RELEASING HIGH-VALUE OPEN HEALTH DATA:
Opportunities, Challenges, and Practice Recommendations

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

CHIPTS Methods Seminar, UCLA
Los Angeles, CA
January 7, 2016

Acknowledgements & Disclosures

- Funding from the Robert Wood Johnson Foundation's Public Health Services & Systems Research Program (grant ID #71597 to Martin and Birkhead)
- Collaborators: Grace Begany, Gus Birkhead, Christine Bozlak, Ian Brisette, Ade Dwicaksono, Natalie Helbig, Jennie Law, Weijia Ran
- JSON technical support: Chris Kotfila
- Gus Birkhead, Ian Brisette, 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

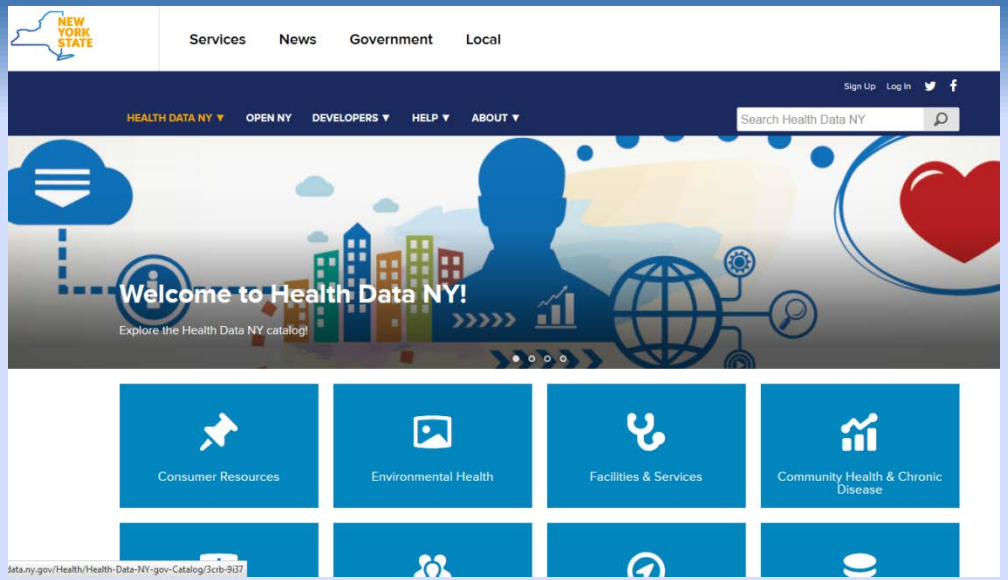
- Study motivation
  - Promises of open data
  - Research gaps
- Findings from aims 1 and 2 of multi-phase study
  - Aim 1: systematic review of data offerings in three open data portals
  - Aim 2: key informant interviews on the opportunities, challenges, and lessons learned from releasing open data
  - Aim 3: pilot study to use open data to evaluate the geospatial relationship between childhood obesity and the built environment
- Translating results to practice
  - Implications for policy and practice
  - Recommendations for improving the value of open data and ensuring sustainability

Study motivation:

promises of open data and research gaps

Open data background

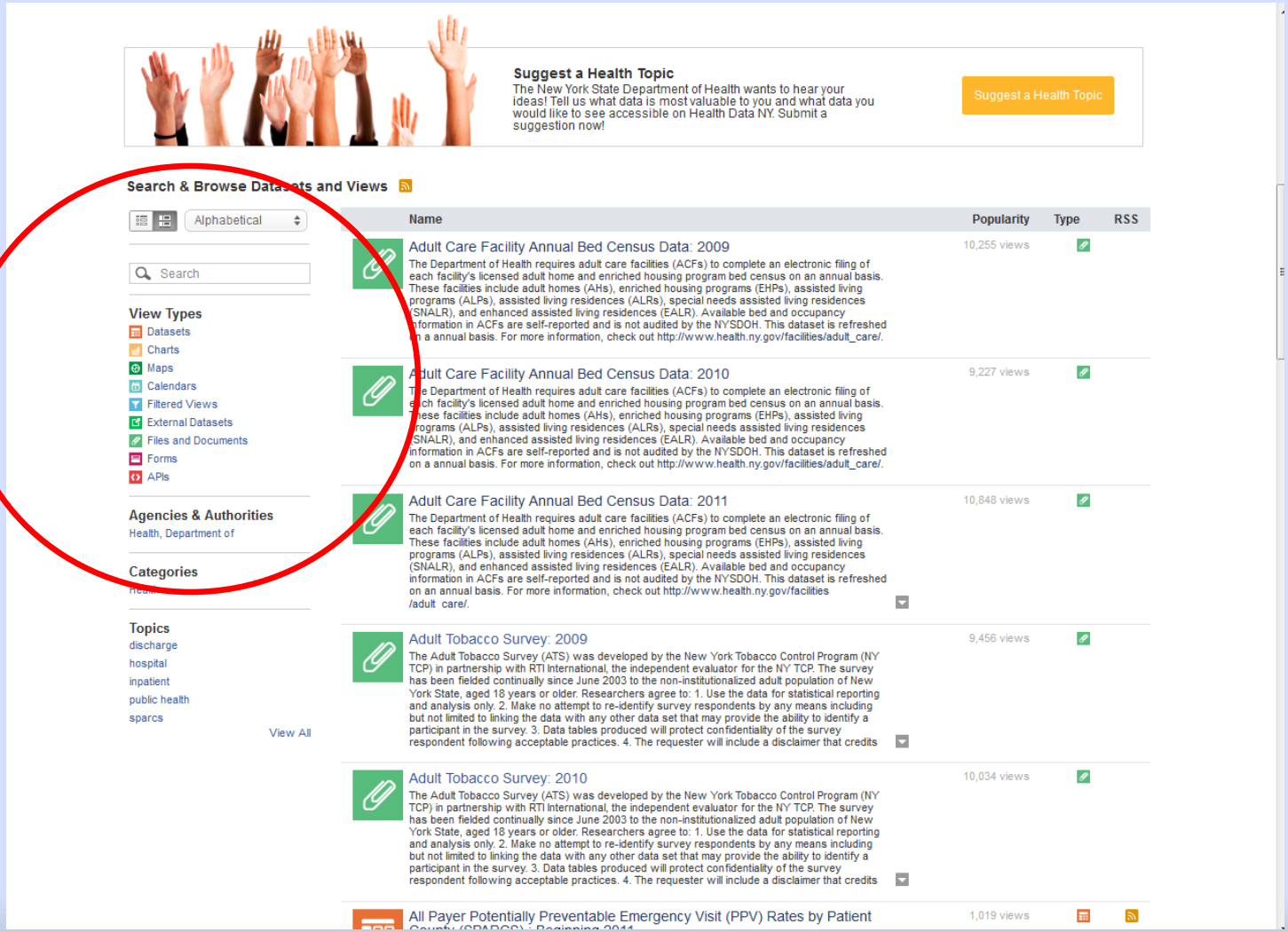
- Thousands of government datasets released on open data platforms at federal, state, and local levels meeting several "openness" criteria
  - Publicly accessible, available in non-proprietary formats, free of charge, unlimited use and distribution rights
- Motivated by government transparency movement,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


