
The Automated Metadata-driven Table Generation Process (TFLGen) at Amgen Inc.

Neil Lyon

Amgen Inc., Uxbridge, UK

The Problem

Existing libraries of standard reporting code are 10+ years old

- They are difficult to use and maintain
- Requirements have been layered on
- Redundant code as functionality has been replaced
- AE tiered table macro has 9554 lines of code, a 60 page user manual, up to 76 input parameters with a log >100,000 lines!

Code is duplicated across programs leading to:

- Duplicated maintenance
- Inconsistency when a change is made in one program but not others

Bug fixes & enhancements are time consuming:

- Teams unnecessarily developing their own solutions

Staff can be hard to move between teams

The Solution

A suite of macros that have a single specific function

- Macros are combined and built upon to form modular macros that can be used to build any table

New problem...

- Lots of new macros that need to be understood

Requirements for a better solution

- Could we simplify the use of these new macros?
- Flexible enough to produce most tables?
- Maximize the re-use of information, reducing duplication of effort?
- Something intuitive?
- Maybe a nice WYSIWYG interface?
- Fully validated to reduce burden of QC?

The Anatomy of a Table

Table 14-2.1.1. Baseline Demographics
(Primary Analysis Set)

	Comparator		AMG-Active	
	10 mg (N = 21)	20 mg (N = 25)	5 mg (N = 24)	10 mg (N = 24)
Sex - n(%)				
Male	13 (61.9)	12 (48.0)	10 (41.7)	13 (54.2)
Female	8 (38.1)	13 (52.0)	14 (58.3)	11 (45.8)
Age (years)				
n	21	25	24	24
Mean (SD)	55.2 (22.5)	53.6 (16.8)	56.1 (19.0)	52.5 (20.4)
Median	56.0	53.0	55.5	50.0
Q1, Q3	37.0, 69.0	44.0, 63.0	43.0, 71.5	40.0, 65.5
Min, Max	21, 88	20, 87	19, 87	21, 90
Age - n (%)				
< 65 years	14 (66.7)	20 (80.0)	15 (62.5)	18 (75.0)
≥ 65 years	7 (33.3)	5 (20.0)	9 (37.5)	6 (25.0)
Age by sex ¹				
Male subjects - n (%)				
< 65 years	8 (61.5)	11 (91.7)	6 (60.0)	10 (76.9)
≥ 65 years	5 (38.5)	1 (8.3)	4 (40.0)	3 (23.1)
Female subjects - n (%)				
< 65 years	6 (75.0)	9 (69.2)	9 (64.3)	8 (72.7)
≥ 65 years	2 (25.0)	4 (30.8)	5 (35.7)	3 (27.3)

N = number of subjects randomized and dosed in the safety analysis set.
¹ Percentages based on number of subjects of each sex.

Program: /home/nlyon/tflgen_training/exercise/tables/program/-dm-sum-pas.sas
Output: t14-02-001-001-dm-sum-pas.rtf (Date Generated: 25AUG15:07:45:42) Source: adam.adae

Titles
Footnotes
File names
Font size
Page orientation

Population and subsets
Treatment column headers
File locations

Statistics
Decimal precision
Labels
By-processing
Blank rows

Output
Specific
Analysis
Level
Table
Shell

Table Shell Information

As metadata in Excel

METADATA	T1	TREATMENTS
label	Sex - n (%)	
statistic	Male	x (x.x)
statistic	Female	x (x.x)
blank_row		
label	Age (years)	
statistic	n	x
statistic	Mean (SD)	x.x (x.x)
statistic	Median	x.x
statistic	Q1, Q3	x.x, x.x
statistic	Min, Max	x, x
blank_row		
label	Age group - n (%)	
statistic	< 65 years	x (x.x)
statistic	≥ 65 years	x (x.x)
statistic	≥ 75 years	x (x.x)

Baseline Demographics

[Headers] (N = xx)	
Sex - n (%)	
Male	x (x.x)
Female	x (x.x)
Age (years)	
n	x
Mean (SD)	x.x (x.x)
Median	x.x
Q1, Q3	x.x, x.x
Min, Max	x, x
Age group - n (%)	
< 65 years	x (x.x)
≥ 65 years	x (x.x)
≥ 75 years	x (x.x)

