PharmaSUG China 2021 - Paper AD-11

Auto-annotation in SDTM aCRF Using R Package and SAS

Teresa Tang, Merck

ABSTRACT

SDTM aCRF is one important component in SDTM submission package, aside CDISC guidance, most of the pharmaceutical MNC established company level SDTM aCRF standard. Thus, to ensure the quality and efficiency on generating the SDTM aCRF based on both CDISC and company level SDTM standard is an essential topic to both sponsor and CRO companies. In this paper, we investigated a novel way to use R package and SAS to extract the Inform eCRF contents and annotations from company level SDTM aCRF standard to build standard master file, based on which study SDTM aCRF can be generated automatically to ensure the quality and efficiency.

INTRODUCTION

SDTM aCRF is one important component in SDTM submission package, and it is also the starting point of SDTM package creation. To generate the SDTM aCRF, most of works rely on manual efforts, as well as the review of SDTM aCRF (as currently Pinnacle 21 dese not support the detailed checks between aCRF annotations against define.xml or xpt datasets). To optimize the process and eliminate manual work as much as possible, it had previously been investigated a lot on how to read the eCRF design towards blank eCRF using Java script[1] or Ghostscript[2]. In this paper, we will introduce a novel way to use R package to interpret blank Inform eCRF to get eCRF design information and by linking with standard master file to generate xfdf file of study annotations which later import into blank eCRF to generate study SDTM aCRF.

FLOWCHART OF SDTM ACRF AUTO-ANNOTATION

To better explain the details of the automation process, we generated the flowchart as below for this Inform EDC SDTM aCRF automation, and the details will be described in the following sections.

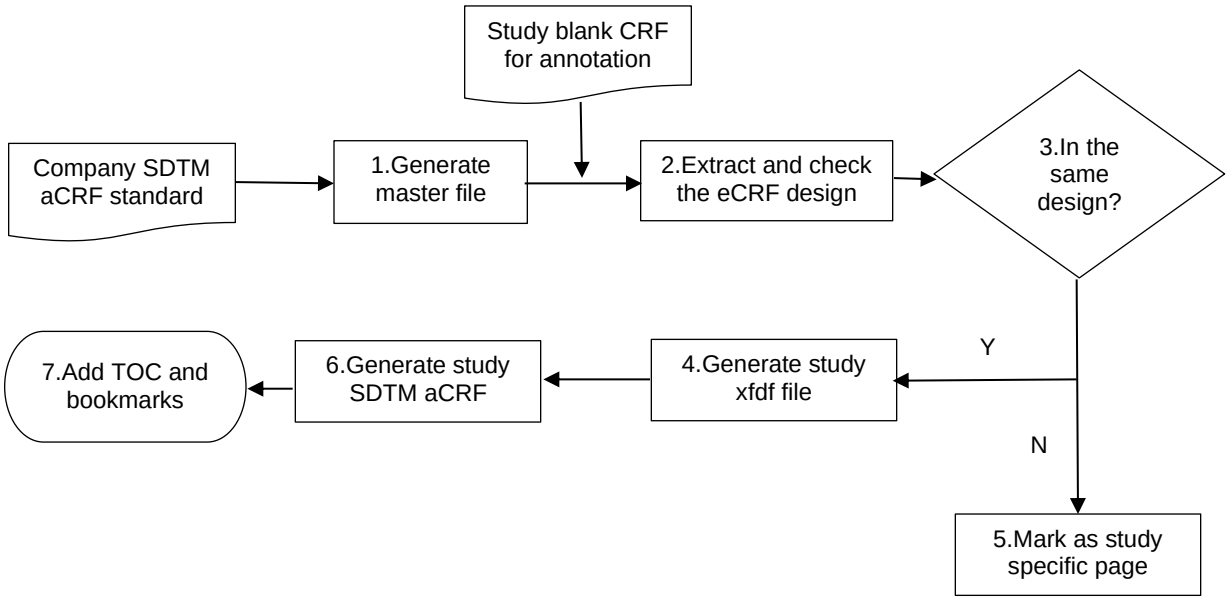

Figure 1. Flowchart of SDTM aCRF auto-annotation

In the above workflow, R package introduced in step 1 and step 2.

DETAILED DESCRIPTION OF SDTM ACRF AUTO-ANNOTATION STEPS

STEP 1 GENERATE MASTER FILE

Use R package **tm** to extract the Inform eCRF contents from company level standard SDTM aCRF.

```
library(tm)
file <- ./acrf.pdf'
Rpdf <- readPDF(control = list(text = "-layout"))
corpus <- VCorpus(URISource(file),
                  readerControl = list(reader = Rpdf))
corpus.array <- content(content(corpus)[[1]])
```

Then replace the carriage return and line feed with the placeholder to make the output more structured and to better identify the key information and structure the array in a data frame.

```
for(i in 1:length(corpus.array)){
  corpus.array[i] <- gsub("[\r\n]", "#####", corpus.array[i])
}


df <- data.frame(corpus.array)
colnames(df) = c("Contents")
```

