PharmaSUG China 2018 – Paper CD-09

Every Second Counts! Save Time on Developing Trial Summary Specification

Mei Chu, PAREXEL International, Taipei, Taiwan
Kyle Chang, PAREXEL International, Taipei, Taiwan

ABSTRACT

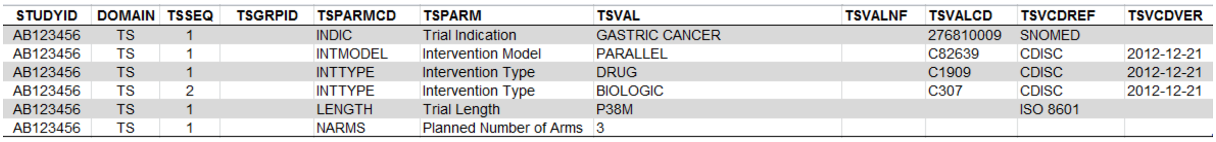
Creation of Trial Summary (TS) domain of CDISC Study Data Tabulation Model (SDTM) is complicated and always a time-consuming challenge for programmers due to the complexity of study design and trial characteristics detailed in protocol, CRF form, study related data or other supporting reference dictionaries and specifications. Apart from digging into above documents, another source is to access to the current summary information of the clinical study, if registered, provided by the ClinicalTrials.gov website. ClinicalTrials.gov is a web-based resource that provides information on publicly and privately supported clinical studies. Throughout the study, the sponsor or principal investigator of the clinical study will provide and update information on the website.

This paper offers an automation solution of SAS macro by which we implement the SAS XML facility to retrieve and convert the Extensible Markup Language (XML) format information from ClinicalTrials.gov website into SAS datasets, and structure them to fit in the electronic submission format, SDTM. By implementing this approach, it is easier to keep the TS domain aligned with the up-to-date public information on the web-based database and could approximately reduce the time and effort spent on the development by half.

INTRODUCTION

Trial Summary dataset is not subject-level data; instead, the purpose of TS dataset is to provide a high-level overview of the key components, planned and actual trial characteristics of a clinical trial, with recording basic information such as study title, trial phase, trial interventions, and trial objectives etc. Trial Summary dataset is in a structure that presents as one record one record for each trial summary characteristic.

TS domain contains below variables: Study Identifier (STUDYID), Domain Abbreviation (DOMAIN), Sequence Number (TSSEQ), Group ID (TSGRPID), Trial Summary Parameter Short Name (TSPARMCD), Trial Summary Parameter (TSPARM), Parameter Value (TSVAL), Parameter Null Flavor (TSVALNF), Parameter Value Code (TSVALCD), Name of the Reference Terminology (TSVCDREF), and Version of the Reference Terminology (TSVCDVER). TSPARMCD and TSPARM are the key components of TS domain. The corresponding attributes are listed in Parameter Value (TSVAL) and may have controlled terminology depending on the value of TSPARMCD. For TSVAL that should be presented with Controlled Terminology, TSVALCD, TSVCDREF, and TSVCDVER must be populated. TSVALCD, if applicable, should be populated with the appropriate dictionary code for the given parameter, with TSVCDREF and TSVCDVER pertaining to the name of the Dictionary and dictionary version respectively, as shown in figure 1 below.



STUDYID	DOMAIN	TSSEQ	TSGRPID	TSPARMCD	TSPARM	TSVAL	TSVALNF	TSVALCD	TSVCDREF	TSVCDVER
AB123456	TS	1		INDIC	Trial Indication	GASTRIC CANCER		276810009	SNOMED	
AB123456	TS	1		INTMODEL	Intervention Model	PARALLEL		C82639	CDISC	2012-12-21
AB123456	TS	1		INTTYPE	Intervention Type	DRUG		C1909	CDISC	2012-12-21
AB123456	TS	2		INTTYPE	Intervention Type	BIOLOGIC		C307	CDISC	2012-12-21
AB123456	TS	1		LENGTH	Trial Length	P38M			ISO 8601	
AB123456	TS	1		NARMS	Planned Number of Arms	3				

