Why "What Data Are Necessary for This Project?" and Other Basic Questions are Important to Address in Public Health Informatics Practice and Research

Brian E. Dixon, MPA, PhD[1,2]; Shaun J. Grannis, MD, MS[2,3]

1 Indiana University, School of Informatics, Indianapolis, IN;
2 Regenstrief Institute, Indianapolis, IN;
3 Indiana University, School of Medicine, Indianapolis, IN

Abstract

Despite the likelihood of poor quality data flowing from clinical information systems to public health information systems, current policies and practices are pushing for the adoption and use of even greater numbers of electronic data feeds. However, using poor data can lead to poor decision-making outcomes in public health. Therefore public health informatics professionals need to assess, and periodically re-evaluate, the quality of electronic data and their sources. Unfortunately there is currently a paucity of tools and strategies in use across public health agencies. Our Center of Excellence in Public Health Informatics is working to develop and disseminate tools and strategies for supporting on-going assessment of data quality and solutions for overcoming data quality challenges. In this article, we outline the need for better data quality assessment and our approach to the development of new tools and strategies. In other words, public health informatics professionals need to ask questions about the electronic data received by public health agencies, and we hope to create tools and strategies to help informaticians ask questions that will lead to improved population health outcomes.

Key Words: ***Public Health Informatics, Systems Analysis, Information Management***

Introduction

Data are the lifeblood of the knowledge economy and health care. Google, Microsoft, and other modern companies compete with innovative solutions that support the generation, analysis, management, and exchange of data. Health care and public health similarly focus efforts on generating, collecting, analyzing, and sharing data about individual patients and populations, respectively. Data are vital to joint activities including surveillance of chronic and communicable disease, population health assessments, and health care policy. Effective practice requires access to representative, complete, and timely data from multiple sources (1, 2).

Unfortunately the quality of the data stored in information systems across industries and organizations is often poor (3). Typical data quality issues encountered include: inaccurate data, inconsistencies across data sources, and incomplete (or unavailable) data necessary for operations or decisions (4). For example, a large bank found that data in its credit-risk management database were only 60 percent complete, which necessitated additional scrutiny by anyone using its data (5). In health care, the completeness of data in electronic health record (EHR) systems has been found to vary from 30.7 to 100 percent (6).

Evidence from the information management literature on the impacts of these issues is sparse, but estimates of impacts include: increased costs ranging from 8-12 percent of organizational revenue, and up to 40-60 percent of a service organization's expenses consumed due to poor data; poorer decisions that take longer to make; lower data consumer satisfaction with information systems; and increased difficulty in reengineering work and information flows to improve service delivery (4). Impacts on health care include ill-informed decisions when humans or machines use poor quality data inputs from EHR systems (7, 8).

The public health literature suggests additional impacts of poor data quality issues for public health agencies and processes. For example, spontaneous reporting rates for infectious diseases range from 9 to 99 percent and have remained relatively unchanged from 1970 – 2000 (9). While some conditions, such as sexually transmitted infections, are reported approximately 80 percent of the time, many conditions (e.g. pertussis, shigellosis) are reported less than half of the time. Timeliness, another attribute of data quality (3), has also been found to be a challenge in public health reporting (10). Delays in the receipt of notifiable disease data (timeliness) and the lack of a complete set of reports (completeness) impact public health agency surveillance processes, including but not limited to the ability of agencies to respond to emerging disease threats.

Although the available evidence suggests that data quality issues are real and can have significant impact on operational processes, personnel, and budgets, too often public health practitioners assume that the increasing flow of electronic data from various information systems used in modern practice are of equal quality. This can lead to suboptimal outcomes beyond those described above, including failed implementations of public health information systems and inefficiencies in data collection and analyses. To ensure the information and knowledge generated from electronic observational clinical data can support effective public health practice, public health agencies must develop effective strategies to accurately characterize electronic clinical data sources both during the initial deployment phase and routinely reassess data characteristics of the operational system on a regular basis. In this paper, we outline the practical needs regarding careful analysis and continuous re-examination of data and their sources. We then describe how our U.S. Centers for Disease Control and Prevention (CDC) Center of Excellence in Public Health Informatics is working with public health informatics stakeholders to develop a framework that will support the operationalization of analyzing data quality continuously within an organization. We conclude with thoughts on the implications for such a framework and our next steps in the development and validation of the framework.

