

## Automated Generation of Clinical Study Reports using SAS® and RTF (Literate programming)

Rajaram Venkatesan, Cognizant Technology Solutions; Julien Sauser, Nestlé Research Center; Carlos Antonio De Castro, Nestlé Research Center

### ABSTRACT

Clinical study report (CSR) is the final milestone in any clinical trial. A CSR typically consists in Table of Contents (TOC), background of study, interpretation of results, tables, listings and figures (TLF). The process of populating TLFs, updating TLF numbers, cross-references and incorporating results in a CSR are laborious and potentially error prone. An 'automated' technique is presented in this paper which uses SAS® and literate programming concept to generate a CSR document populated with specified TLFs, text and cross-references. CSR written using literate programming reduces significant amount of time post database lock (DBL). Literate programming efficiently combines the text and the statistical analysis in one single SAS® source file which can be run immediately after un-blinding the data to automatically produce the CSR in Microsoft (MS) Word and PDF format. The method can also be used to produce reports that need to be updated on an ongoing basis such as Data Quality Review report as the single source document can be reused and updated multiple times without applying any change. The first draft of the CSR or any other document can then be ready in hours as opposed to days post un-blinding.

### INTRODUCTION

Nestlé regularly conducts clinical trials to ensure the safety of their products and to investigate health benefits that their products stimulate and/or support. The CSR is a key document for Nestlé Research and Development as it is usually the final deliverable. Nestlé's CSR integrates outputs (tables, listings and figures) as well as interpretation, explanations and recommendation in one document.

