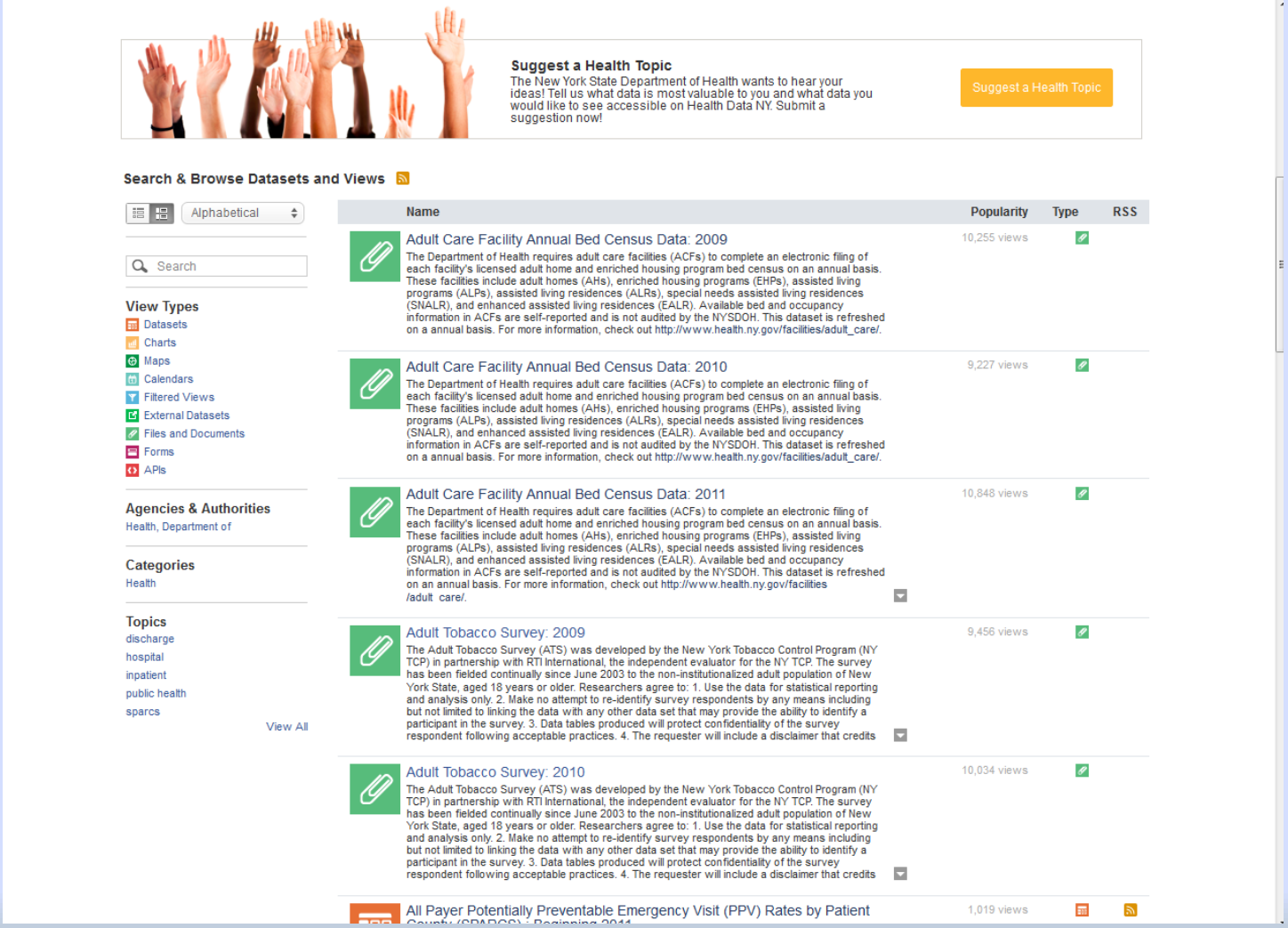
- New source of information for public health research
  - Martin, Helbig, Birkhead *J Public Health Manag Pract* 2014
- Motivated by government transparency movement, including President Obama's memorandum on open government
- Thousands of government datasets released on open data platforms at federal, 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

