Secondary Use of EHR: Data Quality Issues and Informatics Opportunities

Taxiarchis Botsis[a,b], Gunnar Hartvigsen[a,c], Fei Chen[b], Chunhua Weng[b]

[a] *Department of Computer Science, University of Tromsø, Tromsø, Norway*
[b] *Department of Biomedical Informatics, Columbia University, New York, U.S.A.*
[c] *Norwegian Centre for Integrated Care and Telemedicine, University Hospital of North Norway, Tromsø, Norway*

Abstract

Given the large-scale deployment of Electronic Health Records (EHR), secondary use of EHR data will be increasingly needed in all kinds of health services or clinical research. This paper reports some data quality issues we encountered in a survival analysis of pancreatic cancer patients. Using the clinical data warehouse at Columbia University Medical Center in the City of New York, we mined EHR data elements collected between 1999 and 2009 for a cohort of pancreatic cancer patients. Of the 3068 patients who had ICD-9-CM diagnoses for pancreatic cancer, only 1589 had corresponding disease documentation in pathology reports. Incompleteness was the leading data quality issue; many study variables had missing values to various degrees. Inaccuracy and inconsistency were the next common problems. In this paper, we present the manifestations of these data quality issues and discuss some strategies for using emerging informatics technologies to solve these problems.

Introduction

Electronic health records (EHR) have become a pervasive healthcare information technology. They replaced paper-based systems in many healthcare organizations and garnered rich health data, which hold great value for reuse. As The American Medical Informatics Association (AMIA) stated at its website, (http://www2.amia.org/inside/initiatives/healthdata/): "*Secondary use of health data can enhance healthcare experiences for individuals, expand knowledge about disease and appropriate treatments, strengthen understanding about the effectiveness and efficiency of our healthcare systems, support public health and security goals, and aid businesses in meeting the needs of their customers*". Retrospective analysis of health data holds promise to expedite scientific discovery in medicine and constitutes a significant part of clinical research. Currently, secondary use of clinical data is still at its early stage [1]. National initiatives have been created to facilitate widening use of EHR to support clinical research in the United States [2].

This paper reports our first-hand experience with some data quality issues in a survival analysis study for pancreatic cancer. We first describe our data source and methods for case identification and research variable extraction. Then we identify the major data quality issues and their manifestations. We discuss the potential applications of the emerging health informatics technologies to mitigate these data quality issues.

Data Source and Methods

The Columbia University Medical Center's clinical data warehouse is the data source of this study (http://ctcc.cpmc.columbia.edu/rdb/index.html). This warehouse has been in operation since 1994 and has accumulated health data for more than 2.7 million of patients seen at The NewYork Presbyterian Hospital. Since 2002, a comprehensive controlled clinical vocabulary called the Medical Entities Dictionary (http://med.dmi.columbia.edu/) has been used to integrate data of various semantic representations from heterogeneous hospital information systems for the clinical data warehouse. Our 3-step procedure to identify the cases of pancreatic cancer in our data warehouse is described as follows:

Step 1. We used the 9th version of International Classification of Diseases, Clinical Modification (ICD-9-CM) and its codes corresponding to the "malignant neoplasm of pancreas" (157.0-157.9) to identify all the patients with ICD-9 diagnoses during the period of (01//01/1999-01/30/2009). The pathology reports, radiology reports, clinical notes, laboratory tests, discharge summaries, as well as the drug registry and administrative files were extracted for further analysis.

Step 2. We queried the pathology reports to exclude patients who did not have adequate documentation about pancreatic cancer diagnoses in the reports. Initially, we applied an SQL query using variations of 'pancreas' key term; subsequently, we manually reviewed the query output to either exclude non-malignancies or filter out pancreatic tumors that were not diagnosed as primary lesions.

Step 3. We divided the remaining patients into groups for endocrine and exocrine neoplasms. Each group was further classified by disease subtype standards, e.g. the WHO Classification of Epithelial Tumors of the Exocrine Pancreas.

After identifying a cohort of patients with pancreatic cancer, we manually abstracted and automatically extracted specific pathologic characteristics from these patients' pathology reports, such as the size, location, and differentiation of tumors, lymph node metastasis, as well as the specifications of related health conditions such as chronic pancreatitis. Various EHR data elements were reviewed to collect other study variables. For example, we abstracted information about metastasis at diagnosis and progression of disease from radiology reports. We abstracted personal medical history, personal habits (e.g., smoking and alcohol consumption), as well as family medical history from clinical notes. We also queried laboratory tests tables to extract biochemistry at diagnosis (aspartate aminotransferase-AST, alanine aminotransferase-ALT, alkaline phosphatase-ALP, albumin, total bilirubin) and tumor markers preoperatively (CA19-9, carcinoembryonic antigen-CEA). Furthermore, we used the drug registry to extract chemotherapy regimens and used administrative files to extract patient demographics (e.g., birth date, gender, race and ethnicity). Also, discharge summaries and hospitalization archives served as an extra data source for filling the missing values of the aforementioned study variables. The tumor stage for all the patients was manually annotated using standard parameters. Manual review of the free text patient information was performed to ensure the accuracy of information extraction. We also applied the three common measurements of data quality, as specified below:

Incompleteness – missing information;

Inconsistency – information mismatch between various or within the same EHR data source;

Inaccuracy – non-specific, non-standards-based, inexact, incorrect, or imprecise information.