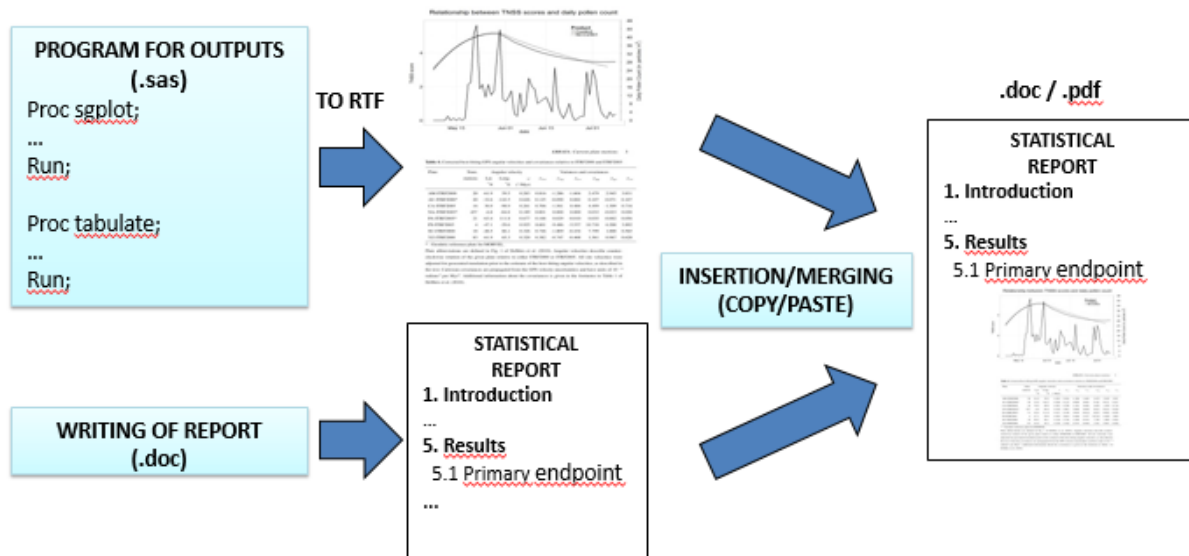


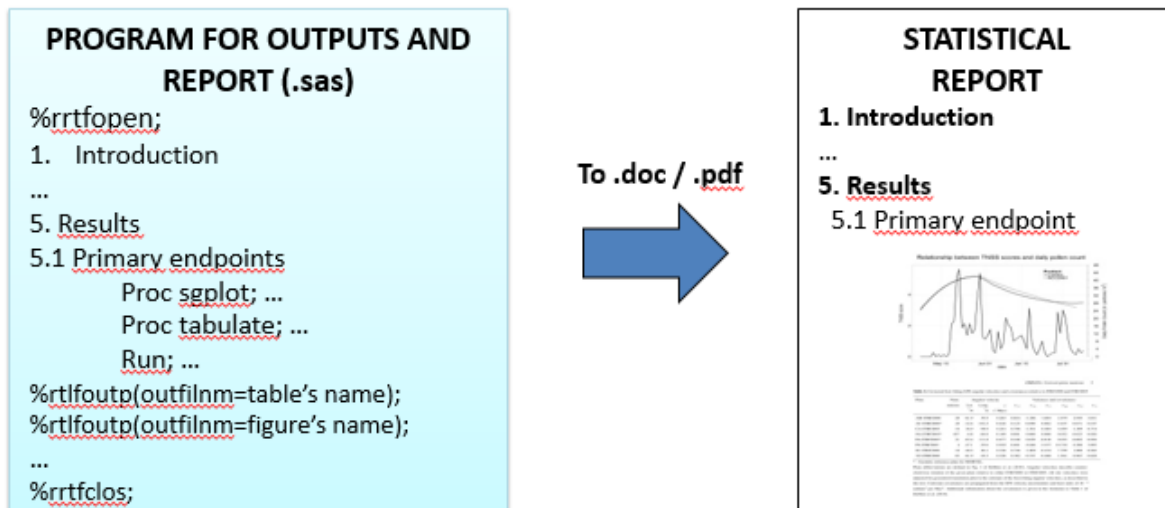
Figure 1 Traditional way of writing CSR

The writing of the report is done similar to how research papers would be developed and it's the reason why the biostatistics team adopted a literate programming approach using SAS® and Rich Text Format (RTF).

The idea is to intermingle text and SAS® code in a single file to create a statistical report.

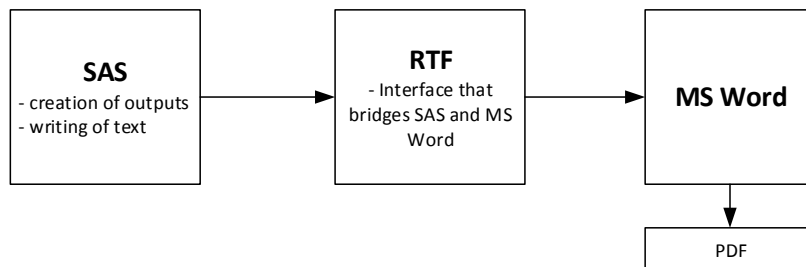
We apply literate programming and reproducible research principles within a SAS® environment. These are achieved by

- a) Structuring and writing the program(s) in an easily readable fashion
- b) Embedding executable code within the report for transparency and traceability.



**Figure 2** Literate programming approach

The SAS® programs that create the statistical outputs and write the statistical report are exported to an RTF file. In order to do this, dedicated macros (Table 1) were created to export both text and SAS® outputs to RTF. This file is then converted to MS Word and PDF files.



**Table 1** List of macros to be used to create CSR

Macro	Purpose
URTFTAGS	Useful RTF tags can be used within data step or macro
RRTFOPEN	Initializes RTF document
RTITLEPG	Creates Title page of CSR template
RSGNTRPG	Creates Signature page of CSR template
RSUMRYPG	Creates Summary page of CSR template
RACKNWDG	Creates Acknowledgements page of CSR template
RTOCPAGE	Creates TOC page of CSR template
RHEADING	Inserts headings
RPARAGRP	Insert paragraphs
RLISTITM	Insert bullet and numbering
RTLFOUTP	Inserts Tables, Listings and Figures created by standard macros
RRTFCLOS	Closes RTF document which is initiated by RRTFOPEN

The numbering of sections, figures and tables are automated. Style, formats and captions can easily be integrated within the outputs and the written text. Special characters and symbols are written through the UNICODE system.