NEW YORK STATE
Services
News
Government
Local
Sign Up
Log In
HEALTH DATA NY
OPEN NY
DEVELOPERS
HELP
ABOUT
Search Health Data NY
Welcome to Health Data NY!
Explore the Health Data NY catalog!
Consumer Resources
Environmental Health
Facilities & Services
Community Health & Chronic Disease

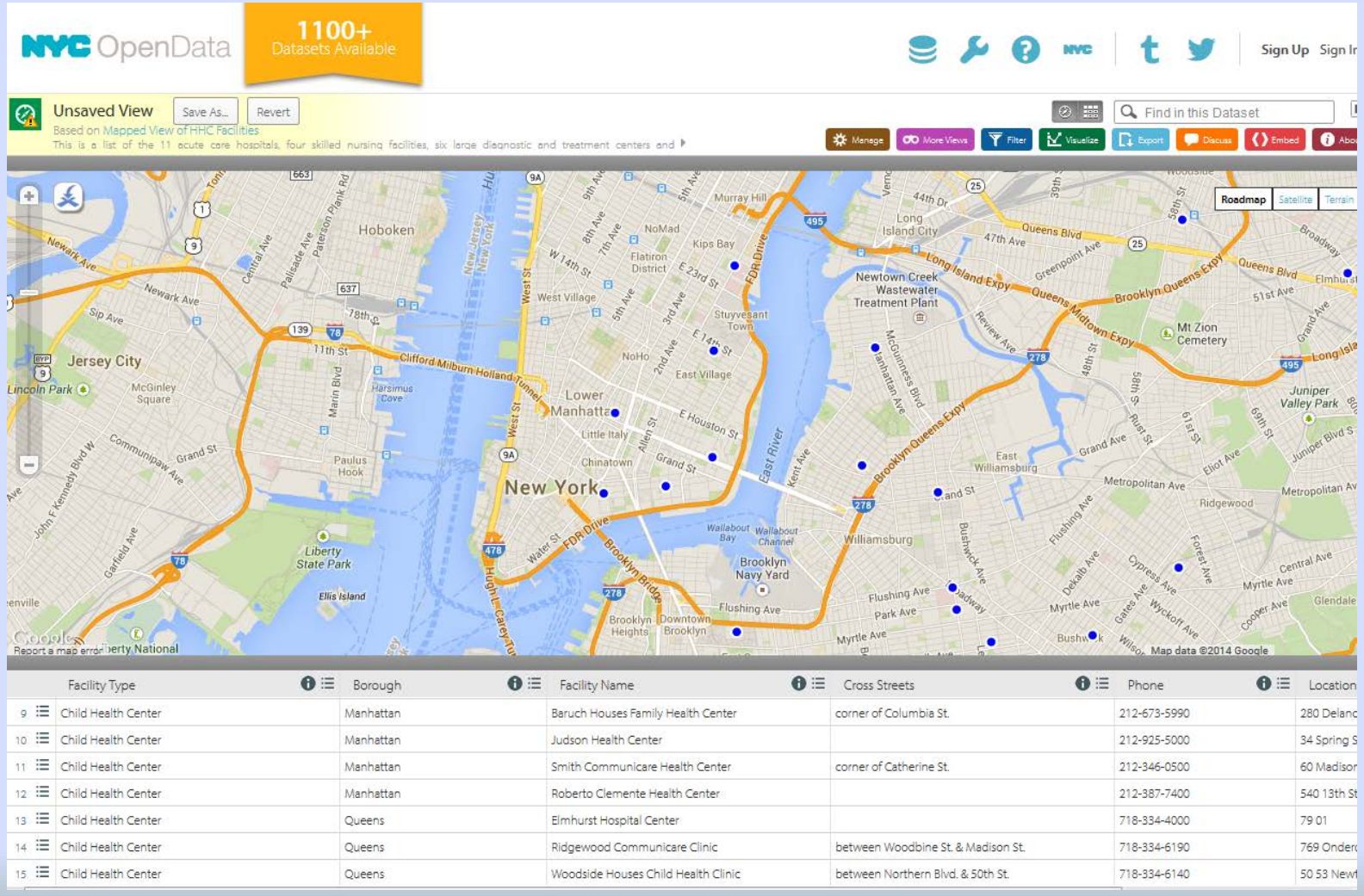

HealthData.gov
Home
Data
Blog
Q & A
Ideas
Developers
Search entire site
Only 1 week left to apply to HHS Entrepreneurs!
Last week to apply for HHS Entrepreneurs!
Six all new projects and we're looking for the best talent to come into government and solve critical problems in health care and government.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