Figure 1. Trial Summary Domain Specification Example

As the study designs vary across studies, the trial summary parameters of one study may be different from other studies. However, there is a minimum set of required or expected parameters that could cover the essential scope and content of the study. In addition, in order to reflect the whole picture of the study, conditionally required parameters and permissible parameters, if applicable, should be also included in the TS dataset. With the extended list of parameters, the TS dataset gives full coverage of the clinical trial. Table 1 shows the list of parameters provided in CDISC SDTM Implementation Guide 3.2 Appendix C1. For clear definition and additional information, please refer to SDTM-IG 3.2 Appendix C1.

TYPE	TSPARMCD
Required	ACTSUB, ADAPT, ADDON, AGEMIN, AGEMAX, DCUTDTC, DCUTDESC, FCNTRY, HLTSUBJI, LENGTH, NARMS, OBJPRIM, OUTMSPRI, PLANSUB, RANDOM, REGID, SENDTC, SEXPOP, SPONSOR, SSTDTC, STOPRULE, STYPE, TBLIND, TCNTRL, TITLE, TPHASE, TTYPE
Conditionally Required	TDIGRP, TINDTP, CURTRT, TRT, RANDQT, PCLAS, INTMODEL, INTTYPE
If Applicable	OBJSEC, COMPTRT, INDIC, STRATFCT, OUTMSSEC, OUTMSEXP, SDMDUR, CRMDUR

Table 1. The List of Trial Summary Parameter Short Name

In current process of developing TS dataset, it is normal to manually grab the comprehensive information by digging into protocol, CRF form, and study related data. In industry, however, every effort is meant to be made to simplify the process and reduce the working time. Good news is that the information for most of above TSPARMCD can be found in ClinicalTrials.gov website, and is archived together in XML format.

This paper introduces a SAS macro executed to automatically extract all information of the clinical trial on ClinicalTrials.gov website, map the information to corresponding TSPARMCD and TSVAL, and convert into TS mapping specification.

OVERVIEW FLOWCHART

The following flowchart outlines the process we will go through the creation of trial summary specification.

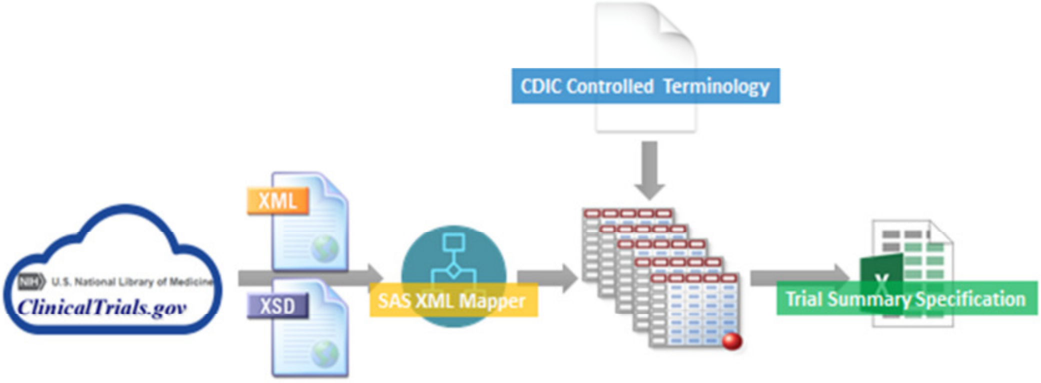

CDIC Controlled Terminology
NIH U.S. National Library of Medicine
ClinicalTrials.gov
XML
XSD
SAS XML Mapper
Trial Summary Specification

Figure 2. Overview Flowchart

PREPARATIONS BEFORE THE MACRO

To read and convert XML data, an XML file needs to combine an XML Schema to allow machines to carry out rules defined in the XML schema. XML Schemas are the languages which express the constraints and content for defining the data structure, shared vocabularies, and semantics of XML documents.

