MSD
INVENTING FOR LIFE

Key performance differences between R and SAS in real-life scenarios

Michal Lewandowski & Shuai Wu

Statistical Programming

October, 2022

Disclaimer

The opinions expressed in this presentation and on the following slides are solely the presenters and not necessarily those of MSD.

Introduction

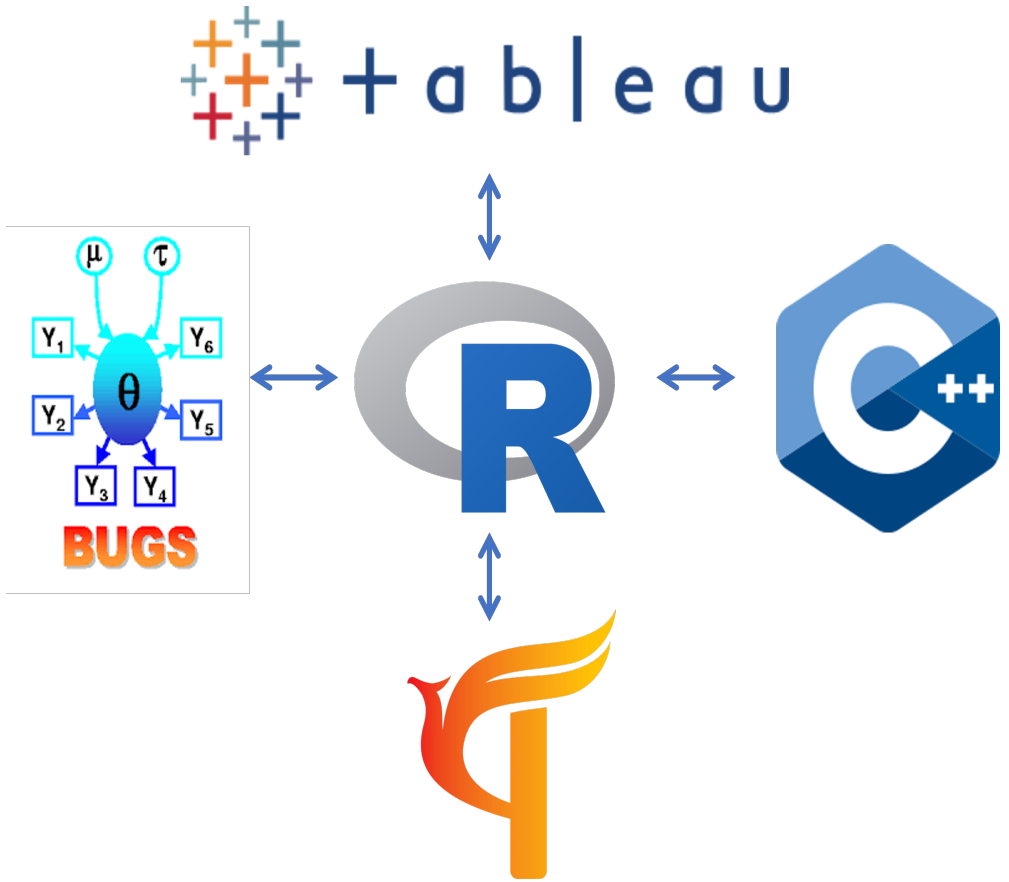
Before the plan was born

Before the plan was born

The dramatical rise of R in the pharmaceutical industry

FREE

Cutting-edge



Comparison of SAS and R at first glance

Attribute	R	SAS
Licensing cost	+	-
Suitability	-	+
Flexibility & innovation	+	-
Validity	-	+
Computing ability	?	?

Research & findings

Prerequisites, implementation and results

Prerequisites

Testing environment	R	SAS
Access	Remote	
OS	Enterprise Linux Server 7 64bit	
App	RStudio Workbench	SAS Base
Version	4.1.1	9.4
Processor	Multi-core	
RAM	47 GB	?

