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

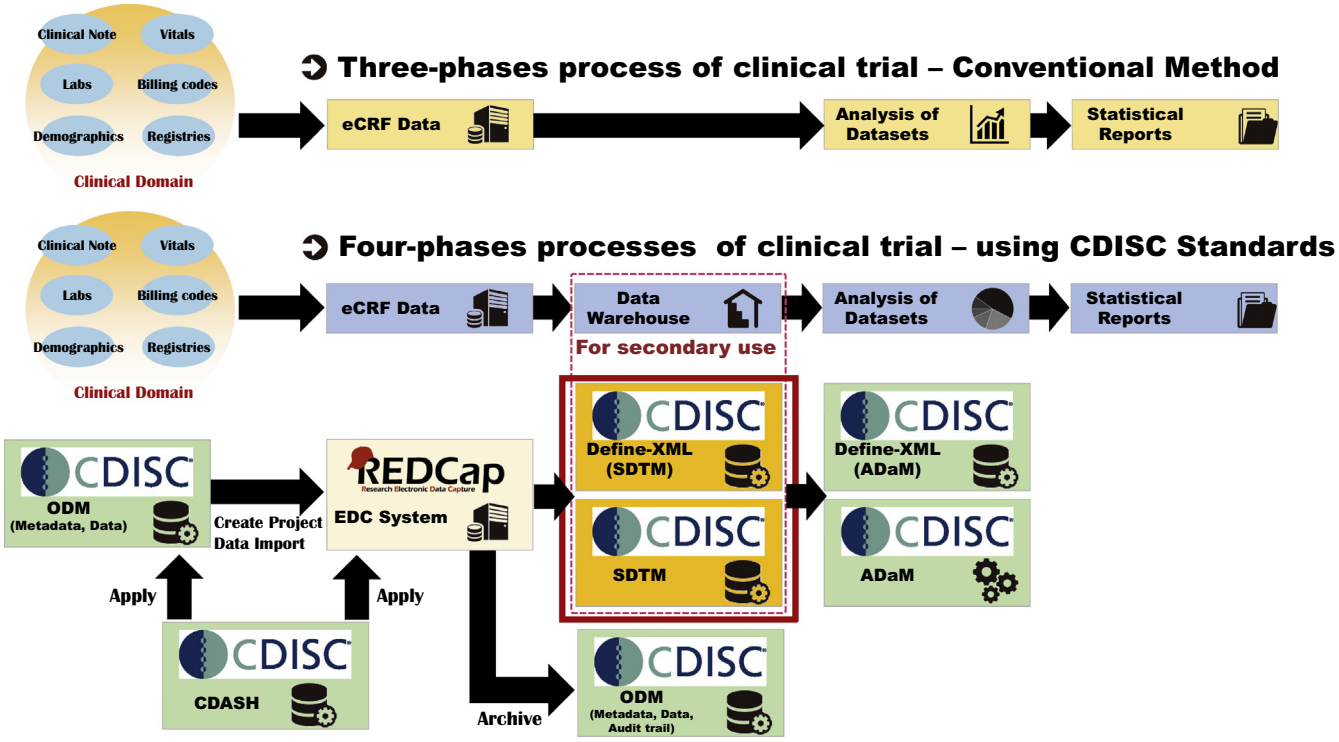

Fig. 1. Process of clinical trials and the relation to CDISC Standards. A clinical trial is usually conducted in three phases. However, there are issues associated with studies that reuse past clinical trial data, as the table and variable names of the databases vary with each clinical trial. To avoid this, it may be useful to modify the process to four phases.

Table 1
Example of SDTM mapping of adverse events.

Data collected	CDASH Variable name	SDTM Variable
Adverse event number	aespid	AESPID
Adverse event name	aeterm	AETERM
Start date	aestdat	AESTDTC
End date	aeendat	AEENDTC
Severity	aesev	AESEV
Seriousness	aeser	AESER

using outside tools, by mapping information stored in the Field Annotation, even when the data must be pivoted to accommodate the SDTM format.

2. Purpose

We developed a tool called "REDCap2SDTM" that can be used to analyze mapping information stored in the Field Annotations of REDCap in R [44], and generate SDTM data and the required xml file (named "Define-XML v2.0") from REDCap dynamically. This study aimed to evaluate whether this tool could be used to convert clinical research data stored in REDCap to SDTM, including conversion to a vertical data structure, and to generate the define.xml file (Define-XML v2.0) dynamically using metadata stored in the Field Annotation of REDCap using a simulated clinical trial without performance or scalability testing.

3. Materials and methods

3.1. Overview of REDCap2SDTM

Each eCRF data field in REDCap can include defined metadata using the Field Annotation function, and REDCap2SDTM executes conversion of data stored in REDCap to SDTM format dynamically, using this metadata in R. Additionally, this tool incorporates a function that generates the XML file required by SDTM, namely "define.xml (Define-XML v2.0)," which lists the metadata.

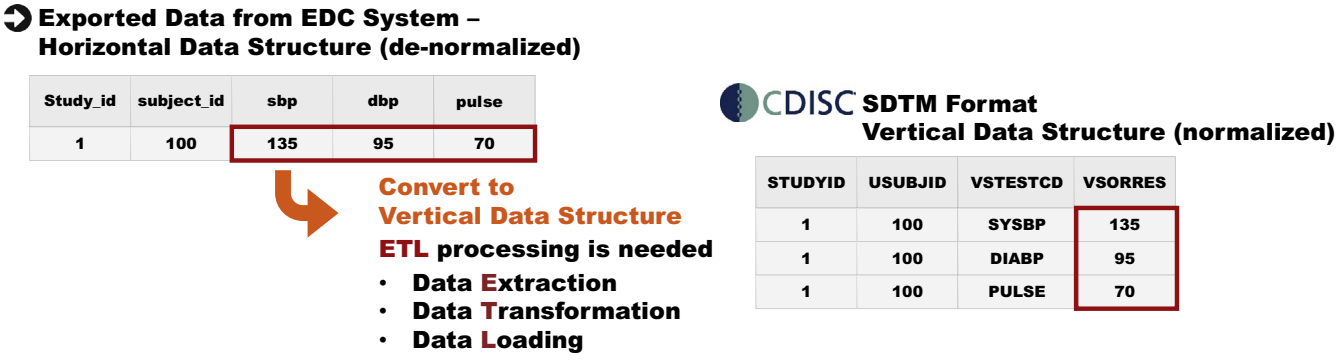

Fig. 2. Example of mapping vital signs onto SDTM (systolic pressure [sbp], diastolic blood pressure [dbp], pulse rate [pulse]). It is often impossible to map the field names of eCRFs directly onto SDTM variables, and conversion to a vertical data structure is imperative (e.g., various vital signs, laboratory values).

To develop and evaluate REDCap2SDTM, we used REDCap version 6.5.20 installed at Osaka University, CRAN R package version 3.2.3, and the R-dependent packages "stringr," "xlsx," "parsedate," and "dplyr." "Pinnacle 21 Community" [45,46] is a software package specifically designed to create FDA submission files. We used Pinnacle 21 Community edition version 2.1.0 to generate the define.xml file.

