

CONFIDENTIAL

Use R to Track and Highlight the Changes Between Two Versions of Data to Have More Efficient Data Review

Disheng Huang
Zhonggai Li

Agenda

Introduction	2 mins
Method and Applications	13 mins
Challenges and Limitations	3 mins
Summary	2 mins
Q & A	5 mins

Introduction

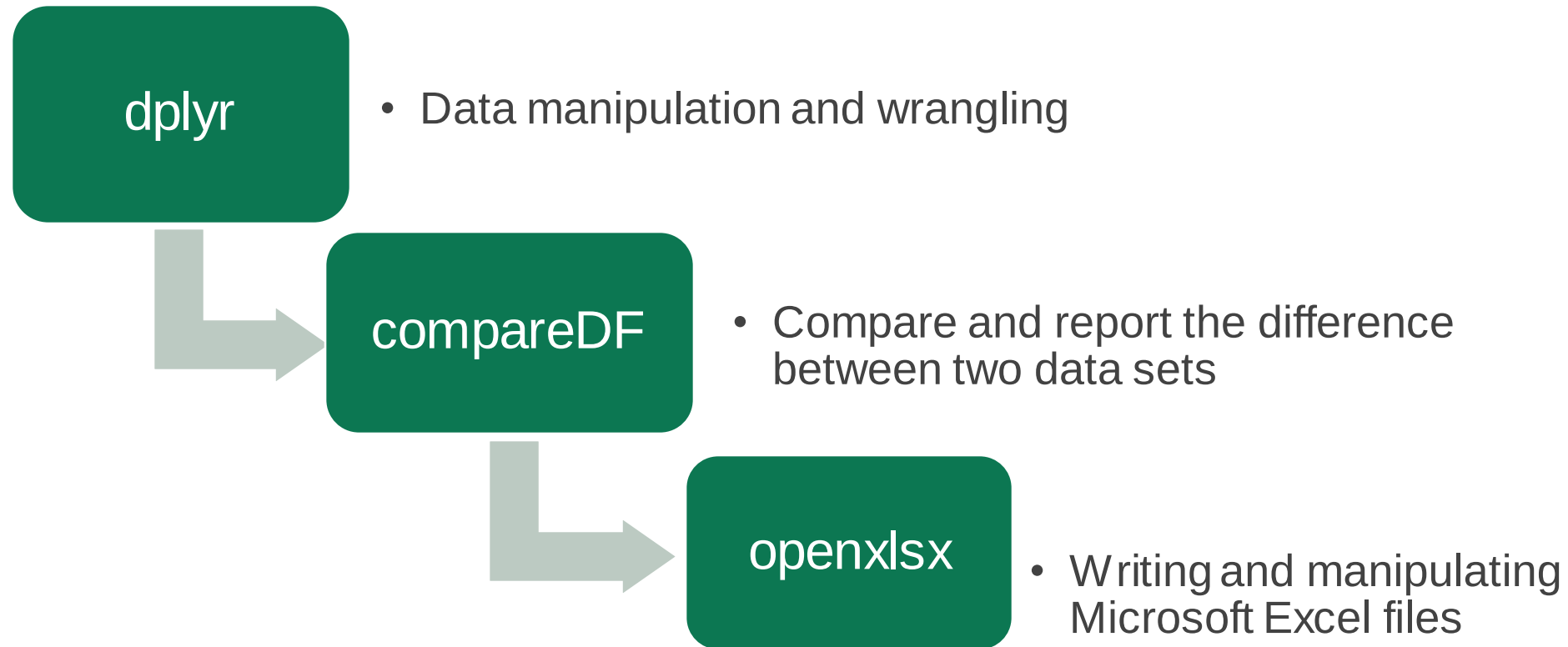
Why do we need to track and highlight the changes between two versions of data?

- In phase 2/3 trials, performing the subject level data review is **quite time consuming and error prone**.
- When the database lock or snapshot is coming, reviewing the data can be needed on **daily basis**. Manually reviewing all the patient data every day is nearly **impossible**.
- If there is a tool that could **identify the change(s)** between two versions/snapshots of data, it could not only reduce the reviewing time, but also reduce the chance of overlooking changes.

Method and Applications

Method

- Why use R?
 - R is open-source software, and it has a very strong community. There are **over 19200** packages available as of Feb 15, 2023. The **three packages utilized** are :



Case 1: Record Added

Data set 1

SUBJID	SEX	AGE
A1	F	21
A2	M	56
A3	F	34



Data set 2

SUBJID	SEX	AGE
A1	F	21
A2	M	56
A3	F	34
A4	F	67

Output

File Generation Date: 2023-02-12			
SUBJID	chnng_type	SEX	AGE
A4	+	F	67

There is a new record inserted. The font is highlighted in **green** and the background is in yellow.

Case 2: Record Deleted

Data set 1

SUBJID	SEX	AGE
A1	F	21
A2	M	56
A3	F	34



Data set 2

SUBJID	SEX	AGE
A1	F	21
A2	M	56

Output

File Generation Date: 2023-02-12			
SUBJID	chnge_type	SEX	AGE
A3	-	F	34

There is a record deleted. The font is highlighted in red and the background is in yellow.

Case 3: Records Changed

Data set 1

SUBJID	SEX	AGE
A1	F	21
A2	M	56
A3	F	34



Data set 2

SUBJID	SEX	AGE
A1	M	21
A2	M	56
A3	F	43

Output

File Generation Date: 2023-02-14			
SUBJID	chnge_type	SEX	AGE
A1	+	M	21
A1	-	F	21
A3	+	F	43
A3	-	F	34

There are two data points changed. New records are highlighted in **green** and previous records are highlighted in **red** and the backgrounds for those changes are in yellow.

Applications

- Track Change of Patient Profiles
- Track Change of Data Cleaning Listings
- Track Change of Efficacy Smart Listings

Challenges and Limitations

Challenges and limitations

- `compareDF::compare_df` needs to know the **unique keys** to compare the data sets. The unique keys should be the same in both data sets.

```
compareDF::compare_df(df_new = df2, df_old = df1,  
group_col = c("var1", "var2"), stop_on_error = FALSE)
```

- `compareDF::compare_df` can only compare two data sets with the **same number of columns and same column names**. But we could summarize the change of columns separately.

Summary

Summary

- The tool allows comparison of two data set with same set of columns/variables for their record/values changes
- Using R to build the track change files to highlight the changes between two version of data sets is very **convenient**.
- The outputs are intuitive and **easy to review**.
- This method can **reduce the clinical data reviewing time** dramatically as well as **improve the quality for data review**.

Question and Answer



Sumitomo Pharma

Sumitomo Pharma Oncology