Design

Clinical & non-clinical applications

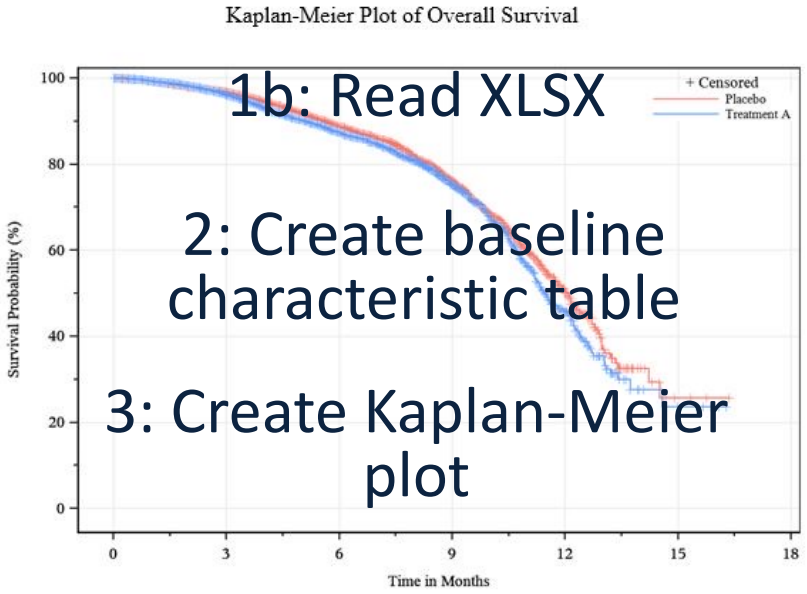
Simple & complex tasks

Everyday scenarios

And we like graphs…

4 Scenarios

1a: Read a CSV and write XLSX



1b: Read XLSX

2: Create baseline characteristic table

3: Create Kaplan-Meier plot

Metrics

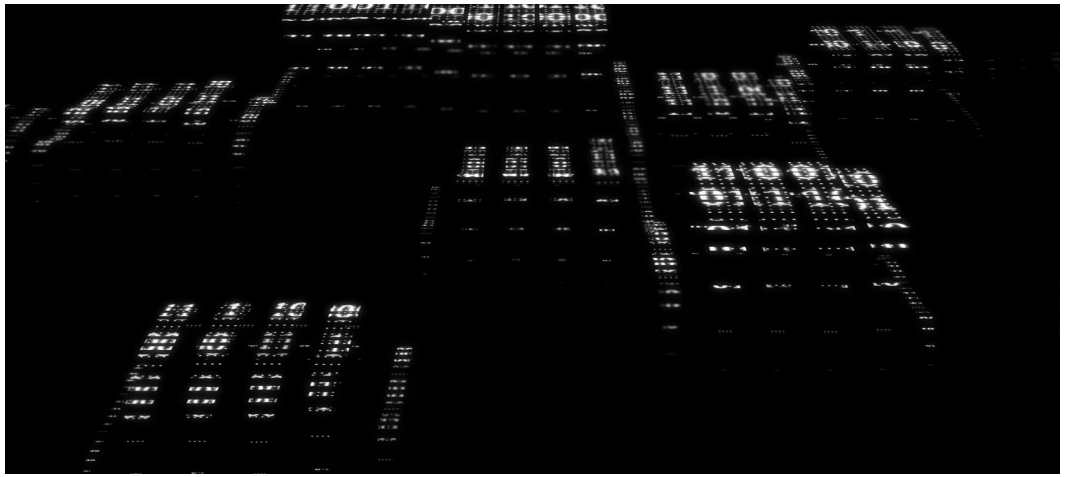
User time

System time

Elapsed time

Memory

Scalability



Testing rules

Produce the same output using the same input

Utilize available in-house tools

Optimize our code

Repeat each scenario 20 times

SAS approach

SAS to import CSV:

proc import
data step
Import Wizard

SAS to import XLSX:

proc import
libname statement with xlsx engine
DDE (Dynamic Data Exchange)
Import Wizard
data step won't work!

SAS to export XLSX:

proc export
ods output (ODS HTML for XLS)
DDE

For table and graph creation, two specialized macros from Macro Library

How to get the quantitative metrics?

`options fullstimer;` and save a log

How to depict them?

Read logs and compute summary for each scenario, run and file in a dataset to be later used for plotting

R approach – evalR package

R to import CSV:

Read.csv function

R to import XLSX:

Read_excel(readxl) function

R to export XLSX:

write_xlsx(writexl) function

Table creation:

r2rtf package developed by MSD

Graph creation:

ggplot2 package

How to get the quantitative metrics?

A small package evalR was designed to get the metrics for multiple R expressions:

Memusedby function:

- System.time function to calculate three kinds of time
- Profmem(profmem) function to profile all memory usage by each R expression

Timepassby function:

-invoke the memusedby function to get the metrics of multiple R expressions and return a dataframe containing all metrics for each scenario.

Alignment

Patient Characteristics

	Placebo		Xanomeline Low Dose		Xanomeline High Dose		Total	
	n	(%)	n	(%)	n	(%)	n	(%)
Participants in population	3,343		2,210		4,447		10,000	
Gender								
Male	1,852	(55.4)	1,186	(53.7)	2,463	(55.4)	5,501	(55.0)
Female	1,491	(44.6)	1,024	(46.3)	1,984	(44.6)	4,499	(45.0)
Age category								
<65	345	(10.3)	229	(10.4)	428	(9.6)	1,002	(10.0)
65-80	2,208	(66.0)	1,460	(66.1)	3,024	(68.0)	6,692	(66.9)
>80	790	(23.6)	521	(23.6)	995	(22.4)	2,306	(23.1)
Age (Years)								
Participants with data	3,343		2,210		4,447		10,000	
Mean	74.6		74.5		74.5		74.5	
SD	7.5		7.5		7.4		7.5	
Median	75.0		75.0		75.0		75.0	
Range	51.0 to 89.0		51.0 to 89.0		51.0 to 89.0		51.0 to 89.0	
Race								
White	3,015	(90.2)	2,014	(91.1)	4,035	(90.7)	9,064	(90.6)
Black or African American	38	(1.1)	22	(1.0)	43	(1.0)	103	(1.0)
American Indian or Alaska Native	290	(8.7)	174	(7.9)	369	(8.3)	833	(8.3)
Ethnicity								
Hispanic or Latino	157	(4.7)	94	(4.3)	191	(4.3)	442	(4.4)
Not Hispanic or Latino	3,186	(95.3)	2,116	(95.7)	4,256	(95.7)	9,558	(95.6)
BMI Group								
<25	1,788	(53.5)	1,160	(52.5)	2,322	(52.2)	5,270	(52.7)
25-30	1,250	(37.4)	847	(38.3)	1,693	(38.1)	3,790	(37.9)
>=30	305	(9.1)	203	(9.2)	432	(9.7)	940	(9.4)
BMI (kg/m^2)								
Participants with data	3,343		2,209		4,447		9,999	
Mean	24.6		24.7		24.7		24.7	
SD	4.1		4.0		4.1		4.1	
Median	24.5		24.7		24.7		24.7	
Range	13.7 to 39.4		13.8 to 40.1		13.7 to 39.7		13.7 to 40.1	

