Published in final edited form as:

Policy Polit Nurs Pract. 2015 August ; 16(0): 117–124. doi:10.1177/1527154415603358.

Challenges Associated With Using Large Data Sets for Quality Assessment and Research in Clinical Settings

Bevin Cohen, MPH, MPhil[1], **David K. Vawdrey, PhD**[2], **Jianfang Liu, PhD, MAS**[1], **David Caplan, BS**[3], **E. Yoko Furuya, MD, MS**[4], **Frederick W. Mis, PhD**[3], and **Elaine Larson, RN, PhD, CIC, FAAN**[1]

[1]Columbia University School of Nursing, New York, NY, USA

[2]Department of Biomedical Informatics, Columbia University, New York, NY, USA

[3]Department of Information Services, New York-Presbyterian Hospital, New York, NY, USA

[4]Department of Medicine, Columbia University, New York, NY, USA

Abstract

The rapidly expanding use of electronic records in health-care settings is generating unprecedented quantities of data available for clinical, epidemiological, and cost-effectiveness research. Several challenges are associated with using these data for clinical research, including issues surrounding access and information security, poor data quality, inconsistency of data within and across institutions, and a paucity of staff with expertise to manage and manipulate large clinical data sets. In this article, we describe our experience with assembling a data-mart and conducting clinical research using electronic data from four facilities within a single hospital network in New York City. We culled data from several electronic sources, including the institution's admission-discharge-transfer system, cost accounting system, electronic health record, clinical data warehouse, and departmental records. The final data-mart contained information for more than 760,000 discharges occurring from 2006 through 2012. Using categories identified by the National Institutes of Health Big Data to Knowledge initiative as a framework, we outlined challenges encountered during the development and use of a domain-specific data-mart and recommend approaches to overcome these challenges.

Keywords

informatics; outcomes measurement; research methodology

Reprints and permissions: sagepub.co.uk/journalsPermissions.nav

Corresponding Author: Bevin Cohen, Columbia University School of Nursing, 630 West 168th Street, New York, NY 10032, USA., bac2116@columbia.edu.

Declaration of Conflicting Interests
The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Introduction

The broad adoption of electronic health records (EHRs) holds great promise for improving coordination and standardization of clinical care and ultimately health outcomes for patients (Blumenthal, 2009). Another benefit of EHR adoption is the availability of vast amounts of treatment and outcome data available electronically for purposes secondary to direct patient care. Such data may be valuable for assessing the clinical effcacy, effectiveness, and cost-effectiveness of preventive and therapeutic interventions, as well as for investigating epidemiologic questions such as identifying risk factors for disease and tracking trends over time (Miriovsky, Shulman, & Abernethy, 2012; Toh & Platt, 2013). Nonetheless, assembling electronic data from multiple unlinked sources and processing the data into a format suitable for research present major challenges. Hence, while huge volumes of patient- and institution-level data are now being collected electronically, they are not optimally used for quality improvement or comparative effectiveness, clinical, or health services research.

Over the past decade, the new discipline of data science has emerged to develop methods for using *big data*, including new and extensive data production and storage capabilities, powerful analytic and computational technologies, improved interoperability between systems, and governance frameworks to protect data security and facilitate sharing (Committee on the Analysis of Massive Data, Committee on Applied and Theoretical Statistics, Board on Mathematical Sciences and Their Applications, Division on Engineering and Physical Sciences, & National Research Council National Research Council, 2013; Dahr, 2013; Herman et al., 2013; Murdoch & Detsky, 2013). To address the challenges of building, utilizing, and maintaining large data sets for clinical research, the National Institutes of Health created the Big Data to Knowledge (BD2K) initiative and named its first Director for Data Science in 2013 (Ohno-Machado, 2014).

