

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
State University of New York**

**Keeneland PHSSR Conference
Lexington, KY
April 22, 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

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

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

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

Suggest a Health Topic

www.health.ny.gov/prevention/prevention_agenda/2013-2017/

HealthData.gov

Only 1 week left to apply to HHS Entrepreneurs!

Last week to apply for HHS Entrepreneurs!

Search the Data

Recent Datasets

Recent Blog Entries

Medicare

Medicaid

Epidemiology

Treatments

Population Statistics

NYC OpenData 1100+ Datasets Available

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





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](#)

Search & Browse Datasets and Views



Alphabetical

View Types

-  Datasets
-  Charts
-  Maps
-  Calendars
-  Filtered Views
-  External Datasets
-  Files and Documents
-  Forms
-  APIs

Agencies & Authorities

Health, Department of



















Categories

Health

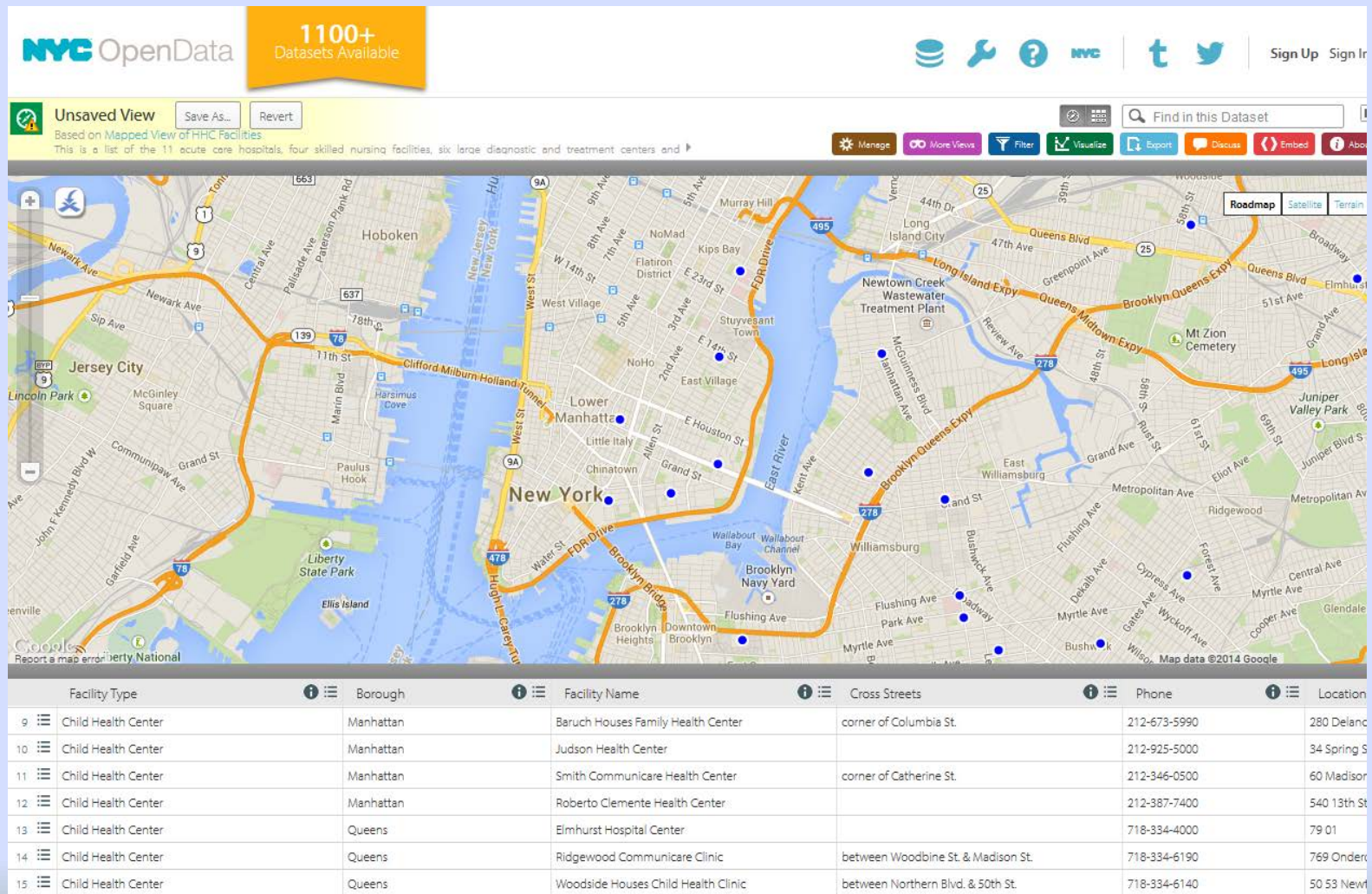
Topics

discharge
hospital
inpatient
public health
spars

[View All](#)