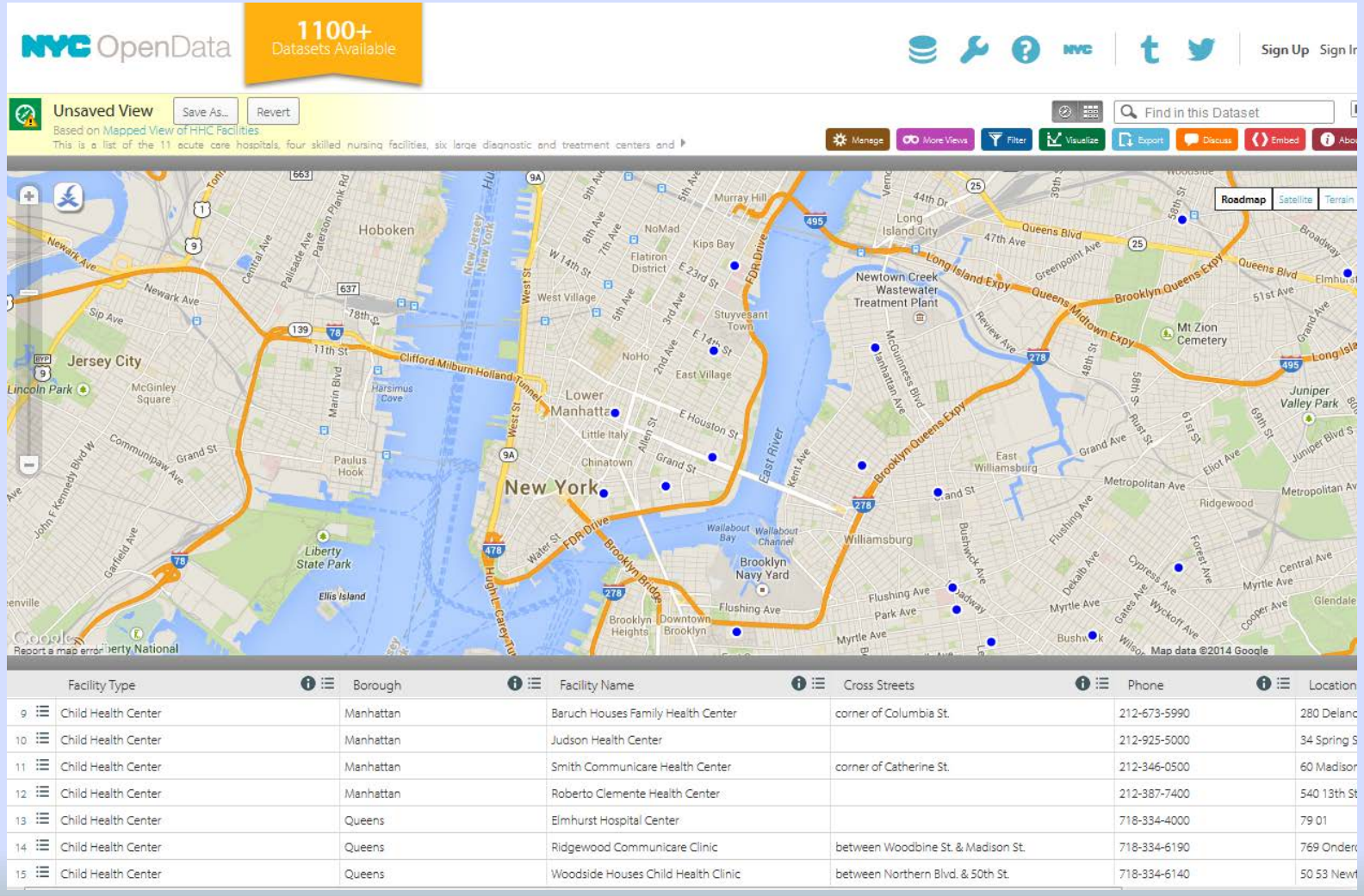

HealthData.gov
Home
Data
Blog
Q & A
Ideas
Developers
Only 1 week left to apply to HHS Entrepreneurs!
Last week to apply for HHS Entrepreneurs!
Six all new projects and we're looking for the best talent to come into government and solve critical problems in health care and government. Apply today! Read 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 engines to locate data objects

