Root Causes Underlying Challenges to Secondary Use of Data

Jessica S. Ancker, MPH, PhD,[1,2] Sarah Shih, MPH,[3] Mytri P. Singh, MPH,[3] Andrew Snyder, MPA,[3] Alison Edwards, MStat,[1,2] Rainu Kaushal, MD, MPH,[1,2] with the HITEC investigators

[1]Division of Quality and Medical Informatics, Departments of Pediatrics and Public Health, Weill Cornell Medical College, New York, NY; [2]Health Information Technology Evaluation Collaborative (HITEC); [3]Primary Care Information Project (PCIP), New York City Department of Health and Mental Hygiene, New York, NY

Abstract

Although one potential benefit of electronic information systems is the opportunity for secondary use of data, it is often challenging in practice to reuse data. We identify challenges to the secondary use of electronic data from a web-based project management system, and trace these challenges to their root causes. Data quality issues arose from: differential incentives for integrity of different data; software flexibility that allowed a single task to be documented in multiple ways; variability in documentation practices; variability in use of standardized vocabulary; and changes in project procedures and system configuration over time. These issues are very similar to the issues that pose challenges for secondary uses of clinical and operational data for research, public health, and quality improvement. We conclude that secondary use of operational data requires an in-depth understanding of the primary workflow processes that produced the data, as these processes lead to data integrity issues.

Introduction

Although electronic information systems are developed to assist in day-to-day operational tasks, they offer the additional promise of data reuse for secondary purposes such as research, quality improvement, and public health.[1-7] Ambulance dispatch calls, retail pharmacy sales of both prescription and over-the-counter drugs, employee absentee rates, and emergency department visit data are all examples of electronic data collected for operational purposes that have been used successfully for syndromic surveillance.[5-7] For example, pharmacy sales can indicate the onset of community influenza activity before it appears in laboratory data.[7] The secondary use of health data is an active area of public policy discussion,[1-2, 8-9] particularly in light of the federal electronic health record (EHR) incentive program designed to increase adoption of EHRs.[10] For example, clinical data could assist in identifying patients eligible for pharmaceutical clinical trials, providing a potential revenue source for the sustainability of EHRs.[11]

Nevertheless, data collected for one purpose are rarely ideally suited for secondary use. Data are frequently of variable quality, and missing data may be common. Manual processing may be needed to assess quality and standardize data formats for analysis.[6] Lack of data standards, or inconsistent application of them, may make it difficult or impossible to analyze data without advanced natural language processing techniques.[3-4]

As part of a series of research and quality improvement projects, we began examining data from a project management system being used to track the progress of electronic health record (EHR) implementations by a regional extension center. This system, hosted by Salesforce (Salesforce.com, Inc., San Francisco, CA), contains information about several thousand clinicians and practices that are receiving EHR implementation support from the Primary Care Information Project (PCIP) at the New York City Department of Health and Mental Hygiene.[12]

As we examined the project management data, we encountered a variety of data quality issues reminiscent of larger issues in secondary use. In this paper, we identify and describe these challenges, trace them to their root causes, and place them in context of similar issues in the literature on secondary use.

Background

The Primary Care Information Project is an initiative of the New York City Department of Health and Mental Hygiene with the mission of improving the delivery of health care in ambulatory settings through promoting adoption and use of EHRs in New York City. PCIP purchases EHR software licenses on behalf of eligible providers, subsidizes maintenance and support costs for 2 years, manages implementation processes in cooperation with the EHR vendor, and provides additional post-go-live EHR training and support with a focus on quality improvement. In 2010, the Fund for Public Health in New York won a federal regional extension center (REC) award and established the Regional Electronic Adoption Center for Health (REACH), a program under PCIP.

Since early 2007, a web-based project management system product by Salesforce.com has been used to track implementations. The project management system is used routinely by multiple PCIP teams. For example, outreach staff collect information about clinicians potentially interested in implementing an EHR, and document ongoing contacts with them. In addition, members of the implementation team use the system as they launch the EHR implementation for each small practice, collect additional descriptive information, and document key project milestones. These implementation staff capture a variety of descriptive information about each practice in structured and free-text data fields, attach documents to the record, and use free-text fields to write notes about telephone calls, questions, unresolved problems, and to-dos. Some of the many milestones recorded in the database in structured format include the date the contract was signed, the date of the so-called "kickoff call" at which the project plan was agreed upon, the dates of EHR and practice management system training sessions, and the EHR go-live date. After the EHR implementation process, a team of quality improvement staff use the same database to document training and assistance provided to clinicians and office staff.