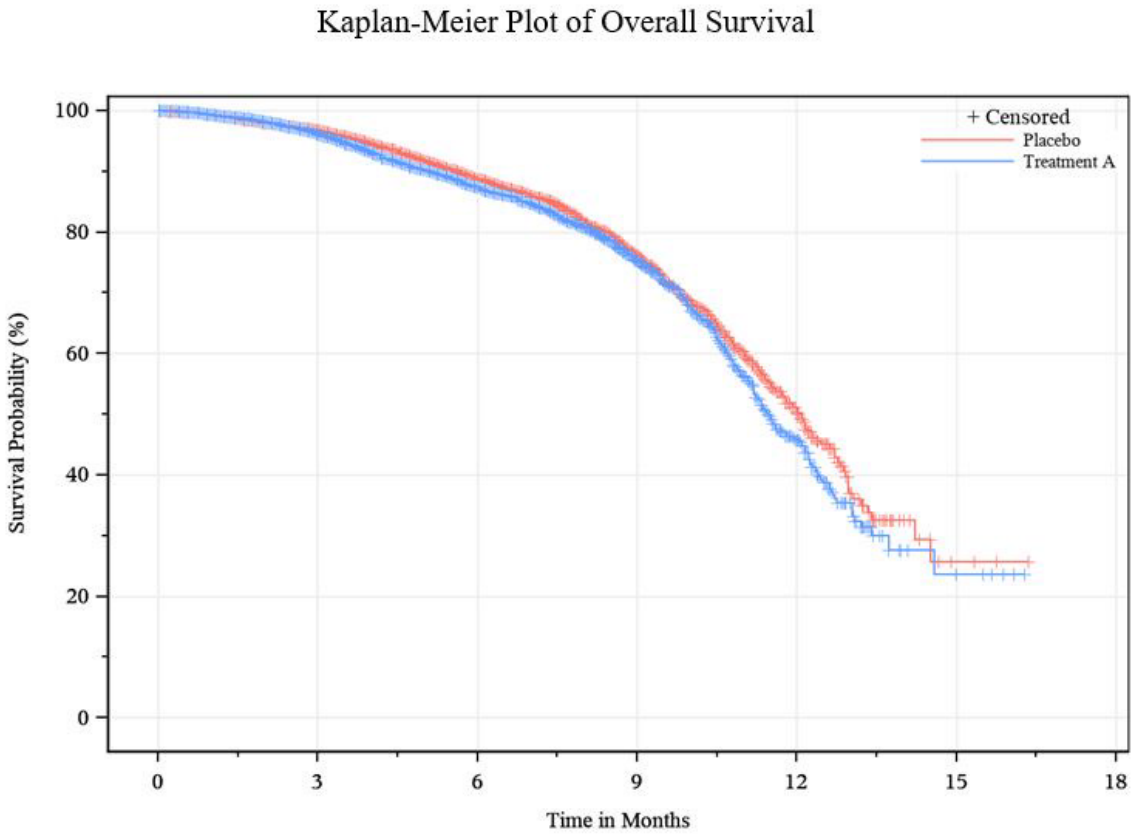
Striving for alignment

- R package made basing on existing SAS macros
- Different default values
- Not enough functionality