Descriptive statistics for incompleteness and qualitative observations for the other two measurements are presented below. The results of the extraction/abstraction process fed both the calculation of descriptive statistics and the formation of specific observations. Particularly, the discrepancies between or within the various EHR elements (measurement of inconsistency) were identified by matching the extraction/abstraction output from two sources (or within the same source) for a single parameter and a number of randomly selected patients e.g. the SQL query output in the drug registry and the manual review of clinical notes for chemotherapy regimen were compared.

Results

Using the ICD-9-CM codes for pancreatic malignancies (157.0-157.9), 3068 patients were identified in the CUMC clinical data warehouse for the reported period (01/01/1999 – 01/30/2009). However, after querying the pathology reports for these patients, we found that 1479 (48%) patients did not have corresponding diagnoses or disease documentation in the pathology reports. Among the remaining 1589 (52%) patients, incompleteness in the key study variables that define the disease stage (e.g. tumor size and extension beyond pancreas, lymph node and distant metastasis for exocrine tumors) further reduced the size of our cohort to 522 (17%) patients, which included 98 patients with endocrine pancreatic cancer, 218 with early stage (resectable) exocrine pancreatic ductal adenocarcinoma and 206 with late stage exocrine pancreatic ductal adenocarcinoma[1]. Significant information incompleteness was observed in many of the study variables so that variables of more than 50% incompleteness were excluded from further analysis. For example, incompleteness of family history of cancer for exocrine pancreatic adenocarcinomas was 56% and 52% for the early and late stage respectively. Table 1a and 1b show the degree of incompleteness (= the percentage of patients with incomplete information in each group of our cohort) for the study variables of the survival analysis. For endocrine pancreatic tumors (Table 1a), the degree of information incompleteness was between 0% (age, gender, functional status and surgery) and 44% (tumor markers). The degree of incompleteness was higher in the later stage ductal adenocarcinomas (Table 1b), with many of the selected variables having more than 50% missing values.

The values of some study variables had to be manually inferred by combining extracted data from various EHR data sources. For example, disease stage and progression are two indispensible variables in a survival analysis, but both of them were not explicitly documented in the EHR and had to be manually inferred from other key variables. Even if disease progression was documented, often the information was not explicitly available and required backward

[1] The selection was based on the AJCC TNM system and the WHO classification system criteria for the exocrine and the endocrine pancreatic tumors correspondingly.

comparison of the most recent with the previous radiology reports using deep knowledge of the international standards and guidelines for defining disease stage and disease progression. Moreover, it was difficult to define time parameters for certain events, e.g. the timing of disease progression. As an important characteristic of dynamic patient phenotype and a crucial factor in survival analysis, temporality was often not captured accurately. Except for "time to critical event" (e.g., death or censoring) and "date of diagnosis", which were commonly documented in a straightforward manner, it was difficult to determine the exact period for medical interventions or events, e.g. the duration of chemotherapy treatments.

Table 1a – Degree of incompleteness for some of the study variables for the endocrine pancreatic tumors

Variables	Endocrine
Necrosis	20%
Number of Mitoses	21%
Lymph Node Metastasis	28%
Perineural/Lymphovascular Invasion	15%
Differentiation	38%
Size	6%
Chronic Pancreatitis	14%
Smoking- Alcohol	27%-29%
History of Other Cancer	35%
Family History of Cancer	39%
Tumor Markers	46%

Table 1b – Contrast of degrees of incompleteness for some of the study variables between the early and late stage ductal adenocarcinomas

Variables	Early	Late
Lymph Node Metastasis	1%	88%
Differentiation	3%	49%
Localization	0%	76%
Tumor Size	2%	86%
Smoking- Alcohol	37%-41%	46%-48%
Chronic Pancreatitis	0%	92%
History of Other Cancer	17%	28%
Biochemistry Labs	6%-9%	13%-23%
Tumor Markers	24%	29%-35%
Chemotherapy	0%	26%
Family History of Cancer	56%	52%

We also observed that information inconsistency occurred either between different EHR data sources or within the same EHR data source. For example, some chemotherapy regimens were documented in the clinical notes but not in the drug registry. However, there was evidence that the patient was treated exclusively in our institution so that their treatment information should be documented in the drug registry. Also, in a few cases, pancreatitis was diagnosed as being chronic in the pathology reports but was reported as being only acute in the clinical notes. Such inconsistencies across different data sources revealed multiple inconsistent entries about the same health problem in different components of the EHR, which could be made by the same or different clinician(s). Uncoordinated or redundant data entries into different data sources in EHR could not only cause information discrepancies but also form big barriers to selecting reliable data sources for secondary use of EHR data. Furthermore, information inconsistency within the same data source was also observed. Some patients received two different ICD-9-CM codes for their diagnoses of diabetes, both 250.01 and 250.02 for type-1 and type-2 respectively.