Name	Popularity	Type	RSS
 Adult Care Facility Annual Bed Census Data: 2009 The Department of Health requires adult care facilities (ACFs) to complete an electronic filing of each facility's licensed adult home and enriched housing program bed census on an annual basis. These facilities include adult homes (AHs), enriched housing programs (EHPs), assisted living programs (ALPs), assisted living residences (ALRs), special needs assisted living residences (SNALR), and enhanced assisted living residences (EALR). Available bed and occupancy information in ACFs are self-reported and is not audited by the NYSDOH. This dataset is refreshed on an annual basis. For more information, check out http://www.health.ny.gov/facilities/adult_care/ .	10,255 views		
 Adult Care Facility Annual Bed Census Data: 2010 The Department of Health requires adult care facilities (ACFs) to complete an electronic filing of each facility's licensed adult home and enriched housing program bed census on an annual basis. These facilities include adult homes (AHs), enriched housing programs (EHPs), assisted living programs (ALPs), assisted living residences (ALRs), special needs assisted living residences (SNALR), and enhanced assisted living residences (EALR). Available bed and occupancy information in ACFs are self-reported and is not audited by the NYSDOH. This dataset is refreshed on an annual basis. For more information, check out http://www.health.ny.gov/facilities/adult_care/ .	9,227 views		
 Adult Care Facility Annual Bed Census Data: 2011 The Department of Health requires adult care facilities (ACFs) to complete an electronic filing of each facility's licensed adult home and enriched housing program bed census on an annual basis. These facilities include adult homes (AHs), enriched housing programs (EHPs), assisted living programs (ALPs), assisted living residences (ALRs), special needs assisted living residences (SNALR), and enhanced assisted living residences (EALR). Available bed and occupancy information in ACFs are self-reported and is not audited by the NYSDOH. This dataset is refreshed on an annual basis. For more information, check out http://www.health.ny.gov/facilities/adult_care/ .	10,848 views		
 Adult Tobacco Survey: 2009 The Adult Tobacco Survey (ATS) was developed by the New York Tobacco Control Program (NY TCP) in partnership with RTI International, the independent evaluator for the NY TCP. The survey has been fielded continually since June 2003 to the non-institutionalized adult population of New York State, aged 18 years or older. Researchers agree to: 1. Use the data for statistical reporting and analysis only; 2. Make no attempt to re-identify survey respondents by any means including but not limited to linking the data with any other data set that may provide the ability to identify a participant in the survey; 3. Data tables produced will protect confidentiality of the survey respondent following acceptable practices; 4. The requester will include a disclaimer that credits	9,456 views		
 Adult Tobacco Survey: 2010 The Adult Tobacco Survey (ATS) was developed by the New York Tobacco Control Program (NY TCP) in partnership with RTI International, the independent evaluator for the NY TCP. The survey has been fielded continually since June 2003 to the non-institutionalized adult population of New York State, aged 18 years or older. Researchers agree to: 1. Use the data for statistical reporting and analysis only; 2. Make no attempt to re-identify survey respondents by any means including but not limited to linking the data with any other data set that may provide the ability to identify a participant in the survey; 3. Data tables produced will protect confidentiality of the survey respondent following acceptable practices; 4. The requester will include a disclaimer that credits	10,034 views		
 All Payer Potentially Preventable Emergency Visit (PPV) Rates by Patient County (SPARCS): Beginning 2011	1,019 views		

Capabilities to interact directly with data in the platform



Challenges and resources for developers



A screenshot of the HealthData.gov website. The header includes the "HealthData.gov" logo and a navigation bar with links: Home, Data, Blog, Q & A, Ideas, and Developers. Below the navigation bar is a search bar. The main content area is titled "Developer's Corner" and features an article by Steven Randazzo titled "HealthGrades Leverages CMS Data to Rate Hospitals in New Report". The article includes a small photo of the author and text about a report on hospital performance. To the right of the article is a "Developer's Corner" sidebar with a brief description of the site's purpose and a list of recent blog entries, including "HealthGrades Leverages CMS Data to Rate Hospitals in New Report" and "HealthData.gov 1.1 Patch Notes".