The last step is to save it in the csv file.

```
write.csv2(df,"./acrf.csv")
```

And use SAS to interpret the csv file to get the eCRF form and question information of each standard SDTM aCRF page.

```
data temp1;
    set temp;
    length temp1 $30000;
    /*using regular expression to identify placeholder of the question*/
      ExpressionID = prxparse("/#{10}\s*\d*[.]\d*/");
      start = 1;
      stop = lengthn(x);
      /*using prxnext to find the next matched string*/
      /*using do loop to find all the matches*/
      /*using prxnext in the loop to move to next round searching*/
      call prxnext(ExpressionID, start, stop, x, position, length);
        do while (position > 0);
           found = substr(x, position, length);
           put found= position= length=;
           temp1=substr(x, start);
           call prxnext(ExpressionID, start, stop, x, position, length);
           output;
        end;
run;
```

Also extract the associate form information to generate the questions listed in forms, output as below.

pageno	panel	quesnum	question
6	oicf	1	Is the subject going to participate?
6	oicf	1	Additional consent type
6	oicf	1	Was consent withdrawn?
6	oicf	1	Record identifier
6	oicf	1	Is the consent part of the main or subsequent consent?
7	dov	1	Date of visit
7	dov	2	Not done
8	dem	1	Birth date
8	dem	2	Sex
8	dem	3	Ethnicity 1
8	dem	4	Ethnicity 2
8	dem	5	Race
9	sentry	1	Are all inclusion criteria fulfilled and are all exclusion criteria

Output 1. Output from SAS interpreted csv file

Extract the annotations from standard SDTM aCRF in xfdf file and use SAS to extract all the important components from xfdf file.

Create the master file by associating the form and questions with annotations and their associated components in the company level standard SDTM aCRF (see example below).

page	panel	question_numbe	question	annotation
7	dem	1	Birth date	DM = Demographics
7	dem	1	Birth date	BRTHDTC
7	dem	2	Sex	SEX
7	dem	3	Ethnicity 1	ETHNIC
7	dem	4	Ethnicity 2	ETHNIC2 in SUPPDM
7	dem	5	Race	RACE
7	dem	5	Race	RACENC in SUPPDM where RACE = ""
7	dem	5	Race	RACEOTH in SUPPDM where RACE = "OTHER"

Figure 2. Example of the master file

crf_page	annotation	style	rect	color	question	question_number	panel
7	DM = Demographics	font-size:14.0pt;text-align:l...	18.480700,40.569500,36.490800,206.737000	#BFFFFF	Birth date	1	dem
7	BRTHDTC	font-size:9.0pt;text-align:lef...	58.318500,438.283000,71.684500,496.760000	#BFFFFF	Birth date	1	dem
7	SEX	font-size:9.0pt;text-align:lef...	91.089300,358.882000,104.455000,384.211000	#BFFFFF	Sex	2	dem
7	ETHNIC	font-size:9.0pt;text-align:lef...	127.948000,357.928000,141.314000,397.183000	#BFFFFF	Ethnicity 1	3	dem
7	ETHNIC2 in SUPPDM	font-size:9.0pt;text-align:lef...	161.431000,353.424000,174.237000,466.030000	#BFFFFF	Ethnicity 2	4	dem
7	RACENC in SUPPDM where RACE = ""	font-size:9.0pt;text-align:lef...	299.640000,370.370000,315.119000,550.855000	#BFFFFF	Race	5	dem
7	RACE	font-size:9.0pt;text-align:lef...	204.639000,394.561000,218.005000,426.130000	#BFFFFF	Race	5	dem
7	RACEOTH in SUPPDM where RACE = "OTHER"	font-size:9.0pt;text-align:lef...	327.640000,368.992000,343.119000,579.627000	#BFFFFF	Race	5	dem

Output 2. XFDF file output

Columns	**Explaining**
CRF_PAGE	CRF page number
ANNOTATION	Value in the annotation box
STYLE	Annotation style of font size, align method, color, etc.

Columns	Explaining
RECT	Axis of the annotation rectangle
COLOR	Color of the annotation background
QUESTION	Question text on eCRF on that page
QUESTION_NUMBER	Question serial number on eCRF
PANEL	Form name on eCRF

Table 1. Columns explanation to XFDF file output

STEP 2 EXTRACT AND CHECK THE ECRF DESIGN

The study blank eCRF information can be extracted also using R package and use SAS to interpret the csv file to get the eCRF form and question information together with page number, the process is quite similar as step 1 for standard aCRF, and this output will be used later in the merge step to check the standard form and questions to perform auto annotations.