As a result, the database contains descriptive records for individual people, as well as a complex set of longitudinal records for healthcare organizations. Currently, the database contains information about more than 2500 healthcare providers, 600 small private physician practices, and 30-plus community health centers, as well as 4 hospital out-patient departments at various stages of EHR implementation.

Methods

We began examining the project management database for several purposes. First, we were interested in studying the challenges associated with EHR implementation among PCIP's participating practices and providers (an ongoing study being reported elsewhere). In addition, we had quality improvement goals for improving PCIP project management procedures.

For the ongoing EHR implementation study described above, we identified more than 30 variables in Salesforce with the potential to be relevant to the outcomes under evaluation. We used the Salesforce.com report tools to query the database for these variables, and computed frequencies to determine the rates of missing data. In cases when different variables had an obvious relationship to each other, we computed crosstab frequencies in order to identify inconsistencies and potential errors, such as a situation in which number of provider full-time equivalents (FTEs) was greater than the number of healthcare providers, or when the start date for a project was recorded as occurring after its end date. In addition, during all analyses, we tracked occurrence of any duplicate records (practices occurring in the database more than once). We held a series of weekly team meetings over about 4 months with key informants involved in the data collection to trace the root causes of these data quality issues and, in some cases, to develop data remediation plans.

The current study was part of a larger study of EHR implementation at PCIP being conducted as part of HITEC (the Health Information Technology Evaluation Collaborative), an academic consortium designated by the state of New York as the evaluation entity for health IT projects funded under the Health Care Efficiency and Affordability Law for New Yorkers capital grants program. The study was approved by the Weill Cornell Medical College Institutional Review Board.

Results

We present 4 illustrative data quality issues and their root causes from one data set of small community practices participating in EHR implementations. These data quality issues were selected for presentation because resolving them was critical before the data could be used for secondary purposes, and because they appeared to illustrate more generalizable issues.

DATA QUALITY ISSUE 1: In this data set, 544 small practices had signed a contract to join PCIP. Of these, 430 (79%) had a recorded EHR go-live date, indicating that they had completed their EHR implementation; the remaining 114 (21%) were still in the process of implementation. However, 265 of the 430 were either missing the date upon which implementation started ("kickoff call"), or had inconsistent dates in different data fields of the database.

Primary and secondary uses of these data: The primary use for which these dates were collected was to establish a project plan for a practice, then document its progress. The order of milestone dates was more important than specific times between them; for example, the contract had to be signed before any of the subsequent milestones. However, for secondary use, the dates became important as markers for duration of implementation and its components.

Root causes of data quality problems: In tracing the data quality issue to the root causes, it became clear that different dates had different interested stakeholders, as well as different financial and contractual implications. Specifically, all the project stakeholders needed access to the correct contract signed date because it marked the start of the small practice's two-year software license. As a result, this date as recorded in the database was highly reliable.

By contrast, the kickoff call was originally a process that launched a series of events, and only later was identified as an operational start point that marked the begin date of implementation. As a result, as part of PCIP process improvement, PCIP worked with the EHR vendor to retrospectively capture the kickoff call date in records where it had not originally been captured. This required the EHR vendor staff to double-enter data into their own project management system and into the Salesforce system, leading to the potential for inconsistent data. During this retrospective data entry process, documentation practices varied, with some of the staff documenting the actual event date and others documenting the originally planned kickoff date, which was not always corrected if the kickoff call date was rescheduled.

In addition, the Salesforce database was constructed in such a way that there were two fields in that reflected the kickoff call date (the date field, and a "stage history" field). Although this flexibility was meant to provide better documentation capabilities for the users, it led to inconsistencies because the fields were not linked, and neither was definitively identified as the gold standard.

Remediation plan: Several members of the implementation team manually reviewed the dates, supporting documents, and free text notes in each practice's electronic record to determine when the "kickoff call" had actually occurred. The resulting gold standard list was subsequently used to correct the data in the database.

DATA QUALITY ISSUE 2: Of the 544 small practices, 31 (5.7%) were documented to have had a previous EHR before joining PCIP, 236 (43.4%) indicated they did not have an EHR, and the remaining 277 (50.9%) had missing data.

Primary and secondary uses of these data: These data were collected as part of an application form that assessed the practice's eligibility for the PCIP program as well as its perceived readiness for the new technology. The perceived readiness questions included questions about previous exposure to EHRs and other technologies. The secondary use of these data was as an indicator of a practice's experience with technology, which might correlate with the speed or ease of the implementation.

Root causes of data quality problems: The missing data problem originated in several changes in the department's procedures pertaining to the recruitment of new practices, which led to corresponding configuration changes in the electronic systems.

