PharmaSUG 2020 - Paper AI-242

Automate your Safety tables using Artificial Intelligence & Machine Learning

Roshan Stanly, Ajith Baby Sadasivan, Limna Salim

Genpro Research

ABSTRACT

As part of the FDA submissions, it has been a common practice to create tables that present the output of statistical analysis of trial data. Some of these tables display the statistics of subject's safety parameters such as adverse events, laboratory results, vital signs etc. This paper explores the possibility of automating safety table (eg. Demographics, disposition, change from baseline etc.) generation using Natural Language Processing and Machine Learning algorithms. The proposed software framework uses Angular JS, SAS®, R and Python® for automating safety table generation.

INTRODUCTION

For automating the safety tables, we assume that standardized table shells and ADaM datasets will be provided to the framework. The system has standardized templates for most of the safety tables which will vary depending upon the study design (e.g. single arm, multiple arm, cross over etc.). As a first step, you must select the table shells from the various templates in your library. Once you feed the shell, this tool automatically extracts its contents. The contents will be classified as Titles, Headers, Parameters & Sub-Parameters, Statistics, Footnotes etc. This is performed using a table extraction tool called Camelot. The extracted contents will then be stored into a CSV file. Once the table contents are extracted, a map file is created using a semi-supervised machine learning model. This map file contains a mapping from the ADAM datasets to the parameters that has been extracted from the table shell. The extracted CSV file, map file and the ADaM data sets are then passed on to a standard macro written in SAS to generate the final table in rtf format. Please note that the automation can be performed only for standardized table shells offered by the tool. If the shells are very complex, then the tool needs to be further customized.

SHELL PROCESSING

As a first step, this tool uses Camelot to extract the contents of the mock shell. Camelot is a Python library that makes it easy for anyone to extract tables from PDF files. The main advantages of using Camelot for extraction is that it gives power to tweak table extraction unlike other tools. We can also export the output to multiple formats like JSON, Excel, CSV and HTML.

As mentioned already, mock shells should follow the structure of standardized templates supported by the tool . Listed below are two different templates for demographics table and a demonstration on how the processed output looks like.

Figure 1. Template for Table 1

Header	Header 1 N=XX n (%)	Header 2 N=XX n (%)	Header 3 N=XX n (%)	Header 4 N=XX n (%)	Column for total N=XX n (%)
Categorical Variable 1					
Categorical Value 1	xx (xx.x)	xx (xx.x)	xx (xx.x)	xx (xx.x)	xx (xx.x)
Categorical Value 2	xx (xx.x)	xx (xx.x)	xx (xx.x)	xx (xx.x)	xx (xx.x)
Categorical Variable 2					
Categorical Value 1	xx (xx.x)	xx (xx.x)	xx (xx.x)	xx (xx.x)	xx (xx.x)
Categorical Value 2	xx (xx.x)	xx (xx.x)	xx (xx.x)	xx (xx.x)	xx (xx.x)
Categorical Variable 3					
Categorical Value 1	xx (xx.x)	xx (xx.x)	xx (xx.x)	xx (xx.x)	xx (xx.x)
Categorical Value 2	xx (xx.x)	xx (xx.x)	xx (xx.x)	xx (xx.x)	xx (xx.x)
Categorical Value 3	xx (xx.x)	xx (xx.x)	xx (xx.x)	xx (xx.x)	xx (xx.x)
Categorical Value 4	xx (xx.x)	xx (xx.x)	xx (xx.x)	xx (xx.x)	xx (xx.x)
Categorical Value 5	xx (xx.x)	xx (xx.x)	xx (xx.x)	xx (xx.x)	xx (xx.x)
Categorical Value 6	xx (xx.x)	xx (xx.x)	xx (xx.x)	xx (xx.x)	xx (xx.x)
Categorical Value 7	xx (xx.x)	xx (xx.x)	xx (xx.x)	xx (xx.x)	xx (xx.x)
Continous variable 1					
Mean ± SD (n)	xx.x ± xx.xx (x)	xx.x ± xx.xx (x)	xx.x ± xx.xx (x)	xx.x ± xx.xx (x)	xx.x ± xx.xx (x)
Median (min, max)	xx.x (xx, x)	xx.x (xx, x)	xx.x (xx, x)	xx.x (xx, x)	xx.x (xx, x)

