Journal of Biomedical Informatics 70 (2017) 65–76

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

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

A pragmatic method for transforming clinical research data from the research electronic data capture "REDCap" to Clinical Data Interchange Standards Consortium (CDISC) Study Data Tabulation Model (SDTM): Development and evaluation of REDCap2SDTM

Keiichi Yamamoto [a,*,1], Keiko Ota [a,1], Ippei Akiya [b], Ayumi Shintani [a,c,1,2]

[a] REDCap Group, Department of Medical Innovation, Osaka University Hospital, Osaka 565-0871, Japan
[b] DataDriven, Inc., Tokyo 135-0064, Japan
[c] Department of Clinical Epidemiology and Biostatistics, Graduate School of Medicine, Osaka University, Osaka 565-0871, Japan

ARTICLE INFO

Article history:
Received 25 July 2016
Revised 29 April 2017
Accepted 4 May 2017
Available online 6 May 2017

Keywords:
Clinical research informatics
Clinical trial
Clinical research data warehouse
Data standards
New drug application

ABSTRACT

The Clinical Data Interchange Standards Consortium (CDISC) Study Data Tabulation Model (SDTM) can be used for new drug application studies as well as secondarily for creating a clinical research data warehouse to leverage clinical research study data across studies conducted within the same disease area. However, currently not all clinical research uses Clinical Data Acquisition Standards Harmonization (CDASH) beginning in the set-up phase of the study. Once already initiated, clinical studies that have not utilized CDASH are difficult to map in the SDTM format. In addition, most electronic data capture (EDC) systems are not equipped to export data in SDTM format; therefore, in many cases, statistical software is used to generate SDTM datasets from accumulated clinical data. In order to facilitate efficient secondary use of accumulated clinical research data using SDTM, it is necessary to develop a new tool to enable mapping of information for SDTM, even during or after the clinical research. REDCap is an EDC system developed by Vanderbilt University and is used globally by over 2100 institutions across 108 countries. In this study, we developed a simulated clinical trial to evaluate a tool called REDCap2SDTM that maps information in the Field Annotation of REDCap to SDTM and executes data conversion, including when data must be pivoted to accommodate the SDTM format, dynamically, by parsing the mapping information using R. We confirmed that generating SDTM data and the define.xml file from REDCap using REDCap2SDTM was possible. Conventionally, generation of SDTM data and the define.xml file from EDC systems requires the creation of individual programs for each clinical study. However, our proposed method can be used to generate this data and file dynamically without programming because it only involves entering the mapping information into the Field Annotation, and additional data into specific files. Our proposed method is adaptable not only to new drug application studies but also to all types of research, including observational and public health studies. Our method is also adaptable to clinical data collected with CDASH at the beginning of a study in non-standard format. We believe that this tool will reduce the workload of new drug application studies and will support data sharing and reuse of clinical research data in academia.

Abbreviations: CDISC, Clinical Data Interchange Standards Consortium; SDTM, Study Data Tabulation Model; CDASH, Clinical Data Acquisition Standards Harmonization; EDC, electronic data capture; ODM, Operational Data Model; ADaM, Analysis Data Model; FDA, Food and Drug Administration; PMDA, Pharmaceuticals and Medical Devices Agency; CSR, clinical study report; eCRF, electronic case report form; ETL, extract, transform, and load; SNOMED, Systematized Nomenclature of Medicine; LOINC, Logical Observation Identifiers Names and Codes; EAV, Entity Attribute Value; API, Application Programming Interface; SHARE, Shared Health and Research Electronic library.

* Corresponding author at: 1-2-7 Asahi-cho, Abeno-ku, Osaka 545-8585, Japan.
E-mail address: yamamoto.keiichi@med.osaka-cu.ac.jp (K. Yamamoto).
[1] Present address: Center for Drug and Food Clinical Evaluation, Osaka City University Hospital, Japan.
[2] Present address: Department of Medical Statistics, Osaka City University, Graduate School of Medicine, Japan.

http://dx.doi.org/10.1016/j.jbi.2017.05.003
1532-0464/

1. Introduction

To improve the efficiency of clinical research, it is critical to streamline its various phases, from protocol formulation through data collection, tabulation, analysis, and reporting. The Clinical Data Interchange Standards Consortium (CDISC) [1] is a non-profit, global organization that has developed several data standards to streamline clinical research [2]. Clinical Data Acquisition Standards Harmonization (CDASH) [3,4] describes the recommended data collection fields, including demographics and adverse events, common to most therapeutic areas and clinical research phases. The Operational Data Model (ODM) [5] is a vendor-neutral, platform-independent format for exchanging and archiving clinical study data and metadata that can be shared among different software systems. The Study Data Tabulation Model (SDTM) [6,7] is a standard for clinical study data tabulations. The Analysis Data Model (ADaM) [8] defines dataset and metadata standards that support the review of clinical trial statistical analyses derived from SDTM. Define-XML [9,10] transmits metadata for SDTM and ADaM datasets. Some of these standards, such as SDTM, ADaM, Define-XML, and CDISC Terminology [11–13], will be required by the U.S. Food and Drug Administration (FDA) [14]. Additionally, in September 2013 the Pharmaceuticals and Medical Devices Agency (PMDA) in Japan [15] expressed interest in adopting CDISC Standards electronically for new drug application studies and is accepting submissions from October 2016. Thus, CDISC Standards are becoming increasingly popular in the pharmaceutical industry. Additionally, these standards may be used not only for new drug application studies but may also enable data aggregation, data sharing, and secondary use for all types of research using CDISC Standards [2]. A clinical trial is usually conducted in three phases: (1) data collection using an electronic data capture (EDC) system from the clinical domain, (2) using the EDC database to create a dataset suitable for statistical analysis, and (3) conducting statistical analysis for the clinical study report (CSR). Individual clinical trials pose no difficulties in terms of statistical analysis for the CSR if all these steps are followed. In addition, in conventional clinical trials that use the aforementioned three-phase model, the variable name and code list value in the EDC system database may be defined by the investigator for each trial instead of utilizing the standardized data model and terminology. However, there are issues associated with studies that reuse past clinical trial data, as database table and variable names vary among clinical trials. To avoid this, it may be useful to modify the process as follows: (1) collect data using the EDC system, (2) create a clinical research data warehouse that harmonizes the differences between each clinical trial, (3) create a statistical dataset from the clinical research data warehouse, and (4) perform statistical analysis for the CSR. We show these clinical study processes and the relation to CDISC Standards in Fig. 1.

