International Journal of Medical Informatics 108 (2017) 78–84

Contents lists available at ScienceDirect

International Journal of Medical Informatics

journal homepage: www.elsevier.com/locate/ijmedinf

Research Paper

Need of informatics in designing interoperable clinical registries

Majid Rastegar-Mojarad[a,b,*], Sunghwan Sohn[a], Liwei Wang[a], Feichen Shen[a], Troy C. Bleeker[c], William A. Cliby[c], Hongfang Liu[a]

[a] Department of Health Sciences Research, Mayo Clinic, Rochester, MN, USA
[b] Dep. of Health Informatics and Administration, UW-Milwaukee, Milwaukee, WI, USA
[c] Mayo Clinic, Rochester, MN, USA

ARTICLE INFO

Keywords:
Informatics
Clinical registry
Machine learning
Natural language processing

ABSTRACT

Clinical registries are designed to collect information relating to a particular condition for research or quality improvement. Intuitively, informatics in the area of data management and extraction plays a central role in clinical registries. Due to various reasons such as lack of informatics awareness or expertise, there may be little informatics involvement in designing clinical registries. In this paper, we studied a clinical registry from two critical perspectives, data quality and interoperability, where informatics can play a role. We evaluated these two aspects of an existing registry, Gynecology Surgery Registry, by mapping data elements and value sets, used in the registry, to a standardized terminology, SNOMED-CT. The results showed that majority of the values are ad-hoc and only 6 of 91 procedures in the registry could be mapped to the SNOMED-CT. To tackle this issue, we assessed the feasibility of automated data abstraction process, by training machine learning classifiers, based on existing manually extracted data. These classifiers achieved a reasonable average F-measure of 0.94. We concluded that more informatics engagement is needed to improve the interoperability, reusability, and quality of the registry.

1. Introduction

National Institute of Health (NIH) defines registry as "*a collection of information about individuals, usually focused around a specific diagnosis or condition*" utilized for research and quality improvement. Another definition of clinical registry is "*an organized system that uses observational study methods to collect uniform data (clinical and other) to evaluate specified outcomes for a population defined by a particular disease, condition, or exposure, and that serves one or more predetermined scientific, clinical, or policy purposes*." [1]. Clinical registries have been designed for various purposes such as: studying the natural history of disease [2], analyzing clinical outcome of surgery/treatment [3], comparing different treatment methods [4], and measuring quality of care [5]. However many clinical registries have been designed successfully and there are user guides, aimed to assist with designing registries [1], but still there are some caveats on designing clinical registries such as: interoperability. Recently the United States congress approved a bill [6], which requires the U.S. Department of Health and Human Services to make recommendations regarding the structure and scope of clinical data registries. This bill mostly focuses on a set of standards to aid interoperable exchange of information between clinical notes and registries [7] and contains some recommendations about design and structure of clinical registries.

Besides interoperability, we faced some other challenges in designing successful and cost-effective clinical registry, while we were developing an enterprise-wide clinical registry infrastructure at Mayo clinic. We studied several existing clinical registries and noticed that 1) data quality 2) cost of human abstraction 3) lack of interoperability with EMRs and 4) lack of a master data resource are some of challenges in designing a clinical registry.

In this study, we hypothesized that effective informatics engagement in designing clinical registries can lead to cost effective, reusable, and interoperable clinical registries. Informatics '*studies the representation, processing, and communication of information in natural and artificial systems*' [8] and in healthcare domain, informatics defines as "*applying information science, computer technology, and statistical modeling techniques to develop decision support systems for improving both health service organizations' performance and patient care outcomes*" [9]. In the process of designing and implementing a clinical registry, informatics can contribute significantly, at least, in two tasks:

1) Defining data elements and determining the corresponding value sets
2) Collecting data and populating the registry.

The first task is critical for designing a reusable and interoperable

* Corresponding author at: Department of Health Sciences Research, Mayo Clinic, Rochester, MN, USA.

http://dx.doi.org/10.1016/j.ijmedinf.2017.10.004
Received 18 May 2016; Received in revised form 29 September 2017; Accepted 2 October 2017

clinical registry. To ensure interoperability of registry, data elements and value sets should come from a standardized and universal health care terminology [10,11] and the value sets should be comprehensive and cover all possible values for the associated data elements. In biomedical informatics domain, there is a valuable resource called, Unified Medical Language System (UMLS) [12], which integrates and distributes key terminologies and coding standards to assist with creating effective and interoperable systems. One of the common and popular clinical terminologies in the UMLS is Systematized Nomenclature of Medicine-Clinical Terms (SNOMED-CT) [13], which can be used in the first task.

The second task is another area that informatics could play an important role and affect cost of human abstraction and more importantly quality of data. Data collection tool (DCT) [14] and Computer-Assisted Coding (CAC) [15–17] are two informatics tools which can assist abstractors in chart abstraction process and make the process automatic or semi-automatic. Using these tools in the second task, not only reduces the coding burden, but decreases some human errors and inconsistency between resources by following a simple rule in informatics "*one entry of a piece of data, many uses*" [18].