Figure 2. Template for Table 2

	Span head 1			Span head 2		Span head 3		
Header	Header 1 (N=XX)	Header 2 (N=XX)	Header 3 (N=XX)	Header 4 (N=XX)	Header 5 (N=XX)	Header 6 (N=XX)	Header 7 (N=XX)	Total (N=XX)
Continous var 1								
Mean (SD)	xx.x (xx.xx)	xx.x (xx.xx)	xx.x (xx.xx)	xx.x (xx.xx)	xx.x (xx.xx)	xx.x (xx.xx)	xx.x (xx.xx)	xx.x (xx.xx)
Median (Q1, Q3)	xx.x (xx.x)	xx.x (xx.x)	xx.x (xx.x)	xx.x (xx.x)	xx.x (xx.x)	xx.x (xx.x)	xx.x (xx.x)	xx.x (xx.x)
Min, Max	xx, xx	xx, xx	xx, xx	xx, xx	xx, xx	xx, xx	xx, xx	xx, xx
Categorical Var 1								
Value 1	x (xx.x)	x (xx.x)	x (xx.x)	x (xx.x)	x (xx.x)	x (xx.x)	x (xx.x)	x (xx.x)
Value 2	x (xx.x)	x (xx.x)	x (xx.x)	x (xx.x)	x (xx.x)	x (xx.x)	x (xx.x)	x (xx.x)
Value 3	x (xx.x)	x (xx.x)	x (xx.x)	x (xx.x)	x (xx.x)	x (xx.x)	x (xx.x)	x (xx.x)
Value 4	x (xx.x)	x (xx.x)	x (xx.x)	x (xx.x)	x (xx.x)	x (xx.x)	x (xx.x)	x (xx.x)
Categorical var 2								
Male	x (xx.x)	x (xx.x)	x (xx.x)	x (xx.x)	x (xx.x)	x (xx.x)	x (xx.x)	x (xx.x)
Female	x (xx.x)	x (xx.x)	x (xx.x)	x (xx.x)	x (xx.x)	x (xx.x)	x (xx.x)	x (xx.x)

Please note that the number of table headers and number/order of variables in which we are calculating the counts or statistics can change based on the study. The template is only a sample representation of the table. There is no limitation to order/display in which descriptive statistics is presented, but it should follow standard naming convention like n, mean, median etc.

Once the appropriate template is selected, the tool creates a csv file named *contents* after extracting all the necessary information from the mock table shell. Table 1 Contents and Table 2 Contents in the below figure represents the format of the extracted output. Camelot is a powerful tool which will generate the output from the shells in an organized way. Below are the examples for two different demographics tables and its corresponding contents file.

Figure 3. Shell for Table 1

Table 14.1.3.1

Demographics

Characteristics	TRT A N=XX n (%)	TRT B N=XX n (%)	TRT C N=XX n (%)	TRT D N=XX n (%)	Overall N=XX n (%)
Gender					
Male	## (##.#)	## (##.#)	## (##.#)	## (##.#)	## (##.#)
Female	## (##.#)	## (##.#)	## (##.#)	## (##.#)	## (##.#)
Ethnicity					
Hispanic or Latino	## (##.#)	## (##.#)	## (##.#)	## (##.#)	## (##.#)
Not Hispanic or Latino	## (##.#)	## (##.#)	## (##.#)	## (##.#)	## (##.#)
Race					
White	## (##.#)	## (##.#)	## (##.#)	## (##.#)	## (##.#)
Black or African American	## (##.#)	## (##.#)	## (##.#)	## (##.#)	## (##.#)
American Indian or Alaska Native	## (##.#)	## (##.#)	## (##.#)	## (##.#)	## (##.#)
Asian	## (##.#)	## (##.#)	## (##.#)	## (##.#)	## (##.#)
Native Hawaiian or other Pacific Islander	## (##.#)	## (##.#)	## (##.#)	## (##.#)	## (##.#)
Others	## (##.#)	## (##.#)	## (##.#)	## (##.#)	## (##.#)
Unknown/declined	## (##.#)	## (##.#)	## (##.#)	## (##.#)	## (##.#)
Age (years)					
Mean± SD (n)	##.# ± ##.## (##)	##.# ± ##.## (##)	##.# ± ##.## (##)	##.# ± ##.## (##)	##.# ± ##.## (##)
Median (min, max)	##.# (##, ##)	##.# (##, ##)	##.# (##, ##)	##.# (##, ##)	##.# (##, ##)
Height (cm)					
Mean ± SD (n)	##.# ± ##.## (##)	##.# ± ##.## (##)	##.# ± ##.## (##)	##.# ± ##.## (##)	##.# ± ##.## (##)
Median (min, max)	##.# (##, ##)	##.# (##, ##)	##.# (##, ##)	##.# (##, ##)	##.# (##, ##)
Weight (kg)					
Mean ± SD (n)	##.# ± ##.## (##)	##.# ± ##.## (##)	##.# ± ##.## (##)	##.# ± ##.## (##)	##.# ± ##.## (##)
Median (min, max)	##.# (##, ##)	##.# (##, ##)	##.# (##, ##)	##.# (##, ##)	##.# (##, ##)
BMI (kg/m^2)					
Mean ± SD (n)	##.# ± ##.## (##)	##.# ± ##.## (##)	##.# ± ##.## (##)	##.# ± ##.## (##)	##.# ± ##.## (##)
Median (min, max)	##.# (##, ##)	##.# (##, ##)	##.# (##, ##)	##.# (##, ##)	##.# (##, ##)