A screenshot of the NYS Health Innovation Challenge website. The header includes the "DEVELOPER CHALLENGE" logo and navigation links: ABOUT, CHALLENGES, CODE-A-THONS, WINNERS, and SPONSORS. Below the navigation bar is a breadcrumb trail: Home > Challenges > Current Challenges > NYS Health Innovation Challenge. The main content area is titled "NYS Health Innovation Challenge" and features a "Pre-Register" button. To the left of the button is a list of links: Background, Description, Timeline, Evaluation, Requirements, Partners, Resources, Media Collateral, Terms & Conditions, and Pre-Register. To the right of the button is a "Submission Deadline" section with the date "July 31, 2014" and a "Contact" section with the name "Jennifer David". Below these sections is a "Prizes" section listing "First Place \$30,000", "Second Place \$10,000", and "Third Place \$3,000". At the bottom of the page is a "Recent Updates" section with the text "Check out new data sets published by the NYSDOH for this challenge! Submission deadline extended to July 31, 2014!".

Build something awesome with Open Data!

The Socrata Open Data API allows you to programmatically access a wealth of open data resources from governments, non-profits, and NGOs around the world. Click the link below and try a live example right now.

https://data.cityofchicago.org/resource/alternative-fuel-locations.json?fuel_type_code=CNG

App Developers

Looking to use open data as part of your application or your business? Learn how to [get started](#).

Libraries & SDKs


Support for most popular programming languages and platforms.

Need Help?

Struggling with a problem you can't figure out? [Get help fast!](#)


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](#)

Datasets (All Categories)

NYC's Data Catalog

Suggest a Dataset

Nominate and Discuss possible data

Mayor's Management Report

Citywide Performance Scorecard




Ideas

You're brilliant, talented, and full of great ideas, right? Share them! How can we drive better health outcomes through the innovative use of data? How can we improve this site? Let's brainstorm together!

Please enter your idea below:

Content limited to 5000 characters, remaining: 5000



What code is in the image? *

Enter the characters shown in the image.

[Share Idea](#)

Note: Only ideas specifically related to HealthData.gov will be considered. Please do not submit any personally identifiable information such as your email address, name, social security number, or home address. Thanks!

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)
 - ❑ Metadata quality
 - ❑ Five-star open data deployment
 - ❑ 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
- ❑ 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
 - ❑ Blog post to NYSDOH SAS user group to solicit comments
 - ❑ Stakeholder feedback on the Prevention Agenda dashboard
 - ❑ Review of a sample of data-based County Health Assessments
 - ❑ Grant reviewers' feedback

Data collection procedures

- ❑ Extensive pilot-testing of coding guide
 - ❑ 16 data offerings from the three platforms which varied widely (e.g. administrative data vs survey, csv-file vs large SAS-file download, 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

