

# **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, Jennie Law, Weijia Ran**  
*Prepared for 7/29/14 lab meeting and Health Data NY team  
Albany, NY*

# Acknowledgements

- ❑ Project team members
  - ❑ Gus Birkhead
  - ❑ Courtney Burke
  - ❑ Natalie Helbig
  - ❑ Theresa Pardo
  - ❑ Ozlem Uzuner
  
- ❑ Funding from the Robert Wood Johnson Foundation's Public Health Services & Systems Research Program (grant ID #71597 to Martin and Birkhead)

# Agenda

- ❑ Introductions and objectives
- ❑ Open data background
- ❑ Research questions
- ❑ Research methods
  - ❑ Overview
  - ❑ Sampling
  - ❑ Coding instrument
- ❑ (VERY) preliminary findings and next steps
- ❑ Discussion

# Introductions and objectives

- ❑ Introductions
  
- ❑ Objectives
  - ❑ Share research-in-progress
  - ❑ Solicit feedback from NYSDOH on how to make research products more useful for public health practice
  - ❑ Solicit feedback from lab group and RIG colleagues on how to make research products more interesting for academic audiences

# Open data background

- ❑ New source of information for public health research
- ❑ 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

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 engine 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

- 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  
sparks

[View All](#)

Name	Popularity	Type	RSS
 <b>Adult Care Facility Annual Bed Census Data: 2009</b> 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 <a href="http://www.health.ny.gov/facilities/adult_care/">http://www.health.ny.gov/facilities/adult_care/</a> .	10,255 views		
 <b>Adult Care Facility Annual Bed Census Data: 2010</b> 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 <a href="http://www.health.ny.gov/facilities/adult_care/">http://www.health.ny.gov/facilities/adult_care/</a> .	9,227 views		
 <b>Adult Care Facility Annual Bed Census Data: 2011</b> 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 <a href="http://www.health.ny.gov/facilities/adult_care/">http://www.health.ny.gov/facilities/adult_care/</a> .	10,848 views		
 <b>Adult Tobacco Survey: 2009</b> 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		
 <b>Adult Tobacco Survey: 2010</b> 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		
 <b>All Payer Potentially Preventable Emergency Visit (PPV) Rates by Patient County (SPARCS), Beginning 2014</b>	1,019 views		

# Capabilities to interact directly with data in the platform

**NYC OpenData** 1100+ Datasets Available

Unsaved View Save As... Revert

Based on Mapped View of HHC Facilities  
This is a list of the 11 acute care hospitals, four skilled nursing facilities, six large diagnostic and treatment centers and

Manage More Views Filter Visualize Export Discuss Embed About

Find in this Dataset

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



**HealthData.gov**

Home Data Blog Q & A Ideas Developers

### Developer's Corner

#### HealthGrades Leverages CMS Data to Rate Hospitals in New Report

By Steven Randazzo  
On Monday, November 5, 2012 - 9:58am

Recently featured in USA Today, a new report by HealthGrades examines hospital performance at the state level for the first time. The newly released report looks at hospitals from 2005 – 2011 and grades them based on their performance in four categories: Coronary artery bypass graft, heart attack, pneumonia, and sepsis. States with the best performing hospitals were rated higher than average in all four categories. The highest rated states were Arizona, California, Illinois and Ohio and the worst rated states were Alabama, Arkansas, Georgia, Nevada, Oklahoma, the District of Columbia and West Virginia.

Healthgrades analyzed the Centers for Medicare and Medicaid's (CMS) Hospital Compare Data to determine which hospitals had the best/worst performance. Hospital compare includes process of care, mortality, and readmission quality measures.

[Read more »](#)

#### HealthData.gov 1.1 Patch Notes

By David Forrest  
On Wednesday, October 17, 2012 - 11:48am

#### Developer's Corner

HHIC hopes HealthData.gov will become a useful hub for developers using government data to improve health. This Developer Corner will become a space for us to highlight uses of health data and to discuss how developers can improve access to the HealthData.gov data catalog.

There are three parts to the developer corner:

- Seven complementary developer challenges.
- The HealthData.gov API
- The source code for this site.

#### Recent Blog Entries

- HealthGrades Leverages CMS Data to Rate...
- HealthData.gov 1.1 Patch Notes
- Upcoming Digital Health Opportunities...
- Making Information More Accessible, The...
- HDP Challenge Webinar

[View more »](#)

**Health 2.0 DEVELOPER CHALLENGE**

ABOUT CHALLENGES CODE-A-THONS WINNERS SPONSORS

Home > Challenges > Current Challenges > **NYS Health Innovation Challenge**

### NYS Health Innovation Challenge

**Submission Deadline**  
July 31, 2014

**Contact**  
Jennifer David

[Pre-Register](#)

**Prizes**

- First Place \$30,000
- Second Place \$10,000
- Third Place \$3,000

**Recent Updates**

Check out [new data sets](#) published by the NYSDOH for this challenge!  
Submission deadline extended to July 31, 2014!

Partners

## 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](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 dataset, and user feedback

...payers by the patient's county.