Tools to search for data products



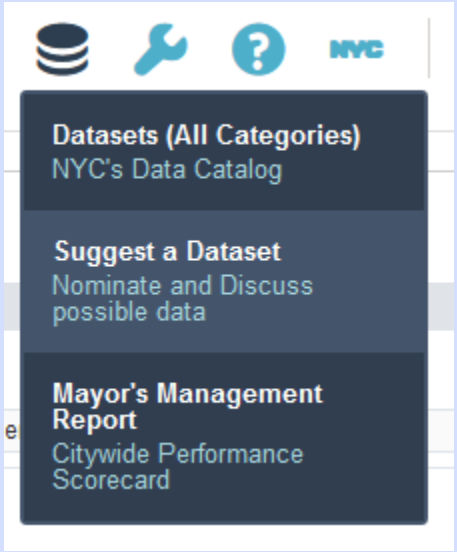
Capabilities to interact directly with data in the platform



	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 datasets and provide user feedback



Research gaps

- Open data are promising but…
  - To what extent are open health data currently ***usable*** and ***fit*** for public health research and practice?
  - How could government agencies improve the ***quality*** of the ***data*** and corresponding ***metadata***?
  - What is the ***perceived value*** of releasing open health data, and what are the ***capabilities*** needed to create a ***successful*** and ***sustainable*** open data ecosystem?
  - How can we develop a robust ***community of practice*** oriented towards using open data for public health research and practice?

Project overview

- Mixed methods study
  - Aim 1: systematic review of data offerings in three open data portals
  - Aim 2: key informant interviews on the opportunities, challenges, and lessons learned from releasing open data
  - Aim 3: pilot study to use open data to evaluate the geospatial relationship between childhood obesity and the built environment

- Overarching goal to provide recommendations for practice
  - How to improve open data and sustain these efforts
  - How to build a robust community of practice oriented around using open data for public health research and practice

Aim 1:

systematic review of data offerings in three open data portals

Collaborators: Gus Birkhead, Natalie Helbig, Jennie Law, Weijia Ran

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
- 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)
  - JL and WR double-coded and compared responses, discussing discrepancies with EGM
  - Interim feedback from NH and GSB
  - 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/

METADATA
Dublin Core® Metadata Initiative
INNOVATION

Five-star open data deployment scheme

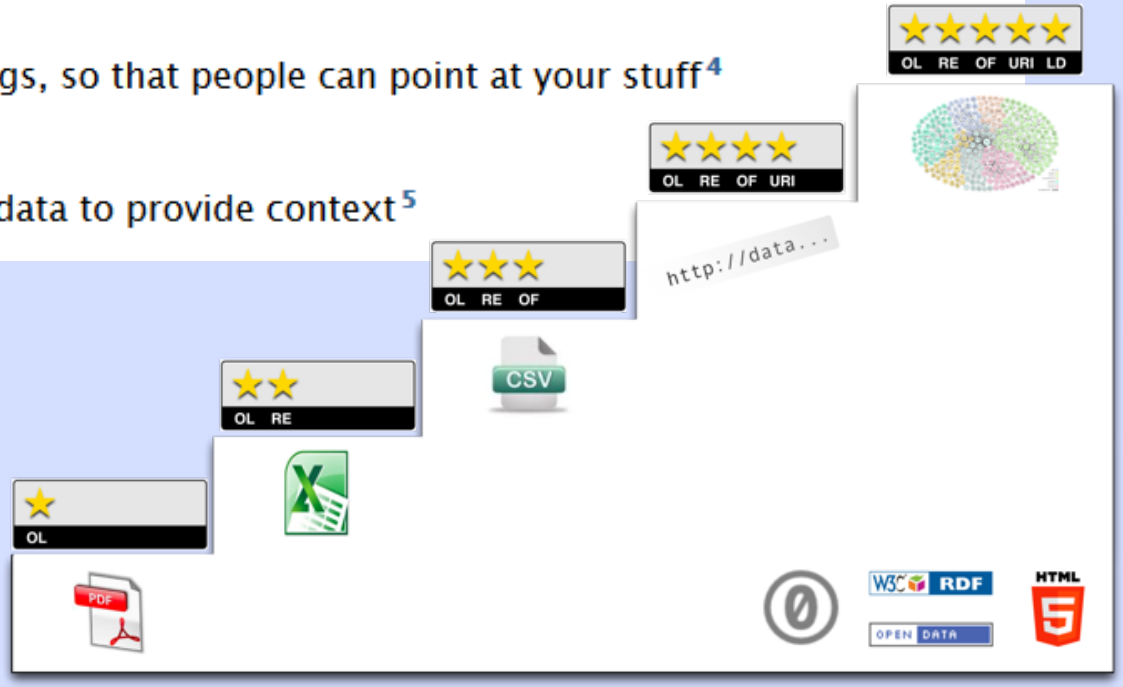
★ make your stuff available on the Web (whatever format) under an open license [1]