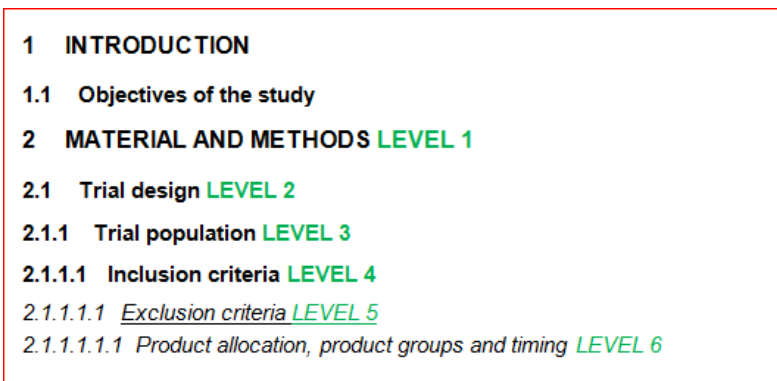
The following 4 main macros are often used for construction of the report:

**%RHEADING**

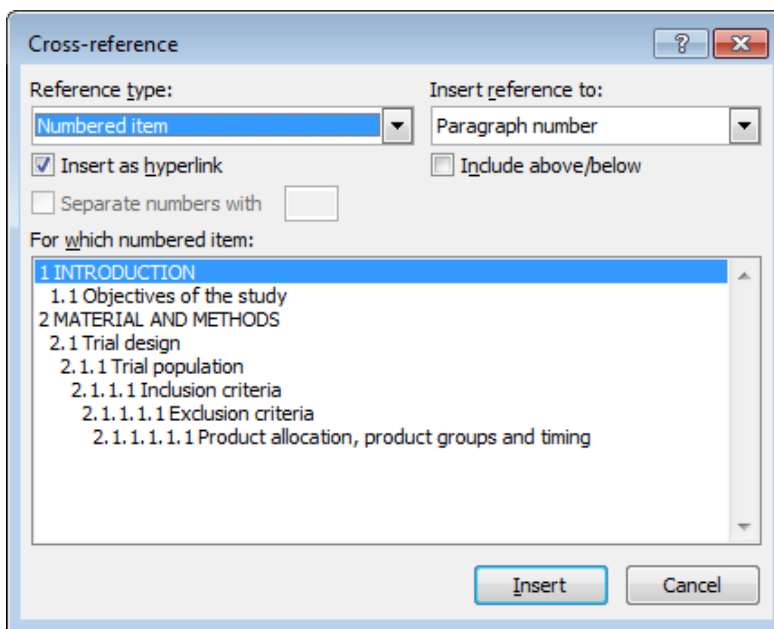
The purpose of the RHEADING is to specify header of the specific section and allows for creation of 6 sub-levels within a section. It also has option to bookmark title for future reference within document. Table 2 summarizes the RHEADING options.

**Table 2 Commonly Used RHEADING Options**

Option	Purpose
TEXT	Specifies header text.
HEADLVL	Specifies heading levels. Option to specify up to 6 levels.



**CROSSREF** Bookmark heading for future reference. Unique alpha numeric value with or without ‘\_’. Maximum 12 characters allowed. CROSSREF is same as Bookmark dialog box in MS Word.



## Sample call

```
%rheading(text=INTRODUCTION);
%rheading(headlvl=2,text=Objectives of the study);
%rheading(text=MATERIAL AND METHODS);
%rheading(headlvl=2,text=Trial design);
%rheading(headlvl=3,text=Trial population);
%rheading(headlvl=4,text=Inclusion criteria);
%rheading(headlvl=5,text=Exclusion criteria);
%rheading(headlvl=6,text=Product allocation$comma product groups and
timing);
```

## Result

<b>1</b>	<b>INTRODUCTION</b>
<b>1.1</b>	<b>Objectives of the study</b>
<b>2</b>	<b>MATERIAL AND METHODS LEVEL 1</b>
<b>2.1</b>	<b>Trial design LEVEL 2</b>
<b>2.1.1</b>	<b>Trial population LEVEL 3</b>
<b>2.1.1.1</b>	<b>Inclusion criteria LEVEL 4</b>
<b>2.1.1.1.1</b>	<i>Exclusion criteria LEVEL 5</i>
<b>2.1.1.1.1.1</b>	<i>Product allocation, product groups and timing LEVEL 6</i>

## %RPARAGRP

RPARAGRP permits the author to write actual text into the report. This is mostly used to describe the analysis being performed and/or to interpret the results that are produced. It has option to specify style, justification, font size and you also use RTF tags from macro URTFTAGS. If more than 2 blank spaces specified then macro keeps 1 space and removes other spaces. Table 3 summarizes the RPARAGRP options.

