International Journal of Medical Informatics 90 (2016) 40–47

Contents lists available at ScienceDirect

International Journal of Medical Informatics

journal homepage: www.ijmijournal.com

Data quality assessment framework to assess electronic medical record data for use in research

Andrew P. Reimer[a,b,*], Alex Milinovich[b], Elizabeth A. Madigan[a]

[a] Frances Payne Bolton School of Nursing, Case Western Reserve University, 10900 Euclid Ave, Cleveland, OH 44106, United States
[b] Cleveland Clinic, 10900 Euclid Avenue, Cleveland, OH 44195, United States

ARTICLE INFO

Article history:
Received 23 October 2015
Received in revised form 7 March 2016
Accepted 18 March 2016

Keywords:
Electronic medical records
Evaluation & assessment
Information storage
Retrieval & integration

ABSTRACT

Introduction: The proliferation and use of electronic medical records (EMR) in the clinical setting now provide a rich source of clinical data that can be leveraged to support research on patient outcomes, comparative effectiveness, and health systems research. Once the large volume and variety of data that robust clinical EMRs provide is aggregated, the suitability of the data for research purposes must be addressed. Therefore, the purpose of this paper is two-fold. First, we present a stepwise framework capable of guiding initial data quality assessment when matching multiple data sources regardless of context or application. Then, we demonstrate a use case of initial analysis of a longitudinal data repository of electronic health record data that illustrates the first four steps of the framework, and report results.
Methods: A six-step data quality assessment framework is proposed and described that includes the following data quality assessment steps: (1) preliminary analysis, (2) documentation–longitudinal concordance, (3) breadth, (4) data element presence, (5) density, and (6) prediction. The six-step framework was applied to the Transport Data Mart—a data repository that contains over 28,000 records for patients that underwent interhospital transfer that includes EMRs from the sending hospitalization, transport, and receiving hospitalization.
Results: There were a total of 9557 log entries of which 8139 were successfully matched to corresponding hospital encounters. 2832 were successfully mapped to both the sending and receiving hospital encounters (resulting in a 93% automatic matching rate), with 590 including air medical transport EMR data representing a complete case for testing. Results from Step 2 indicate that once records are identified and matched, there appears to be relatively limited drop-off of additional records when the criteria for matching increases, indicating the a proportion of records consistently contain nearly complete data. Measures of central tendency used in Step 3 and 4 exhibit a right skewness suggesting that a small proportion of records contain the highest number of repeated measures for the measured variables.
Conclusions: The proposed six-step data quality assessment framework is useful in establishing the metadata for a longitudinal data repository that can be replicated by other studies. There are practical issues that need to be addressed including the data quality assessments—with the most prescient being the need to establish data quality metrics for benchmarking acceptable levels of EMR data inclusiveness through testing and application.

1. Introduction

Medical transport plays an integral role in supporting health care delivery as patients often present to hospitals or clinics that do not provide the necessary services that acutely ill or injured patients require. Patient transfers primarily can be categorized as emergent or non-emergent. A growing body of evidence supports transfer of patients experiencing time sensitive emergencies such as trauma, stroke, or heart attack. However, there is sparse evidence to support the decisions of if, how, and when to transfer non-emergent patients who oftentimes experience poor outcomes [1–3].

Investigating patient transfers have presented multiple challenges due to the many facets involved in moving patients between hospitals and sometimes across health systems. A primary limitation is the lack or accessibility of data that are required to adequately assess the effect of the transfer on patient outcomes. Until recently, most research efforts investigating transferred

* Corresponding author at: Frances Payne Bolton School of Nursing, Case Western Reserve University, 2120 Cornell Rd, Cleveland, OH 44106, United States.
E-mail addresses: axr62@cwru.edu (A.P. Reimer), milinoa@ccf.org (A. Milinovich), elizabeth.madigan@case.edu (E.A. Madigan).

http://dx.doi.org/10.1016/j.ijmedinf.2016.03.006
1386-5056/

patients have remained isolated to individual units or hospitals, producing limited insight and restricting our overall understanding of how transfer influences patient outcome.

The proliferation and use of electronic medical records (EMR) in the clinical setting now provide a rich source of clinical data that can be leveraged to support research on patient outcomes, comparative effectiveness, and health systems research. Particularly, reusing EMR data provides the distinct ability to study patients and interventions in actual clinical practice as they naturally occur [4], facilitating rapid translation of findings back into practice. Most research efforts now include EMR data abstraction to support individual studies, or more generally to support aggregation of large volumes of data in disease specific registries or clinical data repositories. Such is the case for the Transport Data Mart (TDM) [5] that we developed to support comprehensive outcomes research efforts that aggregates patient data across the entire episode of care for a patient who is transported from one hospital to another. However, amassing the large volume and variety of data that robust clinical EMRs provide is only the first stage. Once the appropriate data are identified and aggregated, the suitability of the data for research purposes [6–8] must be addressed.