Tiered Table Template

METADATA	T1	TREATMENTS	STATISTIC	INDENT	DATA_SET	VARIABLE	SORT_SECTION	SORT_DIRECTION
sps_generate_module	table_tiered_001							
section_wrap_flag								
col_width	AUTO							
col_align								
col_label	System Organ Class Preferred Term	[Headers]						
col_big_n_label		(N = xx)						
col_n_pct_label		n (%)						
blank_row								
level_0	Subjects reporting an adverse event	x (x.x)	percent	0				
blank_row								
level_1	System Organ Class	x (x.x)	percent	0	adam.adad	aebodsys	1	descending
level_2	Preferred Term	x (x.x)	percent	1	adam.adad	aeodecod	2	descending

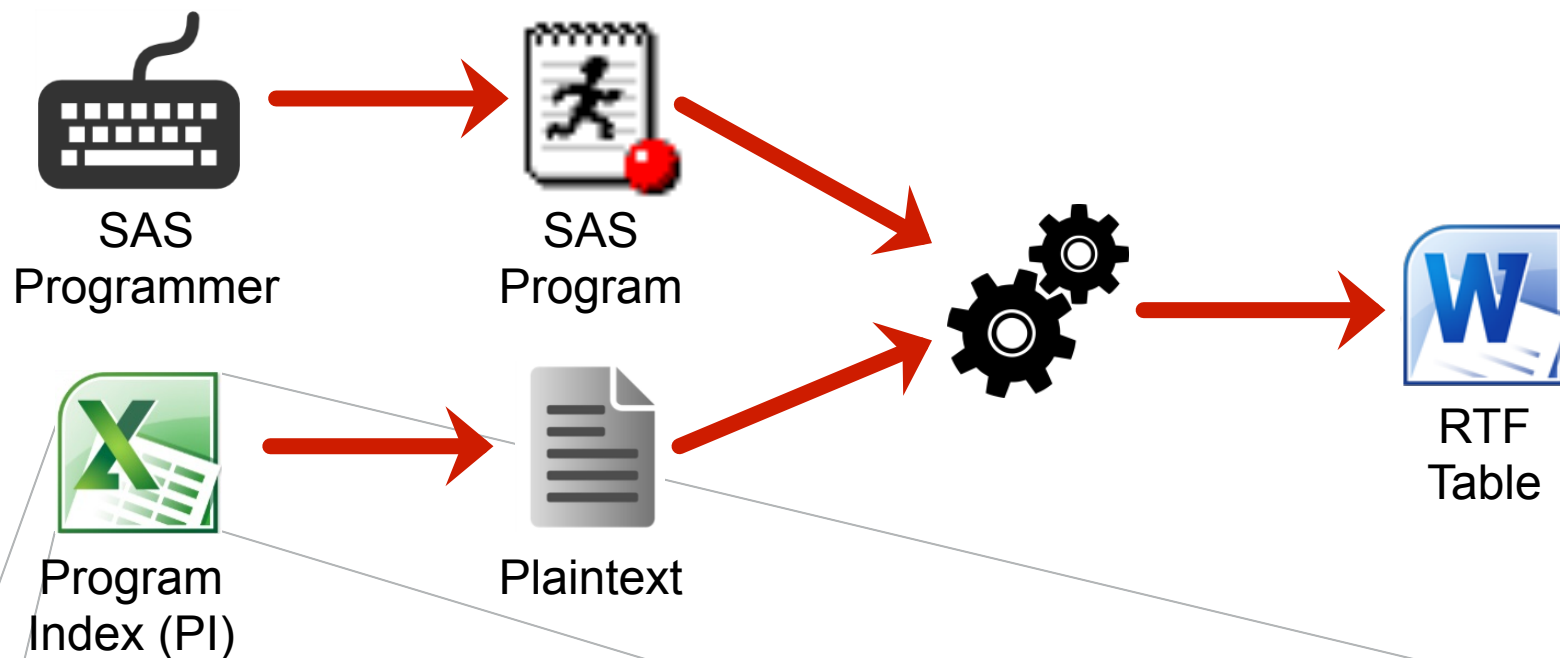
- **METADATA** = Keyword for metadata stored in row
- **T1** = Metadata values + leftmost table column labels
- **TREATMENTS** = Treatment column labels + decimal precision metadata
- **STATISTIC** = Keywords for statistics
- **INDENT** = Indentation for leftmost table column
- **DATA_SET** = Name of dataset containing variable to derive statistics from
- **VARIABLE** = Name of variable to derive statistics from
- **SORT_SECTION/SORT_DIRECTION** = Sorting metadata