★★ 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

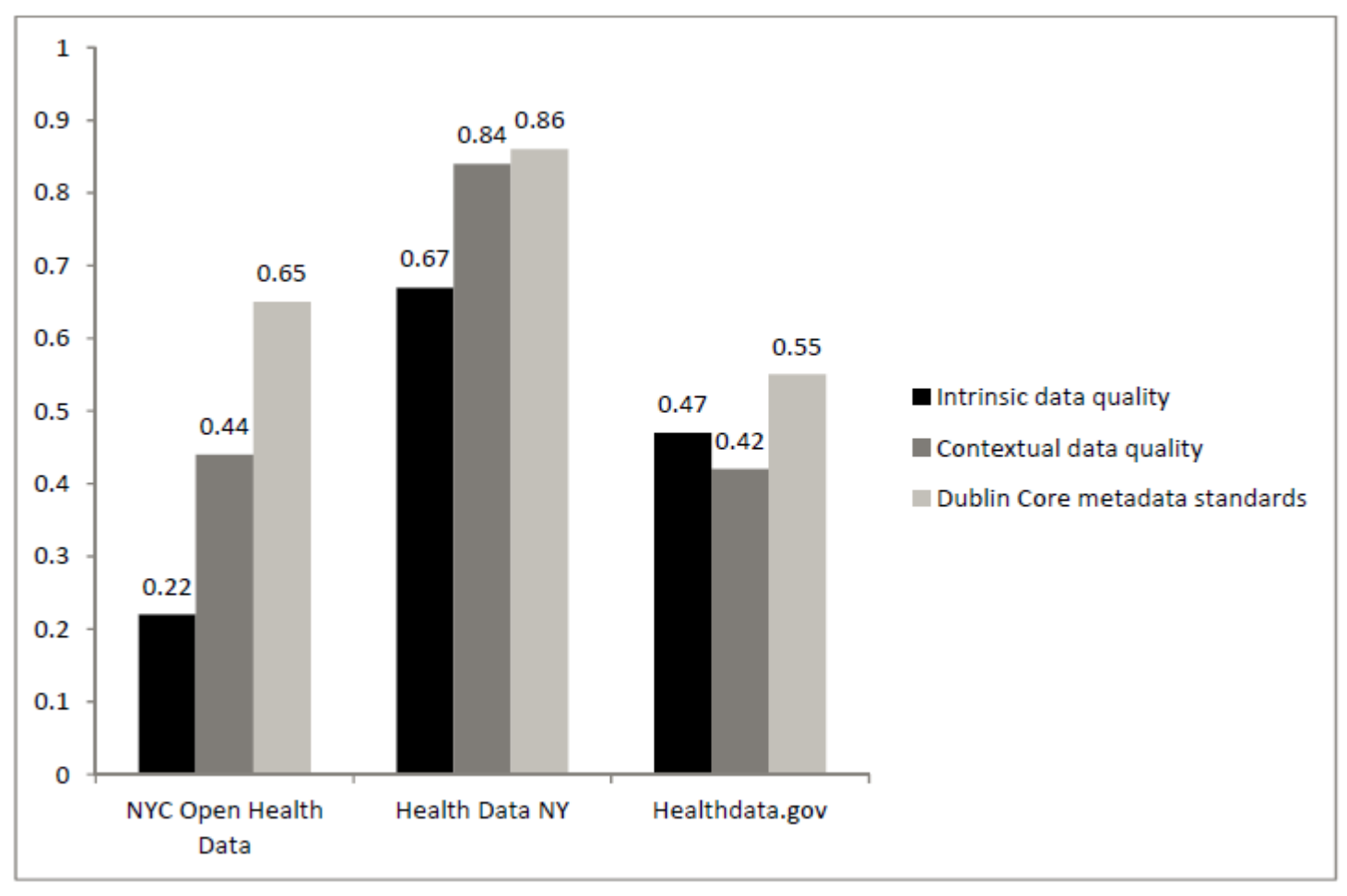
Main findings

- 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)

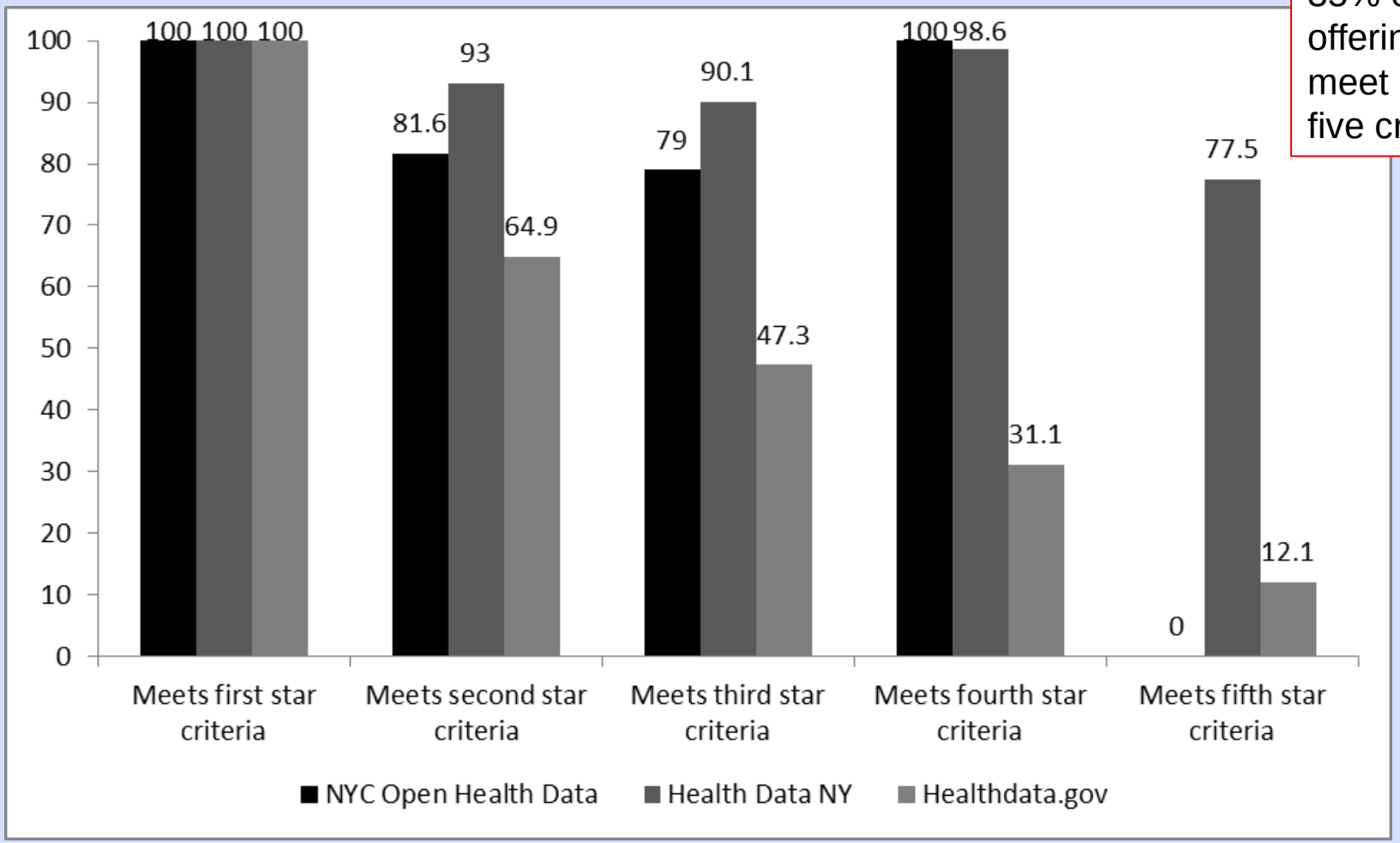
Characteristic	NYC Open Data (city, N=38)[1]	Health Data NY (state, N=71)	Healthdata.gov (federal, N=74)
Data object year			
Historical data,[4] N (%)	12 (31.6)	31 (43.7)	22 (29.7)
Start year, mean (min, max)	2008 (2003, 2012)	2009 (1974, 2014)	2002 (1920, 2014)
Includes multiple years, N (%)	7 (18.4)	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