REDCap2SDTM has three functions, namely, "genMWB," "genDataset," and "genDefine.Spec." "MWB" stands for "Mapping Work Book." Since SDTM uses its own terminology when mapping from REDCap, our tool executes conversion using a prepared table. The "genMWB" function generates a mapping specification template, "Mapping-spec," as an Excel file; this template contains various mapping information including the aforementioned conversion table. The "genDataset" function generates SDTM data. The Pinnacle 21 Community software requires an Excel specification to generate the define.xml file [46], and the "genDefine.Spec" function generates this. In this manuscript, the Excel specification has been named "Define.xml-spec" to differentiate it from the Mapping-spec.

3.2. Procedure for generating SDTM data and define.xml using REDCap2SDTM

The procedure for generating SDTM data and the define.xml file by REDCap2SDTM is as follows:

Step 1. Enter the necessary mapping information in the Field Annotation of the eCRF of REDCap.
Step 2. Obtain data and metadata (i.e., "Data Dictionary" of REDCap [47]), including the Field Annotation from REDCap, and place them in the predefined directory. These files are provided by REDCap as one-click downloads and do not require additional preparation on the part of the user.
Step 3. Generate the Mapping-spec template (i.e., an Excel file) automatically by executing the genMWB function of REDCap2SDTM using the Data Dictionary, and update and complete the Mapping-spec manually
Step 4. Generate SDTM data by executing genDataset of REDCap2SDTM using the Mapping-spec and data obtained from REDCap
Step 5. Generate the Define.xml-spec template (i.e., an Excel file) by executing the genDefine.Spec function of REDCap2SDTM using the Mapping-spec, update and complete the Define.xml-spec manually, and generate the define.xml file by Pinnacle 21 Community software using the Define.xml-spec.

The conceptual diagram of this system is shown in Fig. 3.

3.2.1. How to enter mapping information into the Field Annotation of REDCap

In the first step of the REDCap2SDTM workflow, the user enters mapping information into the Field Annotation using the expression "OID Attribute" of "ItemDef element" of Define-XML v2.0 (e.g., "IT.VS.VSORRES.WEIGHT") [10]. In this way, SDTM data generation is executed using the same mapping information when eCRF fields and SDTM variables correspond one-on-one, as for instance with adverse events, as well as when conversion to a vertical data structure is required.

The basic format described in the Field Annotation is "SDTM:IT.<<domain name>>.<<variable name>>.<<test code>>;." Three examples are shown below:

Example 1 (*SDTM:IT.VS.VSORRES.SYSBP;*). Many domains of SDTM use a vertical (normalized) structure. When the table structure of REDCap uses a horizontal (de-normalized) structure, conversion to a vertical data structure is necessary to solve this mismatch and the above description is needed. Using the example of systolic blood pressure, the value is stored in a variable called VSORRES in the VS domain of SDTM. VSORRES stands for "Vital Sign Result or Finding in Original Units;" VS stands for "Vital Signs," and VSTESTCD stands for "Vital Signs Test Short Name" [7]. The "SYSBP," which is the test code of "Systolic Blood Pressure" defined in SDTM Terminology [9], is set as the VSTESTCD variable in the VS domain. (See Fig. 4).

Example 2 (*SDTM:IT.AE.AETERM;*). Here "AE" stands for "Adverse Events [7]." When creating an eCRF for adverse events with REDCap, the table structure results in one event in one record structure. When the table structure of REDCap is in this vertical (normalized) structure, conversion to a vertical data structure is unnecessary and thus the description uses the above format. If it is possible to map an eCRF data field onto a SDTM variable directly, both the domain name and variable name are set.

Example 3 (*SDTM:IT.LB.LBLOINC.<<LOINC code>>;*). LB stands for "Laboratory Test Results [7]." By setting the LOINC code with the above format, its value will be stored in the LBLOINC variable of the LB domain.

3.2.2. Download study data and metadata from REDCap

The Data Dictionary of REDCap is a specifically formatted spreadsheet of comma-separated values (CSV) containing the metadata used to construct data collection instruments and fields [47]. REDCap allows us to access the data and Data Dictionary (i.e. metadata) using two methods. One is to download them manually with one click as a file from the REDCap web page. The other is to use the web Application Programming Interface (API) of REDCap [48]. To use REDCap2SDTM, both the data and Data Dictionary must be placed in CSV format in a predefined directory prior to conversion.

3.2.3. How to create the Mapping-spec

After entering the mapping information into the Field Annotation, Mapping-spec templates are generated by executing the genMWB function of REDCap2SDTM. Thereafter, additional necessary information is entered manually to complete the template.

The generated SDTM data should be entered using only SDTM Terminology; it is necessary to map from the categorical data with the predefined code list on REDCap. To achieve this, the terms in this simulated trial data and the terms in the SDTM Terminology conversion table are prepared in the Mapping-spec. Using controlled terms defined in the Mapping-spec, SDTM data is generated automatically in accordance with SDTM Terminology. There is also an additional section that allows input of REDCap code lists in the define.xml file for use while the file is being generated.

The Mapping-spec is an Excel file with nine sheets. The sheet names and columns are presented in Table 2. The General sheet contains "STUDYID" of SDTM, and the DatasetMetadata sheet contains metadata relating to the SDTM datasets; thus, it is necessary to generate the define.xml file. In the VariableMetadata sheet, "CODELIST.OID" defines the code list of CDISC, which is necessary to choose manually from the Codelist sheet. The fields "SDTM.DOMAIN," "SDTM.VARIABLE," and "SDTM.TESTCD" contain mapping information generated automatically by genMWB of REDCap2SDTM from the Field Annotation. The "TA," "TE," "TV," "TI," and "TS" sheets contain information necessary for generating the "Trial Design Datasets" of SDTM. TA stands for "Trial Arms," TE for "Trial Elements," "TV" for "Trial Visits," "TI" for "Trial Inclusion/Exclusion Criteria," and "TS" for Trial Summary" [7].