**Table 3 Commonly Used RPARAGRP Options**

Option	Purpose
TEXT	Specifies header text.
STYLE	Specifies paragraph style.
	Following options available
	<i>ITALIC</i>
	<b>BOLD</b>
	<b><i>BOLDITALIC</i></b>
JUST	Specifies paragraph justification/alignment.
	Following options available
	L=Left Justification
	R=Right Justification
	C=Center Justification
	J=Justified Justification

Option	Purpose
FONTSIZE	Changes size of the paragraph.

### Sample call

```
%rparagrp(text = %str(The primary outcome of interest was change in weight
from baseline at month 12.));
%rparagrp(text = %str($CROSSREF(tvs) provides the inferential statistics
results using Analysis of Covariance (ANCOVA) model analysis on change in
weight from baseline at month 12 for full analysis set.));
%rparagrp(text = %bquote($ITALIC(Change in weight from baseline = Baseline
weight + Stratification factor (Gender and Site) + Treatment)$NEWPARA
Analysis was performed using 2-sided test at $SALPHA.=0.05.$NEWPARA
The change from baseline in weight at month 12 was statistically
%sysfunc(ifc(&pvalue lt 0.05, significant,non-significant)) (LS Mean (SE) =
&estimate; 95% CI = &clm; p=&pvalue.));
```

### Result

The primary outcome of interest was change in weight from baseline at month 12.

Table 3 provides the inferential statistics results using Analysis of Covariance (ANCOVA) model analysis on change in weight from baseline at month 12 for full analysis set.

*Change in weight from baseline = Baseline weight + Stratification factor (Gender and Site) + Treatment*

Analysis was performed using 2-sided test at  $\alpha=0.05$ .

The change from baseline in weight at month 12 was statistically non-significant (LS Mean (SE) = 4.15 (3.37); 95% CI = -2.47, 10.76;  $p=0.2186$ ).

### %RTLFOUTP

The purpose of the RTLFOUTP is to inserts TLF created by ODS RTF without page breaks and removed header and footer tags generated by SAS®. It also accepts PNG and JPEG formats from R, SAS®, S-Plus etc. Table 4 summarizes the RTLFOUTP options.

**Table 4 Commonly Used RTLFOUTP Options**

Option	Purpose
DIRPATH	TLF location.
OUTFILNM	Output filename without extension.
GHEIGHT	Adjust figure height in inches. It automatically adjust if height is more than 8 inch. It is applicable to PNG or JPEG files.
GWIDTH	Adjust figure width in inches. It automatically adjust if width is more than 6.5 inch. It is applicable to PNG or JPEG files.
GARATIO	Adjust figure aspect ratio in percentage. It is applicable to PNG or JPEG files.