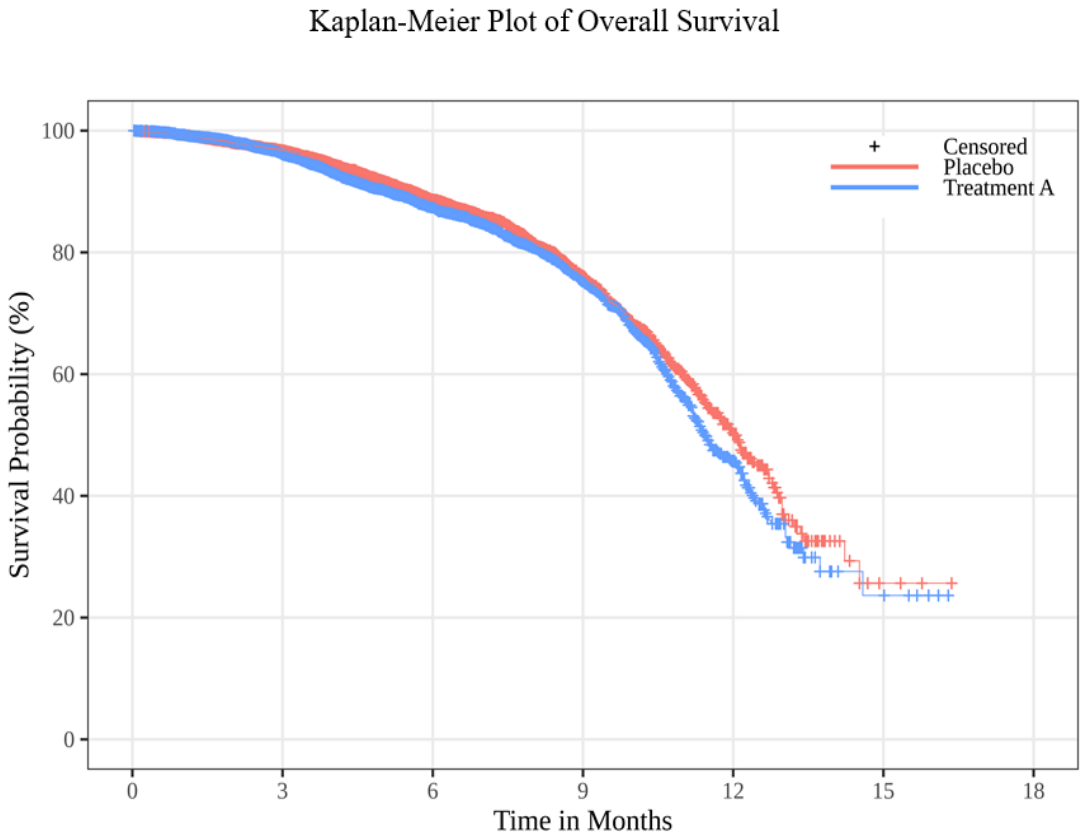
Patient Characteristics

	Placebo		Xanomeline Low Dose		Xanomeline High Dose		Total	
	n	(%)	n	(%)	n	(%)	n	(%)
Participants in population	3,343		2,210		4,447		10,000	
Gender								
Male	1,852	(55.4)	1,186	(53.7)	2,463	(55.4)	5,501	(55.0)
Female	1,491	(44.6)	1,024	(46.3)	1,984	(44.6)	4,499	(45.0)
Age category								
<65	345	(10.3)	229	(10.4)	428	(9.6)	1,002	(10.0)
65-80	2,208	(66.0)	1,460	(66.1)	3,024	(68.0)	6,692	(66.9)
>80	790	(23.6)	521	(23.6)	995	(22.4)	2,306	(23.1)
Age (Years)								
Participants with data	3,343		2,210		4,447		10,000	
Mean	74.6		74.5		74.5		74.5	
SD	7.5		7.5		7.4		7.5	
Median	75.0		75.0		75.0		75.0	
Range	51.0 to 89.0		51.0 to 89.0		51.0 to 89.0		51.0 to 89.0	
Race								
White	3,015	(90.2)	2,014	(91.1)	4,035	(90.7)	9,064	(90.6)
Black or African American	38	(1.1)	22	(1.0)	43	(1.0)	103	(1.0)
American Indian or Alaska Native	290	(8.7)	174	(7.9)	369	(8.3)	833	(8.3)
Ethnicity								
Hispanic or Latino	157	(4.7)	94	(4.3)	191	(4.3)	442	(4.4)
Not Hispanic or Latino	3,186	(95.3)	2,116	(95.7)	4,256	(95.7)	9,558	(95.6)
BMI Group								
<25	1,788	(53.5)	1,160	(52.5)	2,322	(52.2)	5,270	(52.7)
25-30	1,250	(37.4)	847	(38.3)	1,693	(38.1)	3,790	(37.9)
>=30	305	(9.1)	203	(9.2)	432	(9.7)	940	(9.4)
BMI (kg/m^2)								
Participants with data	3,343		2,209		4,447		9,999	
Mean	24.6		24.7		24.7		24.7	
SD	4.1		4.0		4.1		4.1	
Median	24.5		24.7		24.7		24.7	
Range	13.7 to 39.4		13.8 to 40.1		13.7 to 39.7		13.7 to 40.1	