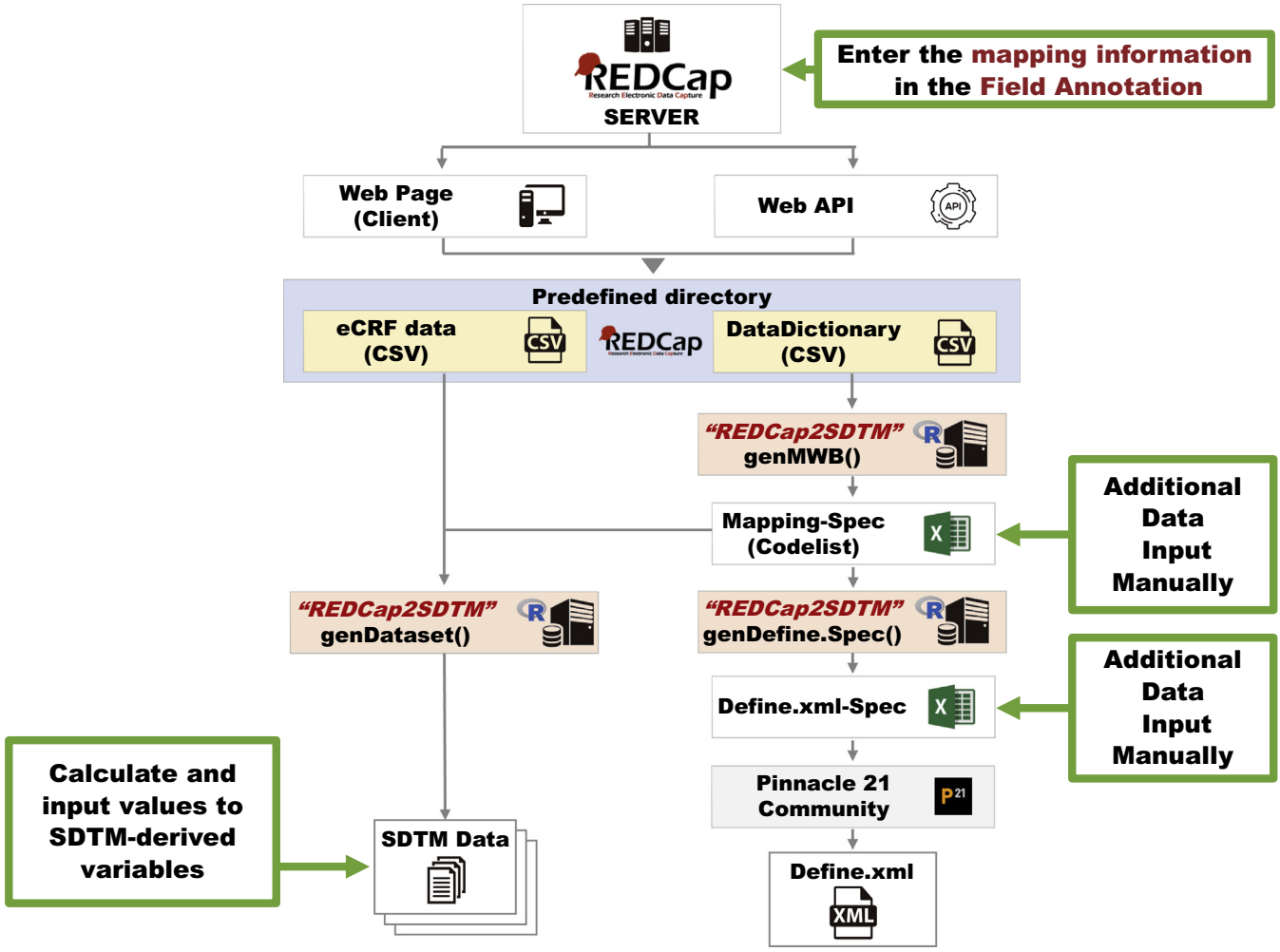

Fig. 3. Conceptual diagram of conversion from REDCap to SDTM. Enter the necessary mapping information in the Field Annotation, and obtain data and metadata (i.e., "Data Dictionary" of REDCap). REDCap2SDTM has three functions: "genMWB," "genDataset," and "genDefine.Spec." The "genMWB" function generates a mapping specification template, "Mapping-spec," which contains various mapping data as an Excel file. The "genDataset" function generates SDTM data. We used Pinnacle 21 Community software to generate the define.xml file. The Pinnacle 21 Community software requires an Excel specification file (i.e., Define.xml-spec) to generate the define.xml file. The "genDefine.Spec" function generates the "Define.xml-spec."

Based on the "SDTM Implementation Guide" [7], the following are input manually.

(1) "Description," "Class of Domain," "Dataset Structure," and "Key Variables" of the Dataset Metadata sheet. These Dataset Metadata should be input manually based on the "SDTM Implementation Guide" [7]. For example, for vital signs, Description is input "Vital Signs"; Class of Domain is input "FINDINGS"; Dataset Structure is input "One record per vital sign measurement per visit per subject"; and Key Variables is input "STUDYID, USUBJID, VSTESTCD, VSDTC, VISITNUM, and VSPOS."
(2) CDISC Controlled Terms corresponding to REDCap Code list of the VariableMetadata and Codelist sheets (see Fig. 5)
(3) Information concerning "Trial Design Datasets," such as "Trial Arm," "Trial Element," "Trial Visit," "Trial Inclusion/Exclusion Criteria," and "Trial Summary" should be input into the TA, TE, TV, TI, and TS sheets, respectively. Since REDCap data does not include the data or the information regarding Trial Design Datasets, it should be input manually.

3.2.4. Generation of SDTM data

SDTM data is generated as an R-data frame in the R-runtime environment by executing the genDataset function of REDCap2SDTM, with calculation and input of values to SDTM-derived variables. For example, "Sequence Number Variable (i.e., --SEQ)" is input, and conversion of the original unit of the laboratory tests to the standard unit is executed manually. For exporting the SDTM files to the outside from the R-runtime environment, built-in functions of R, such as "write.csv" function [49], are useful.

3.2.5. Generation of the define.xml file

The Pinnacle 21 Community software requires an Excel specification file (i.e., Define.xml-spec) to generate the define.xml file. The Define.xml-spec is generated by executing the genDefine.Spec function of REDCap2SDTM. After adding the necessary information and completing the Define.xml-spec, the Pinnacle 21 Community software uses it to generate the define.xml file. The information that is necessary for manual input is each variable and metadata element concerning the "Value Level Metadata," such as "Origin," "Significant Digit," "Format," "Method," and "Comment." These metadata are defined based on "SDTM Implementation Guide" [7] and "Define-XML Specification Version 2.0" [10].

3.3. Protocol of simulated clinical trial for evaluation of the tool

A simulated single-arm, unblinded, single-center clinical trial with aortic aneurysm as the disease of interest was used to evalu-