### Example

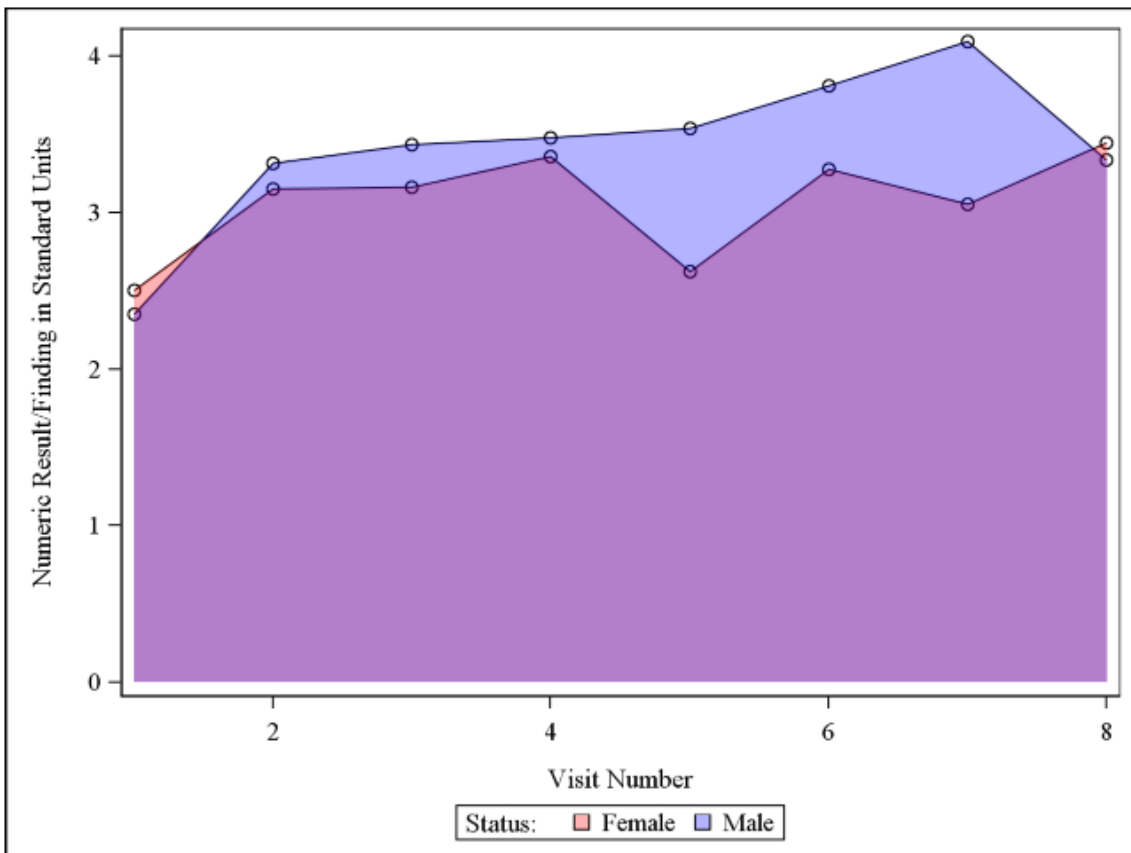
```
%rheading(headlvl=5, text=%str(Via Area Under the Curve));  
%rparagrph(text=%str(Similar to what was presented in the  
previous section, this AUC analysis represents  
the difference between the gender rather than  
the differences between term and pre-term infants));  
proc sgplot; /*code producing f2_fatAUCsex image*/ run;  
%rtlfoutp(OUTFILNM=f2_fatAUCsex);
```



#### 2.2.1.1.4 Via Area Under the Curve

Similar to what was presented in the previous section, this AUC analysis represents the difference between the gender rather than the differences between term and pre-term infants

**Area Under the Curve**



**Figure 7: Area Under the curve for Fat (split by sex)**

%RLISTITM

The purpose of the RLISTITM is to insert bullet and numbering into RTF. It has option to specify different numbering and bullet styles, sub-levels, nesting levels, indentation, font size and you also use RTF tags from macro URTFTAGS. If more than 2 blank spaces specified then macro keeps 1 space and removes other spaces. Table 5 summarizes the RLISTITM options.