Alignment

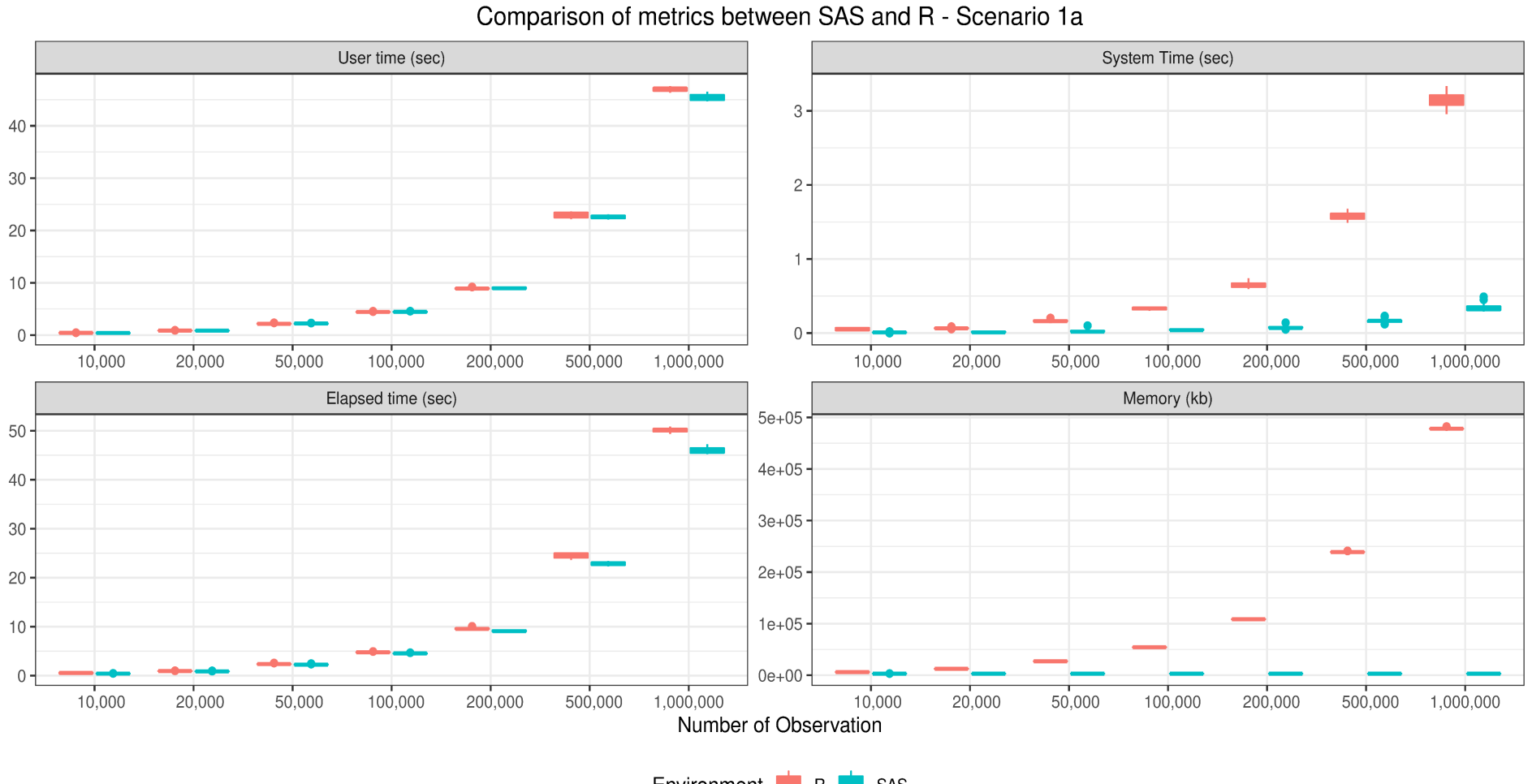


Database cut-off date: 13OCT2022
Data source: edited file from
http://hedwig.mgh.harvard.edu/biostatistics/sites/default/files/public/education/tmp_sample_data.csv