The Importance of Understanding Data Needs in Public Health

Developing methods and operational practices for assessing data quality requires an understanding of the various public health business processes, as well as the context in which those processes occur. Business processes are sometimes referred to as "use-cases." To

understand a particular business process or use-case, one must ask questions about what, where, when, why, and how data are collected, stored, shared, and used to support the activities performed by public health and health care professionals involved in the business process. An example of a public health business process is syndromic surveillance.

Syndromic surveillance detects initial manifestations of disease before diagnoses (clinical or laboratory) are established (11-13). Data and information in syndromic surveillance systems come from a variety of sources, including hospital emergency department visits, ambulatory clinic visits, school absenteeism, poison control centers, and over-the-counter medication sales (2). In syndromic surveillance, agencies have an inherent need to understand the quality of the data from emergency departments, schools, and a variety of other sources to assess whether the data are able to accurately describe emerging public health threats. Poor quality data will lead epidemiologists down dead end pathways including false-alerts or missed alerts, which will waste scarce public health resources.

Initially syndromic surveillance focused solely on patients' chief complaints, the primary reason for their visit to a health care provider, as they presented in the emergency department (ED). Chief complaints were found to be generally representative, complete, sensitive, specific, and reliable enough to detect emerging outbreaks including the start of the influenza season and bioterrorism events (2). Subsequent research demonstrated that grouping chief complaints into syndromes and adding discharge diagnosis information into syndromic surveillance analysis improves validity and reliability (14-16). The current effort to improve the meaningful use (MU) of EHR systems incentivizes ED and ambulatory providers to submit chief complaint data to public health (17, 18). Before leveraging chief complaint data from all EHR source systems, the quality of new sources must be assessed for completeness, validity, and reliability. For example, chief complaints reported by specialty providers, like thoracic surgeons, may not be as useful for detecting emergent threats as those from the ED or full-spectrum primary care clinics. Data from specialty providers likely are predominantly grouped into a single syndromic category and therefore likely to lack specificity and sensitivity, an issue under investigation now by the Indiana State Department of Health and other health departments across the U.S.

In addition to syndromic surveillance, MU is driving health care providers to increasingly submit electronic laboratory reporting (ELR) and immunization data to public health agencies. This will dramatically increase the number of data sources providing electronic data to public health agencies over the next 5 to 10 years. For example, in Indiana where we have steadily increased the number of providers utilizing ELR to report communicable disease cases to the state health agency over the last 10 years, we now have over 200 discrete data feeds from laboratory information systems, and expect that number to increase substantially. Continuously analyzing data quality across those feeds is a necessary but challenging task. ELR monitoring is especially challenging given the frequency with which laboratories update and append new tests to their catalogues (19, 20).

Beyond meaningful use, public health agencies must collect, manage, and utilize data and information across a number of program areas. Each program area may impose unique requirements on data captured through a general or shared collection method. Recently the state health agency in Indiana sent a letter to several hospital laboratories highlighting the lack of guarantor information (e.g., name and contact information for the person financially responsible for health care expenses) when submitting ELR information to the state's communicable disease

program. The letter originated not from the communicable disease program director but a different program within the agency responsible for following up on case reports for minors. Some public health officials perceived the traditional patient contact information in the ELR messages as insufficient and decided to instead use guarantor information which typically contains parental or legal guardian's contact information. However, these decisions were not shared with the data providers nor laboratory directors until after the state had performed an analysis to determine which hospitals were sending the guarantor information and which ones were not. A lack of communication about data needs and uses across program areas within the agency resulted in confused laboratory directors contacting vendors and data partners about the letters from the state. Furthermore, it's important to note that the laboratory information systems (LIS), which are commonly used to process lab results, do not readily capture patient guarantor information. Instead, this is often managed by enterprise billing software. Consequently, supplying guarantor information likely requires non-traditional integration of two information systems, which may require new processes outside of the typical ELR information flow.

