International Journal of Medical Informatics 94 (2016) 271–274

Contents lists available at ScienceDirect

International Journal of Medical Informatics

journal homepage: www.ijmijournal.com

Validating the extract, transform, load process used to populate a large clinical research database

Michael J. Denney (MA)[a], Dustin M. Long (PhD)[b], Matthew G. Armistead (BS)[a], Jamie L. Anderson (RHITCHTS-IM)[c], Baqiyyah N. Conway (PhD)[d,*]

[a] Biomedical Informatics, West Virginia Clinical and Translational Science Institute, Morgantown, WV, USA
[b] Department of Biostatistics, West Virginia University, Morgantown, WV, USA
[c] Department of Health Information Management, West Virginia University Healthcare, Morgantown, WV, USA
[d] Department of Epidemiology, West Virginia University, Morgantown, WV, USA

ARTICLE INFO

Article history:
Received 9 December 2015
Received in revised form 8 July 2016
Accepted 13 July 2016

Keywords:
Correctness
Clinical data warehouse
Electronic health record
Extract transform load
Informatics

ABSTRACT

Background: Informaticians at any institution that are developing clinical research support infrastructure are tasked with populating research databases with data extracted and transformed from their institution's operational databases, such as electronic health records (EHRs). These data must be properly extracted from these source systems, transformed into a standard data structure, and then loaded into the data warehouse while maintaining the integrity of these data. We validated the correctness of the extract, load, and transform (ETL) process of the extracted data of West Virginia Clinical and Translational Science Institute's Integrated Data Repository, a clinical data warehouse that includes data extracted from two EHR systems.
Methods: Four hundred ninety-eight observations were randomly selected from the integrated data repository and compared with the two source EHR systems.
Results: Of the 498 observations, there were 479 concordant and 19 discordant observations. The discordant observations fell into three general categories: a) design decision differences between the IDR and source EHRs, b) timing differences, and c) user interface settings. After resolving apparent discordances, our integrated data repository was found to be 100% accurate relative to its source EHR systems.
Conclusion: Any institution that uses a clinical data warehouse that is developed based on extraction processes from operational databases, such as EHRs, employs some form of an ETL process. As secondary use of EHR data begins to transform the research landscape, the importance of the basic validation of the extracted EHR data cannot be underestimated and should start with the validation of the extraction process itself.

1. Introduction

The widespread adoption of Electronic Health Records (EHR) offers great potential for clinical translational research through reuse of the data. As federal funding agencies heavily incentivize this reuse of EHR data, the conduct of clinical research will be greatly affected. A major caveat however is that EHR systems were not designed to be used for research. While it may be debated whether "data shall only be used for the purpose for which they were collected" [1] or whether that data simply needs to meet the criteria of "fitness for use," [2] EHR data were collected to support healthcare clinical decision making and not for research purposes. Unless its data are carefully validated for such repurposing, the integrity of the research results generated from it may be questionable at best.

A critical step in ensuring the validity of research is making sure the data are 'correct.' Correctness is one of the five dimensions of data quality put forth by Weiskopf and Weng in assessing the fitness of EHR data for its reuse for research. Their meta-analysis evaluated how 60 studies assessed correctness in the reuse of EHR data. For example, the definition of correctness suggested by Hogan and Wagner is summarized as the "proportion of data elements present that are correct." Weiskopf and Weng found that the most common method used for assessing correctness was comparison of EHR data to some gold standard [3].

* Corresponding author.
E-mail address: bnconway@hsc.wvu.edu (B.N. Conway).

http://dx.doi.org/10.1016/j.ijmedinf.2016.07.009
1386-5056/

The purpose of this study was to validate the correctness of the West Virginia Clinical and Translational Science Institute (WVCTSI)'s Integrated Data Repository (IDR) data elements. In this study we evaluated the IDR using the EHR as the gold standard in order to validate the correctness of the extract, transform and load (ETL) process used in migrating the data from the EHR sources to the IDR target. To do this, we used a two-step process in which we randomly selected data from a subset of patients and compared them to the EHR databases from which they were extracted.

2. Materials and methods

The WVCTSI IDR is a comprehensive clinical data warehouse, first deployed in June 2012. Currently, it contains observations on approximately 2 million patients, that is, information such as lab tests, medications, diagnoses and procedures as well as demographic data including but not limited to patient age, race and gender. The IDR contains over 250 million observations, captured from records of both inpatient and outpatient visits. The IDR uses the widely-employed database model of the i2b2 (Informatics for Integrating Biology and the Bedside) platform to store data. The i2b2 platform was designed by Partners Health System in conjunction with Harvard University faculty as part of an NIH-supported effort to develop a scalable informatics framework for translational research. This framework has been adopted by many major research institutions [4] and has become a standard tool used to support cohort discovery, clinical trial recruitment and hypothesis generation.