BD2K identified seven major obstacles associated with using *biomedical big data*. They are (a) locating data and software tools; (b) accessing data and software tools; (c) standardizing data and metadata; (d) extending policies and practices for data and software sharing; (e) organizing, managing, and processing biomedical big data; (f) developing new methods for analyzing and integrating biomedical data; and (g) training researchers who can use biomedical big data effectively. The purpose of this article is to describe these seven obstacles and recommend methods for overcoming them, using our experience as a multidisciplinary team developing and utilizing a large research data-mart in the domain of infection control and prevention.

Methods

In 2007, our research team received funding from the National Institute of Nursing Research to investigate the financial costs associated with antimicrobial resistance in hospitals (National Institute of Nursing Research, 2007). To address the aims of the project, we amassed a large data-mart encompassing medical, billing, and demographic information of all patients discharged from four hospitals within a single academically affiliated health-care network from 2006 through 2008. The data-mart contained information for more than

319,000 discharges culled from several electronic sources, including the institution's admission-discharge-transfer system, cost accounting system, EHR, clinical data warehouse (CDW), and departmental records (Apte, Neidell, et al., 2011). Although the data-mart was created to address specific aims related to the cost of care for patients with antimicrobial resistant infections, the project resulted in a novel, comprehensive data source that investigators eventually used throughout the institution to answer a variety of clinical and epidemiological research questions (Apte, Landers, Furuya, Hyman, & Larson, 2011; Jeon, Furuya, Berman, & Larson, 2012a; Jeon, Furuya, Smaldone, & Larson, 2012b; Jeon, Neidell, Jia, Sinisi, & Larson, 2012c; Landers et al., 2010; Neidell et al., 2012; Patel, O'Toole, & Larson, 2012; Pogorzelska-Maziarz, Furuya, & Larson, 2013; Wolfe, Cohen, & Larson, 2014). To maintain, update, and make the data available for future research, funding was renewed in 2013 with additional comparative effectiveness aims and a broader focus on developing policies and procedures for data sharing and stewardship (National Institute of Nursing Research, 2012). The data-mart was expanded to include all patient discharges from 2006 through 2012, totaling more than 760,000.

Results

Table 1 provides our experience with the seven challenges identified in BD2K, an overview of the issues we faced with each challenge, and recommended approaches to overcoming each challenge.

Locating Data and Software Tools

A major component of the data acquisition phase was determining, through conversations with source data experts, clinical collaborators, and manipulations of sample data sets, which data elements could feasibly be obtained from each data source, which variables would not be available at all from the institution's electronic data sources, and what the limitations of each data element would be. Ultimately, we identified 22 classes of data relevant to our project that were located in four source systems (Table 2). In addition, more than 30 reference tables such as ICD-9 codes, lists of clinical units, and codebooks for antibiotic codes or organism codes of culture data had to be assembled and incorporated into the data-mart.

To enhance identification of an institution's data sources that can potentially be used for clinical research, we recommend creation of an inventory of electronic systems containing data that are available to researchers. The inventory should include information about the type of data found in each system, how frequently they are updated, when the data began being populated in the system, and who is responsible for granting access.

Accessing Data and Software Tools

In creating our data-mart, we had to understand and navigate institutional policies and identify key individuals who could provide permission and sponsor access to data. Although the original grant submission included letters from the hospital's chief information officer, chief quality officer, and the director of quality and outcomes research indicating their support for the project and permission to utilize institutional data, it was also essential to identify specific data *stewards*, who were responsible for entering, maintaining, or monitoring the use of various data elements.

To address privacy and data use concerns, our institution formalized and standardized the process for requesting access to clinical data for the purposes of research. A central committee was created to triage, prioritize, approve, and monitor data requests. In 2013, the committee was established and began meeting weekly to review requests and assign them to the appropriate technical team. The creation of this process ensured that researchers such as our group would have a single point of contact to request and obtain data. The central committee adopted a transparent process, eliminating the need for researchers to establish connections with disparate approvers for each individual data source.

