Keeneland Conference
Hosted by the National Coordinating Center PHSSR
April 7, 2014
Lexington, KY

Open Data Priorities: Aligning Public Health Researchers' Needs with Agencies' Organizational Capacities to Release Data

Erika G. Martin PhD MPH[1], Natalie Helbig PhD MPA[2], Guthrie S. Birkhead MD MPH[3]

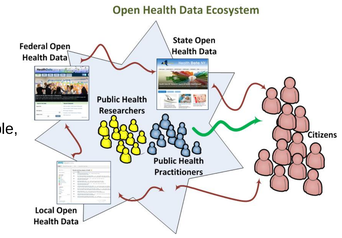
[1] Rockefeller College of Public Affairs & Policy, University at Albany-State University of New York (SUNY); Nelson A. Rockefeller Institute of Government-SUNY; [2] Center for Technology in Government, University at Albany-SUNY; [3] New York State Department of Health; School of Public Health, University at Albany-SUNY

Contact: emartin@albany.edu

Supported by the New York State Health Foundation and the Robert Wood Johnson Foundation

OPEN DATA FOR PUBLIC HEALTH SERVICES AND SYSTEMS RESEARCH (PHSSR)

- **Open government data is a new resource that could potentially accelerate PHSSR**
  - Motivated by President Obama's Memorandum on Transparency and Open Government (2009)
- **Features of open government data:**
  - Free of charge
  - Available to the public (often through data portals)
  - Accessible in multiple formats, including API-enabled for developers
  - Unlimited and unrestricted use and distribution
- **Researchers are part of emerging "open data ecosystem"**
  - Metaphor to describe the set of interdependencies among people, data, technology, and innovation in a particular context
- **Open data platforms have the potential to build health departments' capabilities to serve more PHSS researchers**
- **Government agencies have little guidance on what types of data and how to release data that are usable for researchers**



RESEARCH QUESTIONS

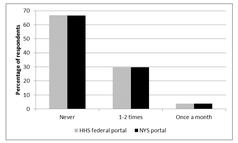
- **What characteristics do health researchers consider in determining an open dataset's fitness for use?**
- **What are the practical challenges of obtaining and using open health data?**
- **What opportunities do open data platforms provide for health researchers?**

ANALYTIC METHODS

- **Data were collected as part of a workshop on open health data in New York State**
  - Health policy leaders discussed: using data-driven research to advance public health in NYS; the past, present, and future of open health data in NYS; and the potential for open data to improve the health of NYS residents
  - Afternoon break-out sessions brought together academic researchers and NYSDOH practitioners to explore their research priorities and data needs
- **Pre- and mid-workshop surveys collected information about participants' awareness and use of open data portals, past collaborations with NYSDOH, and workshop goals**
  - Pre-workshop survey collected information about participants' background, specialty areas, and quantitative data analysis techniques
  - Mid-workshop survey collected information about what participants learned and/or found useful for teaching or research, and planned uses of the information
- **Two facilitated focus group discussions examined research priorities and data wish lists**
  - Seven groups of academic researchers and NYSDOH practitioners, arranged based on stated areas of interest, geographic diversity, and methodological diversity (from pre-workshop surveys)
  - First session: research priorities over next two years, extent to which relevant data is currently available, barriers to obtaining the data
  - Second session: desired data (data types, data sources, data categories, specific variables), known sources of these data, keywords to describe data, how data could be collected if not yet available
- **Paper survey and focus group transcripts reviewed to identify major themes and concepts**
  - Researchers' conversations compared to speakers' prepared remarks to identify: data needed by health researchers and for what purposes; barriers to obtaining and using health data; and potential solutions and opportunities for both researchers and NYSDOH

FINDINGS: LOW USE OF OPEN DATA PORTALS AND LIMITED PREVIOUS COLLABORATIONS WITH DEPARTMENT OF HEALTH

- **Open data ecosystem has not yet fully integrated public health researchers and NYSDOH practitioners**
  - Half of survey respondents reported no previous collaborations with NYSDOH; only a few reported "multiple engagements"
  - Two-thirds had not used federal or state open data portals prior to hearing about the workshop



FINDINGS: CHARACTERISTICS OF IDEAL OPEN HEALTH DATA

- **Participants expressed interest in datasets that were: geocoded, longitudinal, and at small area granularity**
- **Participants discussed the importance of linking multiple types of data and sources:**
  - Across data "types" (e.g. claims, surveillance, electronic health records)
  - Across data "sources" (e.g. multiple insurance claims or multiple clinics; all-payer data)