In the early years of the EHR implementation program, physician practices completed a paper application form, which was sometimes input into the database by project manager but other times was scanned and attached as a PDF to the project management record, where it could be consulted by any PCIP staff member. However, later, the department implemented an online application form linked directly to the Salesforce database, so that the questionnaire answers automatically populated the database. As a result, our initial attempt to export the questionnaire answers for analysis revealed large quantities of missing data in the structured fields.

An additional challenge was that in several cases, the PDF was not linked to the practice's electronic project management record but rather to the electronic record for the practice employee who had completed the questionnaire. This was most likely because at the time the application was submitted, an electronic record had not yet been created for the practice.

Remediation plan: A student intern retrieved the PDFs where available and manually input the questionnaire data into the appropriate fields of the database.

DATA QUALITY ISSUE 3: In our initial data query, we identified several practices with the same name but different PCIP-assigned ID numbers, as well instance in which as the same ID number was assigned to practices with different names.

Primary and secondary uses of these data: The PCIP ID number was originally assigned to track each practice in terms of their service contract and linked this contract to their name as originally entered. For secondary use, the PCIP ID number became the way that all entries about any practice were linked for tracking and trending over time.

Root causes of data quality issues: Most PCIP staff tended to use the practice name as its identifier, rather than the PCIP ID number. Although this did result in the ad hoc development of a standardized vocabulary of practice names, practice names still occasionally varied, especially for newly enrolled practices. Over time, some practices changed their names, merged, split, or closed, leading to duplicate records. Duplicate PCIP ID numbers occurred in a very small number of cases, most of which were traced to preliminary contacts with practices that did not end up enrolling with PCIP and that were associated with an almost entirely empty electronic record.

Remediation plan: Manual review of records successfully disambiguated all of the cases.

DATA QUALITY ISSUE 4: A database field entitled "number of providers" for a practice yielded a different number than was produced by a count of individual provider records linked to the practice record.

Primary and secondary uses of these data: For primary use, the number of providers was helpful in developing the project plan as well as tracking completion of milestones such as provider training. For secondary use, the number of providers was collected as a potential predictor of implementation time for the entire EHR implementation.

Root causes: The "number of providers" field originated from the value on the application questionnaire, which was either self-reported by the practice or estimated by a PCIP outreach team member. At best, it represented an estimate of practice's staffing level before joining PCIP for very rough planning purposes. However, the electronic database records associated with the individual providers reflected the actual number of EHR software licenses issued upon joining PCIP. This number was determined to be more reliable, as external stakeholders (in this case the software vendor) needed to know the number of software licenses.

Remediation plan: No remedial actions were taken, but we determined to use the provider count for future analyses rather than the "number of providers" field.

Discussion

A large project management data system used in a consistent fashion for 4 years provided a rich data set for secondary uses including research and quality improvement. Nevertheless, early experiences using this data for secondary purposes revealed considerable variability in quality and integrity of the data. In our exploration of root causes, we determined that these variations in data quality arose from:

- ***Differential incentives for the accuracy of the data***. Data were documented consistently if they had had financial or contractual implications and were of interest to external stakeholders such as lawyers, the software vendor, or the clinician clients of PCIP, whereas data being used solely for internal purposes showed more variability.
- ***Flexibility in system software that allowed multiple routes to documenting the same tasks***. For example, two structured fields were available for documenting a particular milestone date, and the application questionnaire was accepted as both PDF and structured data.
- ***Variability in documentation practices among different personnel documenting the same task.*** For example, a particular date field could be used to document either the scheduled date of an event or the actual date of that event.
- ***Variability in use of standardized vocabulary,*** specifically, the internally developed standardized vocabulary of practice names.
- ***Changes in project procedures and electronic system configuration over time,*** as when a paper questionnaire was replaced with an electronic version.

A larger issue linking all of these observations was that our secondary use of data, which required aggregating historical data within each practice and also across practices, required a different and generally higher degree of data integrity than was required for the original primary use. Staff members could successfully manage EHR implementation even with imperfect database data because this database was only one source of information: project managers were also immersed in a rich ongoing stream of information from meetings, telephone calls, e-mail, site visits, and paper documents. In addition, the sequential nature of project management meant that pieces of data might be relevant only for short periods of time, limiting the impact of any inaccuracies or missing data in the database. Finally, in this decentralized system, a single project manager took responsibility for a single practice throughout the implementation process. Idiosyncratic ways of entering data thus had no serious impact, as a single person was both the source of data input and the audience for that data.