Standardizing Data and Metadata

Despite having structured data entry fields in the EHR, some data elements important for clinical research are not always systematically or accurately recorded. For other types of information, such as subjective assessments or changes in patient status, there may not be discrete, coded data entry fields in the EHR, and instead, documentation may be recorded in free-text format, which cannot be readily queried without use of text processing. Although our institution has conducted extensive research in natural language processing, we have observed that extracting discrete parameters from narrative text can be resource-intensive and may not achieve the level of accuracy desired by researchers. For these reasons, we recommend that where possible, clinical research groups desiring to use EHR data coordinate with information technology staff, and most importantly, clinicians documenting in the EHR, to collect important items for research as discrete values. Careful consideration should be given to the increased documentation burden that is often imposed when narrative text in EHRs is pushed toward structured data entry (Cusack et al., 2013; Johnson et al., 2008; Rosenbloom et al., 2008).

Because we were linking data elements across multiple years from disparate institutions, we sometimes encountered illogical discrepancies in some variables (e.g., a large increase or decrease during a given time period in the incidence of certain health-care-associated events such as infection). A considerable amount of validation work was required to determine whether such changes reflected real outcome changes or were artifacts of changes in data definitions, labeling, or coding. In some cases, it was necessary to recode data elements so that they were consistent across time and location. Local terminology management tools and resources such as the Medical Entities Dictionary used at our institution (Cimino, 2000) can be valuable for mapping terms and maintaining semantic consistency of data over time.

Extending Policies and Practices for Data and Software Sharing

In the course of assembling our infection control data-mart, our institution's policies and processes for requesting clinical data for research evolved. Initially, the data manager received access and queried data directly from the CDW. Later, a CDW analyst extracted the data and transferred it to our data manager. In both cases, the data manager needed to work closely with experts who were familiar with the data in source systems to locate the data elements needed, understand any limitations in how those data were collected and stored, and develop queries to extract the data. Often, it took several attempts to acquire an accurate and complete extract due to the complexity of the source data. This caused delays in the development of the research database and was time consuming for the analysts and subject matter experts assisting with data queries, who were providing their efforts in kind.

Developing New Methods for Analyzing and Integrating Data

Determining the accuracy and quality of electronic data prior to using it for clinical research is essential, but performing traditional validity assessments is not always feasible due to the lack of reference standards for many data elements. Even if documentation is accurate and complete, establishing how raw data should be used to create study variables is not always straightforward. Many therapeutic interventions occurring during the hospital stay, such as administration of medications and use of catheters, occur intermittently. Depending on the question, researchers may need to create variables that reflect whether patients ever had the intervention, had the intervention before or after a certain date, or had the intervention before or after a particular clinical event. In some cases, these types of interventions are documented at regular intervals, allowing confirmation of the sequence of events. In other cases, data may be recorded only once per day or once per admission, limiting the ability to establish temporality. Invariably, clinical researchers will need data that are not captured at present in EHRs. Future research should focus on bridging the gap between data collection for clinical care and data collection for research, as timely and complete documentation of nursing assessments are essential for accurate analyses. Automated data acquisition from biomedical devices can address the temporality issue in some cases, such as vital sign collection in intensive care units and medication drip rate changes in infusion pumps. Instead of relying on nursing documentation, which may not capture event occurrence times accurately (Nelson, Evans, Samore, & Gardner, 2005), some institutions have drip rate changes and vital signs recorded automatically in the EHR (Dalto, Johnson, Gardner, Spuhler, & Egbert, 1997; Gardner, Hawley, East, Oniki, & Young, 1991; Vawdrey et al., 2007).

