A Vision for the Systematic Monitoring and Improvement of the Quality of Electronic Health Data

Brian E. Dixon[a,b,c], Marc Rosenman[b,d], Yuni Xia[e], Shaun J. Grannis[b,f]

[a] *School of Informatics and Computing, Indiana University-Purdue University Indianapolis, Indianapolis, IN, USA*
[b] *Center for Biomedical Informatics, Regenstrief Institute, Indianapolis, IN, USA*
[c] *Center for Implementing Evidence-Based Practice, Department of Veterans Affairs, Veterans Health Administration, Health Services Research and Development Service, Indianapolis, IN, USA*
[d] *Department of Pediatrics, School of Medicine, Indiana University, Indianapolis, IN, USA*
[e] *Department of Computer Science, Purdue University, Indianapolis, IN, USA*
[f] *Department of Family Medicine, School of Medicine, Indiana University, Indianapolis, IN, USA*

Abstract

In parallel with the implementation of information and communications systems, health care organizations are beginning to amass large-scale repositories of clinical and administrative data. Many nations seek to leverage so-called Big Data repositories to support improvements in health outcomes, drug safety, health surveillance, and care delivery processes. An unsupported assumption is that electronic health care data are of sufficient quality to enable the varied use cases envisioned by health ministries. The reality is that many electronic health data sources are of suboptimal quality and unfit for particular uses. To more systematically define, characterize and improve electronic health data quality, we propose a novel framework for health data stewardship. The framework is adapted from prior data quality research outside of health, but it has been reshaped to apply a systems approach to data quality with an emphasis on health outcomes. The proposed framework is a beginning, not an end. We invite the biomedical informatics community to use and adapt the framework to improve health data quality and outcomes for populations in nations around the world.

Keywords:

Medical Records Systems, Computerized; Quality Control; Quality Improvement; Information Storage and Retrieval; Medical Informatics.

Introduction

Health care organizations globally are increasingly adopting and using health information and communications technologies (ICT) to collect, store, manage, and analyze data on patients and populations. In the latest surveys of providers in the United States, 26% of physicians and 15% of hospitals indicate usage of a basic electronic health record (EHR) system [1, 2]. A recent analysis of low- and middle-income country data from the Center for Health Market Innovations reveals that of the health programs reporting use of ICT: 42% use technology to extend geographic access to health care, 38% to improve data management, and 31% to facilitate communication between patients and physicians outside the physician's office [3]. The Commonwealth Fund recently reported that ICT adoption and usage among primary care providers is also increasing in 10 high-income nations [4].

In parallel with growing adoption and use of ICT, health care organizations are developing and utilizing data warehouses to integrate and analyze the electronic data they now systematically capture and manage using EHR and other clinical systems [5, 6]. Clinical data warehouses deployed across clinics, hospitals and entire health systems are amassing large data sets ushering in an era of 'Big' medical data.

Given the increasing availability of data warehouses and a shift towards 'Big data,' many countries are developing strategies to extract and analyze large-scale electronic data for a variety of uses including but not limited to bio-surveillance, quality reporting, comparative effectiveness research, translational research and health services research [7, 8]. For example, initiatives such as the Learning Health System advanced by the U.S. Institute of Medicine [9] aim to use large-scale clinical and administrative data sets to develop, implement, and evaluate new interventions aimed at improving health outcomes, enhancing care-delivery efficiencies, and reducing health disparities. However, leveraging such large-scale clinical and administrative data sets assumes that the data are of sufficient quality to enable valid, generalizable conclusions about health outcomes, drug safety, emerging health threats, and efficiency of care-delivery.

Quality of Electronic Health Data

We define data quality (DQ) as a set of dimensions relating to how well data are 'fit-for-use' by data consumers [10, 11]. In other words, data are of high quality if they are fit for their intended uses in operations, decision-making, and planning [12]. Most research about DQ involves four dimensions: accuracy, completeness, consistency, and timeliness. Typical DQ issues encountered include inaccurate data, inconsistencies across data sources, and incomplete (or unavailable) data necessary for operations or decisions [13]. While evidence on the impact of DQ issues is sparse, estimates of impacts include: up to 40-60% of a service organization's expenses consumed as a result of 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 [13].