Furthermore many public health professionals, like data consumers in other health care segments and industries, believe that data is easily and uniformly captured and stored across the spectrum of health care services. However, data are captured for a specific purpose, and the collection of additional data elements is costly. Additional data elements require staff to ask for and then record the information, which translates into additional time and labor. Software systems often must be modified to accommodate the new data fields and new workflow, also a costly process. Therefore data consumers must understand the impact of the cost of data collection on the characteristics of data captured in various environments, like their completeness, when making decisions about secondary use. Public health officials, for example, might benefit from understanding that elements like the provider's phone number and address have little clinical relevance to the physician receiving the results of a lab test. These fields are poorly populated by laboratory information systems. Although these fields can be required according to state and federal regulations, it does not guarantee that they will be complete and available for public health surveillance processes. For example, provider addresses and phone numbers are missing in more than 95% of the electronic laboratory messages sent directly from lab systems to the Indiana Network for Patient Care (INPC), a health information exchange repository that receives lab data from over 70 hospitals and their send out facilities (21). Thus policies to require additional data elements are unlikely to impact data collection processes unless laboratories and hospitals are incentivized to capture the additional data elements needed for public health surveillance processes.

The above set of issues illustrates a clear need for public health agencies to assess and document their data needs and sources. Without clear expectations regarding public health's desired needs and uses of data, information systems will likely receive, analyze, manage, and generate poor quality data, information, and knowledge about population health. One pressing role for public health informatics professionals is clear: they must support agencies' assessment of data needs and sources. This means asking questions about the what, how, and why of existing business processes as well as the data, and their sources, involved in those processes. Informatics professionals should further capture and share information with public health stakeholders regarding what can and cannot be feasibly collected from clinical information systems given existing information and work flows. However, there is a paucity of practical tools and resources for informatics professionals to utilize. In the rest of this article, we describe our approach to

developing a framework to support better assessment of needs and sources to support more efficient use of information systems and resources.

What Our CDC Center of Excellence is Doing to Better Understand Data Needs

The Indiana Center of Excellence in Public Health Informatics (ICEPHI) resolved to develop methods and operational practices for assessing data quality from a myriad of clinical data sources. Our initial attempt involved scheduling meetings with a broad range of stakeholders within the Indiana public health community. We invited public health researchers, state health officials, local health officials, and informatics professionals in both health care and public health settings. While the meetings confirmed the real-world challenges previously described, they were suboptimal with respect to outlining the necessary methods and practices for performing data quality assessment. Given the demands of public health personnel, not surprisingly meetings often only contained a subset of the full stakeholders involved in ICEPHI, so it was not possible to have every stakeholder in attendance at any given meeting. Furthermore, participating stakeholders often failed to provide full details about their information needs. This was due to a variety of reasons, including individuals lacking full details for a particular workflow, individuals who are uncomfortable speaking at meetings, and individuals who primarily work with reports but do not fully understand the provenance of the data that comprise those reports. We transitioned from the meeting strategy when it became clear that this approach would not capture data quality perceptions and needs in a systematic fashion that would meaningfully inform standard methods and practices.

To systematically explore data quality perceptions and needs from a diverse group of stakeholders in public health and informatics, we selected a modified Delphi technique (22). The Delphi technique was chosen because this method can 1) structure group communication and decision making processes and 2) foster consensus from very diverse groups of individuals (23). Delphi studies have proven useful in a range of industries and settings, including medicine, nursing, and informatics (23-25).

The Delphi study is being conducted as the third of three phases in our systematic approach to examining data quality perceptions and needs. In phase one, we convened a diverse set of purposefully selected stakeholders from across academia, state health agencies, and local health departments to define and prioritize a limited set of public health work processes that involve data collection from health care providers and information systems. The majority of our stakeholders were selected from Indiana. However, we further invited select individuals from other states who had previously collaborated with our Center of Excellence.