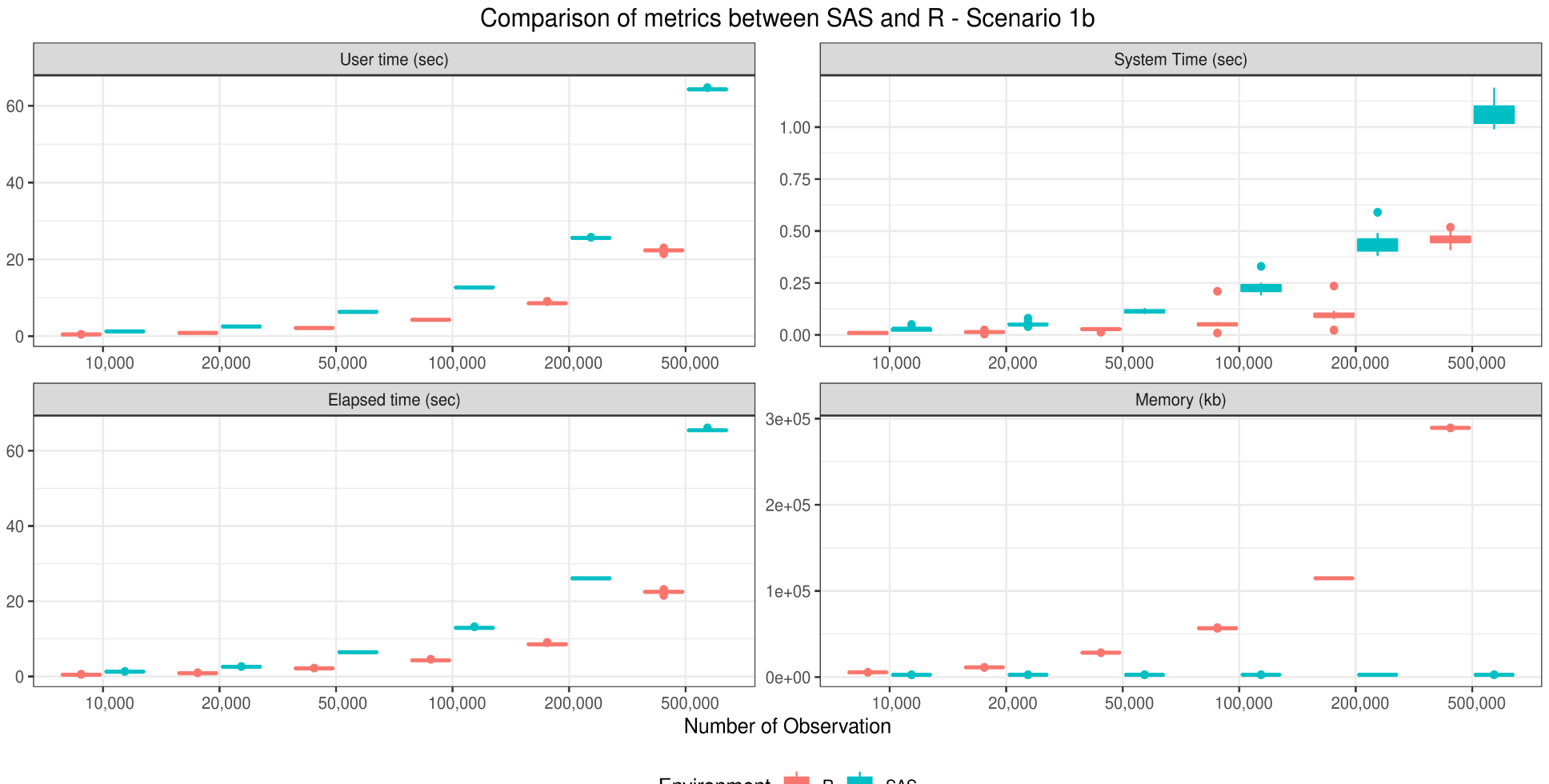


Database cut-off date: 13OCT2022
Data source: edited file from
http://hedwig.mgh.harvard.edu/biostatistics/sites/default/files/public/education/tmp_sample_data.csv

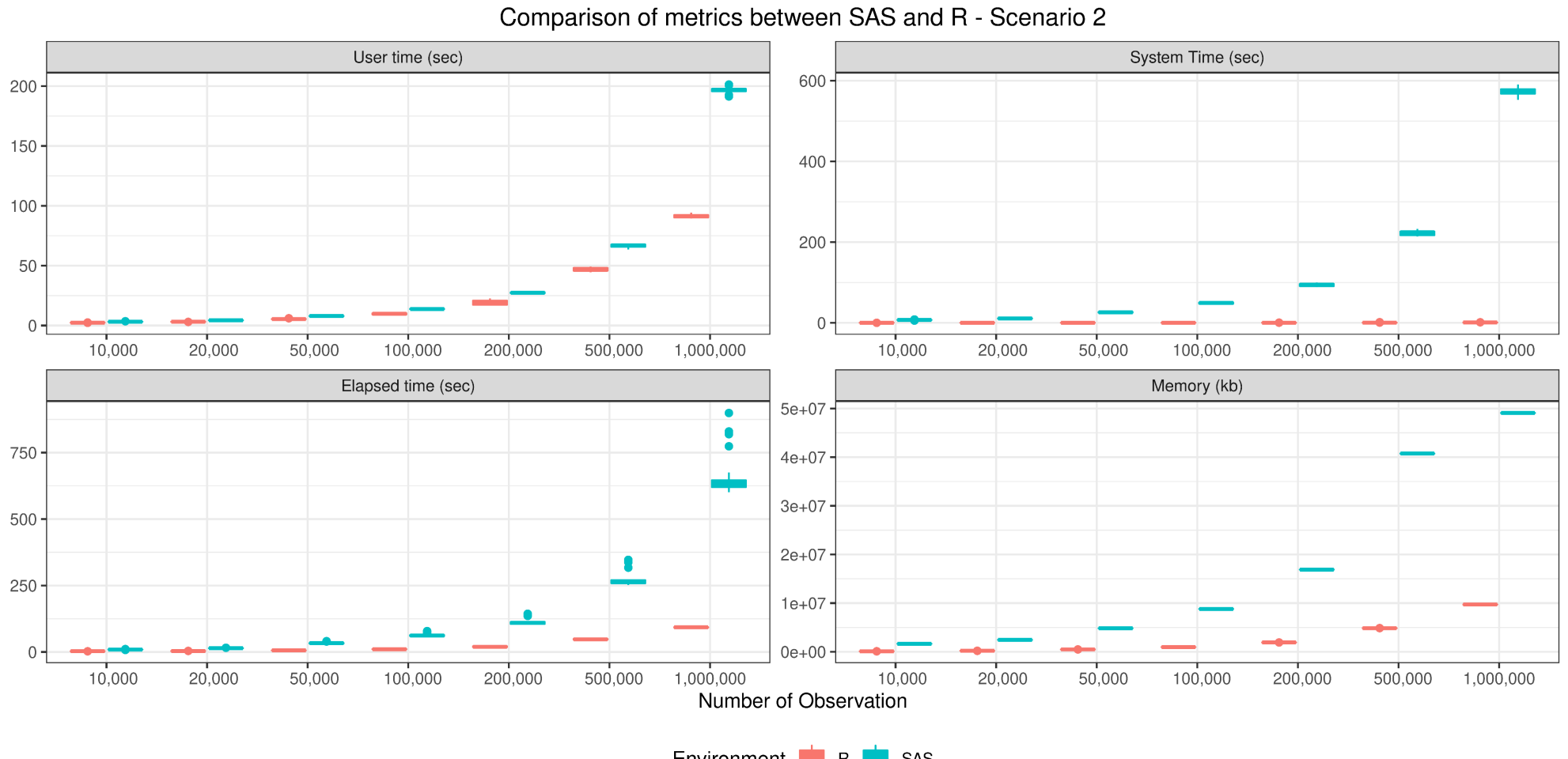
Importing a CSV file and export it as an EXCEL file