The biomedical literature documents numerous examples where the quality of electronic health data is suboptimal. A study by Kaboli et al. compared electronic pharmacy data with the medications actually taken by patients [14], concluding that only 1-in-20 patients had perfect agreement between their computerized medication profile and what they were actually taking. Liaw et al. examined the completeness and accuracy of emergency department information system (EDIS) data for identifying patients with select chronic diseases (e.g., Type 2 diabetes mellitus, cardiovascular disease, chronic obstructive pulmonary disorder). The researchers found that information on the target chronic diseases was missing from EDIS discharge summaries in 11-20% of cases [15]. Furthermore, an audit of discharge summaries could only confirm up to 61% of diagnoses found in a query of the EDIS for the target conditions. This has implications for correctly identifying patients for research or quality measurement or to prevent readmissions.

Dixon et al. analyzed electronic laboratory data for positive cases of communicable disease transmitted from multiple hospital information systems [16]. The researchers found low completeness for a number of data fields critical to public health surveillance processes. Further, McGinnis et al. found that completeness and accuracy vary even when the electronic data sources are managed by the same health system [17]. The study compared two databases maintained by the U.S. Department of Veterans Affairs (VA). Each database extracted data from laboratory tests performed by the VA. However, the databases differed in their completeness and accuracy of data for certain lab tests in a cohort of HIV positive veterans.

Finally, a review by Chan et al. examined the quality of electronic data used for quality measurement [18]. Completeness of electronic data varied "substantially across studies" ranging from 0.1% - 51% for blood pressure; 10% - 38% for smoking status. Omission rates for medication lists varied from 27% in oncology to 53% for primary care patients. Missing data ranged between 24% and 38% for LDL; 3% and 31% for blood pressure; and 5% and 23% for HbA1c.

The studies mentioned here are only a sample of the documented examples of suboptimal data quality in EHRs and other health ICT systems. The studies range across use cases in biomedical informatics, spanning quality measurement to population health to translational informatics studies.

Despite a growing body of evidence pointing to suboptimal quality of the clinical and administrative data being collected, managed and used by various health ICT systems, a comprehensive framework for methodologically examining electronic health DQ is lacking. We believe that to effectively realize the vision set forth by the IOM and other nations, which seek to leverage growing volumes of Big Data to improve clinical and health system outcomes, the biomedical informatics community must embrace and advance a common framework for describing, measuring, and analyzing the quality of electronic health data.

In the remainder of this paper, we propose a novel framework for electronic health data stewardship adapted from the established data quality management literature in business and government. We then describe a vision for applying the framework to advance electronic data and systems toward improved outcomes in every nation.

Review of Data Quality Scholarship

Poor data quality is common and affects all industries and organizations that employ information systems [10]. Therefore, a number of DQ management researchers outside of health care have examined the challenge of data stewardship in business and government. A review by Batini et al. reported that there are 13 distinct methodologies for defining, implementing, and evaluating a DQ strategy for an organization [11]. We briefly review two methodologies: the original framework proposed by Wang [19] and a modern derivation of the same model adopted for large-scale computing [20]. These models were selected as exemplars because they have been used and extended in numerous articles in information systems and computing literature, which suggests that they are more widely accepted than other models.

The first general methodology for DQ was defined as an extension of the principles of Total Quality Management, which seeks to eliminate discrepancies between the output of the system and customers' requirements by reengineering processes. The framework defined a cycle involving four phases. The first phase focused on defining information system functions and DQ requirements. The second phase focused on measuring information quality. The third phase analyzed metrics for causes of error in poor quality. The final phase identifies strategies and techniques (reengineering) to be implemented to address the identified errors.

Total Data Quality Management (TDQM) has served as the basis for a variety of other data quality methodologies in business, education, and government. The recent iteration of TDQM, described by Batini et al. [20], updates the original framework to account for evolutionary changes in the ICT infrastructure since the original model was proposed in the late 1990s. This new framework, called Heterogeneous Data Quality Methodology (HDQM), incorporates techniques for data that are no longer stored in a centralized, enterprise database but are now spread across disparate data sources in a network or cloud infrastructure. HDQM considers three distinct types of data to be particularly relevant to biomedical informatics: structured data, semi-structured data, and unstructured data.