Training Researchers Who Can Use Data Effectively

Researchers may lack skills and expertise related to use of electronic data, be unaware of the technical expertise needed, and not have contact with individuals who can manage large data sets effectively. Furthermore, investigators may struggle with what questions are appropriate and answerable with such data and how to sustain the networks needed for data use and governance. In our experience, identifying a data manager or programmer with the skills required to complete the project presented some difficulties because this type of endeavor had not previously been undertaken by anyone on the research team. During the recruitment and interview process, it was challenging to ascertain whether candidates had the technical abilities needed for the project, both because the scope and methodology were unknown and because our core team of clinicians and researchers were not familiar with specific technologies used in the institution’s information systems. In addition to querying data stored in a variety of formats and linking, processing, and cleaning these data, the data manager was also responsible for performing statistical analyses and working directly with investigators to create data sets for specific research aims. Thus, the data manager needed not only a broad range of programming experience, but analytical expertise as well as a working knowledge of medical terminology. This combination of skills is extremely difficult to find in a single individual, and thus, research groups may need to consider allocating multiple technical resources for projects similar to the one we undertook. Educational training programs in data science and biomedical informatics can prepare individuals to fill such roles in the future.

Discussion

Through our experience, we learned that using electronic data systems, while clearly a required skill for the future success of clinical research and quality assessment, is considerably more complex and challenging than most clinicians and researchers appreciate. The imperative to network and collaborate with informaticians, data modelers, and programmers was clear. Team science is highly relevant in projects such as this. The initial database took approximately three years to assemble. It required the efforts of a full-time data manager or programmer, a half-time project manager, an interdisciplinary team of coinvestigators including two health economists, an infectious disease physician, a nurse epidemiologist, a nurse manager, a director of data analytics and clinical information services, and the in-kind efforts of programmers and administrators of various data sources throughout the hospital and university.

Our original database included patient discharges from 2006 through 2008 and was subsequently updated through 2012. The process of updating the database took approximately two years to complete. Because the data manager had already identified the source of each data element, written the code for data extraction, and worked with study investigators to create and program decision rules for each variable, we anticipated that adding new data would require substantially less effort than the initial database creation. Instead, our team found several unexpected new challenges such as changes in coding practices and data fields, and issues with integrating the old and new data sets, as described earlier.

As Halamka (2014) noted, to make it possible for accountable care organizations to meet their mandate of measuring quality of care for populations, new data resources and expertise are necessary. The burgeoning quantity and increasing access to patients’ health information promises individual researchers the opportunity to investigate an infinite array of health topics using data from within their own institutions, as well as from facilities across the globe. Still, a number of technical, procedural, and data quality issues are barriers to using these data most effectively for research purposes. While the digitization of patient information holds promise for streamlining data collection, allowing for studies to include additional subjects and variables with minimal increase in cost, the process of creating a data set using multiple electronic sources requires a substantial investment of resources to initiate and maintain over time. Conducting this type of research is similar to traditional clinical and community-based research projects in terms of resource intensity, need for collaboration with multiple disciplines and departments, assistance with data collection, and permissions from multiple levels of administrators.

Many of the challenges we faced while creating and using the electronic database are consistent with those reported by others: variations in data definitions or coding over time, inaccurate or inaccessible data elements, access to sufficient informatics and programming expertise to analyze data, and the complexity of linking multiple and varied data sources (Halamka, 2014; Hersh et al., 2013a, 2013b). We believe the lessons from our project can be generalized to guide similar research endeavors in other health-care settings.

As others have previously reported, applications of electronic patient data for research, surveillance, quality improvement, and optimization of patient care are rapidly expanding (Jhung & Banerjee, 2009; Poon et al., 2010; Westra, Delaney, Konicek, & Keenan, 2008). In the field of infection prevention and control, specifically, automated methods of case finding have helped ease the burden of manual data collection and mandatory public reporting to local, state, and federal agencies, allowing clinical staff to focus on other priorities such as education and quality improvement initiatives. Nonetheless, although electronic algorithms have proved valid for some surveillance and research applications such as the identification of bloodstream infections, other types of infection require more nuanced review by experienced clinicians for diagnosis and follow-up (Cato, Cohen, & Larson, 2015).

