Journal of Biomedical Informatics 46 (2013) 830–836

Contents lists available at SciVerse ScienceDirect

Journal of Biomedical Informatics

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

Defining and measuring completeness of electronic health records for secondary use ☆

Nicole G. Weiskopf [a,*], George Hripcsak [a], Sushmita Swaminathan [b], Chunhua Weng [a]

[a] Department of Biomedical Informatics, Columbia University, New York, NY 10032, United States
[b] Department of Computer Science, Columbia University, New York, NY 10027, United States

ARTICLE INFO

Article history:
Received 30 January 2013
Accepted 22 June 2013
Available online 29 June 2013

Keywords:
Data quality
Electronic health records
Secondary use
Completeness

ABSTRACT

We demonstrate the importance of explicit definitions of electronic health record (EHR) data completeness and how different conceptualizations of completeness may impact findings from EHR-derived datasets. This study has important repercussions for researchers and clinicians engaged in the secondary use of EHR data. We describe four prototypical definitions of EHR completeness: documentation, breadth, density, and predictive completeness. Each definition dictates a different approach to the measurement of completeness. These measures were applied to representative data from NewYork–Presbyterian Hospital's clinical data warehouse. We found that according to any definition, the number of complete records in our clinical database is far lower than the nominal total. The proportion that meets criteria for completeness is heavily dependent on the definition of completeness used, and the different definitions generate different subsets of records. We conclude that the concept of completeness in EHR is contextual. We urge data consumers to be explicit in how they define a complete record and transparent about the limitations of their data.

1. Introduction

With the growing availability of large electronic health record (EHR) databases, clinical researchers are increasingly interested in the secondary use of clinical data [1,2]. While the prospective collection of data is notoriously expensive and time-consuming, the use of an EHR may allow a medical institution to develop a clinical data repository containing extensive records for large numbers of patients, thereby enabling more efficient retrospective research. These data are a promising resource for comparative effectiveness research, outcomes research, epidemiology, drug surveillance, and public health research.

Unfortunately, EHR data are known to suffer from a variety of limitations and quality problems. The presence of incomplete records has been especially well documented [3–6]. The availability of an electronic record for a given patient does not mean that the record contains sufficient information for a given research task.

Data completeness has been explored in some depth. The statistics community has focused extensively on determining in what manner data are missing. Specifically, data may be considered to be missing at random, missing completely at random, or missing not at random [7,8]. Datasets that meet these descriptions require different methods of imputation and inference.

The statistical view of missing or incomplete data, however, is not sufficient for capturing the complexities of EHR data. EHR records are different from research data in their methods of collection, storage, and structure. A clinical record is likely to contain extensive narrative text, redundancies (i.e., the same information is recorded in multiple places within a record), and complex longitudinal information. While traditional research datasets may suffer from some degree of incompleteness, they are unlikely to reflect the broad systematic biases that can be introduced by the clinical care process.

There are several dimensions to EHR data completeness. First, the object of interest can be seen as the patient or as the health care process through which the patient was treated; there is a difference between complete information about the patient versus complete information about the patient's encounters. A patient with no health care encounters and an empty record has a complete record with respect to the health care process, but a blank one with respect to the patient. Furthermore, one can measure completeness at different granularities: the record as a whole or of logical components of the record, each of which may have its own requirements or expectations (e.g., demographic patient information versus the physician thought process) [9,10]. Another

* Corresponding author. Address: Department of Biomedical Informatics, Columbia University, 622 W 168th Street, VC-5, New York, NY 10032, United States.
E-mail address: nicole.weiskopf@dbmi.columbia.edu (N.G. Weiskopf).

1532-0464/$ - see front matter © 2013 The Authors. Published by Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.jbi.2013.06.010

dimension of completeness emerges from the distinction between intrinsic and extrinsic data requirements. One can imagine defining minimum information requirements necessary to consider a record complete (which could be with respect to either the patient or the health care process), or one can tailor the measurement of completeness to the intended use. Put another way, we can see completeness in terms of intrinsic expectations (i.e., based a priori upon the content) or extrinsic requirements (based upon the use) [11,12].

The EHR data consumers who define these extrinsic requirements will have different data needs, which will in turn dictate different conceptualizations of a complete patient record. Here, Juran's definition of quality becomes valuable: "fitness for use" [12]. It may be that data completeness does not have a simple, objective definition, but is instead task-dependent. Wang and Strong, for example, in their work developing a model of data quality, define completeness as "[t]he extent to which data are of sufficient breadth, depth, and scope for the task at hand" [13]. In other words, whether a dataset is complete or not depends upon that dataset's intended use or desired characteristics. In order to determine the number of complete records available for analysis one must first determine what it means to have a complete patient record. The quality of a dataset can only be assessed once the data quality features of interest have been identified and the concept of data quality itself has been defined [11].