HDQM follows a three-stage approach to improving DQ. The first phase, state reconstruction, characterizes DQ problems and dimensions. The second phase assesses an organizational unit, measuring DQ dimensions. The final phase selects an optimal improvement process for addressing problems identified through measurement in the second phase.

Framework for Electronic Health Data Quality

While TDQM and HDQM provide models for defining the dimensions of DQ, measuring DQ, and proposing solutions to address DQ deficiencies, these models are largely generic, focus on process, and fail to explicitly relate external outcomes (e.g., improved health) to DQ – a criticism of many informatics studies [21]. In health care, information processes (e.g., ordering medication using a computerized provider order entry system) are intrinsically linked to care processes and outcomes. Therefore we have developed a variant of HDQM that uses a systems approach to link DQ processes and measures to patient and population health outcomes (Figure 1).
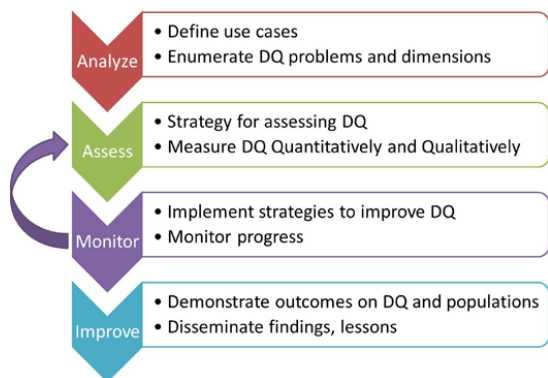
Figure 1 – The Health Data Stewardship (HDS) framework

In the Health Data Stewardship (HDS) framework, four phases iteratively advance the ability to define, characterize, measure, and improve health DQ and health outcomes. The framework further focuses on various dimensions of DQ, including technical (e.g., network outage that prevents an HL7 message from being received), human (e.g., laboratory technician failed to set the abnormal flag for a communicable disease result), and organizational (e.g., hospital A does not permit transmission of mental health consults in continuity of care documents it sends to hospital B), applying a systems approach (people, organizations, informatics) to understanding factors that lead to improved DQ and subsequently better health outcomes. By emphasizing systems and outcomes the model embodies stewardship, an ethos of responsibility for managing resources to ensure sustainability. The model seeks to improve DQ to enable fit-for-purpose across the spectrum of uses envisioned for electronic health data.

In the first phase, use-cases are defined that explicitly model information flow, business processes, and data elements necessary to support activities such as communicable disease reporting. Data element characteristics are defined with respect to their fit with user and system needs, enabling a clear definition of fit-for-use. Pitfalls and challenges to high quality data are also enumerated. Information needs and system requirements can be gathered using a variety of methods, including but not limited to focus groups, user surveys, or structured interviews. Prior work has shown that various use cases in clinical and public health present different data needs. Population surveillance, for example, has a low dependency on complete data on the providers involved in health care transaction; whereas public health reporting processes are dependent on such detail to enable case investigation follow-up. Support for various use cases is a unique aspect of this framework, accommodating known and unforeseen uses of electronic health data.

In the second phase, a strategy is defined for characterizing and measuring DQ. The strategy will largely be dictated by the first phase in which fitness-for-use is defined. Communicable disease epidemiologists, for example, may elect to focus on completeness, timeliness, and accuracy as the three dimensions of DQ most germane to public health surveillance. Metrics for these dimensions are enumerated, and baseline quality of electronic data is measured. Metrics can be quantitative in nature, such as the percentage of complete fields or reports received from physician practices. They may also be qualitative, whereby epidemiologists are asked whether or not they perceive arriving data from EHR systems to be adequate. Mixed methods may also be appropriate for examining quantitative

and qualitative aspects of DQ. The assessment will identify where data are not a good fit-for use.

In the third phase, strategies to improve the quality of electronic health data are defined and implemented. For communicable disease case information, health departments may elect to notify providers where their completeness or timeliness is suboptimal. The health ministry may further request or require care providers to change how they collect or report data. They may also suggest methods or tools to enhance data capture or reporting. This phase also involves monitoring progress towards the improvement of DQ. Phases 2 and 3 allow for iteration, meaning that they can be repeated over time to enhance or incrementally improve health DQ.