Percentages are based on number of subjects in the column header

Figure 4. Table 1 Contents

Table Number	Title 1	Table Header 1	Categories	Category values	Footnote 1
Table 14.1.3.1	Demographics	Characteristics	Gender	Male	Percentages are based on number of subjects in the column header
		TRT A N=XX n (%)	Gender	Female	
		TRT B N=XX n (%)	Ethnicity	Hispanic or Latino	
		TRT C N=XX n (%)	Ethnicity	Not Hispanic or Latino	
		TRT D N=XX n (%)	Race	White	
		Overall N=XX n (%)	Race	Black or African American	
			Race	American Indian or Alaska Native	
			Race	Asian	
			Race	Native Hawaiian or other Pacific Islander	
			Race	Other	
			Age(years)	Mean± SD (n)	
			Age(years)	Median (min, max)	
			Height (cm)	Mean± SD (n)	
			Height (cm)	Median (min, max)	
			Weight (kg)	Mean± SD (n)	
			Weight (kg)	Median (min, max)	
			BMI (kg/m2)	Mean± SD (n)	
			BMI (kg/m2)	Median (min, max)	

Figure 5. Shell for Table 2

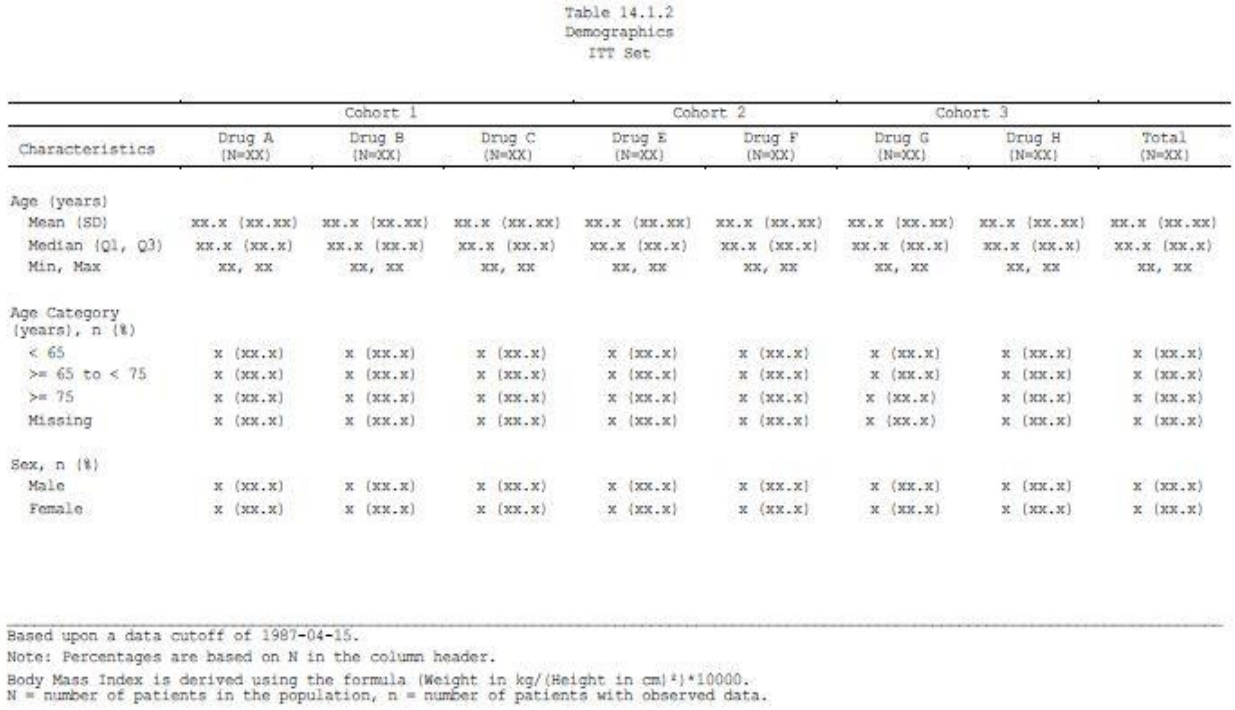

Table 14.1.2
Demographics
ITT Set

Characteristics	Cohort 1			Cohort 2		Cohort 3		
	Drug A (N=XX)	Drug B (N=XX)	Drug C (N=XX)	Drug E (N=XX)	Drug F (N=XX)	Drug G (N=XX)	Drug H (N=XX)	Total (N=XX)
Age (years)								
Mean (SD)	xx.x (xx.xx)	xx.x (xx.xx)	xx.x (xx.xx)	xx.x (xx.xx)	xx.x (xx.xx)	xx.x (xx.xx)	xx.x (xx.xx)	xx.x (xx.xx)
Median (Q1, Q3)	xx.x (xx.x)	xx.x (xx.x)	xx.x (xx.x)	xx.x (xx.x)	xx.x (xx.x)	xx.x (xx.x)	xx.x (xx.x)	xx.x (xx.x)
Min, Max	xx, xx	xx, xx	xx, xx	xx, xx	xx, xx	xx, xx	xx, xx	xx, xx
Age Category (years), n (%)								
< 65	x (xx.x)	x (xx.x)	x (xx.x)	x (xx.x)	x (xx.x)	x (xx.x)	x (xx.x)	x (xx.x)