FINDINGS: CURRENT ENVIRONMENTAL CONDITIONS AND BARRIERS TO USING HEALTH DATA FOR RESEARCH

- **Legal constraints to using data**
  - Lack of clarity and education on how to meet federal HIPAA and IRB regulations
- **Administrative and institutional constraints**
  - Complexity of data agreements and human subjects protections (e.g. IRB, DEAA, MOU)
  - Length of time to obtain datasets from NYSDOH
  - Universities' reluctance to store data on servers to avoid liabilities
- **Technical constraints**
  - Data storage and processing for big data; time to clean and link data; addressing cyber security
- **Financial constraints**
  - Limited NYSDOH staff to prepare datasets and conduct research; funding pressures for researchers
- **Political constraints**
  - Data silos within NYSDOH, leading to challenges when handling data collected by multiple jurisdictions, collaborating across agencies, and working with multiple stakeholders
  - Data ownership and reluctance to release certain data, e.g. cost data
- **Inconsistencies across datasets**
  - Need standardized common data elements and controlled vocabulary ontology for tags and metadata
- **Limited information about what data are available at NYSDOH, behind the firewall**
  - Capacity to search for appropriate datasets and compare data elements across datasets
  - Details on variables, quality, completeness; interactions with practitioners to understand data

FINDINGS: POTENTIAL OPPORTUNITIES FROM OPEN HEALTH DATA AVAILABLE ON *WWW.HEALTH.DATA.NY.GOV*

- *"Things are changing for the better, a person like me who has spent 40 years [in this field], I've never seen a more dramatic and fast change for the better. I look at this and it's amazing."*
- *"If the open healthcare data system is a success… we could answer questions – and I don't yet know what the specific questions are – about cost of health, and the location of health issues, and the prevalence of health issues that we can't easily answer today… knowing what's happening will let us make that [next] policy step." … "It's about getting the big picture through the big data."*

WHAT IS AVAILABLE VERSUS DESIRED: IDENTIFYING THE GAPS

- **Generally, participants did not differentiate between open health data and health data**
  - Open data concept originates from government transparency movements, not researchers
  - Researchers may be thinking of "open science," not government transparency
  - Semantic confusion with term "open": what some participants considered to be "open data" (i.e., ICPSR website or national health interview surveys) are not "open data" from a technical view
- **Participants had an ideal set of data characteristics that they require to answer research questions**
- **Disconnect between how participants categorize topics and how NYSDOH organizes staff and data**
- **Gap between ideal data characteristics and what is feasible to release through an open data portal**

TRANSLATION TO PRACTICE: IDEAS FOR NEW YORK STATE'S OPEN HEALTH DATA TEAM

- **Continue to promote the *Health Data NY* data portal**
  - Promote awareness and community through media attention, social media, webinars, and other events
  - Pay attention to changing researcher environment (e.g., faculty turnover)
- **Create "fit for use" profiles for different types and levels of researchers**
  - Open data not ideal for some researchers (e.g. studies required longitudinal linked data at the individual level), indicating a need to develop strategies for tiered data releases
  - Use site analytics, user surveys, or other feedback mechanisms to prioritize data releases
  - Continue to focus on improving meta data and searchable key terms that resonate with users
- **Connect more closely the "opening health data initiative" with the state's overall vision for providing access to *all* NYSDOH data resources**
  - Provide capability for researchers to "browse" NYSDOH data resources, including those behind the firewall (does not require *access* to actual data)
  - Create a NYSDOH data asset map for external use, including: data collection instruments, data dictionaries, explanation of how raw data get transmitted to NYSDOH data warehouses, and key contacts
- **Continue to foster the development of an open health data ecosystem**
  - Develop engagement strategies with researchers (e.g., forums, sabbatical program, internships)
  - Assess internal capacity and mechanisms for engagement, including existing collaborations
  - Tap into universities' existing resources (e.g., SUNY Health Center for Excellence, service learning courses)

STUDY LIMITATIONS

- **This was a pilot study with several limitations**
  - Non-representative convenience sample
  - Focus group output may be influenced by facilitators, the artificial environment, or one or two individuals dominating the session

AREAS OF FUTURE RESEARCH

- **Findings will be used as background for the mentored research scientist development award**
  - Aim 1: To systematically review a random sample of open health datasets in local, state, and federal portals to enumerate characteristics of open data, its usability for PHSSR, and differences across jurisdictions
  - Aim 2: To conduct semi-structured qualitative interviews with practitioners at NYSDOH and other agencies releasing open health data to assess the capability needed to, and the value of, integrating health researchers into their open health data ecosystem
  - Aim 3: To use available open health data to complete a pilot data linkage project on whether the prevalence of childhood obesity in NYS is associated with aggregate measures of the built environment, and enumerate specific open data challenges experienced with the analysis
  - Aim 4: To synthesize findings from aims 1, 2, and 3 to make recommendations for building a robust community of practice oriented toward the use of open health data for PHSSR

The Nelson A. Rockefeller Institute of Government — The public policy research arm of the State University of New York

Center for Technology in Government

University at Albany — State University of New York

New York State Department of Health

NYS Health Foundation — Improving the state of New York's health

Robert Wood Johnson Foundation

Open Health Data Open Opportunities

For workshop details, presentation videos, and handouts, visit: www.rockinst.org/ohdoo