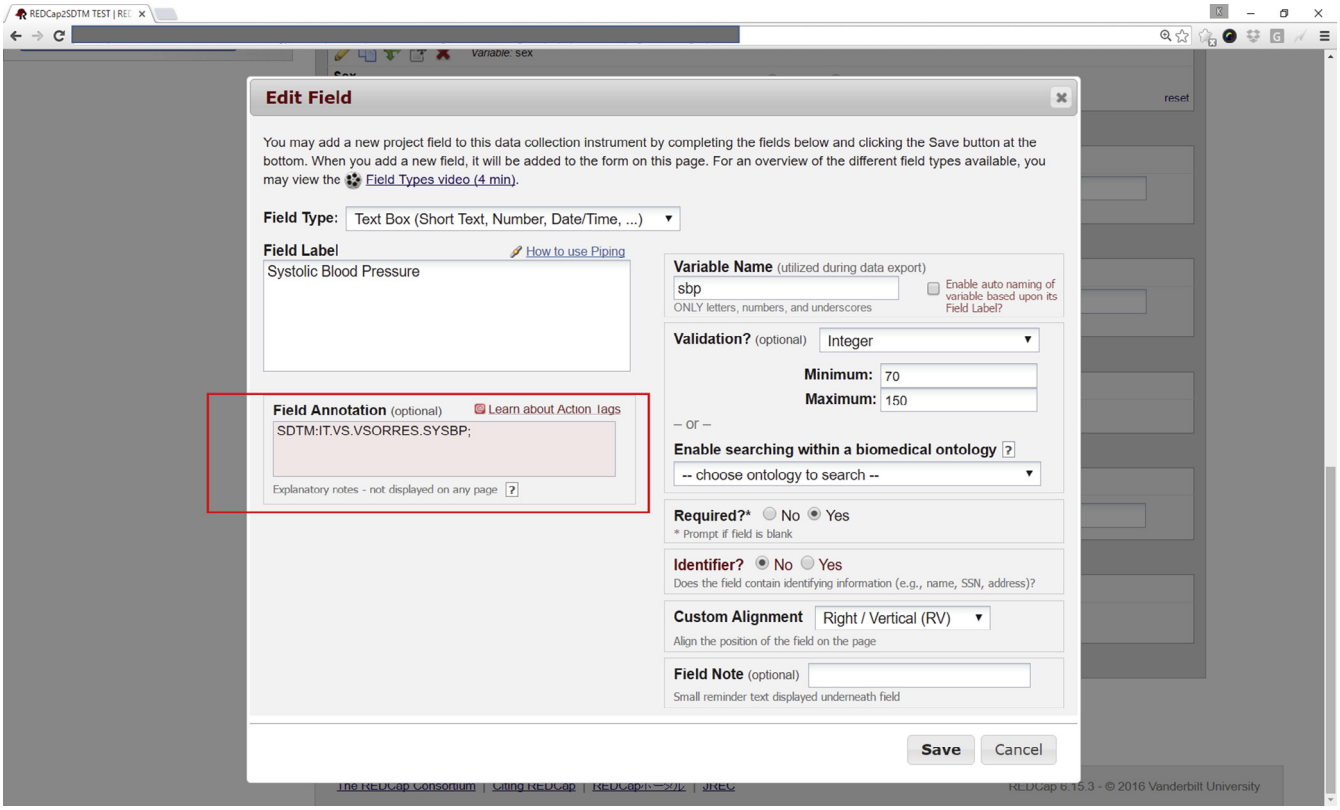

Fig. 4. Image of SDTM mapping information entered into the REDCap Field Annotation box. Mapping information is entered into the Field Annotation using the expression "OID Attribute" of "ItemDef element" of Define-XML v2.0.

ate this tool. The treatment of choice was surgery (laparotomy or thoracotomy, hybrid, and intravascular treatment), and patients were eligible if they were 20 years or older and had an aortic aneurysm. The eCRFs were filled out preoperatively (Visit 1), during surgery (Visit 2), and postoperatively (Visit 3).

Five fictitious patient profiles were created to test the tool. The data collected at each visit are presented in Table 3.

In this simulated clinical trial, vital signs, laboratory values, and questionnaires are output as the horizontal (de-ormalized) structure from REDCap and conversion to vertical data structure is necessary.

4. Results

The R package of REDCap2SDTM, generated SDTM data, the define.xml file, the Data Dictionary, test data, the Mapping-spec file, and the Define.xml-spec file of the simulated clinical trial are attached to this manuscript as appendices.

4.1. Data conversion from REDCap to CDISC SDTM

We inspected whether the mapping information in the Field Annotation of REDCap could be used to dynamically convert the simulated clinical trial data from the REDCap database to SDTM. The Field Annotation function in REDCap serves two purposes: (1) to apply one or more "Action Tags" [50] that control the web page and data using specific variables; and (2) to serve as design note for the project designer, or to denote a specific CDISC mapping. We also confirmed that these two purposes can be used concurrently.

The test was run on a personal computer that was usually used for daily work (OS: Windows 10 Pro 64 bit; CPU: Intel Core i5-5287U 2.90 GHz; RAM: 4.00 GB). By executing REDCap2SDTM, we were able to generate SDTM data files (i.e., CSV files: dm.csv, ho.csv, lb.csv, mh.csv, pr.csv, qs.csv, su.csv, and vs.csv) from REDCap data and metadata of the simulated clinical trial.

The main focus of our evaluation was whether the mapping information set in the Field Annotation using the expression "OID Attribute" of "ItemDef element" of Define-XMLv2.0 (e.g., "IT.VS.VSORRES.WEIGHT") could be used dynamically to perform conversion using the R program. This idea is the key concept of REDCap2SDTM.

Since the Field Annotation has another function, namely "Action Tag", with strings beginning with "@," we decided to define the mapping information strings such that they began with "SDTM:" and terminated with ";." Mapping information itself is thus "IT.<<domain name>>.<<variable name>>.<<test code>>;." For example, "IT.VS.VSORRES.WEIGHT" is interpreted as a value of weight set to VSORRES in the VS domain and WEIGHT is set to VSTESTCD.

As mentioned previously, SDTM requires the use of standardized terms, known as "SDTM Terminology" [11]. For example, according to this terminology, males must be coded as "M" and females as "F." However, since REDCap employs user-defined code lists for data entry, males are coded as "0" and females as "1." In order to bridge this gap, we created a code conversion table (i.e., an Excel sheet) in the Mapping-spec, which permits easy execution of data conversion from REDCap to SDTM using SDTM Terminology.

The conversion results of SDTM, including conversion to a vertical data structure, are presented in Appendix A2, and the Mapping-spec, including the code conversion table, are presented in Appendix A6.

4.2. Generation of the define.xml file

We also investigated whether the define.xml file could be generated dynamically. By executing REDCap2SDTM, we were able to

Table 2
Sheet and column list of mapping-spec.

Sheet name	Column	Autofilled using REDCap metadata	Manual input
General	Parameter		X
	Value		X
DatasetMetadata	Dataset	X	
	Description		X
	Class		X
	Structure		X
	Key.Variables		X
VariableMetadata	field_name	X	
	form_name	X	
	field_type	X	
	field_label	X	
	select_choices_or_calculations	X	
	text_validation_type_or_show_slider_number	X	
	required_field	X	
	field_annotation	X	
	CODELIST.OID		X
	SDTM.DOMAIN	X	
	SDTM.VARIABLE	X	
	SDTM.TESTCD	X	
Codelist	REDCAP.VAR.NAME	X	
	REDCAP.CLI.CODE	X	
	REDCAP.CLI.TERM	X	
	CODELIST.OID		X
	NCI.CODE		X
	NCI.CODELIST.CODE		X
	CODELIST.NAME		X
	SDTM.TERM		X
	DECODED.VALUE		X
TA	STUDYID		X
	DOMAIN		X
	ARMCD		X
	ARM		X
	TAETORD		X
	ETCD		X
	ELEMENT		X
	TABRANCH		X
	TATRANS		X
	EPOCH		X
TE	STUDYID		X
	DOMAIN		X
	ETCD		X
	ELEMENT		X
	TESTRL		X
	TEENRL		X
	TEDUR		X
TV	STUDYID		X
	DOMAIN		X
	VISITNUM		X
	VISIT		X
	VISITDY		X
	ARMCD		X
	ARM		X
	TVSTRL		X
	TVENRL		X
TI	STUDYID		X
	DOMAIN		X
	IETESTCD		X
	IETEST		X
	IECAT		X
	IESCAT		X
	TIRL		X
	TIVERS		X
TS	STUDYID		X
	DOMAIN		X
	TSSEQ		X
	TSGRPID		X
	TSPARMCD		X
	TSPARM		X
	TSVAL		X
	TSVALNF		X
	TSVALCD		X

(*continued on next page*)

Table 2 (*continued*)

Sheet name	Column	Autofilled using REDCap metadata	Manual input
	TSVCDREF		X
	TSVCDVER		X