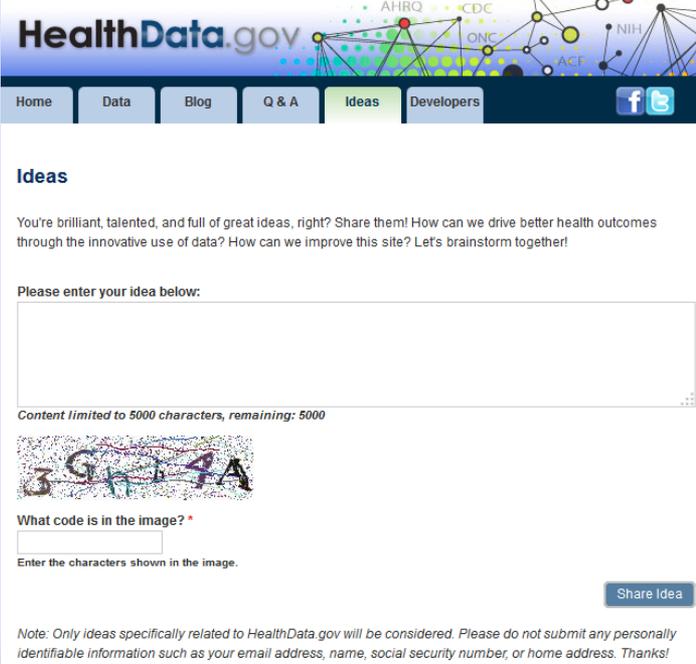
### Suggest a Health Topic

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](#)



- Datasets (All Categories)**  
NYC's Data Catalog
- Suggest a Dataset**  
Nominate and Discuss possible data
- Mayor's Management Report**  
Citywide Performance Scorecard



**HealthData.gov** AHRQ CDC ONC NIH ACE

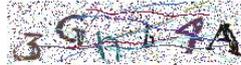
Home Data Blog Q & A **Ideas** Developers

### 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 objects on federal, state, and local platforms
  - ❑ Adapted from Institute of Medicine and Patient-Centered Outcomes Research Institute standards 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=38)
  - ❑ 25% random sample of Health Data NY data objects (N=71, of 308 available)
  - ❑ 5% random sample of Healthdata.gov data objects (N=75, of 1,526 available)
  - ❑ Total of 184 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, forthcoming, *J Public Health Manag Pract*)
  - ❑ 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

# ***Factors that influence the use of governmental data and subsequent health outcomes***

## **Conditions**

### **Data characteristics**

- Populations represented
- Sample size and sampling methods
- Unit of analysis (e.g. individuals, treatment episodes, healthcare facilities)
- Data elements included
- Data collection method (e.g. administrative records, surveys, medical records)
- Study design (e.g. cross-sectional, repeated measures)
- 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

### **Data owner organizational capabilities**

- 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

## **States**

### **Intrinsic data quality**

- Accuracy
- Believability
- Objectivity
- Reputation
- Reliability
- Confidentiality
- Semantic representation

### **Contextual data quality**

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

### **Platform usability**

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

### **Metadata quality**

- Completeness
- Consistency
- Clarity

## **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 decision-making by patients, providers, and policy-makers

Notes: The extent to which states align influence 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.

# 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
- ❑ Documents archived on hard drive
  - ❑ Static documents (e.g. codebooks, dataset downloads)
  - ❑ 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
- ❑ Will use a simplified guide to evaluate platform usability

# 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)

# VERY preliminary findings

- ❑ NYC Open Data (city)
  - ❑ Most originate from the Health and Hospitals Corporation or Human Resources Administration
  - ❑ Many repeated data objects, especially relating to the location of Health and Hospitals Corporation facilities
  - ❑ Data objects presented as maps do not show in browsers (Google Chrome, Mozilla Firefox, Internet Explorer)
  - ❑ Very little provenance about the data objects (e.g. where data came from, how they were collected)

# VERY preliminary findings

- ❑ Health Data NY (state)
  - ❑ Compared to NYC Open Data and Healthdata.gov, standardized format of metadata page provides:
    - ❑ More provenance about the data
    - ❑ Information in a more standardized and user-friendly format
  - ❑ Metadata page often references an external site or data object that provides additional context or details on the data
    - ❑ Very helpful!
  - ❑ APIs reference an nonexistent “About” section, making it difficult to find information about the data

# VERY preliminary findings

- ❑ Healthdata.gov (federal)
  - ❑ Many data artifacts that do not fit the definition of a data object
    - ❑ Examples: collections of PDF documents, legislation
  - ❑ Most difficult platform to navigate to find:
    - ❑ Data provenance (e.g. web links that lead to series of web links)
    - ❑ Data objects (e.g. data object may link to a page with multiple data objects)
  - ❑ Data objects often located on an external agency site, rather than being downloadable from the platform
  - ❑ Inter-agency variation in the fitness and usability of data objects
    - ❑ Less engagement from open data team?

# Planned research products

- ❑ Primary manuscript of main findings
  - ❑ Target: high-impact medical journal or epidemiology journal
  
- ❑ Data collection tools, to post to project webpage
  
- ❑ 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
  - ❑ Targets: *Health Affairs Blog, Frontiers in PHSSR, AJPH*

# Future 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
  - ❑ Potential opportunity to collaborate with Health Data NY team and Socrata on pilot effort to link data within the platform

# Discussion questions

- ❑ What does the Health Data NY team want to learn from this systematic data object review? How can we make findings more useful for the team?
- ❑ How can we make these ideas more interesting for a general academic audience?
- ❑ Are there ideas for additional commentaries or articles based on this research?
- ❑ How can we make our conceptual framework more intuitive for a general public health audience?