



How Python or R Support/Facilitate the Validation Process of TFLs for a Study Development Using Hybrid SAS and R

Linshan Yuan (LYuan7@ITS.JNJ.com)

Weiwei Xu (wxu52@ITS.JNJ.com)

08 Mar 2023

Sebastian Ferreira, *Untitled 1*

Artwork from the National Art Exhibitions of the
Mentally Ill, Inc (NAEMI).

Agenda

- Background & Goals
- Functions development with R implementation
- Application 1: rtf to rtf comparison (regardless language system)
- Application 2: populations counts from all tables in deliverables
- Application 3: counts (at timepoints) among related tables
- Conclusions

Backgrounds & Goals

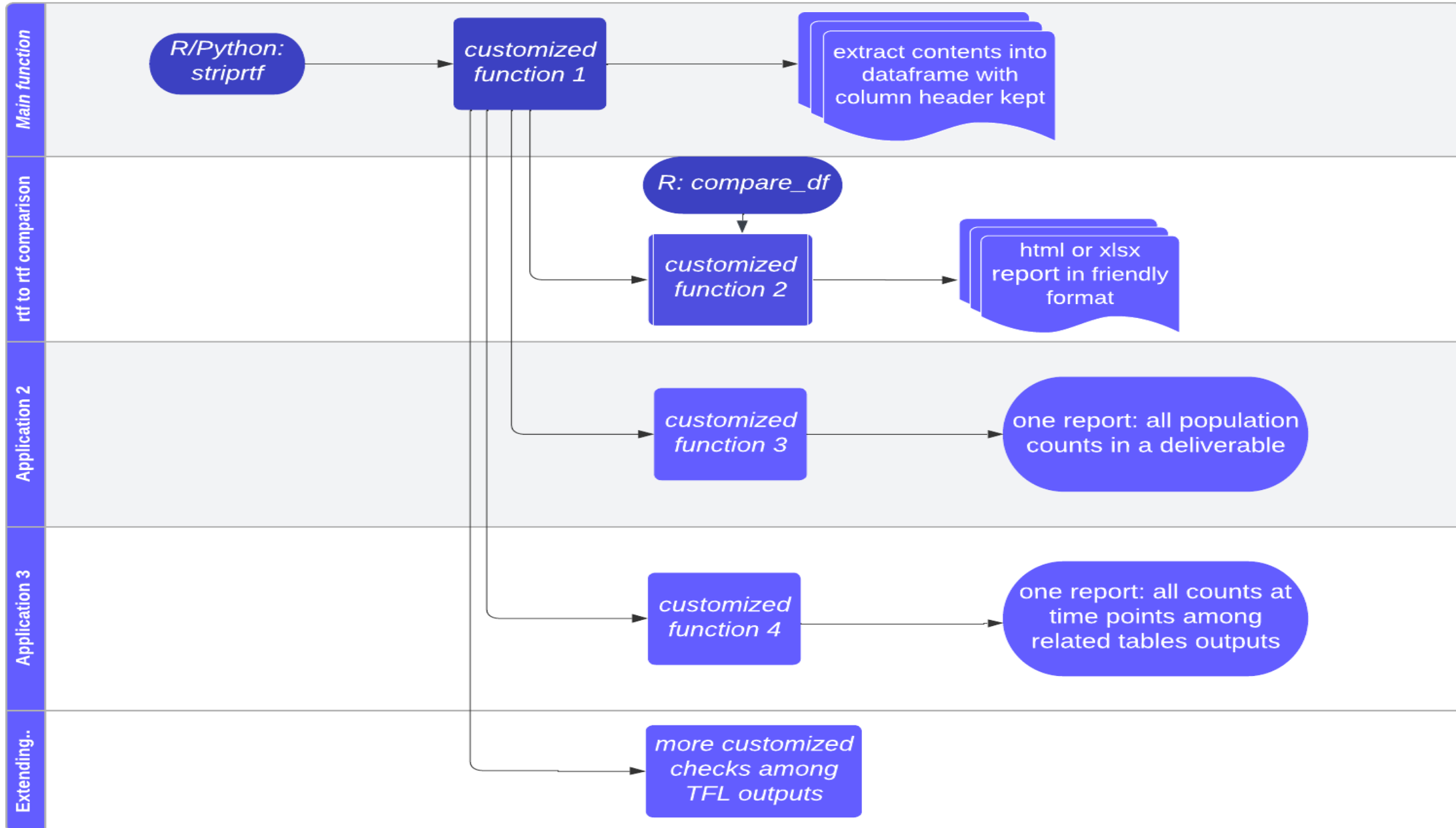
- Background

- ✓ Looking for comparison tool to compare R rtf outputs against SAS rtf outputs
- ✓ Lacking a method to track and monitor the overall quality of deliverable: values checking

- Goals

- ✓ Utilizing R/Python packages to extract high level contents of rtf regardless language system
- ✓ Functions customized to tackle various outputs with any layout
- ✓ Functions defined to convert column headers into column names from different table layouts
- ✓ Functions developed to extract population counts or counts at timepoints among related outputs
- ✓ Functions extendable for specific checks with user input

Functions development with R/Python implementation



Application 1: rtf to rtf comparison

- Case 1: single file comparison between SAS generated outputs