- Download Study XML File and XML Schema File

  Each registered study information can be downloaded in XML format from the ClinicalTrials.gov. To download or display an individual study record in your browser in XML, we can add the URL parameter "`displayxml=true`" to the end of the URL. The XML schema defining the structure of clinical trial study records can be downloaded from the ClinicalTrials.gov website, https://clinicaltrials.gov/ct2/html/images/info/public.xsd.

- Generate and Customize the XML Map File

  The SAS XML Mapper is implemented to analyze the XML schema file, and generate an XML MAP file for the study XML file, which is a set of instructions or definitions of the data structure on how to read XML data. The created XML MAP file will then be used by SAS XML LIBNAME engine to convert the XML data into SAS data sets.

The default length for string data in the original XML map file is 32, which is highly insufficient for the trial summary information. We update it to 32000 to ensure all information is properly read into SAS data sets.

```
<TABLE description="clinical_study" name="clinical_study">
        <TABLE-PATH syntax="XPath">/clinical_study</TABLE-PATH>
    <COLUMN name="official_title">
        <PATH syntax="XPath">/clinical_study/official_title</PATH>
        <DESCRIPTION>official_title</DESCRIPTION>
        <TYPE>character</TYPE>
        <DATATYPE>string</DATATYPE>
        <LENGTH>32000</LENGTH>
    </COLUMN>
</TABLE>
```

CONCEPT AND PROCESS OF THE SAS MACRO

Key features of the macro

a. Follow the list of TSPARMCD. By doing so, developers are allowed to drop the unessential TSPARMCDs or to manually modify or fill in information if a specific TSPARMCD is important to this study.

b. Perform fuzzy matching for TSPARMCD and get the corresponding TSVAL automatically.

c. For the parameter referencing to CDISC controlled terminology, ISO8601 and ISO3166, automatically fill in the appropriate dictionary code, the name of the dictionary and dictionary version from the dictionary files.

d. For TSVAL which exceeds the length limit of 200 characters, automatically split into TSVAL - TSVALn variables.

```
%macro autoTs(numberNCT = NCT12345678,
              verCT     = 2017-06-30,
              pathOut   = /user/specified/output/path/
);
```

Explanation of the macro

- Download Study XML File

  We use the command “curl” to automatically download study specific XML file by using National Clinical Trial (NCT) number in the macro. Apply -k option to turn off the verification for the certificate of the website, and –o option followed with a new file name to rename the downloaded file.

```
curl -k -o &numberNCT..xml
"https://clinicaltrials.gov/ct2/show/NCT12345678??displayxml=true"
```

- Convert XML file with SAS XML/XMLV2 LIBNAME engine

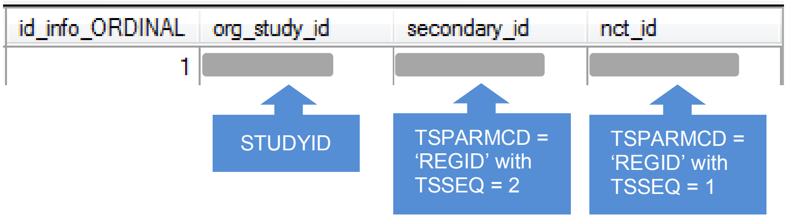
  The SAS XML/XMLV2 LIBNAME engine translates the XML document to SAS data sets based on the created XML Map file. The COPY procedure is then used to restore the series of data sets from the XML document. The naming of created SAS data sets will align with the corresponding TABLE description defined in the XML document.

```
FILENAME ts    "&numberNCT..xml";
FILENAME tsmap "ClinicalTrialsGov.map";

LIBNAME ts XMLV2 XMLMAP=tsmap ACCESS=readonly;

PROC COPY IN=ts OUT=work;
RUN;
```

During macro execution, only data sets with observations will be converted into individual data set and then all of them are aggregated into one data set for later outputting TS specification. Below are the examples of individual data set.

- ID_INFO data set:

  Record the study id, and the registry identifier. Normally there should be two types of registry identifier; one is NCTxxxxxxxx and the other one isxxxx-xxxxxx-xx.