35% of offerings meet all five criteria



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 structured 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

Aim 2:

key informant interviews on the opportunities, challenges, and lessons learned from releasing open data

Collaborator: Grace Begany

Research design overview

- Key informant interviews with practitioners and policymakers in New York State and DHHS
  - Confirmed findings with respondents in Chicago, Washington, and California to assess generalizability
- Main discussion topics:
  - Perceived public value of releasing open health data
  - Policy, management, and technology challenges of developing open data portals and releasing data
- Transcripts analyzed using grounded theory framework to discover and report themes and concepts
  - Systematic method to analyze qualitative data
  - Themes in data tagged with codes; data subsequently re-reviewed to identify concepts and categories, which can then generate theory

Methods: sampling and recruitment

- Purposive sampling, aiming for diversity in key informants' roles and areas of expertise
  - Roles: executive leadership, program directors, data owners, open data staff, legal affairs, independent contractors
  - Areas of expertise: leadership, project management, epidemiology, public health law, information technology
- Focus on early leaders in open data
  - New York State Department of Health
  - U.S. Department of Health and Human Services
  - Non-governmental organizations (e.g. Health Data Consortium)
  - Other innovative states/cities
- Evolving sample, until no new topics or viewpoints emerged
- Final sample: 40 key informants, 32 interviews

Methods: semi-structured interview guide

- Specific questions tailored to respondents (6-8 questions)
- Topics covered
  - **Historical context** (evolution of Health Data NY, Healthdata.gov, etc.; history of open data movement)
  - **Availability of open data** (what data are being released, how data prioritized for release, factors that determine which data to release)
  - **Current and future benefits of releasing open data** (use cases, benefits already realized, long-term visions)
  - **Challenges of releasing open data** (technical, management, political)
  - **Capabilities needed to release data**
  - **Personal interactions with open data sites**
  - **Open data release process** (how data and metadata prepared, processes to de-identify data)
  - **Early leaders in releasing data** (states, cities)
  - **Legal environment for open data** (relevant laws and regulations, expert determination, legal review)

Methods: data analysis

- Interviews transcribed and uploaded into Atlas.ti
- Grounded theory approach to systematically discover and report themes and concepts
- Preliminary coding guide developed, based on review of all transcripts
- 5 transcripts double-coded by EGM & GMB to refine coding
- GMB subsequently coded all transcripts, conferring with EGM throughout
- EGM & GMB re-reviewed coded data to synthesize themes

Main findings

- Wide range of perceived benefits
  - Internal benefits include improved data quality and more efficient public health operations
  - External benefits include health literacy, data-driven improvements in healthcare delivery, community empowerment
  - New users can bring fresh innovative ideas
- Numerous challenges to releasing data
  - Open data not perceived as a “technical issue”
  - Key challenges include resources, cultural resistance, legal and regulatory issues, and data/metadata quality
- General optimism that open data movement will continue
  - Yet success depends on sustained leadership, resources, cultural changes, promoting the use of data, and establishing governance

Perceived benefits of open data

- More efficient public health operations
  - Removing internal data silos
  - Faster internal clearance to publish presentations and reports
  - Fewer Freedom of Information Act requests
  - Reduced volume of repeated queries about specific datasets
  - Using food safety data to prioritize which restaurants to inspect first
- Improved data quality, timeliness, and usefulness
  - End-users may have questions about the data or find errors
  - Data release process may improve metadata
  - Agencies pressured to release more timely data
  - Data can be automatically refreshed, making it more timely
  - Data can be downloaded in different formats
  - Open data portals contain analytic tools for end-users

Perceived benefits of open data

- External researchers can have improved access to data
  - Scientific research beyond what agencies can do in-house
  - Pilot studies
  - Mechanism to develop new collaborations between public health practitioners and academic partners

- Using data to improve healthcare delivery and the built environment
  - Promote data-driven improvements in healthcare delivery
  - Empower local communities to take action on public health issues

Perceived benefits of open data

- Improved health literacy
  - Promote awareness of health issues
  - Improve consumer decision-making (e.g. locating providers, selecting restaurants with fewer health code violations)
  - Increase awareness of the value of public health activities
- Reaching new audiences
- Creating new applications
- Promoting government transparency and fairness