A clinical research data warehouse will simplify extraction of data on specific diseases by the use of specific parameters. The SDTM is a data standard that can be used to effectively create warehouses of this kind. The FDA has already developed such a warehouse, known as the "Janus Clinical Trials Repository" [16], and uses it to extract data on specific diseases or drugs to review new drug applications and advise pharmaceutical companies. This type of clinical research data warehouse could help leverage clinical research study data across studies performed within the same disease area and support data sharing and reuse in academic research [17–19].

Various conversion tools have been created using CDISC Standards for streamlining of clinical research [20–29]. However, most of them aim to exchange protocols, electronic case report forms (eCRFs), or clinical data in ODM format to facilitate EDC setup and data exchange between the EDC system and other external sources such as electronic medical records. To permit the secondary use of clinical data accumulated using the ODM format, it is necessary to set variable names using standardized code lists such as CDISC Terminology [11] and CDASH ODM.XML [3] before the research starts. Otherwise, mapping processes to SDTM after data collection is troublesome [21].

On the other hand, most EDC systems are unequipped to export data in the SDTM format; therefore, in many cases, statistical software is used to generate SDTM datasets from accumulated clinical data [30–32]. Theoretically, data from the EDC system can be easily converted to SDTM if eCRF fields are defined such that they correspond to the SDTM variables. For example, eCRF field names can be assigned according to the CDASH variable names. The CDASH is used in the earlier part of clinical trial data flow and defines a basic set of highly recommended and recommended/conditional data collection fields that are expected to be present on the majority of CRFs. The CDASH data collection fields (or variables) facilitate mapping to the SDTM [4]. When the data collection items correspond to the variables of SDTM, such as with adverse events, the eCRF field names can be set according to the variable names of CDASH (See Table 1). However, most current clinical trials are conducted using the aforementioned three-phase model, and not all clinical research uses CDASH from the setup phase of the study. Clinical research that is begun without correspondence to CDASH is difficult to map to SDTM. Utilizing CDASH or data collection instruments that have been aligned with SDTM improves the ability to map collected data to SDTM, but many sponsors have opted to align their existing data collection standards with SDTM rather than converting to CDASH. Moreover, it is often impossible to map eCRF field names directly onto SDTM variables. Many SDTM domains, particularly the Findings domains, present a vertical data (normalized) structure (one record for each item). However, many EDC systems might hold the data in a horizontal data (de-normalized) structure (one variable for each item) [4]. Therefore, conversion to a vertical data structure is imperative (e.g., various vital signs; Fig. 2), and extraction, transformation, and load (ETL) processing must be conducted outside the EDC system during SDTM data conversion.

In order to facilitate efficient secondary use of accumulated clinical research data using SDTM, it is necessary to develop a new tool to enable mapping information for SDTM, even during or after the clinical research.

REDCap (Research Electronic Data Capture) [33–35] is an EDC system developed by Vanderbilt University that can be used to design eCRFs and edit checking programs without having professional knowledge of the software. The REDCap Consortium, which is a vast support network of collaborators [35], includes over 2100 active institutional partners in 108 countries, and the REDCap application is currently being used globally by over 446,000 users for 317,000 projects in academic clinical research [35]. Concerning CDISC Standards, it became possible to import and export metadata and data in the ODM format in REDCap version 6.12 [36], and REDCap data collection instruments compliant with CDASH, such as "Adverse Events," "Common Identifiers," "Demographics," and "Protocol Deviations," can be shared globally through the REDCap Shared Library [37].

The "Field Annotation" function [38] of REDCap was introduced in version 6.5. This function provides a flexible-use field within the REDCap metadata specification that can be used to store standard data mapping for each variable. This permits mapping of various data standards such as CDISC, Systematized Nomenclature of Medicine (SNOMED) [39], and Logical Observation Identifiers Names and Codes (LOINC) [40] to REDCap data fields. In the case of REDCap, the backend database is relational and the data structure is the Entity Attribute Value (EAV) model [41–43], but the data output format is the horizontal (de-normalized) structure. We believe that it is possible to generate SDTM data efficiently from REDCap