Base Table Template

METADATA	T1	TREATMENTS	STATISTIC	INDENT	DATA_SET	VARIABLE	QUANT	WHERE_CLAUSE	DENOMINATOR	BY_DISPLAY_VARIABLE
sps_generate_module	table_base_001									
section_wrap_flag	N									
col_width	AUTO									
col_align	L									
col_label		[Headers]								
col_big_n_label		(N = xx)								
col_n_pct_label										
blank_row										
label	Sex - n(%)			0						
statistic	Male	x (x.x)	n percent	1	adam.adsl	sexn	1			
statistic	Female	x (x.x)	n percent	1	adam.adsl	sexn	2			
blank_row										
label	Age (years)			0						
statistic	n	x	n	1	adam.adsl	age				
statistic	Mean (SD)	x.x (x.x)	mean stdev	1	adam.adsl	age				
statistic	Median	x.x	median	1	adam.adsl	age				
statistic	Q1, Q3	x.x, x.x	q1 q3	1	adam.adsl	age				
statistic	Min, Max	x, x	min max	1	adam.adsl	age				
blank_row										
label	Age - n (%)			0						
statistic	< 65 years	x (x.x)	n percent	1	adam.adsl			.<age<65		
statistic	≥ 65 years	x (x.x)	n percent	1	adam.adsl			age>=65		
blank_row										
label	Age by sex ¹			0						
by_start	!BY_DISPLAY_VARIABLE! subjects - n (%)			1	adam.adsl	sexn				sex
statistic	< 65 years	x (x.x)	n percent	2	adam.adsl			.<age<65	sexn>.	
statistic	≥ 65 years	x (x.x)	n percent	2	adam.adsl			age>=65	sexn>.	
by_end										

- QUANT** = Quantifying value for categorical statistic
- WHERE_CLAUSE** = SAS clauses for categorical statistics or to limit data
- DENOMINATOR** = SAS clause for alternative denominators
- BY_DISPLAY_VARIABLE** = Decode variable when by-processing

Pre-TFLGen Process Flow



Output Number	Titles	Footnotes	Output Name	Program Name
14-2.1.1	Baseline Demographics{\line}{Primary Analysis Set)	N = number of subjects randomized and dosed in the safety analysis set.{\line} ¹ Percentages based on number of subjects of each sex.	t14-02-001-001-dm-sum-pas.rtf	t-dm-sum-pas.sas
14-6.2.1	Treatment Emergent Adverse Events by System Organ Class and Preferred Term{\line}{Safety Analysis Set)	N = number of subjects randomized and dosed in the safety analysis set.{\line}Coded using MedDRA version &meddrav..	t14-06-002-001-ae-soc-pref-saf.rtf	t-ae-soc-pref-saf.sas

Analysis Level: Populations/Subsets

Populations

Population	Where Clause
(Safety Analysis Set)	saffl = 'Y'
(Primary Analysis Set)	pasfl = 'Y'
(All Randomized Subjects)	n(randdt)

Subsets of Populations

Subset	Where Clause
Subjects with Anemia at Baseline	< blhgb < 10
Male subjects Only	sexn = 1
Female Subjects Only	sexn = 2

- **Population/Subset** = Names of populations and subsets
- **Where Clause** = SAS expressions to describe them

Analysis Level: Treatment Columns

Treatment Column Headers

Header	Span Labels	Treatment	Column Label	Data Set	Variable	Quant	Where Clause	Column Width	Column Alignment
safety		1	Comparator	adam.adsl			trt01an in(1,2)	1	C
safety		2	AMG-Active	adam.adsl			trt01an in(3,4)	1	C
safety		3	Total	adam.adsl			trt01an in(1,2,3,4)	1	C
primary	Comparator	1	10 mg	adam.adsl	trt01an	1		1	C
primary	Comparator	2	20 mg	adam.adsl	trt01an	2		1	C
primary	AMG-Active	3	5 mg	adam.adsl	trt01an	3		1	C
primary	AMG-Active	4	10 mg	adam.adsl	trt01an	4		1	C