In phase two, informatics researchers within ICEPHI compiled a list of data elements required under Indiana state law to be reported to the health department for the various work processes under discussion. The stakeholders were invited to review this list and propose additional data elements they felt would be useful to have in those work processes. The additional data elements were appended to the respective work process lists. These final lists represent an initial, stakeholder-driven view of the data necessary to optimally perform common work processes in public health practice. These processes included: aggregate surveillance for influenza across counties and the state; individual case reporting for meningococcal disease; and ongoing monitoring of diabetes levels at the county and state level.

The third and final phase of our design involves administering the modified Delphi study. We used the work processes, including their full descriptions and list of data elements, from the prior phases to create an online survey instrument to record the opinions and perceptions of a much wider group of public health stakeholders. Our goal is to recruit 20-30 stakeholders with expertise in various aspects of public health practice, those which align with our work processes, and achieve consensus on the need for and use of the data elements specified in phase two. Individuals in the larger stakeholder group will be able to suggest additional data elements each round, and participants will be able to amend their responses based on a review of the groups collective thinking to date (this is a major component of Delphi studies). In addition to helping ICEPHI more robustly specify the work processes and necessary data elements, stakeholders will also be asked to indicate their perceptions of the quality of data they currently receive from existing sources. We will, for example, ask about the timeliness and completeness of patient contact information provided in existing case reports. Here, too, we hope to achieve consensus on the perceived need for specific data elements to be timely or complete since various data sources will possess different timeliness and completeness characteristics.

To date we have completed the first two phases of our project as well as the survey instrument we will use to conduct the third phase. In the coming months, ICEPHI will recruit a sizable group of public health stakeholders and conduct the Delphi study with this group.

Future Directions and Work

The primary goal of our work is to utilize the product of the study to develop and evaluate systems that can improve public health practice in Indiana. We aim to take the results of the Delphi study and begin applying it to our work within the INPC. We will focus on the prioritized work processes and data elements to ensure the INPC is addressing these adequately for the local and state agencies engaged with data exchange activities. Where deficiencies are found, we will work with ICEPHI stakeholders to improve data capture, analysis, and reporting. We will further examine and document the challenges with meeting the consensus-based priorities of public health stakeholders given the state of clinical information systems, MU, and the capacity of the INPC. Our findings and lessons will be shared with the public health informatics community to inform future public health and informatics research, practice and policy in Indiana and beyond.

In addition to local applications, we intend to develop a workbook or other product to provide guidance on data quality issues to other public health entities. The workbook would consist of both a general framework for addressing data quality issues in a setting where electronic data is being received from numerous information systems and templates that agency personnel could complete to identify and prioritize agency business processes and needs. The framework will be largely informed by our Delphi study methods. However, we recognize that agencies will likely lack the ability to replicate a full Delphi study. Recognizing that simple stakeholder meetings are insufficient, the framework will identify strategies for bringing stakeholders together meaningfully to discuss and prioritize items developed initially by agency staff. The templates will guide staff members through the complex issues surrounding data quality and help them generate documents that stakeholders can review and use to establish priorities, policies, and directives. The goal for the workbook is to be practical, and its design and usage will be modeled after other practical informatics tools such as the U.S. Agency for Healthcare Research and Quality (AHRQ) "Health Information Technology Evaluation Toolkit" (26) and the Public

Health Informatics Institute (PHII) "Collaborative Requirements Development Methodology" (27).

Finally, our aims include a comparison of measured data quality perceptions and needs with measured real-world data quality from clinical information systems. Our current project will collect perceived quality of existing ELR, syndromic surveillance, and other public health information system data as well as the perceived needs of those in public health working with ELR, syndromic, and other information system data. ICEPHI also includes an operational health information exchange (the INPC) with hundreds of interfaces to real-world clinical information systems. We aim to measure the quality of the data (e.g., completeness, timeliness) flowing into the INPC. We can then compare the real-world quality of the data with the perceived quality and needs of public health stakeholders. Such evidence will provide insight on where we are in developing public health informatics systems in relation to where public health practitioners need them to be.