>= 65 to < 75	x (xx.x)	x (xx.x)	x (xx.x)	x (xx.x)	x (xx.x)	x (xx.x)	x (xx.x)	x (xx.x)
>= 75	x (xx.x)	x (xx.x)	x (xx.x)	x (xx.x)	x (xx.x)	x (xx.x)	x (xx.x)	x (xx.x)
Missing	x (xx.x)	x (xx.x)	x (xx.x)	x (xx.x)	x (xx.x)	x (xx.x)	x (xx.x)	x (xx.x)
Sex, n (%)								
Male	x (xx.x)	x (xx.x)	x (xx.x)	x (xx.x)	x (xx.x)	x (xx.x)	x (xx.x)	x (xx.x)
Female	x (xx.x)	x (xx.x)	x (xx.x)	x (xx.x)	x (xx.x)	x (xx.x)	x (xx.x)	x (xx.x)

Based upon a data cutoff of 1987-04-15.
Note: Percentages are based on N in the column header.
Body Mass Index is derived using the formula (Weight in kg/(Height in cm)^2)*10000.
N = number of patients in the population, n = number of patients with observed data.

Figure 6. Table 2 Contents

Table Number	Title 1	Title 2	Table Header	Table Header 2	Categories	Category value	Footnote 1	Footnote 2	Footnote 3	Footnote 4
Table 14.1.2	Demographics	ITT Set		Characteristics	Age (years)	Mean (SD)	Based upon a data cutoff of 1987-04-15.	Note: Percentages are based on N in the column header.	Body Mass Index is derived using the formula (Weight in kg/(Height in cm)^2)*10000.	N = number of patients in the population, n = number of patients with observed data.
			Cohort 1	Drug A N=XX n (%)	Age (years)	Median (Q1, Q3)				
			Cohort 1	Drug B N=XX n (%)	Age (years)	Min, Max				
			Cohort 1	Drug C N=XX n (%)	Age Category (years), n (%)	< 65				
			Cohort 2	Drug E N=XX n (%)	Age Category (years), n (%)	>=65 to <75				
			Cohort 2	Drug F N=XX n (%)	Age Category (years), n (%)	>= 75				
			Cohort 3	Drug G N=XX n (%)	Age Category (years), n (%)	Missing				
			Cohort 3	Drug H N=XX n (%)	Sex, n (%)	Male				
				Total	Sex, n (%)	Female				