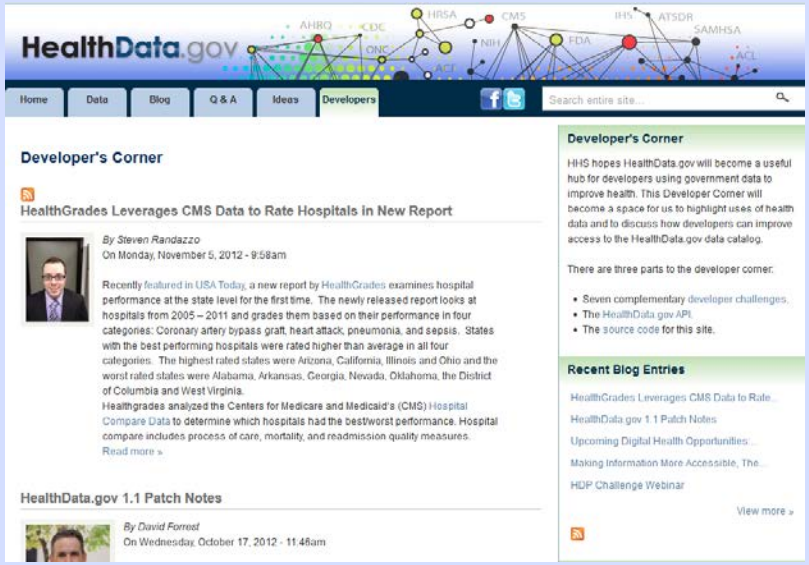


Capabilities to interact directly with data in the platform



	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 Madison
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



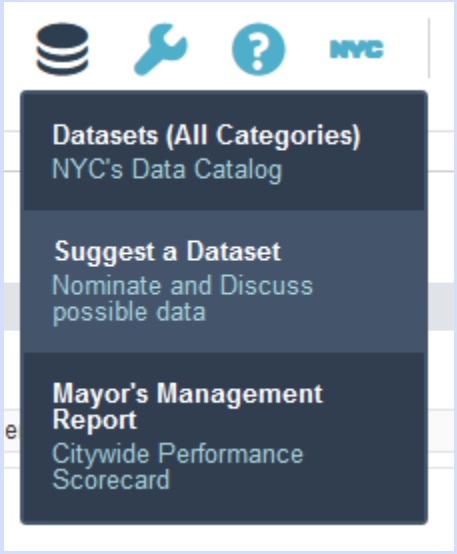
Opportunities to submit ideas for new datasets and provide user feedback

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 offerings 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 offerings randomly sampled from three platforms
  - Healthdata.gov (federal)
  - Health Data NY (state)
  - NYC Open Data (city)
- All data offerings examined with a coding guide to evaluate:
  - Data quality (intrinsic, contextual)
  - Five-star open data deployment
  - Metadata quality
  - Platform usability

Sampling design

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

- Systematic random sampling of data offerings
  - Metadata from platforms scraped into three Excel spreadsheets
  - Excel-based random number generator assigned 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 health department 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

Categories of questions

- Descriptive information
- Intrinsic data quality
- Contextual data quality
- Adherence to Dublin Core international metadata standards
- Consistency with five-star open data deployment scheme

Dublin Core international metadata standards

The Elements

Term Name: contributor	
URI:	http://purl.org/dc/elements/1.1/contributor
Label:	Contributor
Definition:	An entity responsible for making contributions to the resource.
Comment:	Examples of a Contributor include a person, an organization, or a service. Typically, the name of a Contributor should be used to indicate the entity.
Term Name: coverage	
URI:	http://purl.org/dc/elements/1.1/coverage
Label:	Coverage
Definition:	The spatial or temporal topic of the resource, the spatial applicability of the resource, or the jurisdiction under which the resource is relevant.
Comment:	Spatial topic and spatial applicability may be a named place or a location specified by its geographic coordinates. Temporal topic may be a named period, date, or date range. A jurisdiction may be a named administrative entity or a geographic place to which the resource applies. Recommended best practice is to use a controlled vocabulary such as the Thesaurus of Geographic Names [TGN]. Where appropriate, named places or time periods can be used in preference to numeric identifiers such as sets of coordinates or date ranges.
References:	[TGN] http://www.getty.edu/research/tools/vocabulary/tgn/index.html
Term Name: creator	
URI:	http://purl.org/dc/elements/1.1/creator
Label:	Creator
Definition:	An entity primarily responsible for making the resource.
Comment:	Examples of a Creator include a person, an organization, or a service. Typically, the name of a Creator should be used to indicate the entity.
Term Name: date	
URI:	http://purl.org/dc/elements/1.1/date