In the fourth phase, improvements to both electronic health data quality and health outcomes are emphasized. Better quality data in communicable disease case reports, for example, may translate into improved efficiencies for the health ministry. Improved data accuracy may produce fewer false positive and false negative signals [22], which in turn leads to more effective use of ministry resources (people, funding) in response to emerging health threats identified by the enhanced surveillance system and associated processes. Linkages between improvements in DQ and health outcomes will be defined using theoretical models in prior research as well as novel associations that can be validated through observational studies. The key to this framework is that health outcomes are not afterthoughts but defined *a priori* and linked to use-case driven DQ given the importance data play in clinical and population health decision-making processes. Outcomes and evidence will also be disseminated in this phase to both stakeholders (e.g., health minister, physician leadership) as well as the broader biomedical informatics community. Dissemination may advance research on improving health outcomes, or it may more narrowly focus on methods and tools for improving the quality of electronic health data.

Discussion

Given the increasing reliance on health ICT to deliver care around the globe, there is a clear need for the biomedical community to address the challenge of electronic DQ. In this paper we have proposed a novel framework for defining, assessing, monitoring, and improving both the quality of electronic health data and subsequent outcomes for patients and populations. The model is derived from prior work in electronic commerce and government that defined methods and processes for enumerating DQ and improving processes that collect, manage, and query data in large-scale information systems. However, the framework described here extends beyond existing models by providing a connection to health outcomes and by using a systems approach to address technical, human, and organizational challenges that affect quality.

Now is the time for a framework to catalyze change. Numerous examples in the biomedical literature demonstrate that data from health ICT have a wide range of DQ challenges including incomplete data, lack of currency (timeliness), and inaccuracy. These are serious issues that need to be addressed given the growing number of health care system stakeholders and services that rely on electronic data to inform healthcare and policy decisions. Poor decisions made at any level of the health system may negatively impact patient and population health.

The HDS framework offers a vision for future research and development in the area of electronic health DQ. Existing knowledge has come from projects that sought to use health data for particular purposes and found the data to be unfit for the target purpose. Future work using the framework is likely to occur in the following areas: 1) application across multiple use cases; 2) discovery of new DQ dimensions and methods; 3) development of automated tools to support monitoring and improvement in clinical and public practice; and 4) evidence on how DQ improvements lead to better health outcomes.

Applying the framework

A first step is applying the framework across the spectrum of biomedical informatics. Use cases in various health system domains will need to be articulated then defined. Some use cases have been explored in the existing literature, such as quality performance and public health reporting [16, 18], yet there are additional use cases to consider. Furthermore, the biomedical informatics community will need to fully define the purpose of use for electronic health data and identify users' information needs within each use case. Systematic reviews may offer insight into DQ issues raised in efforts that have used electronic data. Additional methods, including user surveys, focus groups or interviews, may be desired to collect information needs across a hospital network, region, or nation. Initial efforts in public health informatics by Kirbiyik and Dixon serve as examples of this type of activity [23, 24].

After developing use cases, researchers should assess intrinsic quality of electronic health data. We anticipate that existing, common dimensions and measures of quality (e.g., accuracy, timeliness and completeness) extracted from the TDQM literature will be readily applicable for many health system use cases. This has been done previously in a limited number of studies. For example, timeliness has been used as a key metric for evaluating the quality of a surveillance system in public health [25, 26]. Further assessment of existing dimensions and metrics will enable refinement of estimates and could drive consensus regarding the quality challenges that should be prioritized for resolution by the informatics community.

Innovation in DQ dimensions and methods

We further hypothesize that an important opportunity for innovation lies in developing additional dimensions and measures of quality that may emerge as useful or appropriate for evaluating electronic health data. Researchers exploring health system use cases will likely discover novel or adapted metrics for assessing DQ through the application of the proposed framework. For example, the informatics community may consider the use of two emerging metrics from the public health surveillance community: information entropy and change point analysis [27, 28].

Entropy measures uncertainty and characterizes the density of the information contained in a message (e.g., chief complaint, physician note, qualitative laboratory result). Recently Grannis et al. examined the information entropy of chief complaints entered into Emergency Department information systems used in over 100 hospitals [28]. The analysis revealed that information content varied by surveillance syndrome, yet intra-organizational entropy was consistent. The researchers also found that a shift in entropy can detect events such as the implementation of a new hospital registration system.