Conclusion

The quality of data from information systems varies over time and across systems, and it is often poor. Using poor data from clinical information systems can lead to poor decisions regarding patient care and public health policy. Despite an abundance of poor data flowing through electronic systems, current policies and practices, knowingly or unintended, assume equal quality across all sources therefore pushing for the adoption and use of more electronic data feeds to public health. Public health officials, epidemiologists, informaticians, and other professionals who recognize the unequal quality of data sources lack practical tools and resources to adequately assess quality to improve the utilization of electronic data sources that feed public health information systems.

Our CDC funded Center of Excellence in Public Health Informatics will continue to examine the quality of data and their sources. We aim to develop novel, practical approaches for continuously assessing data quality to improve the decisions and work processes involved in modern public health practice. Our findings and lessons learned should contribute to future research and policies that will improve data quality, utilization of available data sources, and the effectiveness of public health information systems. In other words, ICEPHI will ask questions about data and information sources that will lead to improved population health processes and outcomes. This is an important role for all public health informatics centers and professionals to perform within their agencies and communities.

"The art and science of asking questions is the source of all knowledge." Thomas Berger

Acknowledgements

The work of the Indiana Center of Excellence in Public Health Informatics reported here is supported by a grant (1P01HK000077-01) from the U.S. Centers for Disease Control and Prevention. The contents of this paper are solely the responsibility of the authors and do not necessarily represent the official views of the Centers for Disease Control and Prevention.

Conflict of Interest

The authors declare that they have no real or apparent conflicts of interest.

Corresponding Author

Brian E. Dixon, MPA, PhD
Assistant Professor of Health Informatics, Indiana University School of Informatics
Research Scientist, Regenstrief Institute
410 W. 10th St., Suite 2000
Indianapolis, IN, 46202
Email: bedixon@iupui.edu

References

1. Buehler JW, Hopkins RS, Overhage JM, Sosin DM, Tong V. Framework for evaluating public health surveillance systems for early detection of outbreaks: recommendations from the CDC Working Group. MMWR Recomm Rep. 2004;53(RR-5):1-11. Epub 2004/05/07.
2. Lombardo JS, Buckeridge DL, editors. Disease Surveillance: A Public Health Informatics Approach. Hoboken: John Wiley & Sons; 2007.
3. Wang RY, Strong DM. Beyond Accuracy: What Data Quality Means to Data Consumers. Journal of Management Information Systems. 1996;12(4):5-34.
4. Redman TC. The Impact of Poor Data Quality on the Typical Enterprise. Communications of the ACM. 1998;41(2):79-82.
5. Bailey JE, Pearson SW. Development of a tool for measuring and analyzing computer user satisfaction. Manaement Science. 1983;29(5):530-45.
6. Hogan WR, Wagner MM. Accuracy of data in computer-based patient records. J Am Med Inform Assoc. 1997;4(5):342-55. Epub 1997/09/18.
7. Hasan S, Padman R. Analyzing the effect of data quality on the accuracy of clinical decision support systems: a computer simulation approach. AMIA Annu Symp Proc. 2006:324-8. Epub 2007/01/24.
8. Stein HD, Nadkarni P, Erdos J, Miller PL. Exploring the degree of concordance of coded and textual data in answering clinical queries from a clinical data repository. J Am Med Inform Assoc. 2000;7(1):42-54. Epub 2000/01/21.
9. Doyle TJ, Glynn MK, Groseclose SL. Completeness of notifiable infectious disease reporting in the United States: an analytical literature review. Am J Epidemiol. 2002;155(9):866-74. Epub 2002/04/30.
10. Effler P, Ching-Lee M, Bogard A, Ieong MC, Nekomoto T, Jernigan D. Statewide system of electronic notifiable disease reporting from clinical laboratories: comparing automated reporting with conventional methods. JAMA. 1999;282(19):1845-50. Epub 1999/11/26.
11. Lober WB, Karras BT, Wagner MM, Overhage JM, Davidson AJ, Fraser H, et al. Roundtable on bioterrorism detection: information system-based surveillance. J Am Med Inform Assoc. 2002;9(2):105-15. Epub 2002/02/28.
12. Buehler JW, Whitney EA, Smith D, Prietula MJ, Stanton SH, Isakov AP. Situational uses of syndromic surveillance. Biosecur Bioterror. 2009;7(2):165-77. Epub 2009/07/29.