http://dublincore.org/documents/dces/

METADATA INNOVATION
Dublin Core® Metadata Initiative

Five-star open data deployment scheme

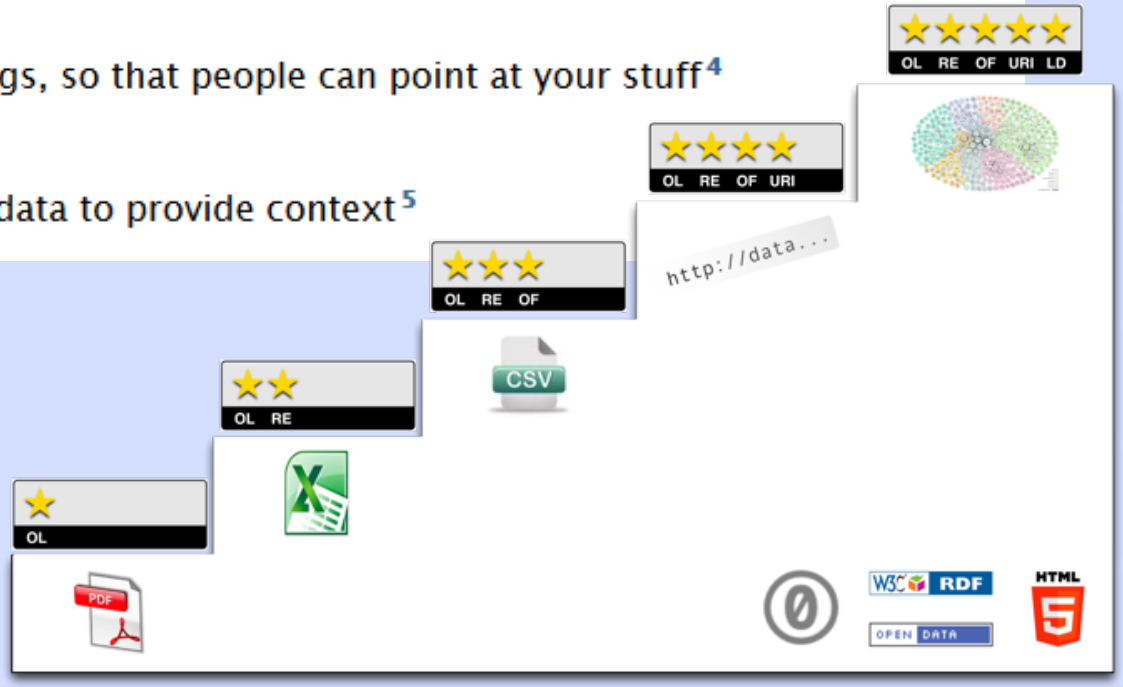
★ make your stuff available on the Web (whatever format) under an open license [1]

★★ make it available as structured data (e.g., Excel instead of image scan of a table) [2]

★★★ use non-proprietary formats (e.g., CSV instead of Excel) [3]

★★★★ use URIs to denote things, so that people can point at your stuff [4]

★★★★★ link your data to other data to provide context [5]



http://5stardata.info/

OL = **O**n**L**ine
RE = can be **RE**used
OF = **O**pen **F**ormats
URI: **U**niform **R**esource **I**dentifier
LD = can **L**ink **D**ata

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, application programming interface (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 uniform resource identifier (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 resource descriptive framework (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

- 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 offerings from the three platforms which varied widely (e.g. administrative data 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 all offerings coded