Multiple interpretations of EHR completeness, in turn, may result in different subsets of records that are determined to be complete. The relationships between research task, completeness definition, and completeness findings, however, are rarely made explicit. Hogan and Wagner offer one of the most widely used definitions: "the proportion of observations that are actually recorded in the system" [5]. This definition does not, however, offer specific measures for determining whether a record is complete. Neither does it account for the possibility that completeness may be task-dependent. What proportion of observations should be present? Which observations are desired? Are there any other considerations beyond simple proportion? Furthermore, observations are complex, nested concepts, and it must be determined what level of detail or granularity is needed or expected. In order of increasing detail, one could record a visit that occurred, the diagnoses, all the symptoms, a detailed accounting of the timing of all the symptoms, the clinician's thought process in making a diagnosis, etc.

In the sections below, we enumerate four specific operational and measurable definitions of completeness. These definitions are not exhaustive, but they illustrate the diversity of possible meanings of EHR data completeness. We ran the definitions against our clinical database in order to demonstrate the magnitude of completeness in the database and to illustrate the degree of overlap among the definitions.

2. Materials and methods

Previously, we conducted a systematic review of the literature on EHR data quality in which we identified five dimensions of data quality that are of interest to clinical researchers engaged in the secondary use of EHR data. Completeness was the most commonly assessed dimension of data quality in the set of articles we reviewed [3]. Based upon this exploration of the literature on EHR data quality, consideration of potential EHR data reuse scenarios, and discussion with stakeholders and domain experts, we describe four prototypical definitions of completeness that represent a conceptual model of EHR completeness. Further definitions of completeness are possible and may become apparent as the reuse of EHR data becomes more common and more use cases and user needs are identified.

Fig. 1 presents a visual model of the four definitions of completeness, which are described further in Section 2.1. In this model of EHR data, every potential data point represents some aspect of the patient state at a specific time that may be observed or unobserved as well as recorded or unrecorded. The longitudinal patient course, therefore, can be represented as a series of points over time that may or may not appear in the EHR.

2.1. Definitions

2.1.1. Documentation: a record contains all observations made about a patient

The most basic definition of a complete patient record described in the literature is one where all observations made during a clinical encounter are recorded [5]. This is an objective, task-independent view of completeness that is, in essence, a measure of the fidelity of the documentation process. Assessments of documentation completeness rely upon the presence of a reference standard, which may be drawn from contacting the treating physician [14], observations of the clinical encounter [15], or comparing the EHR data to an alternate trusted data source—often a concurrently maintained paper record [16–19]. Documentation completeness is also relevant to the quality measurements employed by the Centers for Medicare & Medicaid Services [20].

In secondary use cases, however, the data consumer may be uninterested in the documentation process. Instead, completeness is determined according to how well the available data match the specific requirements of the task at hand, meaning that completeness in these situations is more often subjective and task-dependent. While documentation completeness is intrinsic, the following three definitions of completeness are extrinsic and can only be applied once a research task has been identified.

2.1.2. Breadth: a record contains all desired types of data

Some secondary use scenarios require the availability of multiple types of data. EHR-based cohort identification and phenotyping, for example, often utilize some combination of diagnoses, laboratory results, medications, and procedure codes [21–23]. Quality of care and clinician performance assessment also rely upon the presence of multiple data types within the EHR (the relevant data types vary depending upon clinical area) [20,24–27]. More broadly, researchers interested in clinical outcomes may require more than one type of data to properly capture the clinical state of patients [28,29]. In the above cases, therefore, a complete record may be one where a breadth of desired data types is present. It is important to note that the absence of a desired data type in a record does not necessarily indicate a failure in the clinical care process or in the recording process. Rather, it may be that a data

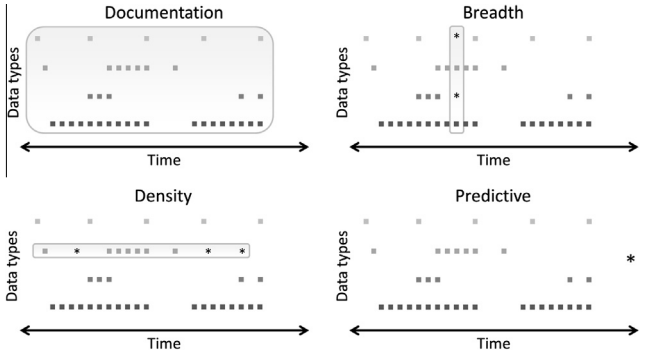

Fig. 1. An EHR completeness model. Each square point denotes an observed and recorded data point, stars are unobserved but desired data points, and the boxes indicate all data points that are required for a given task.