- Case 2: single file comparison between SAS output and R output
- Case 3: folder to folder comparison

Production Side (SAS) vs QC (SAS)

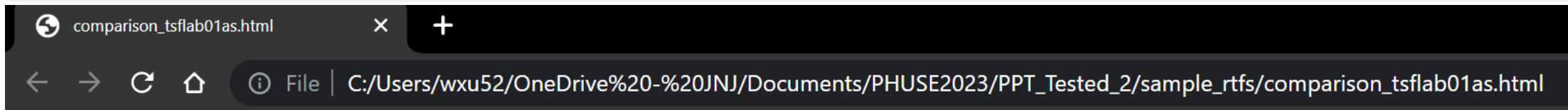
TSFLAB01: Number of Subjects with Post-Baseline Markedly Abnormal Hematology Laboratory Values; Safety Analysis Set		
	Placebo	Active
Analysis Set: Safety	28	56
Total number of subjects with marked abnormality	7 (25%)	8 (14.3%)
Hematocrit (decreased)		
N	27	55
Subjects with any abnormal value	0	0
Subjects with > 1 abnormal value	0	0
Hematocrit (elevated)		
N	27	55
Subjects with any abnormal value	0	0
Subjects with > 1 abnormal value	0	0
Hemoglobin (decreased)		
N	27	55
Subjects with any abnormal value	0	0
Subjects with > 1 abnormal value	0	0
Hemoglobin (elevated)		
N	27	55
Subjects with any abnormal value	2 (7.4%)	5 (9.1%)
Subjects with > 1 abnormal value	0	0
Leukocytes (decreased)		
N	27	55
Subjects with any abnormal value	5 (18.5%)	2 (3.6%)
Subjects with > 1 abnormal value	2 (7.4%)	0
Leukocytes (elevated)		
N	27	55
Subjects with any abnormal value	0	0
Subjects with > 1 abnormal value	0	0
Platelets (decreased)		
N	27	55
Subjects with any abnormal value	0	1 (1.8%)
Subjects with > 1 abnormal value	0	0
Platelets (elevated)		
N	27	55
Subjects with any abnormal value	0	0
Subjects with > 1 abnormal value	0	0
Note: N under each parameter refers to the count of non-missing post-baseline measurements within the parameter.		
TSFLAB01.SAS		