- **Header** = Name of header
- **Span Labels** = Labels above multiple treatment group columns
- **Treatment** = Number showing treatment group column order
- **Column Label** = Label for each treatment group column
- **Data Set** = Containing data that describes each treatment group
- **Variable/Quant** = Describes each treatment group
- **Where Clause** = Describes treatment group using SAS expressions
- **Column Width** = Width in inches
- **Column Alignment** = Centre, left or right

Analysis Level: File Locations

Table Template Locations

Template Location	UNIX Directory
therapeutic_area	/home/therapeutic_area/docs/templates
product	/home/therapeutic_area/product/docs/templates
protocol	/home/therapeutic_area/product/protocol/docs/templates

- **Template Location** = Name given to template locations
- **UNIX Directory** = UNIX paths for each specified location

General Information

Item	Description	Value
table_program_dir	File location for table programs	/home/therapeutic_area/product/protocol/analysis/tables/program
table_metadata_dir	File location for table metadata sets	/home/therapeutic_area/product/protocol/analysis/tables/metadata
table_bpdf_dir	File location for table BPDFs	/home/therapeutic_area/product/protocol/analysis/tables/validation
table_output_dir	File location for table outputs	/home/therapeutic_area/product/protocol/analysis/tables/output
subject_var	Name of the subject/patient variable	usubjid
custom_init_macro_dir1	First location searched for custom macros	/home/therapeutic_area/product/protocol/analysis/utilities
custom_init_macro_dir2	Second location searched for custom macros	/home/therapeutic_area/product/utilities
adam_dir	File location for ADaM data sets	/home/therapeutic_area/product/protocol/analysis/statdata/adam

- **Item** = Keyword for metadata stored in the row
- **Description** = To let users know what the keywords refers to
- **Value** = Value of the metadata