13. Lazarus R, Kleinman KP, Dashevsky I, DeMaria A, Platt R. Using automated medical records for rapid identification of illness syndromes (syndromic surveillance): the example of lower respiratory infection. BMC Public Health. 2001;1:9. Epub 2001/11/28.
14. Chapman WW, Dowling JN, Wagner MM. Classification of emergency department chief complaints into 7 syndromes: a retrospective analysis of 527,228 patients. Annals of emergency medicine. 2005;46(5):445-55. Epub 2005/11/08.
15. Fleischauer AT, Silk BJ, Schumacher M, Komatsu K, Santana S, Vaz V, et al. The validity of chief complaint and discharge diagnosis in emergency department-based syndromic surveillance. Acad Emerg Med. 2004;11(12):1262-7. Epub 2004/12/04.
16. Reis BY, Mandl KD. Syndromic surveillance: the effects of syndrome grouping on model accuracy and outbreak detection. Annals of emergency medicine. 2004;44(3):235-41. Epub 2004/08/28.
17. The American Recovery and Reinvestment Act of 2009, House of Representatives, 111th Congress Sess. (2009).
18. HIMSS. The American Recovery and Reinvestment Act of 2009: Summary of Key Health Information Technology Provisions 2009 January 25, 2010 [cited 2010 January 25]. Available from: http://www.himss.org/content/files/HIMSS_SummaryOfARRA.pdf.
19. Vreeman DJ. Keeping up with changing source system terms in a local health information infrastructure: running to stand still. Studies in health technology and informatics. 2007;129(Pt 1):775-9. Epub 2007/10/04.
20. Vreeman DJ, Stark M, Tomashefski GL, Phillips DR, Dexter PR. Embracing change in a health information exchange. AMIA Annu Symp Proc. 2008:768-72. Epub 2008/11/13.
21. Dixon BE, McGowan JJ, Grannis SJ. Electronic Laboratory Data Quality and the Value of a Health Information Exchange to Support Public Health Reporting Processes. AMIA Annu Symp Proc. 2011:Forthcoming.
22. Gordon TJ. The Delphi Method. The Millennium Project; [cited 2011 August 4]; Available from: http://www.millennium-project.org/FRMv3_0/04-Delphi.pdf.
23. Keeney S, Hasson F, McKenna HP. A critical review of the Delphi technique as a research methodology for nursing. International journal of nursing studies. 2001;38(2):195-200. Epub 2001/02/27.
24. Pare G, Sicotte C, Jaana M, Girouard D. Prioritizing the risk factors influencing the success of clinical information system projects. A Delphi study in Canada. Methods of information in medicine. 2008;47(3):251-9. Epub 2008/05/14.
25. Poon EG, Jha AK, Christino M, Honour MM, Fernandopulle R, Middleton B, et al. Assessing the level of healthcare information technology adoption in the United States: a snapshot. BMC Medical Information Decision Making. 2006;6:1. Epub 2006/01/07.
26. Cusack CM, Poon EG. Health Information Technology Evaluation Toolkit. Rockville, MD: Agency for Healthcare Research and Quality; 2007 [June 10, 2009]; Available from: http://healthit.ahrq.gov/portal/server.pt/gateway/PTARGS_0_1248_807442_0_0_18/AHRQ_Evaluation%20Toolkit.pdf.
27. Collaborative Requirements Development Methodology (CRDM) Walkthrough. Public Health Informatics Institute; 2011 [cited 2011 September 8]; Available from: http://www.phiicrdm.org/methodology.