Big Data

Key Points

- New techniques and attitudes will need to be developed to deal with both volume and variable quality of data likely to be received by PHAs over the coming years.
- Support for "big data" brings a number of considerations to PHAs which most likely require new ways of thinking about information and information processing, including consideration of cloud computing, use of HIEs as intermediaries, and distributed query rather than traditional data submission.
- To best leverage activities around them, PHAs need to invest in trained informatics and analysis staff to develop relationships, and monitor and participate in local and national activities.
- An agency-wide information/informatics vision is critical in an era of increasingly digital healthcare data and increasing volumes, variety, and velocity of all data sources.

As more and more data flows in and around the health and healthcare ecosystem, public health agencies (PHAs) need to be prepared to receive, process, and analyze more and more data, some of it qualitatively different than current data streams. The tools and skills necessary to do this may differ from current tools and training. Just because there is *more* data does not mean it is necessarily *better* data. New techniques and attitudes will need to be developed to deal with higher volumes, diverse data sources, and variable data quality.

"Big data" usually refers to the rapid growth in volume, variety, and velocity of data available to an industry or field. For public health:

- The increased **volume** comes as healthcare data is increasingly made digital, and digital data can more easily be sent to public health automatically based on business rules. For instance, various studies have shown or estimated a two- to five-fold increase in reportable condition reporting based on automated data reporting from EHRs or Health Information Exchanges (HIEs).[1]
- The expanded **variety** of data will in part come of new sources of healthcare data, but also from data sources related to other determinants of individual and population health: economic, retail, transportation, public safety, environmental, even recreational sources of data will take on greater importance to PHAs as we strive as a nation to improve population health outcomes and reduce disparities.[2] Social media has the potential to provide timely

[1] J. Marc Overhage, Shaun Grannis, and Clement J. McDonald, *A Comparison of the Completeness and Timeliness of Automated Electronic Laboratory Reporting and Spontaneous Reporting of Notifiable Conditions*, Am J Public Health. 2008;98(2):344-350.

[2] *For the Public's Health: Investing in a Healthier Future*, Institute of Medicine, 2012.

insights into public concerns and awareness (or the lack of). Many of these new sources may be unstructured data, or use standards that are not currently part of the health information arena.

- The increased **velocity** of data comes in part from the automatic flows of digital data, but also how fast it is being created and must be processed to meet public or other expectations.
- Lastly, the **variable quality** relates to the reality that data coming from varied sources and created for varied purposes will not necessarily meet public health's usual standard for structure, completeness, or other factors. Ignoring such data streams, however, could mean not having a complete picture of community risks or assets, or situational awareness about emerging events.

The sheer volume of data has the potential to overwhelm PHAs – their staff, networks, and systems. As older systems mature, their data sets grow and grow; rarely is older data removed, purged, or archived as patient-oriented systems become lifelong repositories and environmental systems accumulate historical data for more accurate trends. Twenty years ago these systems were new and their data sets were often built on a go-forward basis. Now, these ever-increasing databases continue to expand, resting on the foundation of years of routine, successful operation.

Support for "big data" brings a number of considerations to PHAs which most likely require new ways of thinking about information and information processing, including:

- **Evolving server platforms:** As data sets become bigger and bigger, traditional data centers may become too expensive or too inflexible to expand (and potentially contract) as data needs change. Cloud computing allows for computing services and capacity to expand (or shrink) according to users' needs with little impact on the users' experience. Cloud computing services are configured on special, network-accessible platforms so that end users are shielded from the technical issues related to their physical configuration. With cloud-based resources, PHAs can purchase a flexible quantity of computing services and not worry about provisioning, operation, or availability. This will be an important consideration moving forward.
- **Management of a proliferation of data sources:** As more data originates in electronic form, so, too, do the number and variety of data sources. The administrative overhead and technical