Challenges to releasing data

- Human resources
  - Reductions in public health workforce
  - Limited ability to reassign grant-funded staff to open data activities
  - Different technical skills required to release open data
- Cultural resistance
  - Breaking down data silos
  - New business model for creating and publishing data
- Legal and regulatory issues
  - Complex set of overlapping federal and state laws and regulations
  - Only data owners have authority to release data

Challenges to releasing data

- Data and metadata quality
  - Need high quality and timely data, with clear metadata
  - Tension between maintaining value and minimizing disclosure risks
  - Lack of standard definitions for data elements limits interoperability
  - Agencies relying on local partners to collect data have less control over data quality

- Technical
  - Extracting data from legacy systems
  - Demand for improved open data platform software, e.g. more sophisticated analytic capabilities, more user-friendly interfaces, enhanced methods to automatically update data
  - Technical errors when uploading data to portals

Challenges to releasing data

- Knowledge gaps
  - Understanding goals and activities of open data teams
  - How to use open data platform technology
  - Methods to appropriately de-identify data, maintain confidentiality, and perform expert determinations
  - Identifying different end-users and their data needs

- Addressing the needs of diverse end-users
  - Need to train end-users to use the platform to discover data, conduct analyses, and interpret findings appropriately
  - How to meet needs of multiple end-users with different demands and skills

Successful strategies/lessons learned

- Executive leadership and “high-level champions” critical
- Devote sufficient resources to develop platforms and ensure sustainability
- Develop a strategy to overcome cultural resistance
  - Understand organizational culture
  - Establish buy-in by meeting with staff , working *with* data owners, and identifying “early win” datasets
  - Provide ongoing status reports to show impact

Successful strategies/lessons learned

- Develop processes to facilitate legal review
  - Gain knowledge of de-identification methods
  - Create transparent legal review process
  - Foster understanding about legal considerations

- Think strategically about improving impact
  - Understand audience and tailor data products
  - Start small with 5-10 “high-interest” datasets that are easy to release to demonstrate value and create a tipping point
  - Use continuous feedback to improve value and prioritize future datasets to release

- Don’t reinvent the wheel

Limitations

- Case selection only includes early innovators
- Data are *perceptions* of key informants
- Over-representation of key informants with positive attitudes about releasing open data
- Potential researcher bias
- Output is description of potential benefits, challenges, and lessons learned– not a representative range of beliefs

Translating results to practice:

how to increase the value of open health data and make the movement sustainable

Implications for policy and practice

- Government agencies have little guidance on how to release open data for different user communities
- Open data are only valuable when used
- A fledgling open data ecosystem is emerging, with many opportunities to shape its future and improve data portals, data quality and usability, and data release strategies
- Although the current policy climate supports the open data movement, need to demonstrate return on investment
- Sustained effort on improving the usability and quality of open data is necessary for improving their value for public health

Preliminary recommendations

- Improving the ***quality*** and ***usability*** of open data
  - Actively engage consumers to understand end users, including their desired data, format, and platform functionalities
  - Focus on standardizing data elements with consistent definitions, aspiring for interoperability
  - Create high-quality and standardized metadata for end users
  - Make data more discoverable by posting to open data portals and using key words to facilitate searching
  - Improve usability by making data readily available in different open formats, e.g. csv instead of SAS or Access
  - Continue to develop improvements in open data platform software to provide analytic capabilities to users and facilitate data uploads
  - Invest in technologies and staff training to assess disclosure risks, to maintain value when de-identifying data

Preliminary recommendations

- Increasing the ***impact*** of open data
  - Start small, with "high value" datasets
  - Release data with public health messaging; promote through public affairs
  - Continue to catalyze the open data movement and "disruptive innovation" through events such as the Health Datapalooza and code-a-thon challenges
    - DHHS Office of the National Coordinator already plays an important role
    - ASTHO and NACCHO could play a role in targeting public health practitioners
  - Continue conversations about how to improve data quality and design data systems that consider future data publication needs
  - Don't reinvent the wheel– talk to other jurisdictions; learn about their platform software, metadata forms, legal review processes, etc.; and adapt their methods
  - Publicize use cases from early leader jurisdictions

Preliminary recommendations

- Ensuring the ***sustainability*** of open data
  - Strong leadership is critical to create a vision, acquire resources, and maintain focus on open data
  - Need to change culture around data sharing
  - Work closely with data owners to get buy-in and improve data products
  - Create an open data handbook to communicate a vision and establish transparent governance
  - Develop standardized processes, e.g. metadata templates and expert determination forms
  - Commit sustained investments in human resources and technology
  - After establishing a new open data site, move it from a "special project" to a program area to make it a routine public health activity
  - Collect ongoing feedback to continuously improve open data and communicate early successes to agency staff and the public

Questions?

Email:

emartin@albany.edu

For additional project information:

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

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, 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*

Aim 3:

pilot study to use open data to evaluate the geospatial relationship between childhood obesity and the built environment

Collaborators: Gus Birkhead, Christine Bozlak, Ian Brisette, Ade Dwicaksono

Research design overview

- Utilize open resources to assess which characteristics of the built environment are associated with school district-level indicators of student overweight and obesity in New York
  - Student Weight Status Category Reporting System recently used by media to highlight geographical disparities in childhood obesity

- "Use case" to demonstrate whether open data can be used for public health research, and to document technical difficulties of using open data for linkage projects

New York-based open datasets used

- Variables
  - Outcome: proportion of school-aged children who are obese/overweight
  - Measures of built environment
  - Control variables: demographics, socioeconomic status

- Population and unit of analysis
  - School districts
  - All New York State districts, excluding New York City

New York-based open datasets used

- Data sources, selected because they contained relevant variables and could be merged at the school district level
  - Student Weight Status Category Reporting System (NYS Dept. of Health)
  - Student Report Card Database (NYS Education Dept.)
  - Food Service Establishment Inspection Data (NYS Dept. of Health)
  - Retail Food Store Data (NYS Dept. of Agriculture and Markets)
  - Farmers Markets in New York State (NYS Dept. of Agriculture and Markets)
  - EPA Smart Location Dataset (US Environmental Protection Agency)