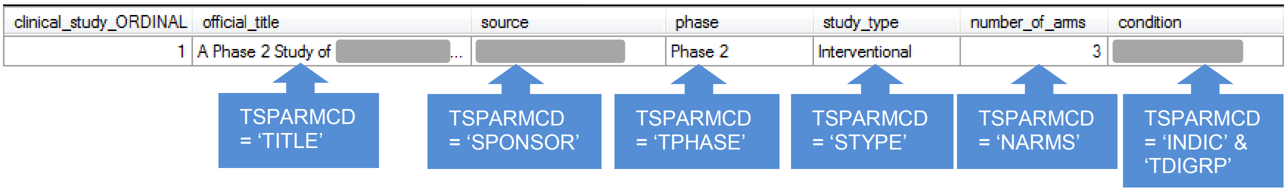


id_info_ORDINAL	org_study_id	secondary_id	nct_id
1	[redacted]	[redacted]	[redacted]

STUDYID
TSPARMCD = 'REGID' with TSSEQ = 2
TSPARMCD = 'REGID' with TSSEQ = 1

- CLINICAL_STUDY data set:

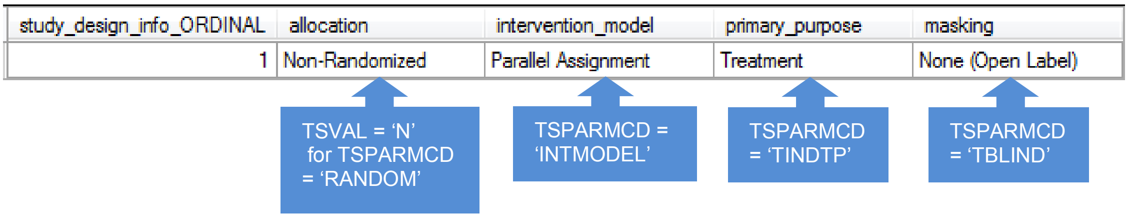
  Record the study title, sponsor, trial phase, study type, number of arms, and the indication.



clinical_study_ORDINAL	official_title	source	phase	study_type	number_of_arms	condition
1	A Phase 2 Study of [redacted]...	[redacted]	Phase 2	Interventional	3	[redacted]

- STUDY_DESIGN_INFO data set:

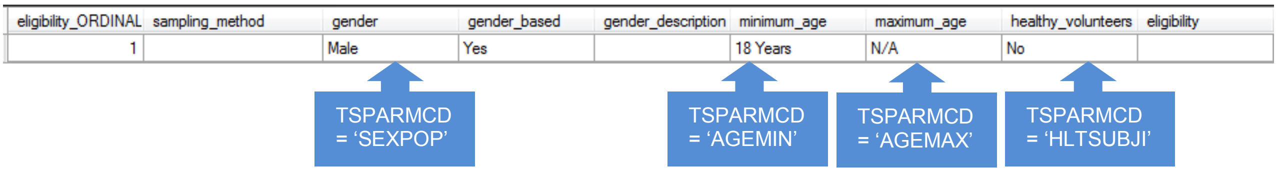
  Information about randomization, masking, intervention approach, and study purpose is recorded in this data set.



study_design_info_ORDINAL	allocation	intervention_model	primary_purpose	masking
1	Non-Randomized	Parallel Assignment	Treatment	None (Open Label)

- ELIGIBILITY data set:

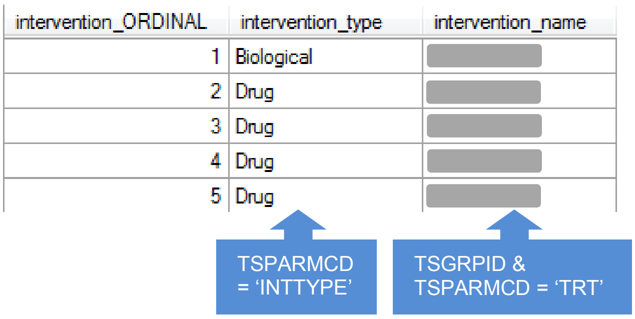
  The criteria for age, gender, and subject's health condition are recorded in ELIGIBILTY data set.