Information inaccuracy was also frequently observed. It was reflected as poor granularity of the diagnosis terms or disease classification codes and inadequate or non-standardized documentation of disease status or treatment details. Consequently, such information could not satisfy the information needs of a survival analysis study. For example, the non-specific ICD-9-CM code for diabetes (250 for diabetes mellitus) was often used. Also, the patient treatment plan was often sketchy with inadequate temporal information. Some study variables (e.g., chemotherapy cycles) were hard to infer because of the inaccuracy of the base variables (e.g., chemotherapy treatment information). Furthermore, in some patient cases, the endocrine tumor grade was also not defined following the WHO classification system guidelines.

The above problems can be exacerbated since EHR users tend to copy and paste information [3], which can propagate the errors.

Discussion

Semantic representational variations among data collected by different EHR systems are typical in many healthcare organizations. A clinical data warehouse aggregates data and greatly facilitates retrospective analysis and data mining [4]. The Columbia University clinical data warehouse equipped with the MED demonstrates the value of a comprehensive controlled clinical vocabulary for integrating heterogeneous data. The current study would be impossible without this valuable data resource. The issues we described above, i.e.,

information incompleteness, inaccuracy, and inconsistency, are not unique to our data warehouse, but are common challenges for many institutions.

It could be argued that a weakness of our study is the lack of descriptive statistics for inconsistency and inaccuracy; however, this is not a simple task since there are many aspects for consideration. If we further analyzed the example of chemotherapy regimen that was mentioned above, we would have observed that there were various ways of registering drugs in EHR, i.e. using: (1) the trade drug name (e.g. Gemzar-Taxotere, in clinical notes), (2) the main compound name (e.g. Gemcitabine-Docetaxel, in drug registry), (3) an acronym substituting the drug names (e.g. GTX, mainly in clinical notes). To accurately and fully check the inconsistencies could be a project itself. Similarly, inaccuracy measurement would require an extensive chart review of each patient case, a rather cumbersome process. Considering the aforementioned we decided to provide some qualitative examples only.

Within a clinical data warehouse, to reduce the health data that is unavailable, inaccessible or incomputable, new technology for storage (e.g. for radiology data) and new methods for natural language processing (e.g. for symptoms or signs recorded on free-text formats) are needed. Various approaches have been suggested for mining clinical data warehouses, such as an extended Structured Query Language (SQL) for manipulating groups of records [5] or text mining tools for the natural language processing of the pathology reports [6]. It should be mentioned though that text mining tools cannot achieve 100% accuracy. Similarly, SQL queries can only assist the researcher in accomplishing part of the tasks, as in our study where SQL queries had to be combined with laborious manual scrutiny of disease-specific information. Therefore, we suggest combining dedicated text mining tools and special post processing to facilitate information retrieval. A dedicated text-mining tool should be based on a source- and domain-specific lexicon. For example, in our case study, the pathology reports could be mined using a lexicon that includes the appropriate pathology terms for pancreatic cancer; this lexicon should be also adjustable to support the mining of other types of notes. Post processing queries could further filter the outcomes and aggregate the values for the disease variables of interest.

Beyond any solutions that may improve the accuracy of information extraction, strategies for improving the quality of collected data are much needed as well. Some solutions were mentioned above (e.g., new tools, better classification systems, etc.); however, their success demands considerable user involvement. The lessons learned from projects that linked EHR with clinical research databases might offer better insight to a more efficient research data capture [7].

The discrepancies of the diagnoses for pancreatic cancer between the ICD-9-CM codes and the pathology notes could be attributed to two possible reasons: (a) *information fragmentation* – e.g., some patients had been initially treated elsewhere so that our institution did not have the longitudinal health records for this subset of transferring patients; and (b) *lack of contextual information in structured disease diagnoses* – e.g., for some patients, the pancreatic tumor was not primary but metastatic, while pathology reports only captured information for the primary tumors. The latter case reveals a problem currently associated with the ICD-9-CM coding system, which is that its classification cannot distinguish primary from metastatic tumors.