Analytic methods

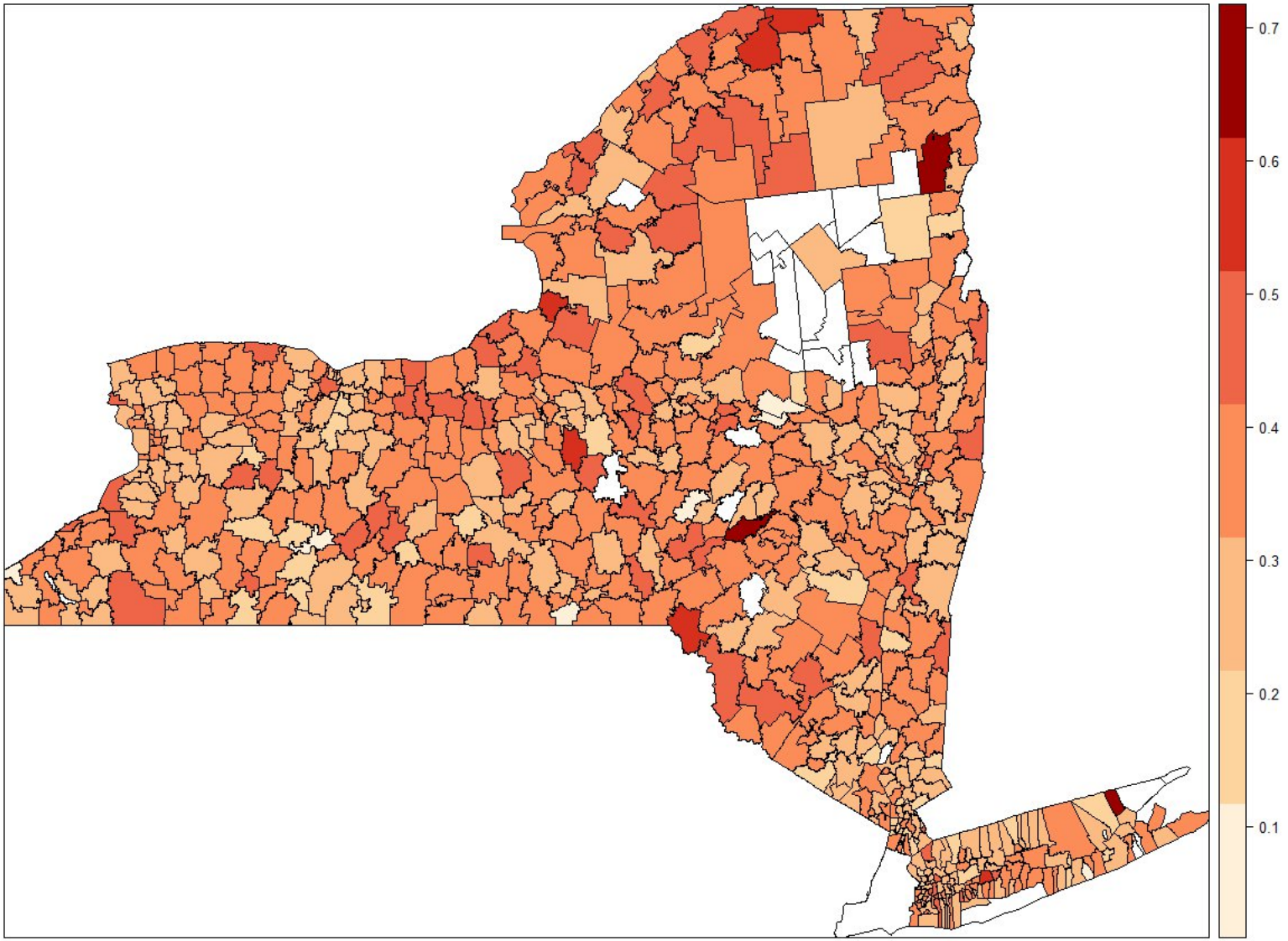
- OLS regressions of determinants of school-district level obesity
  - Separate models for obesity rates among elementary and middle/high school students
  - Sensitivity analyses for % overweight

- Geographically weighted regressions to evaluate how associations varied across local regions