eligibility_ORDINAL	sampling_method	gender	gender_based	gender_description	minimum_age	maximum_age	healthy_volunteers	eligibility
1		Male	Yes		18 Years	N/A	No	

- INTERVENTION data set:

  The INTERVENTION data set provides information related to the intervention type and intervention name. If there are more than one intervention names, the intervention name are treated as the group id, i.e. TSGRPID, and will then be used to tie together the multiple records within the same parameter in the TS data set, e.g., INTTYPE.



intervention_ORDINAL	intervention_type	intervention_name
1	Biological	[redacted]
2	Drug	[redacted]
3	Drug	[redacted]
4	Drug	[redacted]
5	Drug	[redacted]

- Data Preprocessing

  Prior to matching the aggregated data set with TSPARMCDs, as there are some imprecise terms in the metadata, and in order to properly correspond them to suitable TS parameters, these terms need to be translated into precise terminologies in a constructive and systematic way beforehand. Standardizing TSVAL to be Y(Yes) or N(No), and replacing with synonyms or analogous text, such as replacing “Gender” with “Sex” or replacing “enrollment” with “Planned number of subjects”, can effectively improve the results. After the data preprocessing, it can perfectly associate the study trial summary metadata with the TSPARMCD list.

- Harmonization for TSVAL via Fuzzy Matching Approach

  The almost last step is to convert the original text string into well-defined TSVAL via suitable controlled terminology depending on TSPARMCD.

  Planned Country of Investigational Sites (FCNTRY) will be converted to ISO 3166-1 alpha-3 code.

  If TSVAL is null, TSVALNF will be populated based on the ISO 21090 null flavor terminology. TSVCDREF is filled with “ISO 21090” and TSVALNF should be assigned appropriate null flavor value. For example, for TSPARMCD = AGEMAX, and if the study does not specify the maximum age, the appropriate null flavor is PINF, which stands for positive infinity. It should be noted that if the study has an upper limit on the age of study subjects, the TSVAL for AGEMAX should be expressed as an ISO 8601 time duration. For TSPARM correlated to date and duration, such as Data Cutoff Date (DCUTDTC), Trial Length (LENGTH) and Stable Disease Minimum Duration (SDMDUR), the TSVALs will be converted to ISO 8601 date format.

  For those TSPARMCD that should be presented with SDTM controlled terminology, a fuzzy matching method is utilized and the COMPGED function in SAS is implemented to perform the successful association between original text string and the CDISC Submission Value. Trial phase is the only one exception that is mapped between original text string and the CDISC Synonym due to the Roman numerals used in CDISC Submission Value. The COMPGED function calculates the distance between two strings, indicating the similarity of them. The modifier ‘I’ within COMPGED function denotes case insensitive matching. The smaller the COMPGED distance is, the more similar the two strings are. Zero denotes two identical strings. In this way, only the pairs with smallest COMPGED distance per TSPARMCD per TSSEQ will be kept. The resulted CDISC Submission Value and CDISC Code then go to the final TSVAL and TSVALCD. TSVCDVER is populated automatically by the version of SDTM CT.

```
PROC SQL;
    CREATE TABLE ts AS
    SELECT a.*, b.* ,
           CASE WHEN prxmatch("/TPHASE/i", TSPARMCD) THEN
                     COMPGED(b.cdisc_synonym, a.TSVAL, 'i')
                ELSE COMPGED(b.cdisc_submission_value, a.TSVAL, 'i') END AS diff
    FROM dummy_ts AS a LEFT JOIN ct AS b
    ON a.TSPARMCD = b.cdname
    ORDER BY TSPARMCD, TSSEQ, calculated diff;
QUIT;
```

- SAS ODS to Output TS Specification

  The last step of the macro is to implement the SAS ODS Tagsets.ExcelXP destination and the REPORT procedure to output the final data set and generate XML output that conform to the Microsoft XML Spreadsheet which can be open with EXCEL. The sample output is displayed below, and some columns are hidden due to the limited space, i.e., STUDYID, DOMAIN, TSPARM.