The proliferation of publications using EHRs and administrative data sources have helped to further our understanding of the benefits, as well as the limitations of data from these sources Häyrinen, Saranto, & Nykänen, 2008). However, published information that focuses on the technical and logistical challenges of formulating usable research databases from the information stored in electronic patient records is lacking. The development of any system is an iterative process that combines the expertise of the users with the technical skills of the developers. Greater focus on methods of collecting, integrating, processing, and storing electronic patient data for research may help streamline database development for clinical and health services researchers (Bowles et al., 2013).

Acknowledgments

The authors gratefully acknowledge the work of Dr. Mandar Apte and the programmers and administrators who provided access to the data, and Dr. Suzanne Bakken who freely shared her wealth of knowledge.

Funding

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by a grant from the National Institute of Nursing Research, National Institutes of Health (R01NR010822).

Author Biographies

Bevin Cohen, MPH, MPhil, is program director of the Center for Interdisciplinary Research to Prevent Infections at Columbia University School of Nursing.

David K. Vawdrey, PhD, is Vice President of The Value Institute at NewYork-Presbyterian Hospital and assistant professor of clinical biomedical informatics in the Department of Biomedical Informatics at Columbia University.

Jianfang Liu, PhD, MAS, data manager for Columbia University School of Nursing.

David Caplan, BS, is director of data analytics at NewYork-Presbyterian Hospital.

E. Yoko Furuya, MD, MS, is medical director of infection prevention & control at NewYork-Presbyterian Hospital and associate professor of medicine in the Division of Infectious Diseases at Columbia University Medical Center.

Frederick W. Mis, PhD, is manager of IT business solutions at NewYork-Presbyterian Hospital.

Elaine Larson, RN, PhD, CIC, FAAN, is associate dean for research and professor of nursing research at Columbia University School of Nursing.

References

Apte M, Landers T, Furuya Y, Hyman S, Larson E. Comparison of two computer algorithms to identify surgical site infections. Surgical Infections (Larchmont). 2011; 12(6):459–464.

Apte M, Neidell M, Furuya EY, Caplan D, Glied S, Larson E. Using electronically available inpatient hospital data for research. Clinical and Translational Science. 2011; 4(5):338–345. [PubMed: 22029805]

Blumenthal D. Stimulating the adoption of health information technology. New England Journal of Medicine. 2009; 360(15):1477–1479. [PubMed: 19321856]

Bowles KH, Potashnik S, Ratcliffe SJ, Rosenberg M, Shih NW, Topaz M, Naylor MD. Conducting research using the electronic heath record across multi-hospital systems: Semantic harmonization implications for administrators. Journal of Nursing Administration. 2013; 43(6):355–360. [PubMed: 23708504]

Cato KD, Cohen B, Larson E. Data elements and validation methods used for electronic surveillance of health care-associated infections: A systematic review. American Journal of Infection Control. 2015; 43(6):600–605. [PubMed: 26042848]

Cimino JJ. From data to knowledge through concept-oriented terminiologies: Experience with the medical entities dictionary. Journal of the American Medical Informatics Association. 2000; 7(3): 288–297. [PubMed: 10833166]

Committee on the Analysis of Massive Data. Committee on Applied and Theoretical Statistics. Board on Mathematical Sciences. Their Applications, Division on Engineering and Physical Sciences. National Research Council National Research Council. Frontiers in massive data analysis. The National Academies Press; Washington, DC: 2013. Retrieved from http://www.nap.edu/catalog/18374/frontiers-in-massive-data-analysis

Cusack CM, Hripcsak G, Bloomrosen M, Rosenbloom ST, Weaver CA, Wright A, Mamykina L. The future state of clinical data capture and documentation: A report from AMIA's 2011 policy meeting. Journal of the American Medical Informatics Association. 2013; 20(1):134–140. [PubMed: 22962195]

Dahr V. Data science and prediction. Communications of the ACM. 2013; 56(12):64–73.