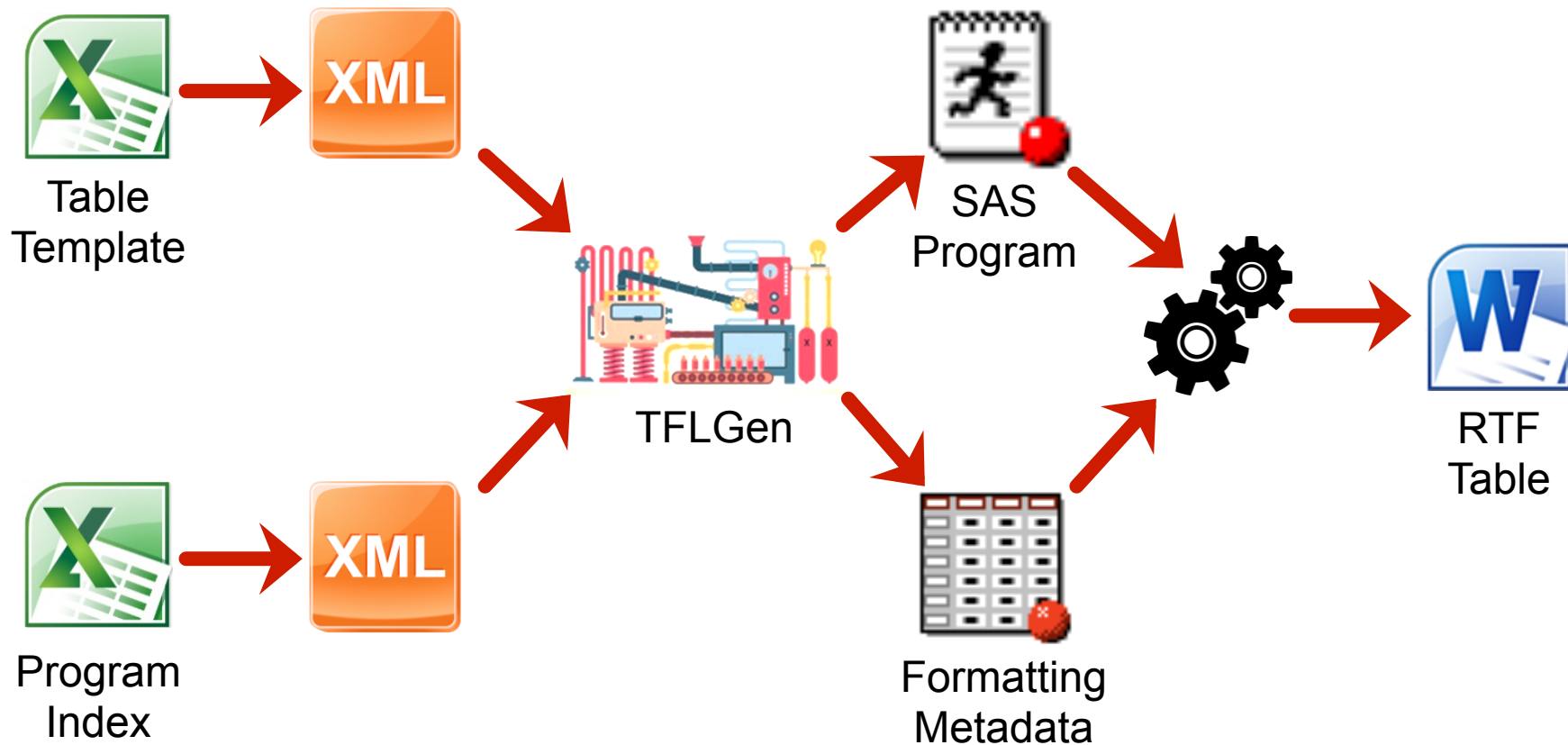
Extending the PI: Linking the Metadata

Linking TFLGen metadata with output level metadata

Population	Subset	Numerator Where	Header	Template Location	Template	Sort Section	Sort Source Treatment	Body Font Height	Orientation
(Safety Analysis Set)	Female Subjects Only	trtemfl = 'Y'	safety	therapeutic_area	t-ae-pt	1	3 2 1	10	Portrait
									Landscape Portrait

- **Population** = Drop-down list created from populations worksheet
- **Subset** = Drop-down list created from subsets worksheet
- **Numerator Where** = Limits all data used for statistic generation
- **Header** = Drop-down list created from headers worksheet
- **Template Location** = Drop-down list from template location worksheet
- **Template** = Name of the template required for the table
- **Sort Section** = Describes table sorting based on numbers
- **Sort Source Treatment** = Describes table sorting based on numbers
- **Body Font Height** = Font size to use in main body of table
- **Orientation** = Landscape or portrait

TFLGen Process Flow



TFLGen Report

TFLGen Metadata Generation Report

PI: /home/nlyon/tflgen_training/exercise/docs/pi/PHUSE_PI_working.xlsm

Mode: CREATE Date: 26-Aug-2015 01:24:20 Requested Template: ALL

Requested TFLs: ALL

TFL Details

(14-6.2.1) t-ae-soc-pref-saf.rtf (Allow Code Generation: Yes)

Template: /home/nlyon/tflgen_training/exercise/docs/templates/phuse-ae-real.xlsm Location Name: protocol

NOTICE: Metadata data set does not currently exist.

NOTICE: Metadata data set requires generation.

NOTICE: SAS program requires generation.

NOTICE: Metadata data set has been generated.

NOTICE: SAS program has been generated.

(14-2.1.1) t-dm-sum-pas.rtf (Allow Code Generation: Yes)

Template: /home/nlyon/tflgen_training/exercise/docs/templates/phuse-dm-real.xlsm Location Name: protocol

ERROR: File ADAM.ADXL.DATA does not exist.

ERROR: [Template, Row 31] DENOMINATOR (sexn = 1) is not valid when applied to the DATA_SET adam.adxl

ERROR: [Template, Row 20] VARIABLE age does not exist in the specified DATA_SET adam.adxl

ERROR: [Template, Row 23] STATISTIC 'mn max' contains the invalid value 'mn'

ERROR: [Template, Row 31] DATA_SET adam.adxl does not exist in the specified library

ERROR: Previous errors prevent metadata data set generation. Please fix these errors and rerun TFLGen.

