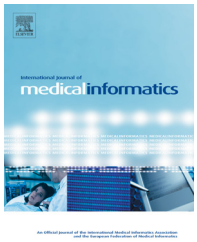
INTERNATIONAL JOURNAL OF MEDICAL INFORMATICS 84 (2015) 76–84

journal homepage: www.ijmijournal.com



Understanding data requirements of retrospective studies

Edna C. Shenvi[a], *Daniella Meeker*[b], *Aziz A. Boxwala*[c,*,1]

[a] *Division of Biomedical Informatics, University of California San Diego, La Jolla, CA, United States*
[b] *RAND Corporation, Santa Monica, CA, United States*
[c] *Meliorix Inc., La Jolla, CA, United States*

ARTICLE INFO

Article history:
Received in revised form
26 September 2014
Accepted 3 October 2014

Keywords:
Data models
Data standards
Queries

ABSTRACT

Background and objective: Usage of data from electronic health records (EHRs) in clinical research is increasing, but there is little empirical knowledge of the data needed to support multiple types of research these sources support. This study seeks to characterize the types and patterns of data usage from EHRs for clinical research.

Materials and methods: We analyzed the data requirements of over 100 retrospective studies by mapping the selection criteria and study variables to data elements of two standard data dictionaries, one from the healthcare domain and the other from the clinical research domain. We also contacted study authors to validate our results.

Results: The majority of variables mapped to one or to both of the two dictionaries. Studies used an average of 4.46 (range 1–12) data element types in the selection criteria and 6.44 (range 1–15) in the study variables. The most frequently used items (e.g., procedure, condition, medication) are often available in coded form in EHRs. Study criteria were frequently complex, with 49 of 104 studies involving relationships between data elements and 22 of the studies using aggregate operations for data variables. Author responses supported these findings.

Discussion and conclusion: The high proportion of mapped data elements demonstrates the significant potential for clinical data warehousing to facilitate clinical research. Unmapped data elements illustrate the difficulty in developing a complete data dictionary.

1. Introduction

Data collected during clinical care can constitute a valuable source of information for secondary use in research studies. Often, this data is used in observational studies; for example, to conduct comparative effectiveness research [1]. Additionally, the data is used to identify patients that might be eligible for prospective studies [2], to populate research data registries [3], and to annotate biospecimens with phenotypic data [4].

The increasing use of electronic health record (EHR) systems and other information systems in clinical practice is increasing the volume of clinical data and provides further opportunities for research. This data, which is in digital form and is codified, also can be much more efficient to use compared to the traditional method of reviewing and abstracting

* *Corresponding author at*: Meliorix Inc., 7919 Avenida Kirjah, La Jolla, CA 92037, United States. Tel.: +1 617 294 9269; fax: +1 435 417 5731.
E-mail address: aziz.boxwala@meliorix.com (A.A. Boxwala).
[1] This research was performed while the author was at UCSD.

http://dx.doi.org/10.1016/j.ijmedinf.2014.10.004
1386-5056/

data from patients' paper medical records or electronic notes (often referred to as *chart review*). In order to facilitate the use in research of data from clinical information systems, most notably from EHRs, many healthcare organizations are employing clinical data repositories (CDRs).

While CDRs are being increasingly employed to support researchers, there is little empirical knowledge of the data needed from clinical databases to support the types of research studies described above. The study described here aims to address this gap by analyzing the data requirements of retrospective observational studies (also known as "chart reviews") published within a one-month period. Our objective was to characterize the data needed for performing such studies, by analyzing the selection criteria of the studies and the types of study data collected. This is a follow-up study to our previous pilot work [5] that mapped data elements from eligibility criteria in smaller number of ambulatory care studies. We have broadened this study in sample size and research settings, and have investigated the types of data used during the study. Furthermore, we attempted to validate our results through a survey of the authors of the published studies.

1.1. *Background*

Many healthcare organizations, primarily academic medical centers, their affiliates, and large health maintenance organizations [6] have implemented CDRs as a tool for researchers. These CDRs draw data mainly from the EHR system, though in many cases, data also are included from other systems such as the billing systems. The data elements that are available in these CDRs are the ones that are commonly recorded as discrete and coded elements in the EHRs such as the patient's demographics, diagnoses, encounters, laboratory test results, medications, and diagnostic and therapeutic procedures. The structure of the clinical data elements in EHRs is very complex, reflecting the nuances of clinical workflows and the operational needs of healthcare organizations. The data are of high dimensionality and often imprecise [7]. Our institution's EHR system, a commercially available product, has several hundred tables in its database. This level of breadth and complexity of the database schema is typical of EHR systems. CDR systems tend to use a less complex data schema, typically containing tens of tables. The choices made in the design of CDR database schemas can impact the granularity of the data elements and the relationships amongst them, and can therefore impact the utility and usability of the CDR for research. For example, problem lists in EHRs are used to document clinical problems including admission diagnoses, discharge diagnoses, and differential diagnoses that are to be ruled-in or ruled-out. CDRs may not consider these variations in their diagnoses list, which can potentially lead to incorrect inclusion or exclusion of patients. EHRs might also record preliminary and final results of diagnostic tests. If the CDRs record only the final results, then studies on preliminary results using the CDR might not be possible.

Another important challenge associated with the design of the CDRs and associated tools is usability, enabling researchers to easily obtain study data. Often designers face tradeoffs between usability and database efficiency. Since many biomedical scientists are not trained in writing database queries, graphical query tools are provided with many CDRs [3,8] to assist researchers in specifying the data to be queried. For example, a cohort discovery tool enables the researchers to compose and execute queries that estimate patient counts matching those queries (due to privacy and regulatory concerns, these tools often prevent the user from obtaining more detailed results such as the patient records). The cohort discovery tools allow the researchers to construct cohort specifications in the form of logical combinations of predicates (inclusion criteria). In order to reduce the complexity of the user interface, not all query predicates can be defined in these tools. As illustrated in Fig. 1, compared to SQL there are limitations on the logical combinations of predicates. Another significant limitation found in some cases is that the predicates cannot be based on aggregate operations (e.g., all patients who have had *two or more* visits in the last year). Many cohort discovery tools [3,9], including the CRIQuET system [10] developed at our institution, share these limitations in the user interface. While these user interfaces might make the tools accessible for users without expertise or training in database queries, it is unclear if the queries constructed with these tools have sufficient expressivity for meeting the data needs of the researchers.

The study we conducted aims to improve the understanding of the data needed in clinical research studies in order to inform the design of schemas for CDRs, the prioritization of data that are needed for research studies, and the design of query tools that are easy to use and sufficiently expressive.

2. Methods

2.1. *Objectives and overview*

The objective of our study was to assess the data requirements for retrospective observational studies. Specifically, we aimed to characterize

1. The clinical data elements needed in these studies, i.e., the data variables.
2. The structure of the queries that have to be executed to obtain the data.

We analyzed patient selection criteria and data variables (which formed the study's data set) for retrospective observational studies. These studies relied upon paper or electronic clinical records to identify patients as the source of the data set. From the full-text manuscripts of a set of observational studies, we extracted the patient selection criteria and the data variables used within the studies. We then mapped the data elements in patient selection criteria and the data variables to data elements in two standards-based data dictionaries. We report the summary statistics of the mappings.

2.2. *Selection of studies*

We obtained a convenience sample of studies by performing a PubMed query for retrospective studies in core clinical journals, published in the month of December 2010 (either in print