From the above figure, we can understand that Drug A, Drug B and Drug C comes under Cohort1, Drug E and Drug F comes under Cohort 2 and Drug G and Drug H comes under Cohort 3. If there are n number of titles in the table, the column names will be labeled as Title 1 to Title n and similarly for footnotes.

MAP FILE GENERATION

Next step is the creation of map file using a semi-supervised machine learning model. Semi-supervised learning is an approach in machine learning that combines a small amount of labeled data with a large amount of unlabeled data during training. Semi-supervised learning falls between unsupervised learning (with no labeled training data) and supervised learning (with only labeled training data). Unlabeled data, when used in conjunction with a small amount of labeled data, can produce considerable improvement in learning accuracy.

In this section, the tool uses the contents file and ADaM datasets as input and generates the map file. This map file contains a mapping from the ADAM datasets to the parameters that has been extracted from the table shell. Below is the map file for the demographics table that we have seen above.

Figure 7. Table 1 Map File

PDF label	ADaM variable	Header Values 1	Population flag	Total Variable
Gender	ASEX	TRT01A	SAFFL	Y
Ethnicity	AETHNIC			
Race	ARACE			
Age(years)	AGE			
Height (cm)	HEIGHTBL			
Weight (kg)	WEIGHTBL			
BMI (kg/m2)	BMIBL			

Figure 8. Table 2 Map File

PDF label	ADaM variable	Header Values 1	Header Values 2	Population flag	Total Variable
Age Category (years), n (%)	AGECAT	COHORT	TRT01A	SAFFL	Y
Age (years)	AGE				
Sex, n (%)	ASEX				

SAS MACRO

There are standardized SAS macros developed to create the final output by taking the contents file, map file and ADaM datasets as input. When this macro executes, it takes the required files from the library and generate the output. Information regarding the titles, footnotes, headers, sub headers and parameters will be extracted from the contents file. The map file is used to check the mapping of all the parameters, headers in corresponding ADaM datasets etc. If the contents file has descriptive statistics like mean, median present for the corresponding pdf label, then proc means procedure will be executed and display the summary statistics based on the layout defined in the contents file. Otherwise counts and percentages will be displayed. Proc template is created in an additional program which is also invoked in these standard macros. The advantage of having proc template in an additional program is that the user can modify the attributes such as font style, appearance of the output etc. according to their requirement if needed. There is an option in the proc template program to specify whether the outputs are draft/final or any other extra information needed in the titles or footnotes in the final rtf output.

FINAL OUTPUT

Below is the sample output generated by the tool for demographics table.

Figure 9. Table 1 Output

Study Name
Sponsor Name

Page 1 of 1
Draft

Table 14.1.3.1
Demographics

Characteristics	TRT A N= 60 n (%)	TRT B N= 0 n (%)	TRT C N= 78 n (%)	TRT D N= 0 n (%)	Overall N= 138 n (%)
Gender					
Male	60 (100%)	0	78 (100%)	0	138 (100%)
Female	0	0	0	0	0
Ethnicity					
Hispanic or Latino	22 (36.7%)	0	13 (16.7%)	0	35 (25.4%)
Not Hispanic or Latino	38 (63.3%)	0	65 (83.3%)	0	103 (74.6%)
Race					
White	50 (83.3%)	0	59 (75.6%)	0	109 (79.0%)
Black or African American	8 (13.3%)	0	17 (21.8%)	0	25 (18.1%)
American Indian or Alaska Native	0	0	0	0	0
Asian	2 (3.3%)	0	1 (1.3%)	0	3 (2.2%)
Native Hawaiian or other Pacific Islander	0	0	0	0	0
Other	0	0	1 (1.3%)	0	1 (0.7%)
Age(years)					
Mean± SD (n)	50.5± 9.83 (60)	0	56.3± 9.75 (78)	0	53.8± 10.18 (138)
Median (min, max)	51 (26, 70)		57 (36, 75)		54 (26, 75)
Height (cm)					
Mean± SD (n)	176.75± 7.5141 (60)	0	177.064± 6.5571 (78)	0	176.928± 6.9646 (138)
Median (min, max)	178 (163, 200)		178 (160, 191)		178 (160, 200)
Weight (kg)					
Mean± SD (n)	99.773± 22.1196 (60)	0	107.249± 17.7759 (78)	0	103.999± 20.0534 (138)
Median (min, max)	96.6 (51.9, 148.1)		107.55 (71.8, 153.4)		104 (51.9, 153.4)
BMI (kg/m2)					
Mean± SD (n)	31.778± 6.2057 (60)	0	34.157± 5.1809 (78)	0	33.123± 5.7508 (138)
Median (min, max)	31.535 (15.96, 45.97)		34.43 (22.38, 47.49)		32.995 (15.96, 47.49)