Typically initial efforts for assessing the quality of EMR data abstracted for research purposes are focused on identifying and validating that the correct patient population was identified and abstracted. One approach, proposed by Faulconer and Lusignan's [9] eight step approach to assessing diagnostic data quality, provides an example of the steps necessary to accurately identify a patient for inclusion in a specific disease registry. However, for transported patients this potentially complicated problem of identifying patients who are transported does not exist. A patient either undergoes transfer or does not, creating a singular inclusion criterion for abstracting that particular episode of care into the TDM. Another approach is the Data Quality Probe method proposed by Brown and Warmington [10] that identifies cases in an EMR system that are not successfully matched (concordance), or contain errors between one item and another, that when applied longitudinally can improve data entry practice and overall quality.

Transported patients present a different problem related to matching the individual encounters across the multiple admissions and discharges that represent the entire episode of care. Capturing the entire episode of care includes linking EMR data from the referring hospital EMR, the transport EMR— if available, and the receiving hospital EMR. While the definition of data completeness can vary depending on the task at hand [11], an overarching assessment of data completeness, or in this case, inclusion and concordance across data sources, must be evaluated to assess the overall integrity of data inclusion and integration within the TDM. Therefore, the purpose of this paper is two-fold. First, we present a stepwise framework capable of guiding initial data quality assessment when matching multiple data sources regardless of context or application. Then, we demonstrate a use case of initial analysis of a longitudinal data repository of electronic health record data that illustrates the first four steps of the framework, and report results.

2. Materials and methods

2.1. Framework

The conceptual framework guiding this study is based on the five dimensions presented by Weiskopf and Weng [6] (completeness, correctness, concordance, plausibility and currency), with a specific focus on guiding deep evaluation of completeness and concordance of variables used in record linkage, and then assessment of the variables required for analysis across multiple data sources. The five dimensions as defined by Weiskopf and Wang are: (1) completeness—"is a truth about a patient present in the EHR?" (2) correctness—is an element that is present in the EHR true?" (3) concordance—"is there agreement between elements in the EHR, or between the EHR and another data source?" (4) plausibility—"does an element in the EHR make sense in light of other knowledge about what that element is measuring?" and (5) currency—"is an element in the EHR a relevant representation of the patient state at a given point in time?"

Concordance becomes particularly important during data quality assessment in a longitudinal data repository. Due to the need to assess several elements of data presence and agreement across multiple records or data sources simultaneously in one distinct step, it is useful to think of concordance as a construct—"*longitudinal concordance*." The construct longitudinal concordance is defined as assessing data element presence, agreement, and source agreement of specified variables across multiple data sources. The first concept, data element presence is defined as the minimum required data elements to facilitate matching data across sources. For example, matching across medical record sources (as illustrated in the case example provided in detail later) would entail identifying the necessary variables required that might include: medical record number, patient name, admission date/time, etc.; while matching across disparate data sources such as Twitter accounts (twitter handle, date/time), to weather data (date/time, location), entails different data to assess for initial feasibility of matching across domains or sources. Then the second concept, data element agreement, is defined as "two or more elements within the target data domains or sources are compared to see if they report the same or compatible information" [6]. Lastly, data source agreement is an extension of data element agreement and is defined as "data from the data domains or sources are compared with data from another domain or source to determine if they are in agreement" [6]. By definition data element presence, data element agreement, and data source agreement are considered as individual data quality assessment methods. For the purpose of longitudinal data quality assessment, they are combined into one methodological step, and can be applied to any sets of variables within and across any two or more data sources. Therefore, the construct *longitudinal concordance* subsumes the dimensions of completeness and correctness within it. The importance of assessing longitudinal concordance can be illustrated by the patient demographics that are present in each episode of care record. The primary questions to be addressed when assessing longitudinal concordance include: (1) are each of the data elements of interest present in each record, and (2) do those same data elements agree across record sources and domains?

Operationalizing the currently proposed conceptual framework requires incorporating four definitions of data completeness [12] (documentation, breadth, density, and prediction) that offer specific assessment measures to develop a standard stepwise approach that can be replicated in other projects. Merging the data quality dimensions with the data quality assessment definitions yields a six step process – due to the addition of a preliminary assessment step for including external data sources (i.e. patient log), and the breakout of data element presence as a discreet and significant assessment step – to conducting data quality assessment for a longitudinal data repository as displayed in Fig. 2 with each step more fully described in the Section 2.4.

2.2. Guiding aim

The guiding aim for this investigation is to identify the total number of patient episodes of care that include data across the entire episode of care for all patients transferred by helicopter (sending hospitalization, transport, and receiving hospitalization). We choose this subgroup because it represents the most restrictive