TSFLAB01: Number of Subjects with Post-Baseline Markedly Abnormal Hematology Laboratory Values; Safety Analysis Set		
	Placebo	Active
Analysis Set: Safety	28	56
Total number of subjects with marked abnormality	9 (32.1%)	8 (14.3%)
Hematocrit (decreased)		
N	27	55
Subjects with any abnormal value	0	0
Subjects with > 1 abnormal value	0	0
Hematocrit (elevated)		
N	27	55
Subjects with any abnormal value	0	0
Subjects with > 1 abnormal value	0	0
Hemoglobin (decreased)		
N	27	55
Subjects with any abnormal value	0	0
Subjects with > 1 abnormal value	0	0
Hemoglobin (elevated)		
N	27	55
Subjects with any abnormal value	3 (11.1%)	5 (9.1%)
Subjects with > 1 abnormal value	0	0
Leukocytes (decreased)		
N	27	55
Subjects with any abnormal value	6 (22.2%)	2 (3.6%)
Subjects with > 1 abnormal value	2 (7.4%)	0
Leukocytes (elevated)		
N	27	55
Subjects with any abnormal value	0	0
Subjects with > 1 abnormal value	0	0
Platelets (decreased)		
N	27	55
Subjects with any abnormal value	0	1 (1.8%)
Subjects with > 1 abnormal value	0	0
Platelets (elevated)		
N	27	55
Subjects with any abnormal value	0	0
Subjects with > 1 abnormal value	0	0
Note: N under each parameter refers to the count of non-missing post-baseline measurements within the parameter.		
QC_TSFLAB01.SAS		

Application 1: rtf to rtf comparison

Case 1: single file comparison between SAS generated outputs

- `library(compareDF), library(striprtf)`
- Function: `extract_table_rmna`



rowname	chng_type	1_	2_	3_
17	+	Subjects with any abnormal value	2 (7.4%)	5 (9.1%)
17	-	Subjects with any abnormal value	3 (11.1%)	5 (9.1%)
2	+	Total number of subjects with marked abnormality	7 (25%)	8 (14.3%)
2	-	Total number of subjects with marked abnormality	9 (32.1%)	8 (14.3%)
21	+	Subjects with any abnormal value	5 (18.5%)	2 (3.6%)
21	-	Subjects with any abnormal value	6 (22.2%)	2 (3.6%)
36	+	TSFLAB01.SAS		
36	-	QC_TSFLAB01.SAS		

Application 1: rtf to rtf comparison

Case 1: single file comparison between SAS generated outputs
- Beyond Compare 4

← → ↺ 🏠 ⓘ File | C:\Users\wxu52\AppData\Local\Temp\20\BC47352.tmp\Report.html

Text Compare
Produced: 20-Feb-23 10:37:34 AM

Mode: Differences
Left file: C:\Users\wxu52\OneDrive - JNJ\Desktop\Temp\tsflab01.rtf
Right file: C:\Users\wxu52\OneDrive - JNJ\Desktop\Temp\qc_tsflab01.rtf

TSFLAB01: Number of Subjects with Post-Baseline Markedly Abnormal Hematology Laboratory Values; Safety Analysis SetPlacebo Active Analysis Set: Safety2856Total number of subjects with marked abnormality7 (25%)8 (14.3%)Hematocrit (decreased)N2755Subjects with any abnormal value00Subjects with > 1 abnormal value00Hematocrit (elevated)N2755Subjects with any abnormal value00Subjects with > 1 abnormal value00Hemoglobin (decreased)N2755Subjects with any abnormal value00Subjects with > 1 abnormal value00Hemoglobin (elevated)N2755Subjects with any abnormal value2 (7.4%)5 (9.1%)Subjects with > 1 abnormal value00Leukocytes (decreased)N2755Subjects with any abnormal value5 (18.5%)2 (3.6%)Subjects with > 1 abnormal value2 (7.4%)0Leukocytes (elevated)N2755Subjects with any abnormal value00Subjects with > 1 abnormal value00Platelets (decreased)N2755Subjects with any abnormal value01 (1.8%)Subjects with > 1 abnormal value00Platelets (elevated)N2755Subjects with any abnormal value00Subjects with > 1 abnormal value00 Note: N under each parameter refers to the count of non-missing post-baseline measurements within the parameter.TSFLAB01.SAS	<>	TSFLAB01: Number of Subjects with Post-Baseline Markedly Abnormal Hematology Laboratory Values; Safety Analysis SetPlacebo Active Analysis Set: Safety2856Total number of subjects with marked abnormality9 (32.1%)8 (14.3%)Hematocrit (decreased)N2755Subjects with any abnormal value00Subjects with > 1 abnormal value00Hematocrit (elevated)N2755Subjects with any abnormal value00Subjects with > 1 abnormal value00Hemoglobin (decreased)N2755Subjects with any abnormal value00Subjects with > 1 abnormal value00Hemoglobin (elevated)N2755Subjects with any abnormal value3 (11.1%)5 (9.1%)Subjects with > 1 abnormal value00Leukocytes (decreased)N2755Subjects with any abnormal value6 (22.2%)2 (3.6%)Subjects with > 1 abnormal value2 (7.4%)0Leukocytes (elevated)N2755Subjects with any abnormal value00Subjects with > 1 abnormal value00Platelets (decreased)N2755Subjects with any abnormal value01 (1.8%)Subjects with > 1 abnormal value00Platelets (elevated)N2755Subjects with any abnormal value00Subjects with > 1 abnormal value00 Note: N under each parameter refers to the count of non-missing post-baseline measurements within the parameter.QC_TSFLAB01.SAS
---	----	---