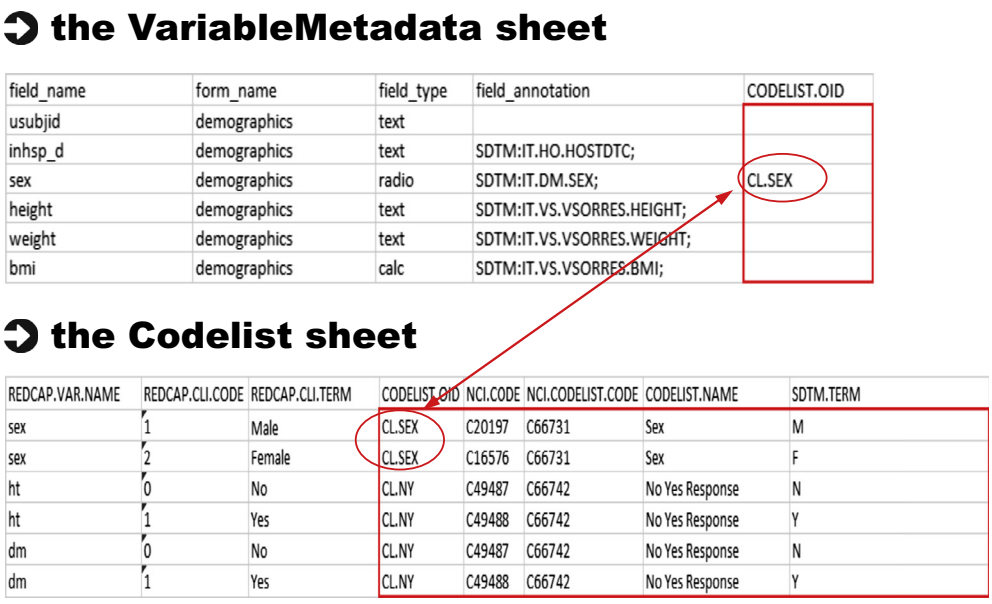

Fig. 5. CDISC Controlled Terms corresponding to the REDCap Code list. In the VariableMetadata sheet, "CODELIST.OID" defines the code list of CDISC; this code list must be chosen manually from the Codelist sheet.

generate the define.xml file (i.e., an xml file) from the REDCap data and metadata of the simulated clinical trial.

In order to generate the define.xml file using Pinnacle 21 Community software, it was necessary to prepare a Define.xml-spec file (i.e., an Excel file) that contained information to be included in the define.xml file. This template was created dynamically using the Mapping-spec; after entering additional information and completing the template, it was used to generate the define.xml file using the Pinnacle 21 Community software.

Fig. 6 shows the define.xml file of vital signs and the Value Level Metadata that executed conversion to a vertical data structure. Additionally, Appendix A3 shows the define.xml file, while Appendix A7 shows the Define.xml-spec.

5. Discussion

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 these data and this file dynamically, without programming, as it only involves entering the mapping information into the Field Annotation and the additional data into the generated the Mapping-spec and the Define.xml-spec. Therefore, the workload for generation of the SDTM data and the define.xml file from the EDC system is substantially reduced.

REDCap processes longitudinal information differently from a basic project with a single-point measurement. REDCap data are acquired as a CSV file, and the single-point measurement data (e.g., height, weight) and multi-point measurement data (e.g., laboratory values) are mixed; handling of these mixed data simultaneously is complicated. We designed REDCap2SDTM as the internal processor; data were divided into several data tables (i.e., R-data frames) based on the "Form Name" of the REDCap Data Dictionary, and mapping to SDTM was executed (see Fig. 7).

In addition, there are variables called "redcap_event_name" that serve as useful key variables for multiple time-point items in the REDCap data. By combining these data tables (i.e., R-data frames) and variables, REDCap2SDTM can deal with multiple-arm, multiple time-point (i.e., longitudinal) clinical studies.

REDCap2SDTM was developed using REDCap version 6.5.20. Since the "Action Tag" of Field Annotation begins with "@," REDCap2SDTM will be available in future versions of REDCap unless the Field Annotation function disappears. REDCap2SDTM can simultaneously extract mapping information from the Field Annotation even though it may include the Action Tag and other metadata related to the REDCap data field. As a result, many different kinds of information might be populated to the Field Annotation. Since this situation is not ideal, we believe it is desirable to prepare another metadata field for SDTM mapping information in future versions.

In addition, the current version of REDCap2SDTM, after generating the SDTM data frame, calculates and inputs values to SDTM-derived variables manually. We think that this work can be automated by integrating a calculation algorithm into REDCap2SDTM.

For utilizing the generated define.xml file, it may be possible to automate database schema design for the secondary use of SDTM. For example, by using metadata such as "Dataset Level Metadata" and "Variable Level Metadata" written in the define.xml file, it may be possible to generate an automatic definition of tables for both clinical research data warehouses and separate databases. Additionally, by using the metadata included in the "Value Level Metadata," it also may be possible to automatically convert SDTM data using a vertical structure (normalized) into a table using a horizontal (de-normalized) structure.

The CDISC Shared Health and Research Electronic (SHARE) [51] is an electronic library for sharing CDISC Standard metadata. It is

Table 3
Data collected during clinical trial simulation.

Data Collection Instrument	Items	Visit		
		Pre-Operative (Visit 1)	During surgery (Visit 2)	Post-Operative (Visit 3)
Demographics	Date of Admission Sex Height Weight BMI (Automated Calculation)	X		
Past Medical History	Hypertension (HT) Diabetes (DM) Hyperlipidemia (HL) Smoking within the past 6 months	X		
Aortic Aneurysm Type	Classification of aneurysms	X		
Lab Test Results	With or without lab test Date of blood collection Total cholesterol (T-cho) Fasting blood sugar (FBS) Serum creatinine (CRE) Urea nitrogen (BUN) AST (GOT) ALT (GPT) Hemoglobin (Hb) White blood cell (WBC)	X		X
Questionnaire	Anxiety Tightening of the chest Depression Difficulty in concentrating Restlessness, tension Hyperkinesis Irritability Exhaustion Fatigue	X		X
In Op	Start date & time of surgery End date & time of surgery Type of Surgery Operative procedure Aortic cross clamping Blood transfusion volume Amount of bleeding Volume of infusion Urine volume		X	
Post OP	Date & time of ICU admission Date & time of arousal Date & time of lower extremity working confirmation Date & time of ICU discharge Spinal disorder			X

possible to obtain the REDCap Data Dictionary using the REDCap API, and to obtain the SDTM metadata and SDTM Terminology using the CDISC SHARE API [51]. By combining these two APIs, it may be possible to reduce the workload required for manual input into the Mapping-spec. For example, by searching for synonyms of SDTM Terminology with the CDISC SHARE API, the terms defined in the REDCap code list may be able to automatically identify appropriate SDTM Standardized terms. In the future, it will be necessary to evaluate the accuracy of this identification process.

We believe that our proposed method is useful as a support tool for clinical research, but it does not have enough features to automatically generate a strict SDTM; future changes are required to realize this goal. Additionally, our simulation trial revealed several issues associated with this tool. In academia, when generating SDTM from EDC systems for secondary use as clinical research data, the specifications of SDTM need not be strictly adhered to. However, in the case of new drug application studies, several issues, discussed below, must be taken into consideration.