Dalto JD, Johnson KV, Gardner RM, Spuhler VJ, Egbert L. Medical information bus usage for automated IV pump data acquisition: Evaluation of usage patterns. International Journal of Clinical Monitoring and Computing. 1997; 14(3):151–154. [PubMed: 9387004]

Gardner RM, Hawley WL, East TD, Oniki TA, Young HF. Real time data acquisition: Recommendations for the Medical Information Bus (MIB). International Journal of Clinical Monitoring and Computing. 1991; 8(4):251–258. [PubMed: 1820414]

Halamka JD. Early experiences with big data at an academic medical center. Health Affairs. 2014; 33(7):1132–1138. [PubMed: 25006138]

Häyrinen K, Saranto K, Nykänen P. Definition, structure, content, use and impacts of electronic health records: A review of the research literature. International Journal of Medical Informatics. 2008; 77(5):291–304. [PubMed: 17951106]

Herman, M.; Rivera, S.; Mills, S.; Sullivan, J.; Guerra, P.; Cosmas, A.; Kim, M. The field guide to data science. Booz Allen Hamilton; McLean, VA: 2013.

Hersh WR, Cimino J, Payne PRO, Embi P, Logan J, Weiner M, Saltz J. Recommendation for the use of operational electronic health record data in comparative effectiveness research. eGEMS (Generating Evidence and Methods to Improve Patient Outcomes). 2013a; 1(1) Article 14. Retrieved from http://repository.academyhealth.org/egems/vol1/iss1/14/.

Hersh WR, Weiner MG, Embi PJ, Logan JR, Payne PR, Bernstam EV, Saltz JH. Caveats for the use of operational electronic health record data in comparative effectiveness research. Medical Care. 2013b; 51(8 Suppl 3):S30–S37. [PubMed: 23774517]

Jeon CY, Furuya EY, Berman MF, Larson EL. The role of pre-operative and post-operative glucose control in surgical-site infections and mortality. PLoS One. 2012a; 7(9):e45616. [PubMed: 23029136]

Jeon CY, Furuya EY, Smaldone A, Larson EL. Post-admission glucose levels are associated with healthcare-associated bloodstream infections and pneumonia in hospitalized patients with diabetes. Journal of Diabetes Complications. 2012b; 26(6):517–521.

Jeon CY, Neidell M, Jia H, Sinisi M, Larson E. On the role of length of stay in healthcare-associated bloodstream infection. Infection Control and Hospital Epidemiology. 2012c; 33(12):1213–1218. [PubMed: 23143358]

Jhung MA, Banerjee SN. Administrative coding data and health care-associated infections. Clinical Infectious Diseases. 2009; 49(6):949–955. [PubMed: 19663692]

Johnson SB, Bakken S, Dine D, Hyun S, Mendoonca E, Morrison F, Stetson P. An electronic health record based on structured narrative. Journal of the American Medical Informatics Association. 2008; 15(1):54–64. [PubMed: 17947628]

Landers T, Apte M, Hyman S, Furuya Y, Glied S, Larson E. A comparison of methods to detect urinary tract infections using electronic data. Joint Commission Journal on Quality and Patient Safety. 2010; 36(9):411–417. [PubMed: 20873674]

Miriovsky BJ, Shulman LN, Abernethy AP. Importance of health information technology, electronic health records, and continuously aggregating data to comparative effectiveness research and learning health care. Journal of Clinical Oncology. 2012; 30(34):4243–4248. [PubMed: 23071233]

Murdoch TB, Detsky AS. The inevitable application of big data to health care. Journal of the American Medical Association. 2013; 309(13):1351–1352. [PubMed: 23549579]

National Institutes of Health. Data science at NIH: BD2K. Jul 23.2015 Retrieved from https://datascience.nih.gov/bd2k.

National Institute of Nursing Research. Distribution of the cost of antimicrobial resistant infections (Project No. 1R01NR010822-01). 2007 Retrieved from http://projectreporter.nih.gov/project_info_description.cfm?aid=8784060&icde=25704072.

National Institute of Nursing Research. Health information technology to reduce healthcare-associated infections (Project No. 6R01NR010822-06). 2012 Retrieved from http://projectreporter.nih.gov/project_info_description.cfm?aid8784060&icde25704072.