As shown in below table, for information that is not provided on the ClinicalTrial.gov, the corresponding TSPARMCD is left blank and the macro automatically highlights the TSVAL in light pink for required TSPARMCD, light blue for conditional required and permissible TSPARMCD.

Only few mandatory TSPARMCDs are missing the corresponding TSVAL either because there is no relevant information provided on ClinicalTrial.gov or because it needs further modification or modified based on the actual data or other dictionary.

The type of parameter will be displayed in the output, keeping the flexibility for the developer to decide if this TSPARMCD is essential for the study. This column can be deleted afterward.

TSSEQ	TSGRPID	TSPARMCD	type	TSVAL	TSVALNF	TSVALCD	TSVCDREF	TSVCDVER
1		ACTSUB	Req		NAV		ISO 21090	
1		ADAPT	Req					
1		ADDON	Req					
1		AGEMAX	Req		PINF		ISO 21090	
1		AGEMIN	Req	P18Y			ISO 8601	
1		COMPTRT	Perm				UNII	
1		CRMDUR	Perm				ISO 8601	
1		CURTRT	cReq				UNII	
1		DCUTDESC	Req					
1		DCUTDTC	Req				ISO 8601	
1		FCNTRY	Req	USA			ISO 3166	
2		FCNTRY	Req	ARG			ISO 3166	
3		FCNTRY	Req	AUS			ISO 3166	
4		FCNTRY	Req	BRA			ISO 3166	
5		FCNTRY	Req	CAN			ISO 3166	
6		FCNTRY	Req	CHL			ISO 3166	
7		FCNTRY	Req	COL			ISO 3166	
8		FCNTRY	Req	FRA			ISO 3166	
9		FCNTRY	Req	DEU			ISO 3166	
10		FCNTRY	Req	MEX			ISO 3166	
11		FCNTRY	Req	ESP			ISO 3166	
1		HLTSUBJI	Req	N		C49487	CDISC	2017-06-30
1		INDIC	Perm	Disease X			SNOMED	
1		INTMODEL	cReq	PARALLEL		C82639	CDISC	2017-06-30
1	TRT A	INTTYPE	cReq	BIOLOGIC		C307	CDISC	2017-06-30
2	TRT B	INTTYPE	cReq	DRUG		C1909	CDISC	2017-06-30
3	TRT C	INTTYPE	cReq	DRUG		C1909	CDISC	2017-06-30
4	TRT D	INTTYPE	cReq	DRUG		C1909	CDISC	2017-06-30
5	TRT E	INTTYPE	cReq	DRUG		C1909	CDISC	2017-06-30
1		LENGTH	Req				ISO 8601	
1		NARMS	Req	3				
1		OBJPRIM	Req					
1		OBJSEC	Perm					
1		OUTMSEXP	Perm					
1		OUTMSPRI	Req	Objective Response Rate (ORR)				
2		OUTMSPRI	Req	Prostate-specific antigen response rate (RR-PSA)				
1		OUTMSSEC	Perm	Radiographic progression-free survival (rPFS)				
2		OUTMSSEC	Perm	Time to response (TTR)				
3		OUTMSSEC	Perm	Duration of response (DOR)				
4		OUTMSSEC	Perm	Time to prostate-specific antigen progression (TTP-PSA)				
5		OUTMSSEC	Perm	Overall Survival (OS)				
6		OUTMSSEC	Perm	Incidence of adverse events (AEs)				
7		OUTMSSEC	Perm	Incidence of serious adverse events (SAEs)				
1		PCLAS	cReq				NDF-RT	
1		PLANSUB	Req	300				
1		RANDOM	Req	N		C49487	CDISC	2017-06-30
1		RANDQT	cReq					
1		REGID	Req	NCT12345678		NCT12345678	ClinicalTrials. GOV	
1		SDMDUR	Perm				ISO 8601	
1		SENDTC	Req	2020-11-18			ISO 8601	
1		SEXPOP	Req	M		C20197	CDISC	2017-06-30
1		SPONSOR	Req	ABC				
1		SSTDTC	Req	2017-12-15			ISO 8601	
1		STOPRULE	Req					
1		STRATFCT	Perm					
1		STYPE	Req	INTERVENTIONAL		C98388	CDISC	2017-06-30
1		TBLIND	Req	OPEN LABEL		C49659	CDISC	2017-06-30