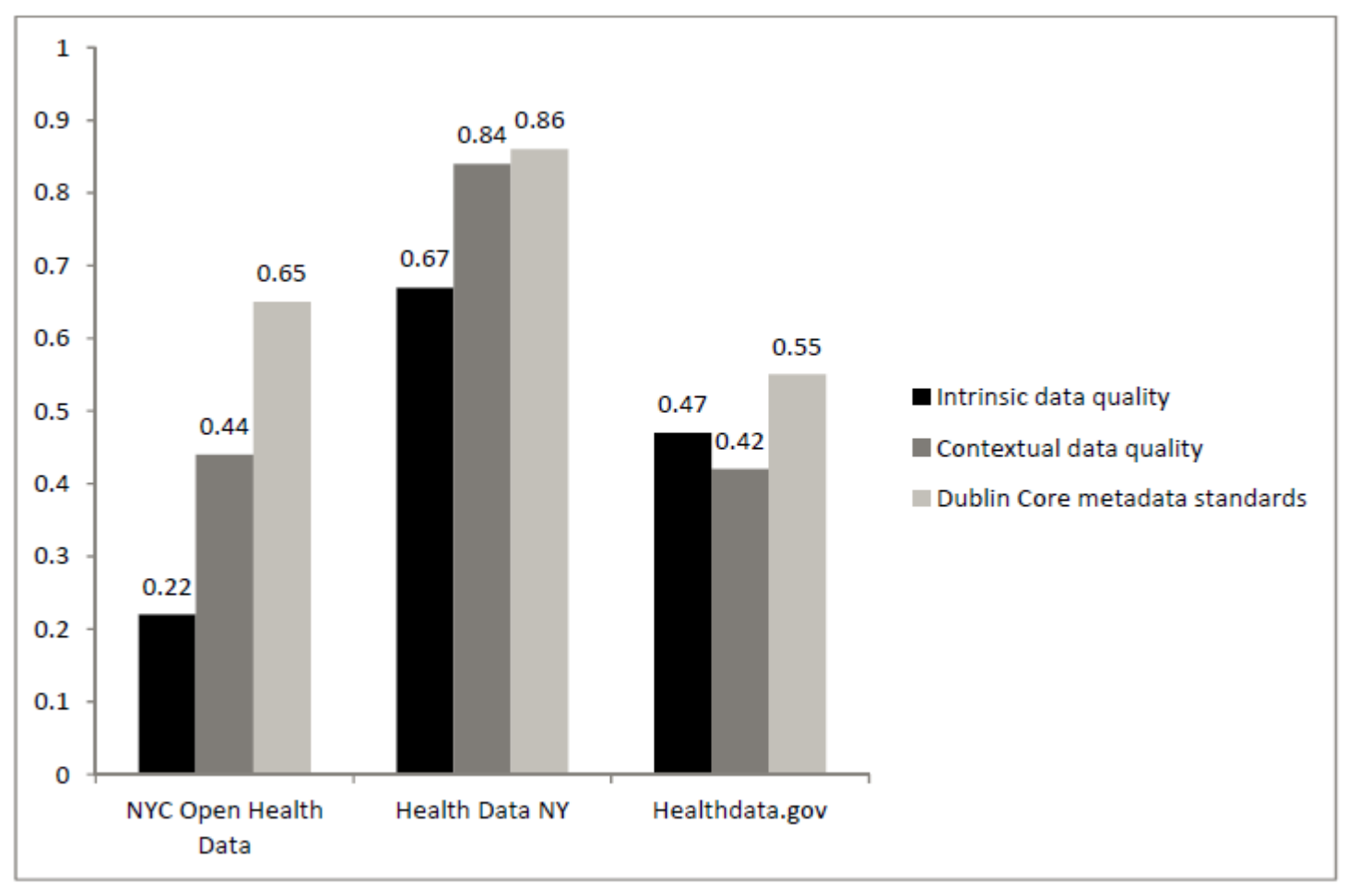
Main findings

- Only one-quarter of open data offerings are tabular datasets
- Most offerings do not contain demographic variables commonly used in public health research
- Health Data NY scored highest on intrinsic data quality, contextual data quality, and adherence to Dublin Core metadata standards
- Gaps in meeting "open data" deployment criteria
  - All offerings met basic "web availability" open data standards
  - Fewer met higher standards of being hyperlinked to other data