ERROR: Previous errors prevent program generation. Please fix these errors and rerun TFLGen.

Report created on 26-Aug-2015 01:24:20 by TFLGen version: 01.02.00

Problem 1: Too Many Templates

Too many near-identical table templates being generated

- Changeable parameters similar to SAS macro variables
- New worksheet on Program Index

METADATA	T1	TREATMENTS	STATISTIC	INDENT	DATA_SET	VARIABLE	SORT_SECTION	SORT_DIRECTION
col_label	System Organ Class{\line} Preferred Term	[Headers]						
col_big_n_label		(N = xx)						
col_n_pct_label		n (%)						
blank_row								
level_0	Subjects experiencing &level_0_text	x (x.x)	n percent	0				
blank_row								
level_1	System Organ Class	x (x.x)	n percent	0	adam.adae	aebodsys	1	descending
level_2	Preferred Term	x (x.x)	n percent	1	adam.adae	aedecod	2	descending

Base Output Name	Parameter	Value	Description
t-ae-soc-pref-fas.rtf	level_0_text	an adverse event	Text for top row of AE table
t-ae-soc-pref-ser-saf.rtf	level_0_text	a serious adverse event	Text for top row of AE table
t-ae-soc-pref-disc-saf.rtf	level_0_text	an adverse event leading to discontinuation	Text for top row of AE table

- **Base Output Name** = Output that parameter to be resolved in
- **Parameter** = Name of parameter to be resolved
- **Value** = String to replace parameter with
- **Description** = To aid recall

Problems 2: TFLGen Too Rigid

Users limited to using data sets directly from a library

- Allow users to enter custom code that runs at start of generated program
- Create work data sets for use within generated program
- Create macro variables that can resolve when program runs

Base Output Name	Custom Code to Include at Start of Program
t-dm-sum-pas.rtf	<pre>*--- Derive BMI ---; data bmi; set adam.adsl; if n(wgt,hgt) = 2 then bmi = wgt / (hgt**2); run;</pre>

- **Base Output Name** = Output that custom code required for
- **Custom Code to Include at Start of Program** = Custom code

Problem 3: Simplifying QC

How to QC programs generated by TFLGen?

- Black text shows data that is displayed
- Red text shows underlying metadata
- Blue highlighting shows changes to metadata since last created

Table 14-2.1.1. Baseline Demographics
(Primary Analysis Set)
Population: pasfl = 'Y'

	Comparator		AMG-Active	
	10 mg adam.adsl. (trt01an=1) (N = xx)	20 mg adam.adsl. (trt01an=2) (N = xx)	5 mg adam.adsl. (trt01an=3) (N = xx)	10 mg adam.adsl. (trt01an=4) (N = xx)
drop_zero_row_flag:N subject_var:usubjid				
Sex - n(%)				
Male	x (x.x)	x (x.x)	x (x.x)	x (x.x)
adam.adsl.(sexn=1)	n percent			
Female	x (x.x)	x (x.x)	x (x.x)	x (x.x)
adam.adsl.(sexn=2)	n percent			
Age (years)				
n	x	x	x	x
adam.adsl.age	n			
Mean (SD)	x.x (x.x)	x.x (x.x)	x.x (x.x)	x.x (x.x)
adam.adsl.age	mean stdev			
Median	x.x	x.x	x.x	x.x
adam.adsl.age	median			
Q1, Q3	x.x, x.x	x.x, x.x	x.x, x.x	x.x, x.x
adam.adsl.age	q1 q3			
Min, Max	x, x	x, x	x, x	x, x
adam.adsl.age	min max			
Age - n (%)				
< 65 years	x (x.x)	x (x.x)	x (x.x)	x (x.x)
adam.adsl.(. <age<65)	n percent			
≥ 65 years	x (x.x)	x (x.x)	x (x.x)	x (x.x)
adam.adsl.(age>=65)	n percent			
Age by sex ¹				
!adam.adsl.sexnl subjects - n (%)				
< 65 years	x (x.x)	x (x.x)	x (x.x)	x (x.x)
adam.adsl.(. <age<65)	n percent(sexn>.)			
≥ 65 years	x (x.x)	x (x.x)	x (x.x)	x (x.x)
adam.adsl.(age>=65)	n percent(sexn>.)			