Production Side (R) vs QC (SAS)

**R_TEFQMG10: Quantitative Myasthenia Gravis (QMG) Total Score: Means and Mean Change
Phase; Seropositive Efficacy (OL) Analysis Set**

	Visit Values							Base
	N	Mean	SD	SE	Med	Min	Max	Mean (SD)
Analysis set:								
Seropositive(OL) Analysis	34							
Placebo/Active	17							
Active /Active	17							
Active (OL)	0							
Placebo/Active								
Baseline	17	15.12	5.760	1.397	15.00	5.0	25.0	
OL Week 4	12	8.58	4.833	1.395	7.00	3.0	17.0	14.83 (6.143)
OL Week 8	9	8.00	5.268	1.756	8.00	0.0	18.0	13.72 (5.274)
OL Week 12	6	9.33	5.750	2.348	7.00	4.0	18.0	12.42 (6.020)
OL Week 16	3	8.33	5.508	3.180	8.00	3.0	14.0	10.83 (4.010)
OL Week 20	1	16.00	-	-	16.00	16.0	16.0	15.00 (-)
OL Week 24	2	10.50	3.536	2.500	10.50	8.0	13.0	12.75 (3.182)
OL Week 36	1	13.00	-	-	13.00	13.0	13.0	15.00 (-)
Active/Active								
Baseline	17	15.29	4.851	1.177	15.00	8.0	28.0	
OL Week 4	13	10.15	3.363	0.933	9.00	4.0	16.0	16.19 (4.994)
OL Week 8	8	9.63	4.173	1.475	9.00	5.0	18.0	15.56 (6.225)
OL Week 12	3	7.00	4.359	2.517	9.00	2.0	10.0	12.83 (4.805)
OL Week 16	2	4.00	4.243	3.000	4.00	1.0	7.0	10.25 (2.475)
OL Week 24	1	0.00	-	-	0.00	0.0	0.0	8.50 (-)
OL Early Termination	1	15.00	-	-	15.00	15.0	15.0	12.00 (-)

Key: OL = open label

Note: Negative change in score indicates improvement.

Note: Baseline for randomized participants is always the double-blind phase baseline. Baseline for participants who were not randomized is the baseline for the double-blind phase Day 1 visit.