Neidell MJ, Cohen B, Furuya Y, Hill J, Jeon CY, Glied S, Larson EL. Costs of healthcareand community-associated infections with antimicrobial-resistant versus antimicrobial-susceptible organisms. Clinical Infectious Diseases. 2012; 55(6):807–815. [PubMed: 22700828]

Nelson NC, Evans RS, Samore MH, Gardner RM. Detection and prevention of medication errors using real-time bedside nurse charting. Journal of the American Medical Informatics Association. 2005; 12(5):390–397. [PubMed: 15802486]

Ohno-Machado L. NIH’s big data to knowledge initiative and the advancement of biomedical informatics. Journal of the American Medical Informatics Association. 2014; 21(2):193. [PubMed: 24509598]

Patel SJ, O’Toole D, Larson E. A new metric of antibiotic class resistance in gram-negative bacilli isolated from hospitalized children. Infection Control and Hospital Epidemiology. 2012; 33(6):602–607. [PubMed: 22561716]

Pogorzelska-Maziarz M, Furuya EY, Larson EL. Risk factors for methicillin-resistant staphylococcus aureus bacteraemia differ depending on the control group chosen. Epidemiology and Infection. 2013; 141(11):2376–2383. [PubMed: 23425708]

Poon EG, Wright A, Simon SR, Jenter CA, Kaushal R, Volk LA, Bates DW. Relationship between use of electronic health record features and health care quality: Results of a statewide survey. Medical Care. 2010; 48(3):203–209. [PubMed: 20125047]

Rosenbloom ST, Denny JC, Xu H, Lorenzi N, Stead WW, Johnson KB. Data from clinical notes: A perspective on the tension between structure and flexible documentation. Journal of the American Medical Informatics Association. 2008; 18(2):181–186. [PubMed: 21233086]

Toh S, Platt R. Is size the next big thing in epidemiology? Epidemiology. 2013; 24(3):349–351. [PubMed: 23549179]

Vawdrey DK, Gardner RM, Evans RS, Orme JF Jr. Clemmer TP, Greenway L, Drews FA. Assessing data quality in manual entry of ventilator settings. Journal of the American Medical Informatics Association. 2007; 14(3):295–303. [PubMed: 17329731]

Westra BL, Delaney CW, Konicek D, Keenan G. Nursing standards to support the electronic health record. Nursing Outlook. 2008; 56(5):258–266. [PubMed: 18922281]

Wolfe CM, Cohen B, Larson E. Prevalence and risk factors for antibiotic-resistant community-associated bloodstream infections. Journal of Infection and Public Health. 2014; 7(3):224–232. [PubMed: 24631369]

Table 1

Approaches for Overcoming Challenges of Working With Electronic Data.