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/>

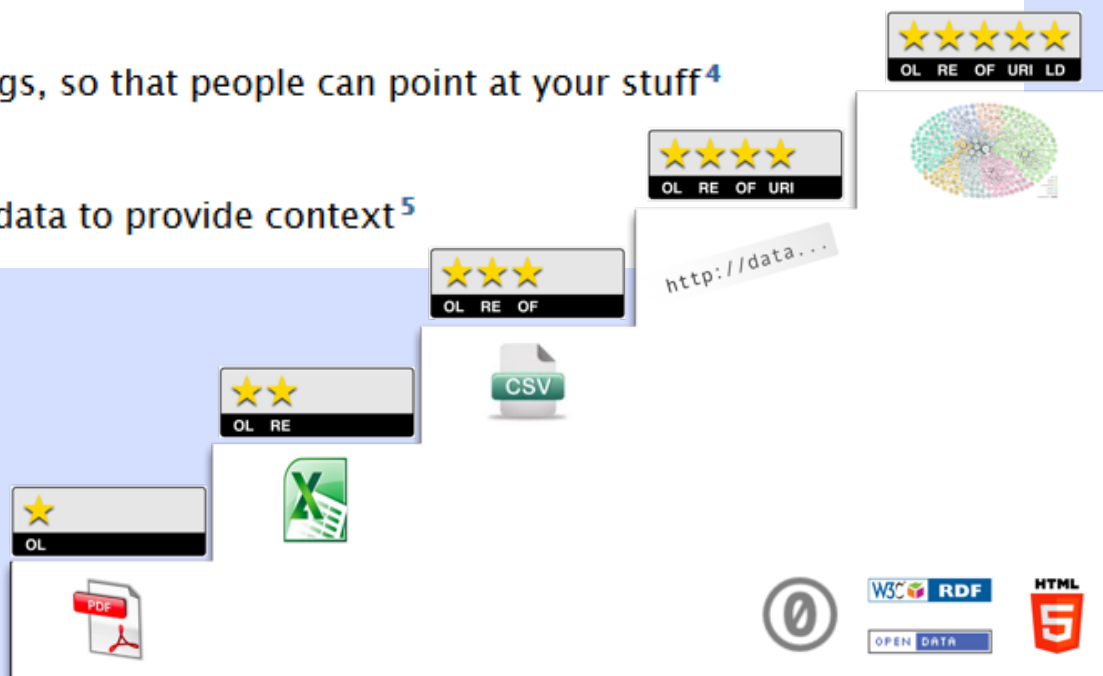


Five-star open data deployment scheme

- ★ make your stuff available on the Web (whatever format) under an open license¹
- ★★ 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 stuff⁴
- ★★★★★ link your data to other data to provide context⁵

<http://5stardata.info/>

OL = OnLine
RE = can be REused
OF = Open Formats
URI: Uniform Resource Identifier
LD = can Link Data



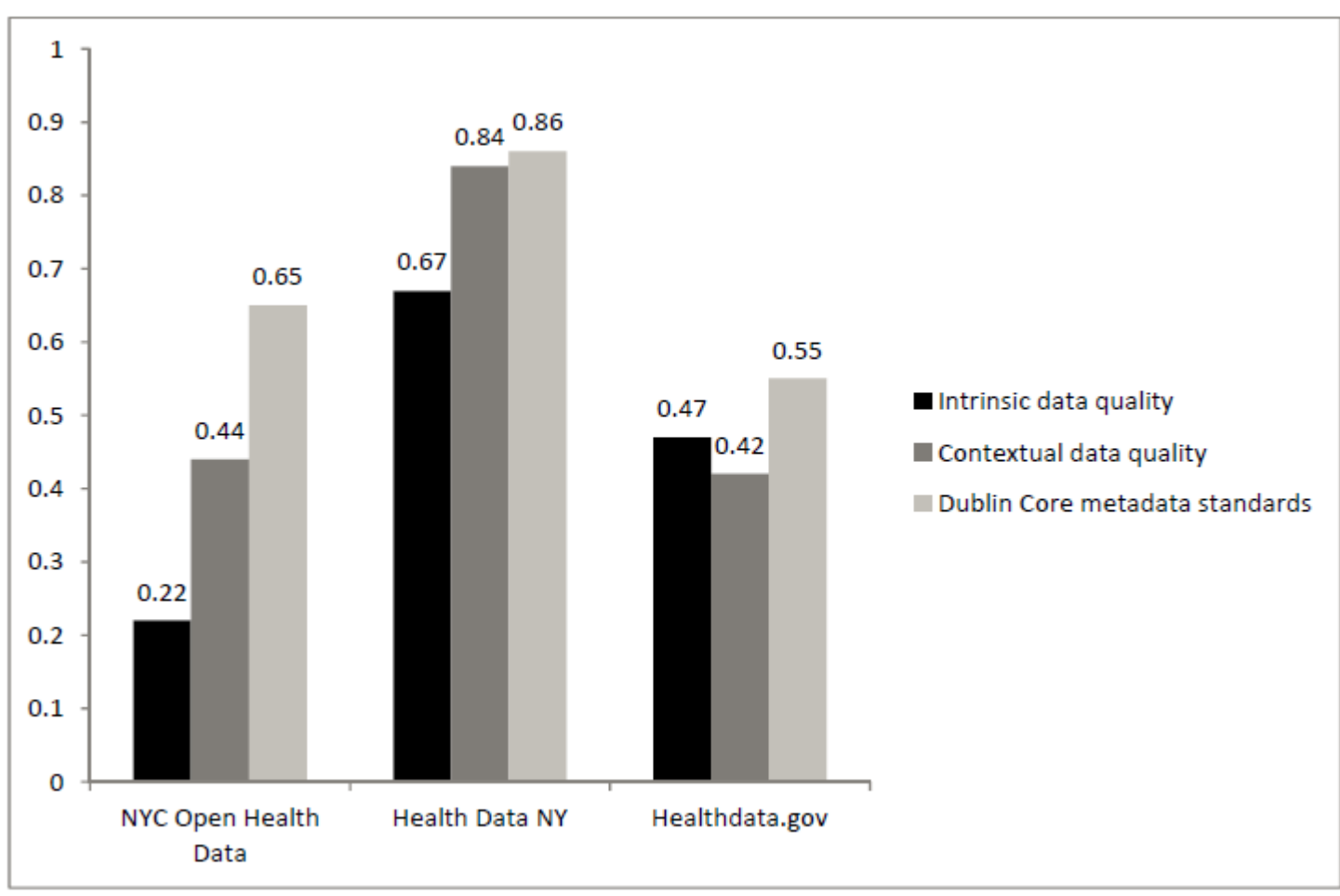
Main findings

- ❑ Only one-quarter of open data offerings are structured 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 to provide context