**Table 5 Commonly Used RLISTITM Options**

Option	Purpose																				
TEXT	Specifies bullet and numbering text. Each bullet and number separated by SPLITCHR.																				
SPLITCHR	Specifies the split character.																				
STYLE	Specifies bullet and numbering style.																				
	Numbering style:																				
	ARABIC Arabic numeral (1, 2, 3)																				
	UROMAN Uppercase Roman numeral (I, II, III)																				
	LROMAN Lowercase Roman numeral (i, ii, iii)																				
	UALPHA Uppercase letter (A, B, C)																				
	LALPHA Lowercase letter (a, b, c)																				
	Bullet style:																				
	EMDASH Em dash (—)																				
	ENDASH En dash (–)																				
	TICK Tick (✓)																				
	SQUARE Square (▪)																				
	DOT Dot (●)																				
	CIRCLE Circle (○)																				
SSTYLE	Specifies numbering style.																				
0	Numbering style with period (.)																				
	If STYLE is																				
	<table border="1"> <thead> <tr> <th>ARABIC</th> <th>UROMAN</th> <th>LROMAN</th> <th>UALPHA</th> <th>LALPHA</th> </tr> </thead> <tbody> <tr> <td>1.</td> <td>I.</td> <td>i.</td> <td>A.</td> <td>a.</td> </tr> <tr> <td>2.</td> <td>II.</td> <td>ii.</td> <td>B.</td> <td>b.</td> </tr> <tr> <td>3.</td> <td>III.</td> <td>iii.</td> <td>C.</td> <td>c.</td> </tr> </tbody> </table>	ARABIC	UROMAN	LROMAN	UALPHA	LALPHA	1.	I.	i.	A.	a.	2.	II.	ii.	B.	b.	3.	III.	iii.	C.	c.
ARABIC	UROMAN	LROMAN	UALPHA	LALPHA																	
1.	I.	i.	A.	a.																	
2.	II.	ii.	B.	b.																	
3.	III.	iii.	C.	c.																	
1	Numbering style with close bracket																				
	If STYLE is																				
	<table border="1"> <thead> <tr> <th>ARABIC</th> <th>UROMAN</th> <th>LROMAN</th> <th>UALPHA</th> <th>LALPHA</th> </tr> </thead> <tbody> <tr> <td>1)</td> <td>I)</td> <td>i)</td> <td>A)</td> <td>a)</td> </tr> <tr> <td>2)</td> <td>II)</td> <td>ii)</td> <td>B)</td> <td>b)</td> </tr> <tr> <td>3)</td> <td>III)</td> <td>iii)</td> <td>C)</td> <td>c)</td> </tr> </tbody> </table>	ARABIC	UROMAN	LROMAN	UALPHA	LALPHA	1)	I)	i)	A)	a)	2)	II)	ii)	B)	b)	3)	III)	iii)	C)	c)
ARABIC	UROMAN	LROMAN	UALPHA	LALPHA																	
1)	I)	i)	A)	a)																	
2)	II)	ii)	B)	b)																	
3)	III)	iii)	C)	c)																	
2	Numbering style within parenthesis																				
	If STYLE is																				
	<table border="1"> <thead> <tr> <th>ARABIC</th> <th>UROMAN</th> <th>LROMAN</th> <th>UALPHA</th> <th>LALPHA</th> </tr> </thead> <tbody> <tr> <td>(1)</td> <td>(I)</td> <td>(i)</td> <td>(A)</td> <td>(a)</td> </tr> <tr> <td>(2)</td> <td>(II)</td> <td>(ii)</td> <td>(B)</td> <td>(b)</td> </tr> <tr> <td>(3)</td> <td>(III)</td> <td>(iii)</td> <td>(C)</td> <td>(c)</td> </tr> </tbody> </table>	ARABIC	UROMAN	LROMAN	UALPHA	LALPHA	(1)	(I)	(i)	(A)	(a)	(2)	(II)	(ii)	(B)	(b)	(3)	(III)	(iii)	(C)	(c)
ARABIC	UROMAN	LROMAN	UALPHA	LALPHA																	
(1)	(I)	(i)	(A)	(a)																	
(2)	(II)	(ii)	(B)	(b)																	
(3)	(III)	(iii)	(C)	(c)																	

Option	Purpose
NESTLVL	Specifies nesting bullet and numbering. Maximum two levels allowed. Only numbers allowed.
RESTART	Restart numbering?
FINDENT	Specifies first-line indent in cm.
INDENT	Specifies left indent in cm.
FONTSIZE	Changes size of the text.

### Sample call

```
%rlistitm(text=INTRODUCTION, style=arabic);
%rlistitm(text=Objectives of the study|
          Trial design|
          Trial population|
          Inclusion criteria,
          style=lroman, nestlvl=1);
%rlistitm(text=MATERIAL AND METHODS|RESULTS, style=arabic);
%rlistitm(text=Description of the product groups,
          style=lroman, nestlvl=1, restart=y);
%rlistitm(text=Disposition of subjects,
          style=lroman, nestlvl=1);
%rlistitm(text=Primary outcome, restart=y, style=arabic);
%rlistitm(text=Secondary outcomes);
```

### Result

1. INTRODUCTION
  - i. Objectives of the study
  - ii. Trial design
  - iii. Trial population
  - iv. Inclusion criteria
2. MATERIAL AND METHODS
3. RESULTS
  - i. Description of the product groups
  - ii. Disposition of subjects
  1. Primary outcome
  2. Secondary outcomes

## CONCLUSION

The developed tool has helped the biostatistics team of Nestlé in 3 axes:

**Efficiency** – By programming the writing of the statistical report, the generation of figures, tables and cross references are automated. In case of data updates, it would suffice to re-run the SAS® programs in order to produce an updated report. No time is wasted with copy and pasting.