Change point analysis techniques characterize the level of change observed in a data stream over time. Sell et al. recently examined chief complaints from 49 emergency departments [29]. Counts in the number of words entered in chief complaint fields were used to examine changes to syndrome categories and hospital information systems. Painter et al. used a Bayesian change point technique to examine syndromic surveillance record counts from a network of hospitals across the United States [30]. The analysis identified outliers, which indicated shifts in system transmission rates or offline status.

Development of informatics tools

Once measures are defined and can assess health data and ICT systems, the focus will shift towards developing strategies and informatics tools to improve health DQ. The workflow associated with chief complaint data entry may need to be re-engineered, or a systems process that transforms data from a local EHR system prior to transmission to a clinical data warehouse may need to be modified. Strategies for improving DQ will be challenging given that changes to human data entry and inter-organizational data exchange activities may require significant time and effort when scaled to regional or national levels. Therefore strategies will need to be carefully planned and implemented, and their impact measured and monitored.

In addition to process redesign, informatics researchers will need to develop ICT tools to support ongoing *in situ* assessment, analysis, and monitoring of DQ over time. We perceive a paucity of off-the-shelf tools to support DQ assessment and monitoring in current real-world health settings. Tools that can easily integrate into an EHR system or other health ICT, especially low resource settings, will be required to enable health information systems personnel to operationalize DQ monitoring and improvement. Not every effort to improve DQ will be academic, so translation of knowledge to operational partners will be critical for sustained improvement.

Development of an evidence base

Last but not least, the community will need to measure short and long-term impacts of DQ interventions on patient and population outcomes. Currently there is little evidence linking DQ improvements to health system improvement. As the community develops its capacity to measure and improve DQ, research must be done to build an evidence base to support the theory of HDS. Studies must be designed appropriately to control for confounders, and promising strategies must be repeated in multiple sites for validation. This can begin within integrated delivery networks and health information exchange initiatives then grow to ministries and across nations. A strong evidence base will encourage adoption of process and technical changes necessary to improve DQ and population health.

The proposed framework represents the first step of a thousand mile journey. While the journey is long, confronting DQ challenges has the potential to significantly improve many aspects of health care delivery and the sub-disciplines in biomedical informatics. We invite feedback on the proposed framework, and we look forward to dialogue with the community on how it can be used to improve the quality of electronic health data.

Conclusions

Health systems must examine electronic DQ in a systematic fashion to realize the goals of national ICT programs. We have proposed a model for discussion within the biomedical informatics community. Together as an international community of scholars, we must work to advance the framework, applying it to information systems at every level of a nation's health in-

formation infrastructure to characterize DQ and develop innovations for improving and sustaining DQ in the long term.

Acknowledgments

References

[1] Rao SR, Desroches CM, Donelan K, Campbell EG, Miralles PD, Jha AK. Electronic health records in small physician practices: Availability, use, and perceived benefits. J Am Med Inform Assoc. 2011;18(3):271-5.

[2] Jha AK, Burke MF, DesRoches C, Joshi MS, Kralovec PD, Campbell EG, et al. Progress toward meaningful use: Hospitals' adoption of electronic health records. Am J Manag Care. 2011;17(12 Spec No.):SP117-24.

[3] Lewis T, Synowiec C, Lagomarsino G, Schweitzer J. E-health in low- and middle-income countries: Findings from the center for health market innovations. Bull World Health Organ. 2012;90(5):332-40.

[4] Schoen C, Osborn R, Squires D, Doty M, Rasmussen P, Pierson R, et al. A survey of primary care doctors in ten countries shows progress in use of health information technology, less in other areas. Health Aff. 2012.

[5] Kaelber DC, Foster W, Gilder J, Love TE, Jain AK. Patient characteristics associated with venous thromboembolic events: A cohort study using pooled electronic health record data. J Am Med Inform Assoc. 2012;19(6):965-72.

[6] Comer KF, Grannis S, Dixon BE, Bodenhamer DJ, Wiehe SE. Incorporating geospatial capacity within clinical data systems to address social determinants of health. Public Health Rep. 2011;126 Suppl 3:54-61.

[7] Shah NH. Translational bioinformatics embraces big data. Yearb Med Inform. 2012;7(1):130-4.