Page 1 of 1

N = number of subjects randomized and dosed in the safety analysis set.

¹ Percentages based on number of subjects of each sex.

Program: /t-dm-sum-pas.sas

Output: t_dm_sum_pas.rtf (Date Generated: 25AUG15:07:44:30) Source: adam.adsl



Problem 4: Programming and Testing

Developing many interconnected macros

- Hard for multiple developers to work on modules
 - Decisions have ramifications for other developers' work
- One programmer oversaw the primary development
 - Areas of commonality identified and distributed amongst team
- Avoid tramping macro parameters
 - Data sets were used to pass information across macros

Testing TFLGen for robustness, reliability and validity

- It must run without problems, generating the expected numbers every time
- Every module is unit tested to ensure that it follows requirements
- Once complete, integration testing was performed
 - Program Index and table templates to test all functions
 - The results were dual-programmed for validity
- Limited production roll-out

Example: Demography

Analysis level metadata (header)

Header	Span Labels	Treatment	Column Label	Data Set	Variable	Quant	Where Clause	Column Width	Column Alignment
primary	Comparator	1	10 mg	adam.adsl	trt01an	1		1	C
primary	Comparator	2	20 mg	adam.adsl	trt01an	2		1	C
primary	AMG-Active	3	5 mg	adam.adsl	trt01an	3		1	C
primary	AMG-Active	4	10 mg	adam.adsl	trt01an	4		1	C

Table shell metadata

[illegible]

Example: Demography

**Table 14-2.1.1. Baseline Demographics
(Primary Analysis Set)**

	Comparator		AMG-Active	
	10 mg (N = 21)	20 mg (N = 25)	5 mg (N = 24)	10 mg (N = 24)
Sex - n(%)				
Male	13 (61.9)	12 (48.0)	10 (41.7)	13 (54.2)
Female	8 (38.1)	13 (52.0)	14 (58.3)	11 (45.8)
Age (years)				
n	21	25	24	24
Mean (SD)	55.2 (22.5)	53.6 (16.8)	56.1 (19.0)	52.5 (20.4)
Median	56.0	53.0	55.5	50.0
Q1, Q3	37.0, 69.0	44.0, 63.0	43.0, 71.5	40.0, 65.5
Min, Max	21, 88	20, 87	19, 87	21, 90
Age - n (%)				
< 65 years	14 (66.7)	20 (80.0)	15 (62.5)	18 (75.0)
≥ 65 years	7 (33.3)	5 (20.0)	9 (37.5)	6 (25.0)
Age by sex ¹				
Male subjects - n (%)				
< 65 years	8 (61.5)	11 (91.7)	6 (60.0)	10 (76.9)
≥ 65 years	5 (38.5)	1 (8.3)	4 (40.0)	3 (23.1)
Female subjects - n (%)				
< 65 years	6 (75.0)	9 (69.2)	9 (64.3)	8 (72.7)
≥ 65 years	2 (25.0)	4 (30.8)	5 (35.7)	3 (27.3)

Page 1 of 1

N = number of subjects randomized and dosed in the safety analysis set.

¹ Percentages based on number of subjects of each sex.

Program: /home/nlyon/tfigen_training/exercise/tables/program/t-dm-sum-pas.sas

Output: t14-02-001-001-dm-sum-pas.rtf (Date Generated: 25AUG15:07:45:42) Source: adam.adae

Conclusion

TFLGen increases consistency, maintainability and reusability

Easily extensible through modularization

Table templates can be re-used across analyses

QC requirements have been minimized

Increases resource portability between teams

TFLGen is expected to produce > 65% of tables

- Single programmer
- 462 out 697 (66%) tables
- only 3 days
- 37 templates

Feedback for TFLGen has been overwhelmingly positive

- Teams reporting quicker table production and QC

Q&A



I would like to thank my colleagues Neville Cope, Jack Fuller, Benno Kurch and Chris St Peter for their invaluable help in designing and coding TFLGen. I would also like to thank David Edwards and Mark Stetz for their patience and support in bringing TFLGen to fruition.