The structure of the eCRFs used in this simulated clinical trial was designed without taking into consideration the conversion to SDTM, and as a result some of the information necessary to generate the SDTM data was unavailable. Therefore, to generate SDTM data automatically and completely, it is necessary to use an eCRF that is designed with the possibility of SDTM conversion in mind. In REDCap, data is created based on each page of the eCRF. For example, if the date and time of a laboratory test and visit information are not present on each eCRF page, automatic conversion is impossible and individual conversion programs are needed for each page. To eliminate this workload, it should be ensured that this information is present on each page of the eCRF. Similarly, units of laboratory tests should be assigned to specific fields in the eCRF as this information is required in the SDTM dataset.

Furthermore, as the method of data conversion evaluated here assumed conversion from one entry field to one variable of one domain in the SDTM, we were unable to convert from one entry field to variables of multiple domains. Further expansion of this method is necessary to overcome this problem. The parser of REDCap2SDTM can be expanded to enable interpretation of multiple SDTM variables into the Field Annotation, such as "SDTM:IT.<<domain name>>.<<variable name>>.<<test code>>, IT.<<domain name>>.<<variable name>>.<<test code>>;." Thus "SDTM:IT.VS.VSDTC, IT.EG.EGDTC;" would indicate that a REDCap variable maps

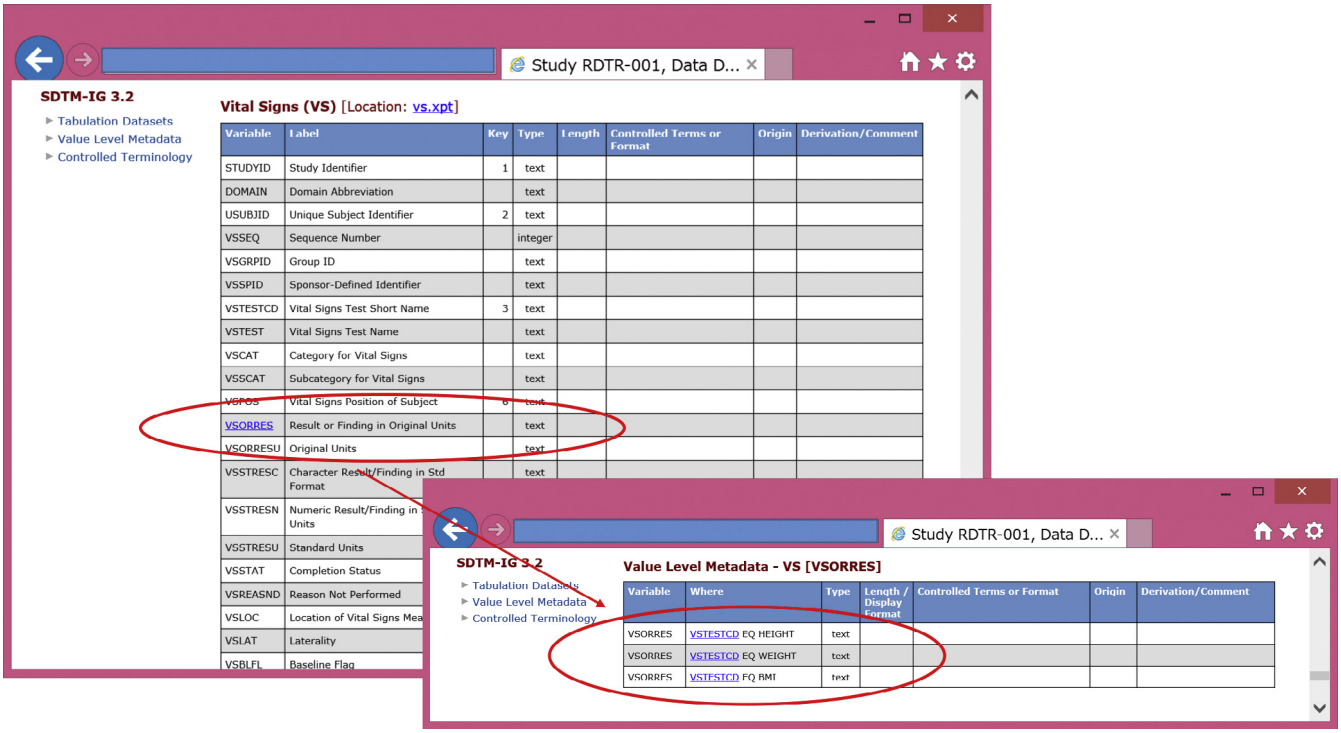

Fig. 6. Image of generated the define.xml file (Vital Signs [VS] and the Value Level Metadata that executed conversion to vertical data structure).

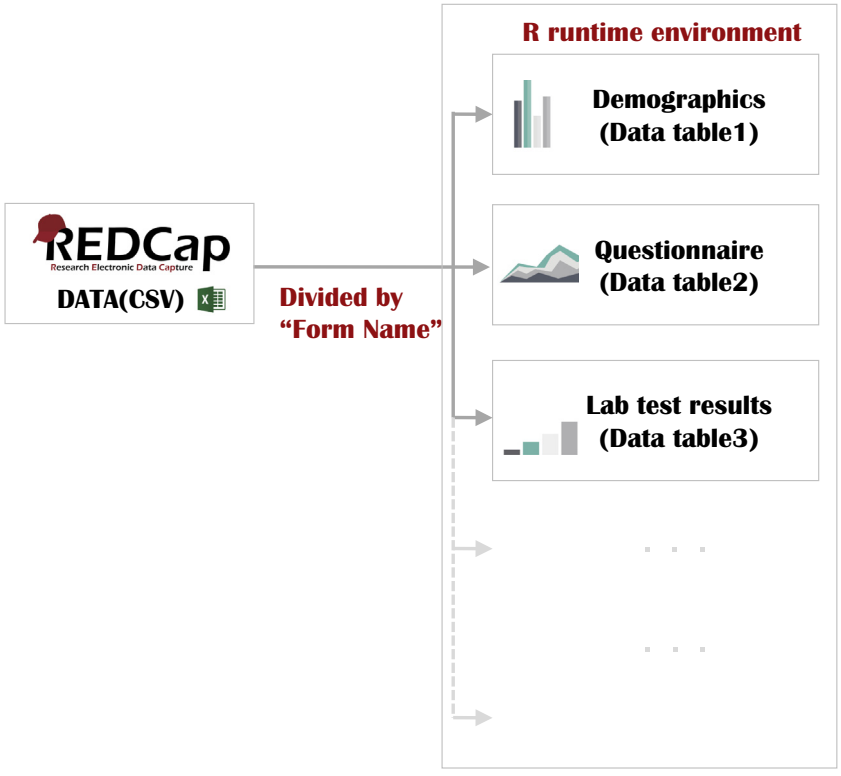

Fig. 7. Data processing image of REDCap2SDTM. With REDCap2SDTM, research data are divided into several data tables (i.e., R-data frames) based on the "Form Name" of the REDCap Data Dictionary. In addition, there are variables called "redcap_event_name" in the REDCap data (i.e., CSV files). REDCap2SDTM is able to deal with multiple-arm, multiple time-point (i.e., longitudinal) clinical studies.

to both the VSDTC Systolic Blood Pressure and EGDTC Electrocardiogram Mean Heart Rate variables. This expansion makes it possible to deal with multi-conversion and to reduce workload.