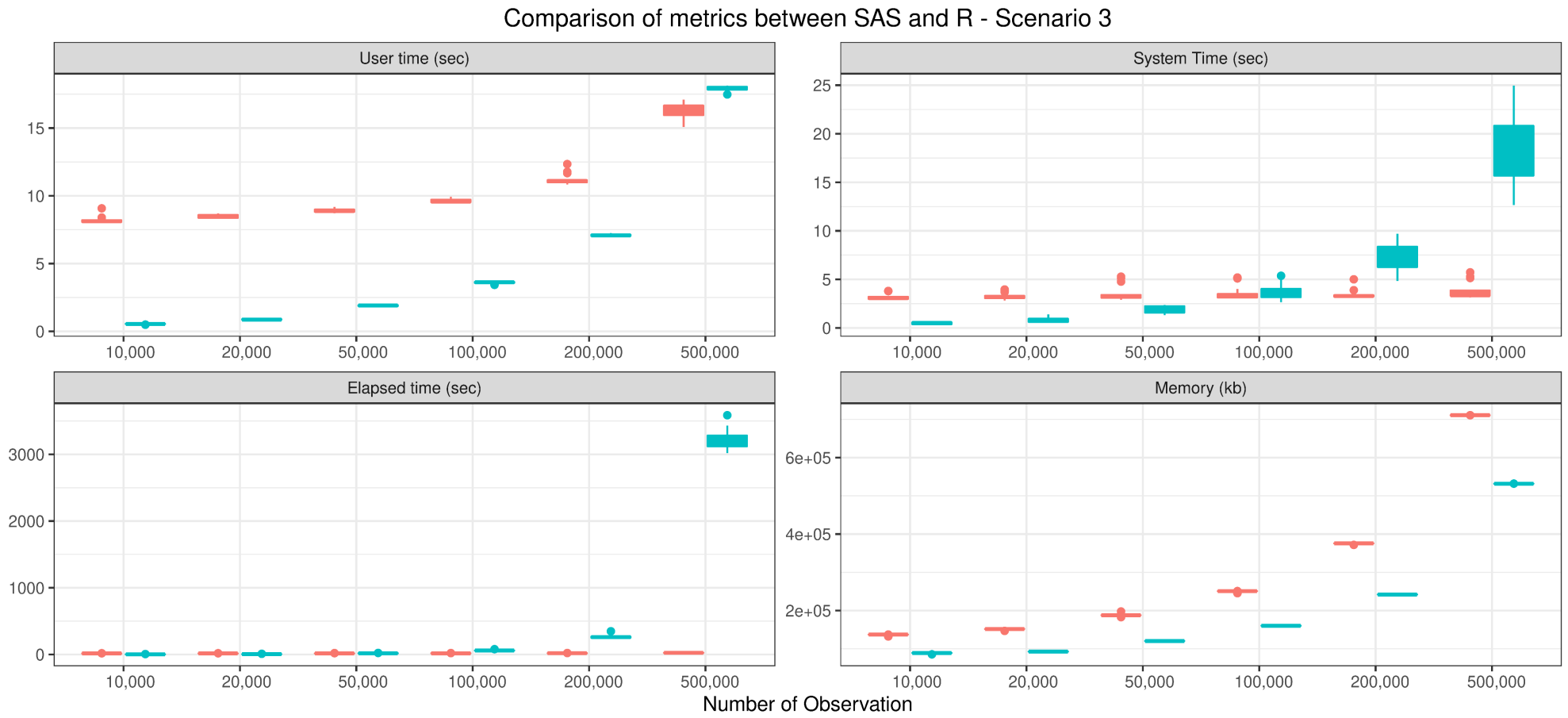
Importing an EXCEL file



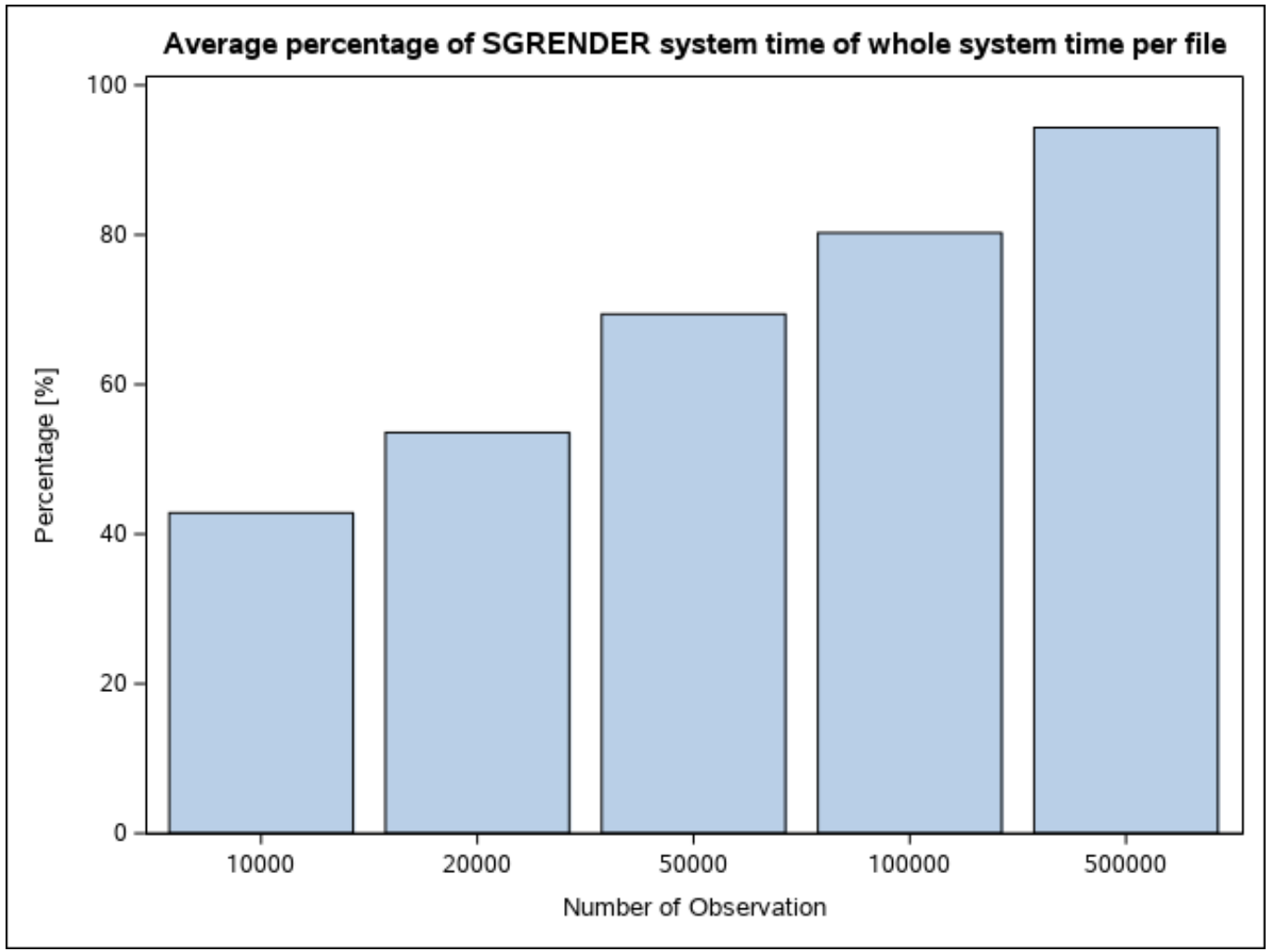
Baseline characteristic table



Kaplan-Meier Plot



Kaplan-Meier Plot



Conclusion

Do you take R or SAS?

Attribute	R	SAS
Cost	+	-
Suitability	-	+
Flexibility & innovation	+	-
Validity	-	+
Time (system, user, elapsed)	+	-
Memory	+ (tabularization and plotting)	+ (data import/export)
Scalability	+ (if you prioritize time)	+ (if you prioritize memory)
Stability	-	+

The prospect of SAS and R

What will replace one technology must not be something similar

If no supersession, SAS and R must be connected tighter

Sources

Common:

System time and user time:
https://en.wikipedia.org/wiki/Time_(Unix)#User_Time_vs_System_Time

Exploring R as a validation tool for statistical programming in clinical trials:
https://www.lexjansen.com/pharmasug/2022/AD/PharmaSUG-2022-AD-076.pdf

FDA Statistical Software Clarifying Statement
https://www.fda.gov/media/161196/download

Concurrent Use of SAS and R in Clinical Trial Reporting – PHUSE SDE 2021
https://phuse.s3.eu-central-1.amazonaws.com/Archive/2021/SDE/US/Virtual%20-%20Autumn/PRE_Americas19.pdf

SAS:

SAS documentation
https://support.sas.com/en/documentation.html

Various threads at https://communities.sas.com

Generating Excel files
https://www.lexjansen.com/pharmasug/2006/Posters/PO13.pdf

Fullstimer metrics description
https://support.sas.com/rnd/scalability/tools/fullstim/index.html

Specifications from MSD BARDS Macro Library for %asr0baseline0characteristic and %kmplot macros

SAS in life science industry:
https://blogs.sas.com/content/hiddeninsights/2021/03/09/jonathan-riches-thomas-keil-sas/

R:

R documentation for function system.time:
https://www.rdocumentation.org/packages/base/versions/3.6.2/topics/system.time

R documentation for function profmem:
https://www.rdocumentation.org/packages/profmem/versions/0.6.0/topics/profmem

Memory usage in R:
http://adv-r.had.co.nz/memory.html#memory-profiling

Functional programming:
https://en.wikipedia.org/wiki/Functional_programming

Documentation for package r2rtf:
https://merck.github.io/r2rtf/

RStudio in life science industry:
https://www.rstudio.com/solutions/pharma/

MSD
INVENTING FOR LIFE

Thank You

MSD BARDS HTA Statistical Programming
E-mails: michal.lewandowski1@msd.com and shuai.wu@msd.com