**sas_tefqmg10: Quantitative Myasthenia Gravis (QMG) Total Score: Means and Mean Change
Phase; Seropositive Efficacy (OL) Analysis Set**

	Visit Values							Base
	N	Mean	SD	SE	Med	Min	Max	Mean (SD)
Analysis set: Seropositive								
Efficacy (OL) Analysis	34							
Placebo/Active	17							
Active/Active	17							
Active (OL)	0							
Placebo/Active								
Baseline	17	15.12	5.760	1.397	15.00	5.0	25.0	
OL Week 4	12	8.58	4.833	1.395	7.00	3.0	17.0	14.83 (6.143)
OL Week 8	9	8.00	5.268	1.756	8.00	0.0	18.0	13.72 (5.274)
OL Week 12	6	9.33	5.750	2.348	7.00	4.0	18.0	12.42 (6.020)
OL Week 16	3	8.33	5.508	3.180	8.00	3.0	14.0	10.83 (4.010)
OL Week 20	1	16.00	-	-	16.00	16.0	16.0	15.00 (-)
OL Week 24	2	10.50	3.536	2.500	10.50	8.0	13.0	12.75 (3.182)
OL Week 36	1	13.00	-	-	13.00	13.0	13.0	15.00 (-)
Active/Active								
Baseline	17	15.29	4.851	1.177	15.00	8.0	28.0	
OL Week 4	13	10.15	3.363	0.933	9.00	4.0	16.0	16.19 (4.994)
OL Week 8	8	9.63	4.173	1.475	9.00	5.0	18.0	15.56 (6.225)
OL Week 12	3	7.00	4.359	2.517	9.00	2.0	10.0	12.83 (4.805)
OL Week 16	2	4.00	4.243	3.000	4.00	1.0	7.0	10.25 (2.475)
OL Week 24	1	0.00	-	-	0.00	0.0	0.0	8.50 (-)
OL Early Termination	1	15.00	-	-	15.00	15.0	15.0	12.00 (-)

Application 1: rtf to rtf comparison

Case 2: single file comparison between SAS output and R output

- `library(compareDF), library(striprtf)`
- Function: `extract_table_rmna`

rowname	chnge_type	1_	2_N	3_Mean	4_SD	5_SE	6_Med	7_Min	8_Max	9_Mean (SD)
1	+	Analysis set: Seropositive(OL) Analysis	34							
1	-	Analysis set: Seropositive Efficacy (OL) Analysis	34							
11	+	OL Week 20	1	16.00	-	-	16.00	16.0	16.0	15.00 (-)
11	-	OL Week 20	1	16.00	-	-	16.00	16.0	16.0	15.00 (-)
13	+	OL Week 36	1	13.00	-	-	13.00	13.0	13.0	15.00 (-)
13	-	OL Week 36	1	13.00	-	-	13.00	13.0	13.0	15.00 (-)
20	+	OL Week 24	1	0.00	-	-	0.00	0.0	0.0	8.50 (-)
20	-	OL Week 24	1	0.00	-	-	0.00	0.0	0.0	8.50 (-)
21	+	OL Early Termination	1	15.00	-	-	15.00	15.0	15.0	12.00 (-)
21	-	OL Early Termination	1	15.00	-	-	15.00	15.0	15.0	12.00 (-)

Application 1: rtf to rtf comparison

Case 3: folder to folder comparison

- library(compareDF)

- Function: extract_table_rmna

› This PC › Documents › PHUSE2023 › PPT › folder2folder › base

Name	Status	Date modified	Type	Size
 tefcp02.rtf	✓	16-May-18 3:59 PM	Rich Text Format	57 KB
 tefcp02a.rtf	✓	16-May-18 3:59 PM	Rich Text Format	53 KB
 tefcp02b.rtf	✓	16-May-18 3:59 PM	Rich Text Format	57 KB
 tefcp03.rtf	✓	16-May-18 3:59 PM	Rich Text Format	34 KB
 tefcp04.rtf	✓	16-May-18 4:00 PM	Rich Text Format	22 KB

› This PC › Documents › PHUSE2023 › PPT › folder2folder › comp

Name	Status	Date modified	Type	Size
 tefcp02.rtf	✓	06-Jun-18 9:11 AM	Rich Text Format	57 KB
 tefcp02a.rtf	✓	06-Jun-18 9:11 AM	Rich Text Format	53 KB
 tefcp02b.rtf	✓	06-Jun-18 9:12 AM	Rich Text Format	57 KB
 tefcp03.rtf	✓	06-Jun-18 9:12 AM	Rich Text Format	34 KB
 tefcp04.rtf	✓	06-Jun-18 9:12 AM	Rich Text Format	22 KB