In this paper, we studied a clinical registry, Gynecology Surgery Registry, used by the Gynecologic Surgery practice at Mayo Clinic in Rochester, Minnesota. It contains basic encounter information, patient demographics, and various surgical related data elements such as procedures, diagnoses, and co-morbidities. In this study, we focused on one data element, *procedure,* of the registry, which captures and codifies the list of procedures performed in gynecologic surgeries. The study contains two parts. To assess interoperability of the clinical registry, we investigated data elements and their value sets and cross-referenced with a standardized terminology, SNOMED-CT. In the second part, we focused on chart abstraction process and making the process more automatic and error-proof. As CDC, we trained multiple binary classifiers, for each procedure in the registry, to identify whether procedures are reported in clinical notes or not. To find the best set of features and learning model for the classifiers, three classification methods (i.e., Naïve Bayes, Random Forest, and Support Vector Machine (SVM)) and three sets of features (i.e., unigrams, bi-grams, and topics retrieved by Latent Dirichlet Allocation [19]) were evaluated. To obtain more insights, the classifiers are analyzed and reasons for low performance in some of the classifiers are discussed.

In the following sections, we first discuss related work. Then the case study is presented. The results of our analysis are presented next followed by error analysis of the classifiers. Finally, the learned lessons, limitations, and future work are discussed.

2. Related work

Many clinical registries have been developed and studied for different conditions and diseases such as: "Alzheimer's Prevention Registry" [20], "Genome Connect" [21], and "Cancer Genetics Network" [22]. Shahian et al. [23] developed a clinical registry to study readmission measure for coronary artery bypass grafting surgery. McCombs et al. [24] studied and analyzed data from a department of veterans affairs clinical registry to evaluate the risk of long-term morbidity in patients with chronic hepatitis C. Sites et al. [25] illustrated the use of international clinical registry in quality improvement. Nwomeh et al. [26] studied trauma registry that as one of components in trauma care systems. Megan Quinn [27] has studied characteristics of cancer in adolescents using Tennessee cancer registry from 2004 to 2008. This type of publications mentioned or presented the importance role that clinical registries can play in various types of researches [28], but there are not much about how to design a successful clinical registry [29], what main concerns are and how to address those concerns. A publication supported by Robert Wood Johnson Foundation [28] is one of few publication which highlighted shortcomings in designing registry and noted these flaws can limit the role of registry. Gliklich et al. [29] provided a comprehensive user guide to design and develop a clinical registry. Silva et al. developed a standard framework for developing a device registry [30]. In this study, we emphasized the role of informatics in designing clinical registry.

Clinical registries value depends on the quality of their data [31–33]. Data in clinical registries have been compared with administrative claims data in several studies [34–37]. However, none of these studies focused on accuracy of clinical registries [32] or accuracy in data population process. There are three main factors impacting data accuracy 1) errors in original resources [38] 2) missing data [39–41] and 3) human errors [42]. The first one could be fixed to some extent with cross-referencing different resources such as clinical notes, surgery notes, structured data, and lab tests. Missing data issue has been addressed in several studies. Mendelsohn at el. [39] studied and characterized missing data in clinical registries and associated factors. Norris at el. [40] developed a method for handling missing data in a cardiac registry. They merged registry data with administrative data to fill missing data. In this study, we only addressed the third one, human errors in populating process. After assessing the accuracy of data in our case study, we discussed how informatics could improve the accuracy and decrease human involvement in populating process.

To improve the quality of data in clinical registry and decrease the ratio of errors (especially when subjective judgment is involved [15]) in the process of collecting information from medical records (chart abstraction process), CAC could be a useful tool. In general, CAC systems utilize natural language processing and machine learning algorithms to facilitate coding process. Predicting procedure codes from clinical notes or other type of text data has been studied in several domains. Hersh et al. [16] developed a machine learning system to assess the accuracy of predicting procedures codes from emergency room dictations. Using available data in trauma registry data, they trained the logistic regression classifiers with words appeared in the notes as features. Resnik at el. [15] implemented a CAC system that performed strongly relative to human performance. Morris at el. [17] developed an automated coding system called LifeCode which could be accurate as human coders. However, because of ambiguity in some of medical coding rules and guidelines, involving a human abstractor besides CAC system, seems necessary and will improve the accuracy. In addition to increase consistency in coding, a CAC system decreases needed labor and time for the process. In our case study, we implemented a CAC system using natural language processing and machine learning algorithms to investigate the potential use of assisting the human abstractors in the populating process.

3. Case study: gynecology surgery registry

The original database used in Gynecology Surgery Registry was derived from a professional society database in gynecologic cancer, and started to collect Gynecology Surgery data in 1990s at Mayo clinic. The primary goal was to tracking data retrospectively rather than focusing quality of elements. There was no electronic medical record available when the database was started and several modifications have occurred over the ensuing years. These limitations have made the database difficult to systematically collect data and hindered interoperability. Hence, we performed a case study applying informatics to the current Gynecology Surgery Registry at Mayo Clinic with respect to interoperability and data quality. We also evaluated the feasibility of using the current registry data to automatically codify procedures.

Fig. 1 shows a snapshot of one surgical encounter in the registry. The current database contains 10,160 visits from 1/20/1998 to 12/11/2014. For visits (7123 visits) with surgical notes, a human abstractor extracted procedures from the surgical notes. In Fig. 1, for the surgical note appeared in "*Procedure From SIRS*" textbox, a human abstractor identified 6 procedures (each row is one procedure). In the current design of the Gynecology Surgery Registry, each procedure is a combination of three fields: "*Anatomic location*", "*Procedure*", and "*Method*