NIH BD2K challenge	Challenges	Recommendations
Locating data and software tools.	Difficult to identify and establish contact with owner or administrator of each data source. Some data needed for analyses are proprietary and not released for research purposes.	• Discuss software needs with programmers from each data source before choosing which package(s) to purchase. • Make software decisions with frontline programmer(s) who will be helping to deliver the data, not the administrator who may lack hands-on experience with preparing the data. • Discuss specific data sources before developing research protocol. Identify the specific data source and any limitations *a priori*. Obtain a sample of the data if possible to ensure it reflects the intended construct and troubleshoot any quality issues. • Retain this code to make sure subsequent data extractions are done using the same methodology and taken from the same source. Work with the same programmer or team if possible to ensure consistency.
Getting access to the data and software tools.	Lack of clarity regarding the order in which approvals should be obtained (e.g., IRB approval was required prior to institutional data use approval, and vice versa). Reliance on programmers with other obligations for data extractions. Programmers may lack the time or experience to review data for accuracy, requiring multiple iterations of data extraction.	• Consider in advance the physical limitations of data sharing and work with relevant IT departments to establish the most efficient system. • Learn who is responsible for granting permissions to use and access data and discuss with them in advance whether direct access to the data can be provided, or whether the data must be delivered by another programmer. • If programming staff from the source data system must be used, account for how that person's time will be allocated and funded.
Standardizing data and metadata.	Evolving institutional data use policies and procedures. Shifting roles, responsibilities, overlap, and turnover among data administrators. Some variables may not be available due to missing fields, inaccurate recording, or changes in recording practices over time. Sources of the same data may not match. Data delivered in incompatible formats.	• Contribute to process improvement by providing feedback about the experience of using electronic data for research purposes. • Keep abreast of changes in institutional policies and staff.
Extending policies and practices for data and software sharing.	No dedicated support for programmers providing data from existing sources. Inadequate funding for data storage space or multiple software packages. Policies and procedures governing secure data transfer evolve rapidly, making it difficult to remain in compliance.	• Understand current policies, keep abreast of changes, and establish collaborative relationships with data administrators. • Consider how long it will take to gain necessary approvals and account for this in the study timeline.
Organizing, managing, and processing data.	Codebooks describing the origins of each element in the raw data are often not available. Difficult to reconcile old and new coding schemes when changes are made over time. Uncertainty about which data source should be considered the *gold standard* when assessing validity. Changes in data collection and storage procedures over time not always documented.	• Maintain detailed records of how every data element was extracted, regardless of whether the study's data manager or programming staff from the source data performed the queries. • When available, retain old codebooks from source data because these may be overwritten as changes occur over time. • Keep detailed records of the decision rules and methodology used to create each variable.

NIH BD2K challenge	Challenges	Recommendations
Developing new methods for analyzing and integrating data.	Clinical investigators must agree on variable definitions that will be suitable for use in multiple study aims. Missing data are common, and effect of bias on planned analyses must be taken into account.	• Develop phenotyping algorithms to identify conditions that are not directly ascertainable from existing electronic data fields. • Conduct validation studies to determine sensitivity and specificity of various data sources relative to each other and clinician chart review.
Training researchers who can use data effectively.	Skills needed to carry out the project not fully understood *a priori*. Clinical investigators not familiar with the technical aspects of the project. Few data managers have programming, analytical, and clinical expertise.	• Create codebooks that include detailed variable definitions, including detailed descriptions of the data sources and known limitations. • For variables created for a specific purpose or project, retain decision rules and rationale so that future investigators can properly determine whether the variable is relevant and appropriate for their study aims. • Partner with bioinformatics, information technology and other staff to provide appropriate expertise.

Source. National Institutes of Health, 2015.

Note. NIH = National Institutes of Health; BD2K = Big Data to Knowledge; IRB = institutional review board.

Table 2

Data Elements and Sources.

Element	Source
Admit source	Manager of data analytics
Risk of mortality and severity of illness	Manager of data analytics
Discharge disposition	Clinical data warehouse
Demographic data	Clinical data warehouse
Present on admission flag for ICD-9 codes	Manager of data analytics
Blood culture	Clinical data warehouse
Urine culture	Clinical data warehouse
Wound culture	Clinical data warehouse
Respiratory culture	Clinical data warehouse
Urine microscopy	Clinical data warehouse
Location	Clinical data warehouse
Medical insurance	Clinical data warehouse
Discharge address	Clinical data warehouse
ICD-9 procedure codes and dates	Clinical data warehouse
ICD-9 diagnosis code	Clinical data warehouse
Central line documentation	Clinical data warehouse
Medication administration record (MAR)	Clinical data warehouse
Urinary catheterization	Clinical data warehouse
Operating room procedure	Operating room
Operating room anesthesia type	Operating room
Charge details	Manager of data analytics
Diagnostic-related groups (DRGs)	Manager of data analytics
Reference tables[a]	Clinical data warehouse, manager of data analytics

[a] More than 30 different reference tables such as ICD-9 codes, lists of clinical units, and codebooks for antibiotic codes or organism codes of culture data.