[8] Ohno-Machado L. Big science, big data, and a big role for biomedical informatics. J Am Med Inform Assoc. 2012;19(e1):e1.

[9] Institute of Medicine. Digital infrastructure for the learning health system: The foundation for continuous improvement in health and health care: Workshop series summary. Washington, D.C.: The National Academies Press; 2011.

[10] Wang RY, Strong DM. Beyond accuracy: What data quality means to data consumers. Journal of Management Information Systems. 1996;12(4):5-34.

[11] Batini C, Cappiello C, Francalanci C, Maurino A. Methodologies for data quality assessment and improvement. ACM Comput Surv. 2009;41(3).

[12] Kerr KA, Norris T, Stockdale R. The strategic management of data quality in healthcare. Health Informatics J. 2008;14(4):259-66.

[13] Redman TC. The impact of poor data quality on the typical enterprise. Communications of the ACM. 1998;41(2):79-82.

[14] Kaboli PJ, McClimon BJ, Hoth AB, Barnett MJ. Assessing the accuracy of computerized medication histories. Am J Manag Care. 2004;10(11 Pt 2):872-7.

[15] Liaw ST, Chen HY, Maneze D, Taggart J, Dennis S, Vagholkar S, et al. Health reform: Is routinely collected electronic information fit for purpose? Emergency medicine Australasia : EMA. 2012;24(1):57-63.

[16] 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;2011:322-30.

[17] McGinnis KA, Skanderson M, Levin FL, Brandt C, Erdos J, Justice AC. Comparison of two va laboratory data repositories indicates that missing data vary despite originating from the same source. Med Care. 2009;47(1):121-4.

[18] Chan KS, Fowles JB, Weiner JP. Review: Electronic health records and the reliability and validity of quality measures: A review of the literature. Med Care Res Rev. 2010;67(5):503-27.

[19] Wang RY. A product perspective on total data quality management. Communications of the ACM. 1998;41(2):58-65.

[20] Batini C, Barone D, Cabitza F, Grega S. A data quality metholodogy for heterogeneous data. International Journal of Database Management Systems. 2011;3(1):60-79.

[21] Dixon BE, Zafar A, Overhage JM. A framework for evaluating the costs, effort, and value of nationwide health information exchange. J Am Med Inform Assoc. 2010;17(3):295-301.

[22] Cadieux G, Buckeridge DL, Jacques A, Libman M, Dendukuri N, Tamblyn R. Patient, physician, encounter, and billing characteristics predict the accuracy of syndromic surveillance case definitions. BMC Public Health. 2012;12:166.

[23] Kirbiyik U, Gamache RE, Dixon BE, Grannis SJ, editors. Evaluating the variation on public health's perceived field need of communicable disease reports. ISDS Annual Conference; 2012; San Diego, CA: International Society for Disease Surveillance.

[24] Dixon BE, Grannis S. Why asking questions about data and their sources is important to public health informatics practice and research. Online Journal of Public Health Informatics. 2011;3(3).

[25] Overhage JM, Grannis S, McDonald CJ. 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):344-50.

[26] 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.

[27] Kass-Hout TA, Xu Z, McMurray P, Park S, Buckeridge DL, Brownstein JS, Finelli L, Groseclose SL. Application of change point analysis to daily influenza-like illness emergency department visits. J Am Med Inform Assoc. 2012;19(6):1075-81.

[28] Grannis SJ, Dixon BE, Xia Y, Wu J. Using information entropy to monitor chief complaint characteristics and quality. In: Ising A, editor. ISDS Annual Conference Proceedings; San Diego, CA. 2012. p. 90.

[29] Sell J, Mathes R, Paladini M. Detecting changes in chief complaint word count: Effects on syndromic surveillance. In: Ising A, editor. ISDS Annual Conference Proceedings; San Diego, CA. 2012. p. 191.

[30] Painter I, Eaton J, Lober B, editors. Using change point detection for monitoring the quality of aggregate data. ISDS 2012 Conference Proceedings; 2012; San Diego, CA. p. 169.

Address for correspondence

Brian E. Dixon, MPA, PhD, FHIMSS
Research Scientist
Regenstrief Institute, Inc.
410 W. 10th St., Suite 2000
Indianapolis, IN, 46202 USA
bedixon@iupui.edu