The varying degrees of information incompleteness between the early and late stage ductal adenocarcinomas (Table 1b) indicated that EHR probably captures less information for patients with terminal diseases than for patients with less severe diseases. It is likely that severe patients might be transferred to dedicated cancer treatment centers so that their information was not captured in our EHR. As aforementioned, this information fragmentation is a big cause for information incompleteness.

Incompleteness caused by information fragmentation of the healthcare systems (i.e., patients moving between multiple healthcare entities for special referrals or emergency healthcare, with each entity holding partial health records for the patients) could be mitigated using health information exchange (HIE) methods that support information federation across multiple healthcare entities. More national and regional health information exchange networks with broad connectivity among EHR systems should be further developed to improve patient data flow across different healthcare entities.

Information incompleteness due to poor documentation could be attributed to both patients and healthcare providers who did not report or document critical information, e.g. family history or personal habits. In the case of survival analyses, this may be handled by the appropriate imputation methods that fill in the missing values for a set of pre-defined variables. However, other solutions should be applied. To date, secondary use of EHR data is largely focused on ad-hoc data extraction support, rather than on proactive documentation support that

improves the comprehensiveness of health data upfront. Next, we describe how emerging informatics technologies such as personal health records (PHR) and clinical registries could offer potential solutions.

PHR are a new form of health records for engaging individual patients to control access to their own health information [8]. Besides the support for enhanced patients-caregiver communications, PHR could also be a potential solution to many data quality issues. For example, information such as personal habits that was rather incomplete in our study is less likely to be recorded by physicians compared to urgent medical conditions in a limited patient encounter time window. Unambiguously, the implementation and adoption of PHR raise various issues such as data confidentiality and security, usability and user acceptance, and so forth; all these can be barriers to the uses of PHR and should be considered.

Clinical registries are another promising technology to improve data quality. A clinical registry collects data for a specific group of patients, e.g. patients with pancreatic cancer, has a predefined format and can be easily designed to interoperate with EHR system and hence to support patient information exchange and federation.

An alternative to address information incompleteness problem is to define "standard content" for EHR. To our knowledge, there is no community agreed-upon "essential content for EHR" or standard common data elements for EHR. It is unknown how much information is truly sufficient or needed at the point of care for diagnostic decision making and what information is mostly important for physicians during the limited patient visit time. For example, "lymph node metastasis" is practically unimportant for late stage ductal adenocarcinomas given that patients classified in this group will not survive long; this is probably the reason that tumor size was missing in 86% of the late stage cases. Answers to these questions can help doctors to better spare their time for entering only important and necessary data. These are open biomedical informatics research questions that have the potential to improve the efficiency of EHR data and alleviate documentation burdens, as well as to reduce redundancy caused by "copy-paste" errors. More standards for clinical documentation should be developed to address this problem.

Conclusion

PHR, clinical registry, and health information exchange will be the key enabling technologies for improving EHR data quality toward longitudinal health records. With more and more institutions maturing in clinical data warehousing, the next step is to develop new methods for clinical analytics. Advanced or automatic data validation and flexible data presentation tools should be developed to ensure information integrity. Effective strategies for secondary use of EHR data could also be accumulated from case studies and shared with the research community as the best practices.

Acknowledgments

We thank Alla Babina for retrieving data from the CUMC clinical data warehouse. This study was supported by the Research Council of Norway Grant 174934 and the NLM Grant 1R01LM009886-01A1.

References

[1] Prokosch HU, Ganslandt T. Perspectives for medical informatics. Reusing the electronic medical record for clinical research. Methods Inf Med 2009;48: 38-44.

[2] National Center for Research Resources (NCRR). Available at http://www.ncrr.nih.gov/news_&_events/upcoming_events/index.asp#10302009

[3] Hirschtick RE. A piece of my mind. Copy-and-paste. JAMA 2006;295: 2335-6.

[4] Lyman JA, Scully K, Harrison JH, Jr. The development of health care data warehouses to support data mining. Clin Lab Med 2008;28: 55-71, vi.

[5] Johnson SB, Chatziantoniou D. Extended SQL for manipulating clinical warehouse data. Proc AMIA Symp 1999: 819-23.

[6] Xu H, Anderson K, Grann VR, Friedman C. Facilitating cancer research using natural language processing of pathology reports. Stud Health Technol Inform 2004;107: 565-72.

[7] Duftschmid G, Gall W, Eigenbauer E, Dorda W. Management of data from clinical trials using the ArchiMed system. Med Inform Internet Med 2002;27: 85-98.

[8] Tang PC, Ash JS, Bates DW, Overhage JM, Sands DZ. Personal health records: definitions, benefits, and strategies for overcoming barriers to adoption. J Am Med Inform Assoc 2006;13: 121-6.

[9] Interfacing Registries with EHRs. AHRQ Draft Research Report, July 28, 2009.