Case Study

The BioSense program[3] tracks health problems in the United States as they evolve. It provides public health officials with the data, information, and tools needed to better prepare for and coordinate responses to safeguard and improve the health of Americans. Analysis of data through BioSense provides insight into the health of communities across the country. Such data are vital to guide decision making and actions by public health agencies at local, regional, and national levels. Using the latest technology, BioSense 2.0 integrates current health data shared by health departments from a variety of sources to provide insight on the health of communities and the country. This distributed environment, governed jointly by state, local, and federal representatives, provides local and state stakeholders secure data storage space and analytics tools at no cost to them. Most importantly, it provides a collaborative shared environment to advance public health surveillance practice and activities.

http://www.iom.edu/Reports/2012/For-the-Publics-Health-Investing-in-a-Healthier-Future.aspx
[3] Compiled from http://www.cdc.gov/biosense/ and subsidiary pages.

infrastructure required to manage a growing number of data relationships may quickly overwhelm PHAs. HIEs can help reduce this burden by acting as intermediaries for healthcare data collection and transport. But this new intermediation brings with it certain other complexities, including differences in authentication and authorization for data connections, changes in PHA-provider relationships, and new patient matching/linking considerations. Much data of growing interest to public health, however, does not originate in the healthcare sector. Patients and citizens represent new and very different sources of data for PHAs (see *Consumer Engagement* issue brief). Other sources of community health or environmental data, as noted above, represent even more sources. This proliferation of sources requires a comprehensive, agency-wide vision for information management and use. It also requires more tools to automate data processing tasks so staff has time to focus on effective use of the data for decision-making and intervention.

- **Distributed query:** Traditionally, PHAs built operational data stores (ODS) and data warehouses to hold data collected from disparate sources. As more clinical data originates in electronic form, emphasis will likely shift over time to data on demand. Through the use of query "agents," PHAs will send out requests for data from their sources and aggregate data received as responses to these electronic queries. Expectations about timing, completeness, and consistency of data may need to be adjusted to accommodate this new data access paradigm.
- **New analytical requirements:** With more data come additional requirements, training, and technologies to process and analyze data, including an increasing need to focus on data semantics and potentially clinical decision support (CDS) services to aid analysis, even at a population level (see related Research Briefs in this series). PHAs will need to arm themselves with new tools, skills, and techniques to ensure that the needs of "big data" allow sufficient time for analysis.

Some things will remain remarkably the same:

- **Multiple, simultaneous levels of analysis** in that PHAs will continue to move between aggregated population-level data and individual case data, as both are required.
- **Standards** continue to evolve and develop as the needs for interoperable data supersede the tendency to fixate on specific jurisdictional requirements.
- **Data privacy** will continue to be a big part of the national health information landscape as greater opportunities to share data will yield greater risks to its appropriate use.
- **Metadata** will still be required to describe these data resources and guide users how best to use and interpret the information represented.

Data for its own sake is limited in its usefulness. More data, more complex data, and more varied data, will bring potential new opportunities to *share* data more than ever before with traditional data sharing partners and others (like consumers!).

Action Steps for State and Local PHAs

- Set priorities among data projects, and look for opportunities to leverage activities within the agency, jurisdiction, and community, as well as regional and national activities. This requires trained informatics staff members who are able to juggle a tactical and strategic focus within the agency, develop relationships with local colleges, universities, and health systems, and monitor and participate in relevant national and local professional associations and standards development/harmonization activities.
- Identify other sources of data that contribute to a comprehensive understanding of community health. Evaluate these sources for timeliness, completeness, and value in contributing to a more complete picture.
- Establish an agency-wide task force to create an overall information/informatics vision that can guide the acquisition of new data, the processing and management of current data streams, and the IT infrastructure/platforms needed to support ever growing volumes and variety of data.

Leadership Steps for National Agencies and Organizations

- Identify resources, tools, and training opportunities for PHAs to assist with "big data" requirements.
- Assist state and local PHAs in understanding the policy and legal issues that emerge from "big data" activities.

More Information

http://ajph.aphapublications.org/doi/pdf/10.2105/AJPH.2011.300542
https://sites.google.com/site/biosenseredesign/?pli=1
http://wiki.siframework.org/Query+Health

This paper is part of a series of information briefs for local and state public health officials and managers, developed by the Joint Public Health Informatics Taskforce in partnership with HLN Consulting, LLC. The full series of seven briefs can be downloaded at no cost from www.jphit.org.