› This PC › Documents › PHUSE2023 › PPT › folder2folder

Name	Status	Date modified	Type	Size
base	✓	18-Feb-23 10:32 AM	File folder	
comp	✓	18-Feb-23 10:32 AM	File folder	
 comparison_tefcp02.html	✓	18-Feb-23 10:32 AM	Chrome HTML Document	96 KB
 comparison_tefcp02.xlsx	✓	18-Feb-23 10:32 AM	Microsoft Excel Worksheet	13 KB
 comparison_tefcp02a.html	✓	18-Feb-23 10:32 AM	Chrome HTML Document	85 KB
 comparison_tefcp02a.xlsx	✓	18-Feb-23 10:32 AM	Microsoft Excel Worksheet	12 KB
 comparison_tefcp02b.html	✓	18-Feb-23 10:32 AM	Chrome HTML Document	93 KB
 comparison_tefcp02b.xlsx	✓	18-Feb-23 10:32 AM	Microsoft Excel Worksheet	12 KB
 comparison_tefcp03.html	✓	18-Feb-23 10:32 AM	Chrome HTML Document	52 KB
 comparison_tefcp03.xlsx	✓	18-Feb-23 10:32 AM	Microsoft Excel Worksheet	10 KB
 comparison_tefcp04.html	✓	18-Feb-23 10:32 AM	Chrome HTML Document	32 KB
 comparison_tefcp04.xlsx	✓	18-Feb-23 10:32 AM	Microsoft Excel Worksheet	9 KB
 TEMP.csv	✓	18-Feb-23 10:30 AM	Microsoft Excel Comma S...	1 KB

Application 2: population counts from all tables

- Showcase 1: Early Phase Study (28 tables)
- Showcase 2: Phase 3 Study (159 tables)

Application 2: population counts from all tables

Showcase 1: Early Phase Study (28 tables)

```
- library(striprtf)
```

- Function: `extract_pop`

[illegible]

Application 2: population counts from all tables

Showcase 2: Phase 3 Study (159 tables)

- library(striprtf)

- Function: extract_pop

1	Analysis set: Full	1_ 2_Placebo 3_Active\n15 mg/kg q2w 4_Total 62 60 122	tabid tsidem03.rtf
1	Analysis set: Full	1_ 2_Placebo 3_Active\n15 mg/kg q2w 4_Total 62 60 122	tabid tsidem04.rtf
1	Analysis set: Full	1_ 2_Placebo 3_Active\n15 mg/kg q2w 4_Total 62 60 122	tabid tsidem05a.rtf
1	Analysis set: Primary Efficacy Analysis	1_ 2_Placebo 3_Active\n15 mg/kg q2w 4_Total 35 34 69	tabid tsidem05b.rtf
1	Analysis set: Full	1_ 2_Placebo 3_Active\n15 mg/kg q2w 62 60	tabid tsidev01a.rtf
1	Analysis set: Primary Efficacy Analysis	1_ 2_Placebo 3_Active\n15 mg/kg q2w 35 34	tabid tsidev01b.rtf
1	Analysis set: Full	1_ 2_Placebo 3_Active\n15 mg/kg q2w 62 60	tabid tsidev02.rtf
1	Analysis set: Enrolled	1_ 2_Placebo/ \nActive 3_Active/ \nActive 4_Active (OL) 66 61 17	tabid tsids02.rtf
1	Analysis set: All Randomized	1_ 2_Placebo 3_Active\n15 mg/kg q2w 66 61	tabid tsids03.rtf
1	Analysis set: Full (OL)	1_ 2_Placebo/ \nActive 3_Active/ \nActive 4_Active (OL) 5_All Active 29 25 17 71	tabid tsids04.rtf
1	Analysis set: Safety	1_ 2_Placebo/ Active 3_Active/ Active 4_Active (OL) 5_All Active 62 60 17 139	tabid tsids05.rtf
1	Analysis set: Safety (DB)	1_ 2_Placebo 3_Active\n15 mg/kg q2w 62 60	tabid tsiex01a.rtf
1	Analysis set: Primary Efficacy Analysis	1_ 2_Placebo 3_Active\n15 mg/kg q2w 35 34	tabid tsiex01b.rtf
1	Analysis set: All Active	1_ 2_Placebo/ \nActive 3_Active/ \nActive 4_Active (OL) 5_All Active 29 60 17 106	tabid tsiex01c.rtf
1	Analysis set: All Active	1_ 2_Placebo/ \nActive 3_Active/ \nActive 4_Active (OL) 5_All Active 29 60 17 106	tabid tsiex02.rtf