Although the current data set included no health data, the issues we have identified map closely to previously identified problems in clinical data quality that pose challenges for secondary uses such as research and quality reporting. Botsis and colleagues[3] have identified such issues in particularly granular detail in a description of their use of the Columbia University Medical Center clinical data warehouse to conduct a survival analysis of pancreatic cancer patients. Although they were able to use database queries to retrieve information, they also had to do significant manual review and data abstraction, including manual review of free-text notes to ensure the accuracy of the extracted data. Incomplete, inconsistent, and inaccurate data were common; in some patient subsets, important variables had more than 50% missing values. The authors did not do a formal root cause analysis, but were able to identify potential causes. For example, inconsistencies arose when the same data were being entered into different fields of a single EHR,[3] just as we observed in cases when the Salesforce database offered multiple alternatives for documenting the same information. Botsis et al also noted that missing and inconsistent data were common. This may have been because the information needed to document treatment may not have included the types of disease progression events that were of interest from a secondary use perspective. Dates were particularly likely to be missing; the difficulty of accurately interpreting temporal information in clinical data is a well-known problem.[13]

The issues recorded here, their root causes, and potential solutions were not evident from inspection of the database. Rather, they emerged only after intensive and collaborative discussions among researchers and the project managers with primary responsibility for data entry. Explanations for the data quality issues and novel ways of analyzing the data emerged only from in-depth understanding of the daily workflow being documented in the project management system and the history of the organization.

We conclude that researchers interested in secondary use of data must immerse themselves in the workflow processes being documented in order to understand the data and reasons for problems. In addition, organizations that may be interested in secondary uses of data will benefit from close attention to documentation practices, including

incentivizing the documentation of important tasks, eliminating redundancy in data fields, ensuring consistent data definitions, and promoting uniform standards and training for those involved in documentation. As others have noted, "no purely technical solution can overcome the capture of inaccurate information by the user of a clinical information system. As such, nontechnical innovations that help improve the accuracy of recorded information and incentivize consistently accurate data collection are critical to the success of research initiatives that rely on the presence of such data."[1]

Acknowledgments

We are grateful to Maryam Kahn for data entry, and to Anupam Kashyap for consultation on some of the root causes of the data quality issues. This study was supported by the New York State Department of Health (NYS contract number C023699).

References

1. Weiner MG, Embi PJ. Toward reuse of clinical data for research and quality improvement. *Annals of Internal Medicine.* 2009;151(5):359-360.

2. Safran C, Bloomrosen M, Hammond WE, et al. Toward a national framework for the secondary use of health data: An American Medical Informatics Association white paper. *JAMIA.* 2007;14(1):1-9.

3. Botsis T, Hartvigsen G, Chen F, Weng C. Secondary uses of EHR: Data quality issues and informatics opportunities. *AMIA Joint Summits on Translational Science Proceedings.* 2010;2010.

4. Prokosch H-U, Ganslandt T. Reusing the electronic medical record for clinical research. *Methods of Information in Medicine.* 2009;48:38-44.

5. Heffernan R, Mostashari F, Das D, et al. System descriptions: New York City syndromic surveillance systems. *Morbidity and Mortality Weekly Report.* 2004;53 (Suppl):23-27.

6. Heffernan R, Mostashari F, Das D, Karpati A, Kulldorff M, Weiss D. Syndromic surveillance in public health practice, New York City. *Emerging Infectious Diseases.* 2004;10(5):858-864.

7. Increased antiviral medication sales before the 2005-06 influenza season – New York City. *MMWR Morbidity and Mortality Weekly Report.* 2006;55:277-279.

8. Behrman RE, Benner JS, Brown JS, McClellan M, Woodcock J, Platt R. Developing the Sentinel system – a national resource for evidence development. *New England Journal of Medicine.* 2011;364(6):498-499.

9. President's Council of Advisors on Science and Technology. *Report to the president – Realizing the full potential of health information technology to improve health care for Americans: the path forward.* Washington, DC: Executive Office of the President of the United States,;2010.

10. Medicare and Medicaid Programs. *Electronic Health Record Incentive Program, Final Rule*: 75 Federal Register 144 (28 July 2010);2010.

11. Miller JL. The EHR solution to clinical trial recruitment in physician groups. *Health Management Technology.* 2009;http://www.healthmgttech.com/.

12. Mostashari F, Tripathi M, Kendall M. A tale of two large community electronic health record extension projects. *Health Aff (Millwood).* Mar-Apr 2009;28(2):345-356.

13. Hripcsak G, Elhadad N, Chen Y, Zhou L, Morrison F. Using empiric semantic correlation to interpret temporal assertions in clinical texts. *Journal of the American Medical Informatics Association* 2009;16(To):220-227.