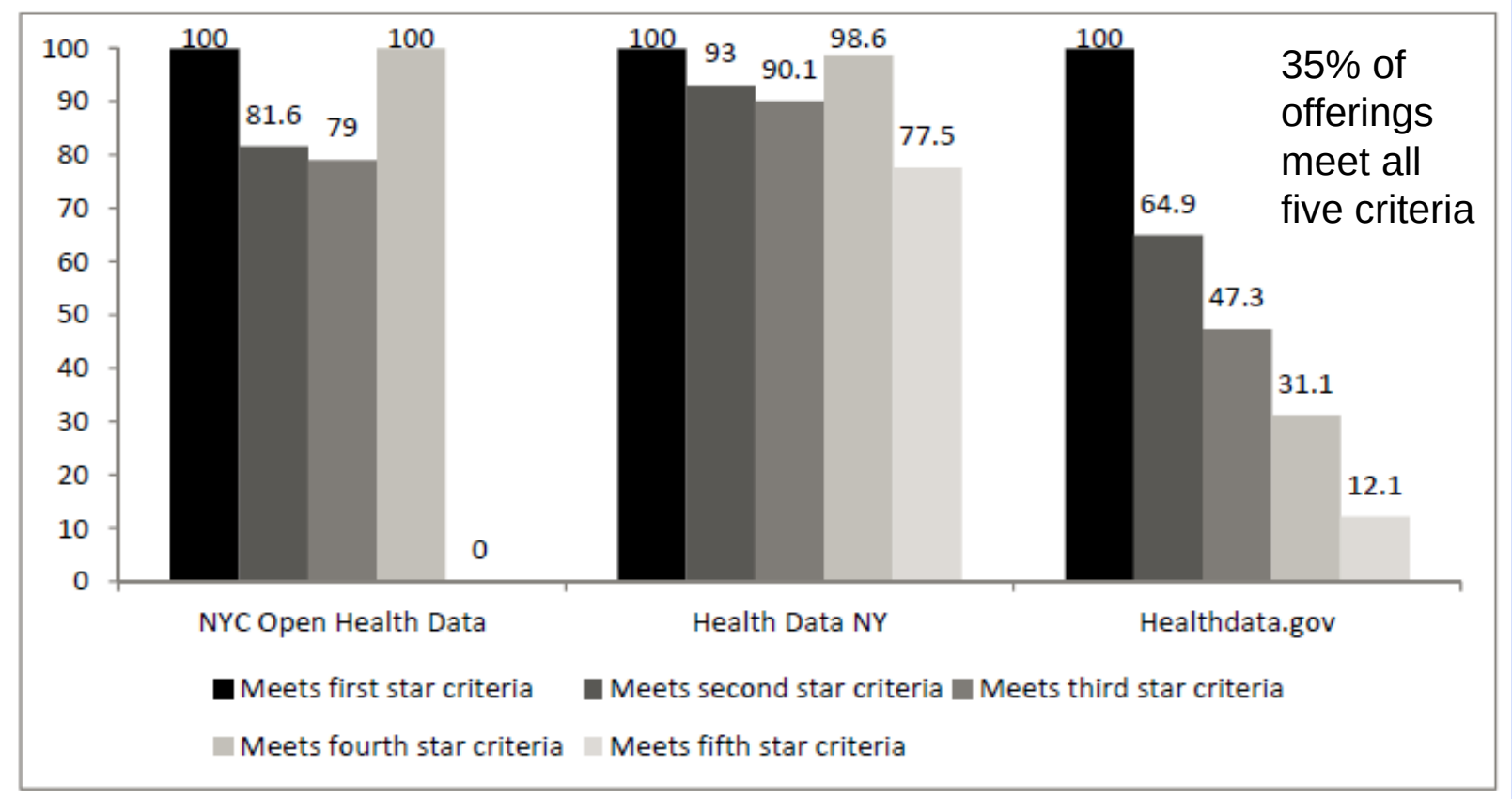
Characteristic	NYC Open Data (city, N=38)[1]	Health Data NY (state, N=71)	Healthdata.gov (federal, N=74)
Primary presentation format in web browser, N (%)			
Table	17 (44.7)	17 (23.9)	12 (16.2)
Chart	--	27 (38.0)	--
Map	9 (23.7)	10 (14.1)	1 (1.4)
External file	1 (2.6)	9 (12.7)	27 (36.5)
Application programming interface	--	2 (2.8)	1 (1.4)
Query tool	4 (10.5)	2 (2.8)	8 (10.8)
Documents about data	3 (7.9)	1 (1.4)	18 (24.3)
Not viewable in a browser[2]	4 (10.5)	3 (4.2)	7 (9.5)
Availability of additional presentation formats, N (%)	11 (29.0)	42 (59.2)	10 (13.5)
Availability of data related to visualizations,[3] N (%)	5 (55.6)	34 (91.9)	1 (100.0)
Ability to view data object in browser, N (%)			
Object is viewable in a browser	28 (73.7)	56 (78.9)	27 (36.5)
Problem with the data access page	5 (13.2)	1 (1.4)	5 (6.8)
Data object is an external file	2 (5.3)	13 (18.3)	21 (28.4)
Data object requires subscription or registration	1 (2.6)	--	6 (8.1)
Data object is only viewable in a proprietary format	1 (2.6)	--	--
Data object not downloadable for other reasons	1 (2.6)	1 (1.4)	15 (20.3)
Ability to download data, N (%)			
Available via platform	--	10 (14.1)	--
Available via data access page	--	--	19 (25.7)
Available from both sources	32 (84.2)	56 (78.9)	23 (31.1)
Not available for download	6 (15.8)	5 (7.0)	32 (43.2)