Application 3: counts (at timepoints) among related tables

- Lab and corresponding shift tables
- Counts at timepoints among ECG tables

Application 3: counts (at timepoints) among related tables

Lab and corresponding shift tables

```
- library(striprtf)
```

- Function: `extract_removepop`

		1_ 2_N	3_Total	4_Low	5_Normal	6_High
5	Avg duration of follow-up (weeks)					
6	Placebo	16.96				
7	Active 15 mg/kg q2w	16.03				
8						
9	Alanine Aminotransferase (Enzyme U/L)					
10						
11	Placebo	58				
15	Total		58	0	52	6
16						
17	Active 15 mg/kg q2w	56				
21	Total		56	0	55	1
22						
23	Alkaline Phosphatase (Enzyme U/L)					
24						
25	Placebo	58				
29	Total		58	6	48	4
30						
31	Active 15 mg/kg q2w	55				
35	Total		55	3	50	2
36						
37	Aspartate Aminotransferase (Enzyme U/L)					
38						
39	Placebo	58				
43	Total		58	0	55	3
44						
45	Active 15 mg/kg q2w	56				
49	Total		56	0	54	2
50						
51	Bilirubin (umol/L)					
52						
53	Placebo	57				
57	Total		57	1	54	2
58						
59	Active 15 mg/kg q2w	56				
63	Total		56	0	52	4
64						

[illegible]

Application 3: counts (at timepoints) among related tables

Counts at timepoints among ECG tables

```
- library(striprtf)
```

- Function: `extract_removepop`

		1	2	3	4	5	6
		Placebo	Active	Active	Active	Active	Active
3	QTcB Interval, Aggregate (msec)						
4	Baseline						
5	N	29	59	17	105		
10	DB Week 12						
11	N	27	32	0	59		
16	DB Week 24						
17	N	28	24	0	52		
22	OL Day 1						
23	N	1	2	0	3		
28	OL Week 12						
29	N	11	4	16	31		
34	OL Week 48						
35	N	0	0	2	2		
40	DB Early Termination						
41	N	0	1	0	1		
46	OL Early Termination						
47	N	1	0	0	1		
52	OL End of Treatment						
53	N	1	1	0	2		
58	Follow-up						
59	N	1	1	0	2		
64	QTcF Interval, Aggregate (msec)						
65	Baseline						
66	N	29	59	17	105		
72	DB Week 12						
73	N	27	32	0	59		
78	DB Week 24						
79	N	28	24	0	52		
84	OL Day 1						
85	N	1	2	0	3		
90	OL Week 12						
91	N	11	4	16	31		
96	OL Week 48						
97	N	0	0	2	2		
102	DB Early Termination						
103	N	0	1	0	1		
108	OL Early Termination						
109	N	1	0	0	1		
114	OL End of Treatment						
115	N	1	1	0	2		
120	Follow-up						
121	N	1	1	0	2		
127							

Conclusions

R/Python packages + R user-defined functions

- provide oversight of deliverable's quality
- improve efficiency of TFL validation process
- can be extended to other areas

The opinions in this presentation are my own and do not necessarily reflect the views and policies of J&J