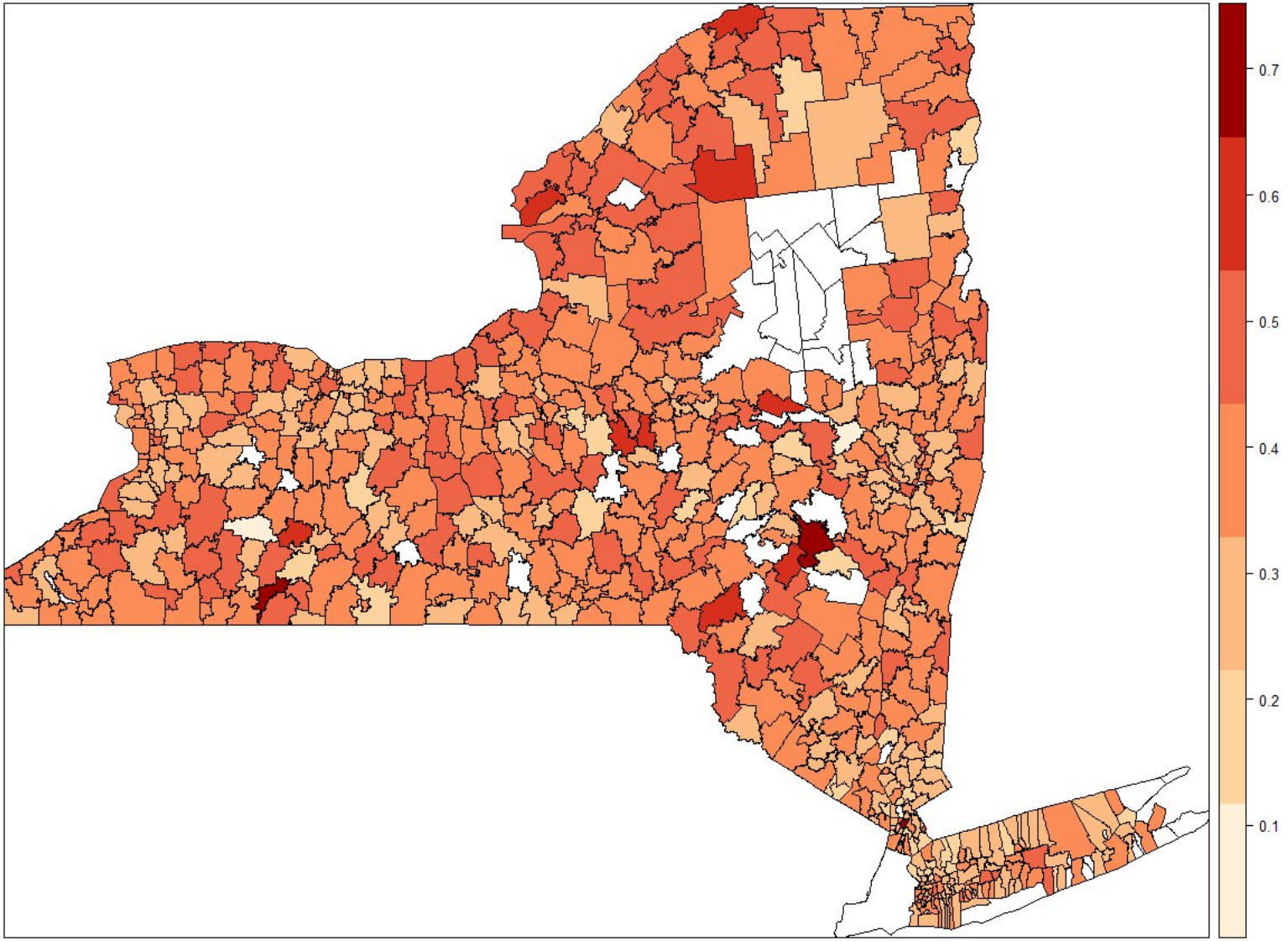
Preliminary findings

- Wide variation in rates of obesity (1.3%-50.0%), with lowest rates among suburban school districts
- Socioeconomic status and urbanicity (rural, suburban, urban) of school districts are key predictors for childhood obesity
- Minimal effects of built environment measures after adjusting for urbanicity
- Existing data may not fully capture all dimensions of built environment, or else suggest a more complex interaction between the built environment and childhood obesity

Proportion of elementary school students who were obese and overweight, 2010-2012



Proportion of middle/high school students who were obese and overweight, 2010-2012



Findings: suitability for research

- Many datasets readily available for public health research
  - Can use data creatively to evaluate multiple dimensions of the built environment (e.g. using restaurant inspections data for fast food availability)
  - Can synthesize data from different domains (health, agriculture, education)
- Challenges consistent with findings from prior aims
  - Lack of standard definitions for data elements severely constrains interoperability and ability to merge by geographic identifier
  - Incomplete metadata, e.g. missing codebooks
  - Data quality, e.g. incomplete addresses, inconsistent location descriptions
  - Data timeliness
  - High level of geographic aggregation limits value
  - Some data not easily discoverable (or available) in open data platforms
  - Data not yet 5-star, e.g. downloadable in multiple non-proprietary formats and with links to provide context
  - Limited usability, e.g. advanced statistical skills required to recode data

Student weight status category reporting system

Advantages

- Available in most accessible format (csv)
- Can be searched and downloaded from Health Data NY
- API makes download process easy and highly customizable

Challenges

- Nonstandard school district identifier variable cannot be linked to other datasets
- Name of school districts in nonstandard format (e.g. abbreviations vs full names)
- No codebook to describe variables
- More recent dataset has more documentation, but unit of analysis is at county level
- School district is high-level aggregation

Student report card database

Advantages

- Rich information, updated regularly
- Past versions available, allowing for trend analysis
- Covers all education-related entities (e.g. counties, school districts, BOCES, schools)

Challenges

- Cannot be discovered in Open NY, although available on NYS Department of Education website
- Only available as an Access database, requiring special procedures to download and process for statistical packages
- Uses an entity/school district ID that cannot be linked to other datasets
- Uses a school district naming system that is not completely consistent with National Center for Education Statistics or Census Bureau

Food services establishment inspection data

Advantages

- Available as comma separated value (csv) file, which is very accessible
- API facility in Health Data NY makes downloading process easy and customizable
- Rich data, containing all inspection results from 2005

Challenges

- Unreliable geocodes
- Many observations have incomplete addresses
- No data dictionaries explaining the variables, including definitions of different establishment types
- Inconsistent restaurant names
- Some geographic areas excluded

Retail food store database

Advantages

- Available as comma separated value (csv) file, which is very accessible
- API facility in Health Data NY makes downloading process easy and customizable

Challenges

- Unreliable geocodes
- Many observations have incomplete addresses
- No data dictionaries explaining the variables

Farmers markets in NYS

Advantages

- Available as comma separated value (csv) file, which is very accessible
- API facility in Health Data NY makes downloading process easy and customizable

Challenges

- Unreliable geocodes
- Many observations have incomplete addresses
- No data dictionaries explaining the variables

EPA smart location dataset

Advantages

- Data at census block level, a very small unit of observation
- Uses standard Census Bureau geographic identifier, which facilitates merging with other data
- Complete and readable data dictionary

Challenges

- Timeliness- last updated in 2010
- Need to use specialized Clip N Ship API to restrict dataset to single state
  - Assumes high proficiency in dataset cleaning and GIS
  - Not ideal for discovering data