Generation of the define.xml file requires manual input of large amounts of information into the Mapping-spec, as all the necessary information is not included in the metadata present in the Field

Annotation. However, this is a problem unique to the generation of the define.xml file, and should not be managed by EDC systems such as REDCap. Examples of data needing manual input are information on where the data originated [7] and a calculation algorithm of "Derived Value" [10]. Similarly, the information in the TA to TS sheets of the Mapping-spec is relevant for the "Trial Design" of SDTM. Trial design datasets (TA, TE, TV, TI, TS) help reviewers of regulatory agencies clearly and quickly grasp the design of a clinical trial, compare the designs of different trials, search a data warehouse for clinical trials with certain features, and compare plans and actual treatments and visits for subjects in a clinical trial [52,53]. Thus, the FDA and PMDA require submission of Trial Design datasets. Since this information is not directly related to data quality, data exchange, or secondary use of clinical research data, it is not regarded as important in academic research. This suggests the existence of a gap between academic clinical research and new drug application reviews performed by regulatory agencies. In the case of conventional three-phase trials, failure to use standardized terminology does not present a problem. However, when clinical research is initiated without using standardized terminology, it is difficult to map the results to SDTM format. If the CDASH data collection instruments of the REDCap Share Library are this problem may be solved, but thus far there are only four instruments. Additionally, there are few affordable tools for converting data collected in a non-standard format into SDTM [30–32]. This may be a reason why researchers are not using standardized terminology at the beginning of research that is not related to regulated or unregulated studies. The work to fill this gap, enhancement of the CDASH data collection instruments, and more widespread use of our method may be a springboard for further extension and leveraging of clinical research process redesign using CDISC Standards and for creating a means for true beginning-to-end automation of clinical research.

6. Conclusion

In conclusion, the evaluation of the REDCap2SDTM tool we created using a simulated clinical trial shows, that this tool can be used to generate SDTM data and the define.xml file from REDCap. Our proposed method is adaptable not only to new drug application studies but to all types of research, such as observational and public health studies. Our method is also adaptable to existing clinical studies in which data exist in ODM or in which CDASH was applied at the start of the research, and to data in a non-standard format. We believe that this tool will reduce the workload of new drug application studies and enable leveraging of clinical research study data across studies within the same disease area, and will thus support data sharing and reuse of clinical research data.

Conflict of interests

The authors declare that they have no conflicts of interest. This work was supported by the Project of Translational and Clinical Research Core Centers of the Japan Agency for Medical Research and Development (AMED), and JSPS KAKENHI Grant No. JP26560270. These funding sources did not influence the conduct of this work.

Acknowledgments

The authors would like to acknowledge Toshiko Seki and Tatuyuki Obana of the REDCap Group of Osaka City University Hospital, and Takuma Ishihara and Keiko Ishibashi of the Department of Medical Statistics of Osaka City University Graduate School of Medicine for their generous support.

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.jbi.2017.05.003.

References

[1] CDISC Strength through Collaboration, Clinical Data Interchange Standards Consortium, 2016. <http://www.cdisc.org/> (accessed 16.12.29).
[2] CDISC Foundational Standards, Clinical Data Interchange Standards Consortium, 2016. <https://www.cdisc.org/standards/foundational> (accessed 16.12.29).
[3] Clinical Data Acquisition Standards Harmonization (CDASH), Clinical Data Interchange Standards Consortium, 2016. <https://www.cdisc.org/standards/foundational/cdash> (accessed 16.12.29).
[4] CDASH v1.1, Clinical Data Interchange Standards Consortium, 2016. <https://www.cdisc.org/sites/default/files/members/standard/foundational/cdash/cdash_std_1_1_2011_01_18.pdf> (accessed 16.12.29).
[5] Operational Data Model (ODM)-XML, Clinical Data Interchange Standards Consortium, 2016. <https://www.cdisc.org/standards/foundational/odm> (accessed 16.12.29).
[6] CDISC Study Data Tabulation Model (SDTM), Clinical Data Interchange Standards Consortium, 2016. <http://www.cdisc.org/sdtm> (accessed 16.12.29).
[7] Study Data Tabulation Model Implementation Guide: Human Clinical Trials Version 3.2, Clinical Data Interchange Standards Consortium, 2016. <https://www.cdisc.org/sites/default/files/members/standard/foundational/sdtmig/sdtmig_20v3.2_20noportfolio.pdf?> (accessed 16.12.29).
[8] CDISC Analysis Data Model (ADaM), Clinical Data Interchange Standards Consortium, 2016. <http://www.cdisc.org/adam> (accessed 16.12.29).
[9] CDISC Define-XML, Clinical Data Interchange Standards Consortium, 2016. <http://www.cdisc.org/define-xml> (accessed 16.12.29).
[10] CDISC Define-XML Specification Version 2.0, Clinical Data Interchange Standards Consortium, 2016. <http://wiki.cdisc.org/download/attachments/3769333/Define-XML-2-0-Specification_comments_Greg_Steffens.pdf?api=v2> (accessed 16.12.29).
[11] CDISC Terminology, National Cancer Institute, 2016. <https://www.cancer.gov/research/resources/terminology/cdisc> (accessed 16.12.29).
[12] CDISC Controlled Terminology, Clinical Data Interchange Standards Consortium, 2016. <http://www.cdisc.org/terminology> (accessed 16.12.29).
[13] CDISC SDTM Controlled Terminology, National Cancer Institute, 2016. <https://evs.nci.nih.gov/ftp1/CDISC/SDTM/SDTM%20Terminology.pdf> (accessed 16.12.29).
[14] U.S. Food and Drug Administration Home Page, U.S. Food and Drug Administration, 2016. <http://www.fda.gov/> (accessed 16.12.29).
[15] Pharmaceuticals and Medical Devices Agency, Pharmaceuticals and Medical Devices Agency (Japan), 2016. <https://www.pmda.go.jp/english/index.html> (accessed 16.12.29).
[16] Janus Clinical Trials Repository, U.S. Food and Drug Administration, 2016. <http://www.fda.gov/ForIndustry/DataStandards/StudyDataStandards/ucm155327.htm> (accessed 16.12.29).
[17] I. Danciu, J.D. Cowan, M. Basford, X. Wang, A. Saip, S. Osgood, J. Shirey-Rice, J. Kirby, P.A. Harris, Secondary use of clinical data: the Vanderbilt approach, J. Biomed. Inform. 52 (2014) 28–35.
[18] J.G. Klann, A. Abend, V.A. Raghavan, K.D. Mandl, S.N. Murphy, Data interchange using i2b2, J. Am. Med. Inform. Assoc. 23 (5) (2016) 909–915.
[19] K. Yamamoto, E. Sumi, T. Yamazaki, K. Asai, M. Yamori, S. Teramukai, K. Bessho, M. Yokode, M. Fukushima, A pragmatic method for electronic medical record-based observational studies: developing an electronic medical records retrieval system for clinical research, BMJ Open 2 (6) (2012) e001622, http://dx.doi.org/10.1136/bmjopen-2012-001622.
[20] P. Bruland, B. Breil, F. Fritz, M. Dugas, Interoperability in clinical research: from metadata registries to semantically annotated CDISC ODM, Stud. Health. Technol. Inform. 180 (2012) 564–568.
[21] F.A. Meineke, S. Staubert, M. Lobe, A. Winter, A comprehensive clinical research database based on CDISC ODM and i2b2, Stud. Health. Technol. Inform. 205 (2014) 1115–1119.
[22] M. Dugas, S. Dugas-Breit, Integrated data management for clinical studies: automatic transformation of data models with semantic annotations for principal investigators, data managers and statisticians, PLoS ONE 28–29 (2) (2014) e90492.
[23] G. Jiang, J. Evans, T.A. Oniki, J.F. Coyle, L. Bain, S.M. Huff, R.D. Kush, C.G. Chute, Harmonization of detailed clinical models with clinical study data standards, Methods Inf. Med. 54 (1) (2015) 65–74.
[24] M. Dugas, ODM2CDA and CDA2ODM: tools to convert documentation forms between EDC and EHR systems, BMC Med. Inform. Decis. Mak. 26 (15) (2015) 40.
[25] V. Huser, C. Sastry, M. Breymaier, A. Idriss, J.J. Cimino, Standardizing data exchange for clinical research protocols and case report forms: an assessment