The IDR currently includes data from two sources, West Virginia University Healthcare's (WVUH)'s EpicCare and Medsite systems. WVUH is a multi-hospital entity, with over sixty affiliated physician practices and clinics, whose largest facility is Ruby Memorial Hospital, a 531 bed tertiary care hospital and Level One Trauma Center. The EpicCare application, from Epic Systems Corporation, provides WVU Healthcare with a full suite of integrated financial and clinical applications. The EpicCare application was implemented in 2008. Prior to that time, WVUH used the Medsite application as its EHR. Medsite was developed in-house and was in full scale usage by staff and clinicians from late 1998 until the Epic EHR implementation (March 2008). Medsite captured and integrated data from WVU Healthcare's inpatient and outpatient registration systems as well as ancillary systems such as laboratory, radiology and cardiology.

The WVCTSI's IDR was developed by an extract, transform and load (ETL) process [Fig. 1]. In this ETL process, data was first extracted from the source systems' databases (in this case, the Medsite and EpicCare applications); second, the extracted data was then transformed to make it accommodate the requirements of the IDR; and, third, once transformed the data was then loaded into the IDR's database. The ETL process was designed and developed entirely by the WVCTSI's Biomedical Informatics staff, including MJD, who were provided access to WVUH's Medsite and EpicCare data via direct Oracle database to Oracle database links. The ETL software was developed using Oracle's PL/SQL programming language and its Integrated Development Environment tool, SQL Developer. The initial ETL development began in early 2011 with Medsite's data and was completed with the first extract, transform and load of EpicCare data in mid-2012. The on-going ETL process is designed to run quarterly against WVUH's EpicCare database.

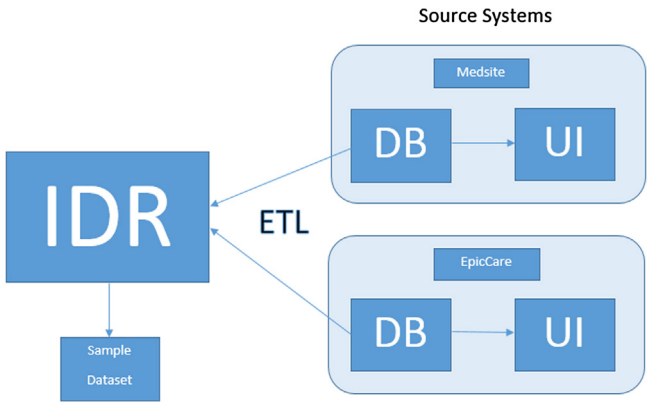

Fig. 1. Overview of ETL Process and Sample Dataset Extraction.
IDR data extracted, transformed, and loaded (ETL) from source systems (Medsite and EpicCare). Sample dataset generated from IDR for comparison with source systems. Comparison done by Health Information Management staff member who used the source system's user interfaces (UI) to validate data contained within their respective databases (DB).

Occasions for error occur in all three steps of the ETL process. For example, during the extraction phase, a field may be extracted incorrectly such as a secondary diagnosis being inadvertently selected as a primary diagnosis. In the transform phase, many opportunities for error exist, as the ETL software makes the source systems' data "fit" the needs of the IDR's display and reporting requirements. Observational data, such as laboratory results, have to be categorized so that they can be used within the ontologies or structured hierarchies of standardized terminologies. So, if the identifying terms for laboratory results are locally developed, they may need to be translated into a standardized terminology such as LOINC (Logical Observation Identifiers Names and Codes). Finally, in the load phase of the ETL process in which the extracted and transformed data is placed in the IDR's data structures, errors can occur that are, in effect, the mirror image of those that might

Table 1
Resolution of initially unmatched observations.

Observation Type	n	Resolution	IDR Correct?	Discrepancy Type
Lab Result	5	EHR used lab IDs while IDR merged these to patient's MRN	Yes	Design Decision
Race	1	Field was empty in EHR at time of ETL but updated before validation	Yes	Timing Issue
Medication	8	Route of admin deliberately not captured by IDR but displayed in EHR	Yes	Design Decision
Diagnosis	1	EHR listed diagnosis twice for same date; IDR considers this just one observation as it occurs on the same date	Yes	Design Decision
Lab Result	1	IDR observation order date (near midnight) confused with EHR collection date (of the following day)	Yes	Design Decision
Diagnosis	1	IDR observation found in EHR after user's account settings modified	Yes	Reviewer Setting
Medication	1	IDR observation found in EHR after user's account settings modified	Yes	Reviewer Setting
Lab Result	1	IDR observation order date (near midnight) confused with EHR collection date (of the following day)	Yes	Design Decision

EHR = electronic health record ID = identification IDR = Integrated Data Repository MRN = medical record number ELT = extract load transform.