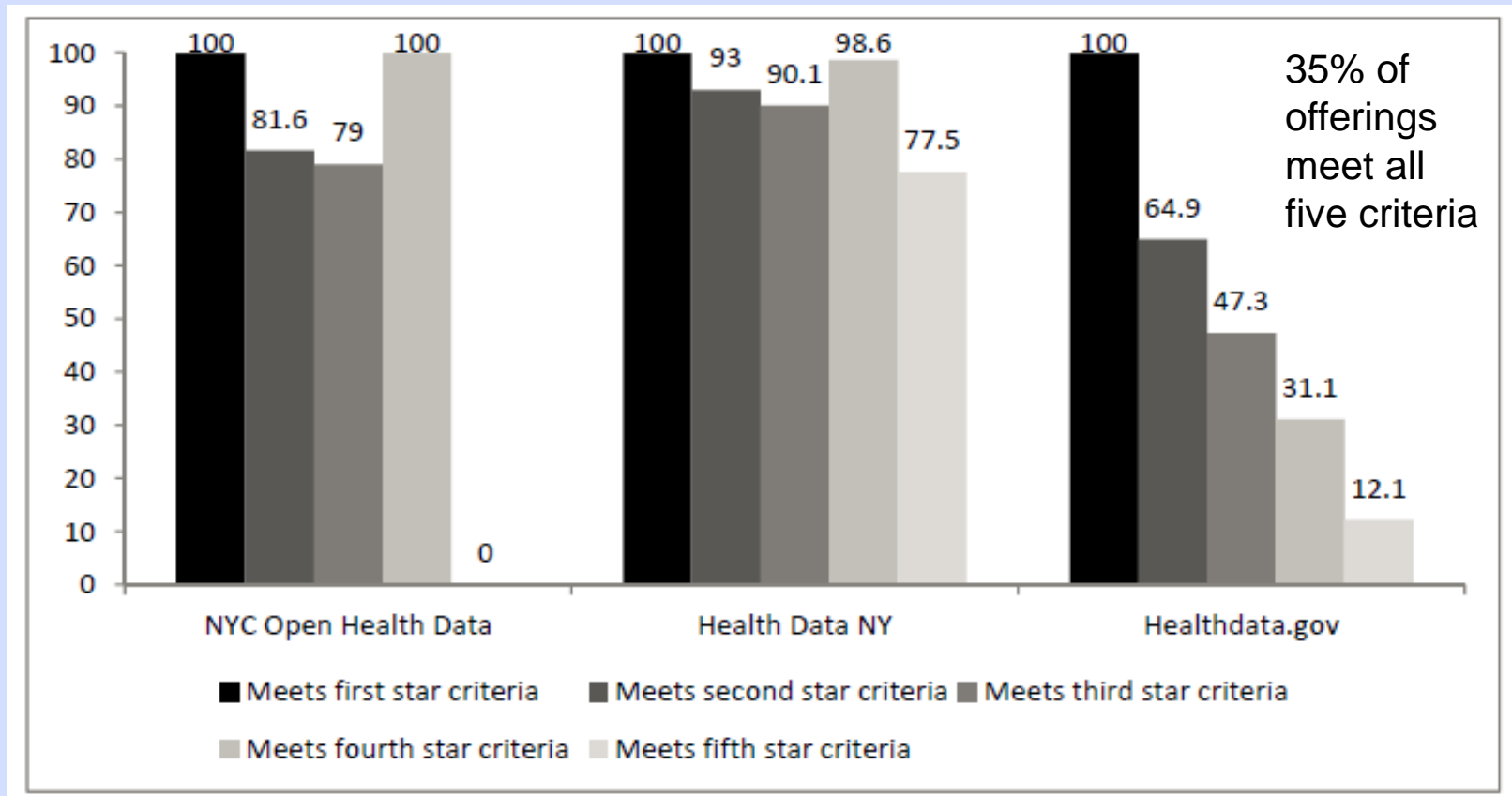
Characteristic	NYC Open Data (city, N=38)¹	Health Data NY (state, N=71)	Healthdata.gov (federal, N=74)
Primary presentation format in web browser, N (%)			
Table	17 (44.7)	17 (23.9)	12 (16.2)
Chart	--	27 (38.0)	--
Map	9 (23.7)	10 (14.1)	1 (1.4)
External file	1 (2.6)	9 (12.7)	27 (36.5)
Application programming 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

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