**Quality** – The standardization/automation of the routine tasks allows the statisticians to focus on important aspects of statistical analysis and interpretation of results which increases the quality of the clinical study reporting. No copy-pasting reduces manually made mistakes in the writing of the report.

**Reproducibility** – Since there is a single source file that integrates programming and writing, the tool provides complete transparency and traceability. Other readers can easily re-run the analyses described, as it is self-documented.

Compared to R and LATEX system that required knowledge in both programming languages, this new reporting system only requires that the user knows how to program in SAS®. That being said, the potential



users would still be limited to statisticians and programmers. This is one of the reasons why a MS Word file was chosen to be the final result, so that other members of the clinical study team can give their input in the final document.

Nestlé's biostatistics team is looking for other applications of the system which would benefit from its reproducible research approach such as automating a report for regular review of eCRF data.

This technique has been successfully implemented on the PC SAS® and SAS® Life Science Analytics Framework platforms.

It should be explored to include mathematical equation features.

## REFERENCES

Knuth DE (1984). "Literate Programming." The Computer Journal, 27(2), 97–111.

Word 2007: Rich Text Format (RTF) Specification, version 1.9. <https://www.microsoft.com>.

## ACKNOWLEDGMENTS

The authors are thankful to Maria Laura Gosoni, and Stephane Collet at Nestlé Research Center and Harshad Kulkarni at Cognizant Technology Solutions for their encouragement and support.

## CONTACT INFORMATION

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

Carlos Antonio De Castro  
Nestlé Research Center  
[CarlosAntonio.DeCastro@rdsg.nestle.com](mailto:CarlosAntonio.DeCastro@rdsg.nestle.com)

Julien Sauser  
Nestlé Research Center  
[Julien.Sauser@rdls.nestle.com](mailto:Julien.Sauser@rdls.nestle.com)

Rajaram Venkatesan  
Cognizant Technology Solutions  
[Rajaram.Venkatesan@cognizant.com](mailto:Rajaram.Venkatesan@cognizant.com)

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

Other brand and product names are trademarks of their respective companies.

## APPENDIX

```

*** RTF Open;
%rtfopen;

*** Title Page;
%rtitlepg( doctitle=Statistical Report,
           trialno=ABC123,
           npdicode=ABC00871780,
           trialtit=%str(Open Label, Multi-center to Assess Damage Done by
McDonalds and KFC) ,
           shorttit=,
           author=Rajaram Venkatesan,
           docdate=,
           version=Draft\0.1);

*** Signature Page;
%rsqnt:rgp(sqntitle=BIOSTATISTICIAN|REVIEWER,
           nametitl=Rajaram Venkatesan$newline Biostatistician|Carlos
Antonio De Castro$newline Biostatistician,

           emailids=rajaram.venkatesan@cognizant.com|carlosantonio.decastro@rdsq.nestl
e.com);

*** Acknowledgements Page;
%racknwdg(acknwdgs=Wendy|Naseer|Adam Smith);

*** Summary Page;
%rsumrypg;

*** TOC Page;
%rtocpage;

*** Level 1 Header;
%rheading(text=INTRODUCTION);

*** Paragraphs;
%rparagr(text=%str(This document is for Statistics Report of ABC123 study.
This could not be approved by any health authority.
This was conducted purely for recreational purpose.));

*** Level 2 Header;
%rheading(headlvl=2,text=Objectives of the study);
%rparagr(text=To assess the impact of McDonalds and KFC product
consumption on daily basis);

*** Level 1 Header;
%rheading(text=MATERIAL AND METHODS);
*** Level 2 Header;
%rheading(headlvl=2,text=Trial design);
*** Paragraphs;
%rparagr(text=%str(This trial is open label, two-treatment clinical
trial.));
*** Level 3 Header;
%rheading(headlvl=3,text=Trial population);

%rheading(headlvl=3,text=Inclusion criteria);