Characteristic	NYC Open Data (city, N=38)[1]	Health Data NY (state, N=71)	Healthdata.gov (federal, N=74)
Data object year			
Historical data,[4] N (%)	12 (31.6)	31 (43.7)	22 (29.7)
Start year, mean (min, max)	2008 (2003, 2012)	2009 (1974, 2014)	2002 (1920, 2014)
Includes multiple years, N (%)	7 (18.4)	38 (53.5)	13 (17.6)
Data update frequency, N (%)			
Daily or Weekly	1 (2.6)	3 (4.2)	--
Monthly	3 (7.9)	8 (11.3)	1 (5.3)
Quarterly, semi-quarterly, or biannually	2 (5.3)	7 (9.9)	5 (26.3)
Annually or biennially	3 (7.9)	50 (70.4)	8 (42.1)
As needed	20 (52.6)	1 (1.4)	--
Not reported	3 (7.9)	1 (1.4)	59 (79.7)
Not updated	6 (15.8)	1 (1.4)	1 (1.4)
Inclusion of demographic variables, N (%)			
Age	2 (5.3)	21 (29.6)	18 (24.3)
Gender	2 (5.3)	13 (18.3)	14 (18.9)
Race/ethnicity	2 (5.3)	8 (11.3)	10 (13.5)
Insurance status	2 (5.3)	20 (28.1)	18 (24.3)
Education	2 (5.3)	10 (14.0)	2 (2.7)
Income	7 (18.4)	5 (7.0)	8 (10.8)
Geographic identifier	17 (44.7)	45 (63.4)	28 (37.8)
Provider and/or health facilities	18 (47.4)	36 (50.7)	24 (32.4)
Size of data object,[5] median (IQR)			
Number of rows	11 (69)	161 (3340)	357 (2011)
Number of columns	6 (4)	18 (8)	11 (17)
Data object hosted on a different platform,[6] % (N)	n/a	n/a	16 (21.6)

Health Data NY scores highest on indices of intrinsic data quality, contextual data quality, and adherence to Dublin Core metadata standards



Gaps in meeting criteria from the five-star open data deployment scheme



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 and vocabulary *(Health Data NY, NYC Open Data)*
- Multiple functions to search for and download data offerings, 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, social media, ways for users to provide comments
- Ability to embed visualizations into external pages *(Health Data NY, NYC Open Data)*

Platform usability: areas for improvement

- Healthdata.gov primarily serves as a search engine
  - All offerings hosted on external webpages, such as CDC
  - Limited interaction with data on the platform
  - Difficult to locate offerings when redirected to other sites
- Technical problems limit functionality
  - Frequent broken links *(Healthdata.gov)*
  - Problems loading map visualizations *(NYC Open Data)*
- No response to our email queries to help desks
- Low visibility on Google searches *(Healthdata.gov, NYC Open Data)*

Limitations

- New York platforms are not nationally representative
- 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
  - Unanticipated finding that most 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 (one aspect of platform usability)
  - Metadata consistency (one aspect of metadata quality)
- Indices need further validation

Implications for policy and practice

- 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 highest by our measures
- 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 standard measures 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

Other PHSSR 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?

Email:

emartin@albany.edu

For additional information on the PHSSR project:

www.publichealthsystems.org/erika-martin-phd-mph-0

For materials from fall 2013 workshop on open health data in New York and links to open data resources:

www.rockinst.org/ohdoo

Open data case studies

(practitioner commentary: Cheryl Wold)

http://stage.chcf.org/programs/marketmonitor/open-data

cheryl@cherylwold.com