of the suitability of the Clinical Data Interchange Standards Consortium (CDISC) Operational Data Model (ODM), J. Biomed. Inform. 57 (2015) 88–99.
[26] G. De Moor, M. Sundgren, D. Kalra, A. Schmidt, M. Dugas, B. Claerhout, T. Karakoyun, C. Ohmann, P.Y. Lastic, N. Ammour, R. Kush, D. Dupont, M. Cuggia, C. Daniel, G. Thienpont, P. Coorevits, Using electronic health records for clinical research: the case of the EHR4CR project, J. Biomed. Inform. 53 (2015) 162–173.
[27] J.C. Camacho Rodriguez, S. Staubert, M. Lobe, Automated import of clinical data from HL7 messages into OpenClinica and tranSMART using mirth connect, Stud. Health. Technol. Inform. 228 (2016) 317–321.
[28] J. Doods, P. Neuhaus, M. Dugas, Converting ODM metadata to FHIR questionnaire resources, Stud. Health. Technol. Inform. 228 (2016) 456–460.
[29] Medical data models, Institute of Medical Informatics, University of Muenster in Germany, 2016. <https://medical-data-models.org/> (accessed 16.7.12).
[30] S Tripathi, SDTM mapping becomes more EASY with SAS Clinical Data Integration!!, 2016. <http://www.lexjansen.com/phuse/2014/pp/PP20.pdf> (accessed 16.12.29).
[31] Formedix Control – Universal Compliance, Formedix, 2016. <http://www.formedix.com/software/formedix-control/> (accessed 16.7.12).
[32] SDTM-ETL for OpenClinica Users, XML4Pharma, 2016. <http://www.xml4pharma.com/SDTM-ETL/SDTM-ETL_OpenClinica.html> (accessed 16.7.12).
[33] P.A. Harris, R. Taylor, R. Thielke, J. Payne, N. Gonzalez, J.G. Conde, Research electronic data capture (REDCap)—a metadata-driven methodology and workflow process for providing translational research informatics support, J. Biomed. Inform. 42 (2) (2009) 377–381.
[34] J.D. Franklin, A. Guidry, J.F. Brinkley, A partnership approach for Electronic Data Capture in small-scale clinical trials, J. Biomed. Inform. 44 (suppl 1) (2011) S103–S108.
[35] REDCap, Vanderbilt University, 2016. <https://projectredcap.org/> (accessed 16.7.12).
[36] REDCap Version 6.12, REDCap support University of Alberta, 2016. <https://labkey.med.ualberta.ca/labkey/wiki/REDCap%20Support/page.view?name=rcv6.12> (accessed 16.12.9).
[37] J.S. Obeid, C.A. McGraw, B.L. Minor, J.G. Conde, R. Pawluk, M. Lin, J. Wang, S.R. Banks, S.A. Hemphill, R. Taylor, P.A. Harris, Procurement of shared data instruments for Research Electronic Data Capture (REDCap), J. Biomed. Inform. 46 (2) (2013) 259–265.
[38] REDCap Using Annotations, Virginia Commonwealth University, 2016. <https://www.youtube.com/watch?v=FaG-hInRBdw> (accessed 16.12.29).
[39] SNOMED CT, U.S. National Library of Medicine, 2016. <https://www.nlm.nih.gov/healthit/snomedct/> (accessed 16.12.29).
[40] LOINC, Regenstrief Institute, 2016. <http://loinc.org/> (accessed 16.12.29).
[41] S.B. Johnson, Generic data modeling for clinical repositories, J. Am. Med. Inform. Assoc. 3 (1996) 328–339.
[42] P.M. Nadkarni, C. Brandt, Data extraction and ad hoc query of an entity—Attribute—Value database, J. Am. Med. Inform. Assoc. 5 (1998) 511–527.
[43] V. Dinu, P. Nadkarni, Guidelines for the effective use of entity-attribute-value modeling for biomedical databases, Int. J. Med. Inform. 76 (2007) 769–779.
[44] The R Project for Statistical Computing, The R Foundation, 2016. <https://www.r-project.org/> (accessed 16.7.12).
[45] Pinnacle 21, 2016. <https://www.pinnacle21.net/> (accessed 16.12.29).
[46] S. Sirichenko, M. DiGiantomasso, T Collopy, Usage of Pinnacle 21 Community Toolset 2.1.1 for Clinical Programmers, 2016. <http://www.lexjansen.com/pharmasug/2016/HT/PharmaSUG-2016-HT04.pdf> (accessed 16.12.29).
[47] In REDCap, what is a Data Dictionary, and how do I use it to change the structure of my database?, Indiana University, 2016. <https://kb.iu.edu/d/bafz> (accessed 16.12.29).
[48] J.R. Alvarez, Using the API through R to automate REDCap exports Some updates, 2016. <http://biostat.mc.vanderbilt.edu/wiki/pub/Main/JoAnnAlvarez/api.pdf> (accessed 16.12.29).
[49] Write CSV in R, R Programming.net, 2016. <http://rprogramming.net/write-csv-in-r/> (accessed 16.12.29).
[50] REDCap now supports Action Tags, REDCap support University of Alberta, 2016. <https://labkey.med.ualberta.ca/labkey/wiki/REDCap%20Support/page.view?name=actiontags> (accessed 16.12.29).
[51] CDISC Shared Health And Research Electronic library (SHARE), Clinical Data Interchange Standards Consortium, 2016. <https://www.cdisc.org/standards/share> (accessed 16.12.29).
[52] Study Data Technical Conformance Guide Version 3.2, U.S. Food and Drug Administration (FDA), 2016. <http://www.fda.gov/downloads/ForIndustry/DataStandards/StudyDataStandards/UCM384744.pdf> (accessed 16.12.29).
[53] Technical Conformance Guide on Electronic Study Data Submissions, Pharmaceuticals and Medical Devices Agency (PMDA) Japan, 2016. <https://www.pmda.go.jp/files/000206449.pdf> (accessed 16.12.29).