Percentages are based on number of subjects in the column header

Figure 10. Table 2 Output

Sponsor Name
Study Name

Page 1 of 1
Draft

Table 14.1.2
Demographics
ITT Set

Characteristics	Cohort 1			Cohort 2		Cohort 3		
	Drug A N= 21	Drug B N= 24	Drug C N= 7	Drug E N= 12	Drug F N= 12	Drug G N= 47	Drug H N= 16	Total N= 138
Age (years)								
Mean (SD)	55.6 (10.67)	54.8 (9.61)	56.7 (15.26)	53.3 (11.51)	52.3 (11.56)	52 (9.95)	55.3 (6.22)	53.8 (10.18)
Median (Q1, Q3)	54(48, 67)	58(48.5, 61.5)	61(41, 70)	55.5(45,62.5)	49.5(42.5, 64)	51(45,59)	56(51,59)	54(46,62)
Min, Max	37, 71	35, 70	33, 73	32, 68	36, 72	26, 75	40, 65	26, 75
Age Category (years), n (%)								
< 65	14 (66.7%)	23 (95.8%)	4 (57.1%)	11 (91.7%)	9 (75.0%)	44 (93.6%)	15 (100%)	120 (87.0%)
>= 65 to < 75	7 (33.3%)	1 (4.2%)	3 (42.9%)	1 (8.3%)	3 (25.0%)	3 (6.4%)	0	18 (13.0%)
>= 75	0	0	0	0	0	0	0	0
Missing	0	0	0	0	0	0	0	0
Sex, n (%)								
Male	21 (100%)	24 (100%)	7 (100%)	12 (100%)	12 (100%)	47 (100%)	15 (100%)	138 (100%)
Female	0	0	0	0	0	0	0	0

Based upon a data cutoff of 1987-04-15.
Note: Percentages are based on N in the column header.
Body Mass Index is derived using the formula (Weight in kg/(Height in cm)*)*10000.
N = number of patients in the population, n = number of patients with observed data.

In proc report, width is specified for each column based on the number of columns in the final output.

If there are any programming notes present in the footnote, those footnotes must be removed from the table shell. The standardized templates from the tool handles all the scenarios for that table.

Line break and page break will be applied in the sas macro based on the different templates.

CONCLUSION

Even though the paper provides examples and illustrations regarding the demographics table, the same can be applied to other safety tables like change from baseline, shift table, summary tables etc. as we can define standard templates for all these tables. Standard macros can be developed for all these safety tables as well. As a result, you can see that most of the safety tables which are not very complex can be automated using this tool.

ACKNOWLEDGEMENT

The content, ideas and recommendations presented in this paper are all developed from experiences in our career. These experiences come through previous companies, various industry leaders, colleagues, mentors, conferences, and direct experience. Any brand and product names are trademarks of their respective companies.

RECOMMENDED READING

- https://www.lexjansen.com/phuse/2016/ad/AD01.pdf
- https://www.lexjansen.com/phuse/2018/ad/AD04.pdf
- https://camelot-py.readthedocs.io/en/master/

CONTACT INFORMATION

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

Roshan Stanly
Genpro Research
9745266416
Roshan.stanly@genproresearch.com

Limna Salim
Genpro Research
8086272364
Limna.salim@genproresearch.com

Ajith Nair
Genpro Research
9847681659
Ajith.nair@genproresearch.com