STEP 3 GENERATE STUDY XFDF FILE

After comparison of extracted eCRF form and questions information between study eCRF and standard aCRF, those consistent parts can be linked with corresponding annotation and related xfdf components based on master file. Then the study aCRF annotation xfdf file can be created for those parts automatically with SAS macro (sample code can be seen in example below).

```
/*create study annotation file merging from standard annoation*/
proc sql noprint;
    create table anno as
    select a.pageno, b.*
    from out.panel_question_xxx as a inner join out.acrf_standard as b
    on a.panel=b.panel and a.question=b.question
    order by pageno, question_number;
quit;

/*create study xfdf file*/
data output;
    length line $ 5000;
    set anno end=last;
    if _n_ = 1 then do;
          line = "<?xml version=""1.0"" encoding=""UTF-8""?>"; output;
          line = "<xfdf xmlns=""http://ns.adobe.com/xfdf/""
xml:space=""preserve"">"; output;
          line = " <annots"; output;
    end;
    line = '><freetext color="' || strip(color) || '" page="' ||
strip(put(pageno, best.)) ||'" rect="'|| strip(rect)||'" rotation="90"
subject="Text Box"'; output;
    line = "><contents-richtext"; output;
    line = '><body style="'||strip(style)||'"'; output;
    if annotation not in ('[NOT SUBMITTED]') then do;
    line = '><p dir="ltr"'; output;
    line = '>'||strip(annotation)||'</p'; output;
    end;
    else do;
    line = '><p'; output;
    line = '>'||strip(annotation)||'</p'; output;
    end;
```

```
    line = "></body"; output;
    line = "></contents-richtext"; output;
    line = "></freetext"; output;
    if last then do;
          line = "></annots"; output;
          line = "></xfdf"; output;
          line = ">"; output;
    end;
run;

data _null_;
    file 'path/test_acrf.xfdf';
    set output;
    put line;
run;
```

The study specific forms which are without standard form association in master file can be addressed later on in a separate file, and they need to be checked later on against the exemptions which are deviations against company standards to ensure the quality.

STEP 4 GENERATE STUDY SDTM ACRF

By importing the xfdf file, a study SDTM aCRF is generated automatically contains all the standard annotations based on eCRF designed forms and questions (see example below).

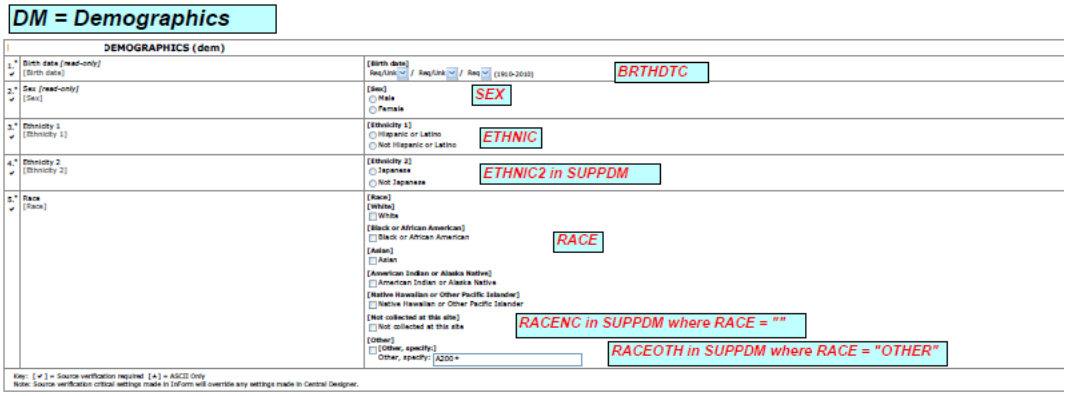

Figure 3. Example of the output aCRF page

This is an automatic step without manual efforts among different studies to reduce manual work to improve both quality and efficiency.

After importing the standard pages, please check and update with study specific pages and include into the company master file after aligned with company standard team.

Before submission, the table of contents and bookmarks also need to be made in SDTM aCRF as well for the hyperlinks of the TOC and bookmarks.

CONCLUSION

To use R package **tm** to extract the Inform eCRF information from company level standard SDTM aCRF and study blank CRF provided us an option in the SDTM aCRF auto-generation. As most of the pharmaceutical companies had already setup R software at their platform, this is to provide a more feasible way to improve the quality and efficiency in the Inform SDTM aCRF generation. After got the

study eCRF design and the company level standard master file, the annotations can be automatically added via xfdf file created by SAS.

REFERENCES

Haiqiang Luo and Yong Cao 2015. “Automatic generating blankcrf.pdf for Rave Study” *PharmaSUG China 2015-60*.

Boxun Zhang, Tyler Kelly 2015. “A Unique Way to Annotate Case Report Forms (CRFs) in PDF, Using Forms Data Format (FDF) Techniques” *PharmaSUG 2015 - Paper BB13*.

ACKNOWLEDGMENTS

I would like to acknowledge Joerg Lehbauer (Merck Germany) for helping us setup the R package in Merck platform and Elaine Zhao who helped me on this logic and programming.

CONTACT INFORMATION

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

Name: Teresa Tang
Enterprise: Merck
Phone: +86 10 59031448
E-mail: teresa.tang@merckgroup.com
Web: www.merckserono.com