```

```

%rheading(headlvl=3,text=Disposition);

*** Paragraph with cross reference;
%rparagr(text=%str(The disposition of the patients in this study is
                    summarized according to completion or withdrawal from
                    the study at data cut-off (DOC) in the following
                    table));

%rparagr(text=%str(A total of 400 patients were randomized in this study,
                    200 patients in the McDonalds group and 200 patients
                    in the KFC group.);

%rparagr(text=%bquote (Almost(77%) patients died on this study. Only 3.5%
                    of patients randomized to McDonalds, and 1% of
                    patients randomized to KFC were continued the study.
                    Other reasons for study discontinuation included
                    informed consent withdrawn and loss to follow up.);

*** Include table which was created already;
%rtlfoutp(outfilnm=tds);

%rheading(headlvl=3,text=Demographic and other patient characteristics);

%rparagr(text=%bquote (The demographic and key baseline disease
                    characteristics of study patients are summarized
                    in the following table. Demographic and baseline
                    characteristics were well balanced between the two
                    treatment groups. The majority of patients were
                    female (63%) and the median age was 62 years.);

*** Include table which was created already;
%rtlfoutp(outfilnm=tdm);

%rheading(text = EFFICACY RESULTS);

data tvs;
  set adam.adsl;
  if ranuni(20160304) le 0.5 then wght12=blwght-round(ranuni(28474)*10);
  else wght12=blwght+round(ranuni(28474)*100);
  chgwght=wght12-blwght;
run;

*** Calculate ANCOVA summary;
%astattst(indsnm=tvs,
          outdsnm=tvrsrslt,
          displbl= null ,
          anlyzvr=chgwght,
          classvr=armno,
          outfrp=full-all,
          stattst=ancova-sex site- -blwght);

data null ;
  set tvrsrslt;
  where statord=3;
  call symputx ('pvalue',pvalue);
  call symputx ('clm',clm);

```

```

        call symputx('estimate', estimate);
run;

proc sql noprint;
    select count(armno) into:col0-:col1 from adam.adsl where full="Y" group
    by armno;
quit;

*** Creating formats for report;
proc format;
    value $strtn '0'="McDonald (N=&col0)"
                '1'="KFC (N=&col1)"
                '0 vs 1'="McDonald vs KFC";
run;

*** Report output dataset;
%rodsopen(uniqueid=itpvs,
          outfilnm=&qoutpath\TLF\tvs.rtf);
%rprintit( indsnm=tvrsrslt,
           sortvrs=varord stator,
           othervrs=statlbl\$strtn.\|l\15|nvalue\|c\5|
                estimate\|c\10|clm\|c\10|pvalue\|c\5,
           cheadlbl="Randomized treatment"~"n"~"LS Mean (SE)"~
                "95% CI"~"P-value");

%rodsclos;

%rheading(headlvl=2, text=Primary analysis);

%rparagr(text = %str(The primary outcome of interest was change in weight
                    from baseline at month 12.$newpara
                    $CROSSREF(tv) provides the inferential statistics
                    results using Analysis of Covariance (ANCOVA) model
                    analysis on change in weight from baseline at month
                    12 for full analysis set.));

%rparagr(text = %bquote($ITALIC(Change in weight from baseline = Baseline
                        weight + Stratification factor
                        (Gender and Site) + Treatment)$NEWPARA

                        Analysis was performed using 2-sided test at
                        $SALPHA.=0.05.$NEWPARA

                        The change from baseline in weight at month 12 was
                        statistically %sysfunc(ifc(&pvalue lt 0.05,
                        significant, non-significant))
                        (LS Mean (SE) = &estimate; 95% CI = &clm;
                        p=&pvalue).));

%rtloutp(outfilnm=tv);

%rheading(headlvl=2, text=Overall survival);

%rparagr(text=%bquote(Overall survival was similar for both treatments (HR
                    1.00, 95% CI 0.80 to 1.25))));

*** Include table which was created already;

```

```
%rtlfoutp(outfilnm=tos);

%rheading(text = SAFETY RESULTS);

*** Level 2 Header;
%rheading(headlvl=2,text=All adverse events);

%rparagrp(text=%bquote (The most common AEs with McDonald, were diarrhoea,
rash, nausea, anorexia and dermatitis acneiform. The
most common AEs with KFC were diarrhoea, rash and
nausea.););

*** Include table which was created already;
%rtlfoutp(outfilnm=tae);

*** Level 1 Header;
%rheading(text=Conclusion);

*** RTF Close;
%rtfclos;
```