TSSEQ	TSGRPID	TSPARMCD	type	TSVAL	TSVALNF	TSVALCD	TSVCDREF	TSVCDVER
1		TCNTRL	Req				CDISC	2017-06-30
1		TDIGRP	cReq				CDISC	2017-06-30
1		TINDTP	cReq	TREATMENT		C49656	CDISC	2017-06-30
1		TITLE	Req	A Phase 2 Study of ……………				
1		TPHASE	Req	PHASE II TRIAL		C15601	CDISC	2017-06-30
1		TRT	cReq	TRT A			UNII	
2		TRT	cReq	TRT B			UNII	
3		TRT	cReq	TRT C			UNII	
4		TRT	cReq	TRT D			UNII	
5		TRT	cReq	TRT E			UNII	
1		TTYPE	Req				CDISC	2017-06-30

Table 2. The Sample Output of Produced Trial Summary Specification

CONCLUSION

This paper introduces a SAS Macro that does most jobs of collecting and cataloging the trial summary information for us simply by downloading the XML-formatted document from the ClinicalTrial.gov web database and executing the automation macro proposed in this paper. And on average, it just takes you less than 2 minutes to execute the macro.

There may be several parameters needed to be further checked and filled in by hand, this macro tackles a lot of works and saves a bunch of effort that we spend on cross-checking between various documents. With this macro, the process of trial summary specification will be significantly facilitated when encountering a bunch of tasks for the development of TS specification.

REFERENCES

Clinical Data Interchange Standards Consortium. Study Data Tabulation Model Implementation Guide: 3.2. 2013. Available at http://www.cdisc.org/sdtm

Liu, Yi; Erskine, Jenny; Read, Stephen, “Summing up - SDTM Trial Summary Domain.” PharmaSUG 2015. Available at https://www.pharmasug.org/proceedings/2015/PO/PharmaSUG-2015-PO10.pdf

Tambascia , Nicola; Eisberg, Anita, “A “How-to Guide” for Trial Summary 3.1.3”, PhUSE SDE Frankfurt 2014. Available at http://www.phusewiki.org/docs/%2309%20-%20A%20how-to%20guide%20for%20Trial%20Summary%203.1.3x.pdf

SAS Institute Inc. “SAS® 9.3 XML LIBNAME Engine: User’s Guide”,
Available at: http://support.sas.com/documentation/cdl/en/engxml/65362/PDF/default/engxml.pdf

Martell, Carol A., “SAS® XML Mapper to the Rescue”, Available at: https://analytics.ncsu.edu/sesug/2007/CC12.pdf

Staum, Paulette; Consulting, Paul W.; Nyack, West. “Fuzzy Matching using the COMPGED Function”,
Available at: https://www.lexjansen.com/nesug/nesug07/ap/ap23.pdf

ACKNOWLEDGMENTS

The special thanks go to Mark Lin of PAREXEL International, for his review, suggestions, comments, and critiques on this paper.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Mei Chu
PAREXEL International
22F, Far Glory International Center, No. 200,
Sec. 1, Keelung Road, Taipei, Taiwan 11071, ROC
Mei.Chu@parexel.com
http://www.parexel.com/

Kyle Chang
PAREXEL International
22F, Far Glory International Center, No. 200,
Sec. 1, Keelung Road, Taipei, Taiwan 11071, ROC
Kyle.Chang@parexel.com
http://www.parexel.com/