

1 CJUG R+SDTM Team Activities

author: CJUG SDTM team, R subteam date: 2017/03/03 autosize: true



“

This presentaion can be found in following URL:

Presentation HTML file: <http://rpubs.com/mokjpn/cjugws2017>

Original R Markdown document: <https://github.com/mokjpn/cjugr/blob/master/Workshop2017.Rpres>

2 CJUG SDTM R subteam

- Objective
- Learn R language environment and related software.
- Discuss how to make data managers' work efficient with R.
- History
- Start from Sep 2015.
- Presentation at CJUG workshop 2016.
- Activity Report on Jun 2016 meeting.
- Presentation at CJUG workshop 2017. (This!)
- Members
- Masafumi Okada, M.K, M.D, M.T, K.N, M.Y

3 Activities

- By plain R
- Import SDTM dataset to R's dataframe.
- Draw some graphics, or make summary table from the imported dataframes.
- By R AnalyticFlow
- Import raw dataset from comma-separated-value text file.
- Merge some tables.

- Calculate some derived values.
- Compose SDTM-like table.
- By Rstudio and R markdown
- Write simple data management report with R markdown format.
- By Rstudio and GitHub
- Version control of R script.
- Co-editing R markdown document

4 Import SDTM dataset to R's dataframe in XPT format.

XPT files can be read by `read.xport()` function of `foreign` package.

```
setwd("~/CJUG/SDTM/20_Work_in_progress/23_HCT-1337")
library(foreign)
# read.xport で xpt ファイルを変数 QS に、データフレームとして読み込み
QS <- read.xport("30_Summary/dataset/QS.xpt")
DA <- read.xport("30_Summary/dataset/DA.xpt")
str(QS)

## 'data.frame': 118 obs. of 18 variables:
## $ STUDYID : Factor w/ 1 level "HCT-1337": 1 1 1 1 1 1 1 1 1 ...
## $ DOMAIN : Factor w/ 1 level "QS": 1 1 1 1 1 1 1 1 1 ...
## $ USUBJID : Factor w/ 59 levels "HCT-13370101",...: 1 1 2 2 3 3 4 4 5 5 ...
## $ QSSEQ : num 1 2 1 2 1 2 1 2 1 2 ...
## $ QSTESTCD: Factor w/ 2 levels "POSTQ","PREQS": 2 1 2 1 2 1 2 1 2 1 ...
## $ QSTEST : Factor w/ 2 levels "Post-dose Calculation test",...: 2 1 2 1 2 1 2 1 2 1 ...
## $ QSCAT : Factor w/ 1 level "Calculation Test": 1 1 1 1 1 1 1 1 1 1 ...
## $ QSORRES : Factor w/ 83 levels "", "102", "104",...: 5 21 20 52 65 76 60 56 49 61 ...
## $ QSSTRESC: Factor w/ 83 levels "", "102", "104",...: 5 21 20 52 65 76 60 56 49 61 ...
## $ QSBLFL : Factor w/ 2 levels "", "Y": 2 1 2 1 2 1 2 1 2 1 ...
## $ VISITNUM: num 2 2 2 2 2 2 2 2 2 2 ...
## $ VISIT : Factor w/ 1 level "Dosing": 1 1 1 1 1 1 1 1 1 1 ...
## $ VISITDY : num 1 1 1 1 1 1 1 1 1 1 ...
## $ EPOCH : Factor w/ 2 levels "SCREENING","TREATMENT": 1 2 1 2 1 2 1 2 1 2 ...
## $ QSDTC : Factor w/ 1 level "..": 1 1 1 1 1 1 1 1 1 1 ...
## $ QSDY : num 1 1 1 1 1 1 1 1 1 1 ...
## $ QSTPT : Factor w/ 2 levels "Post-dose","Pre-dose": 2 1 2 1 2 1 2 1 2 1 ...
## $ QSTPTNUM: num 1 2 1 2 1 2 1 2 1 2 ...
```

5 Import SDTM dataset to R's dataframe in Dataset-XMLformat.

Dataset-XML files can be read by `read.dataset.xml()` function of `R4DSXML` package.

```

setwd("../Define2Validate")
library(R4DSXML)

## Loading required package: XML

## Warning: package 'XML' was built under R version 3.3.2

# read.dataset.xml で Dataset-XML ファイルを変数 CM に、データフレームとして読み込み
CM <- read.dataset.xml("Odm_CM.xml", "Odm_Define.xml")
str(CM)

## 'data.frame': 2 obs. of 6 variables:
## $ DOMAIN : chr "CM" "CM"
## $ CMSEQ : int 1 2
## $ CMTRT : chr "マイスタン錠 5 m g" "クレストール錠"
## $ CMDOSE : int 3 5
## $ CMDOSU : chr "CAPSULE" "mgg"
## $ CMSTDTC: chr "2016-03-16T10:56:40" "2016-03-16T10:56:401"

```

6 Draw some graphics

```

library(tidyr)

## Warning: package 'tidyr' was built under R version 3.3.2

library(dplyr)

##
## Attaching package: 'dplyr'

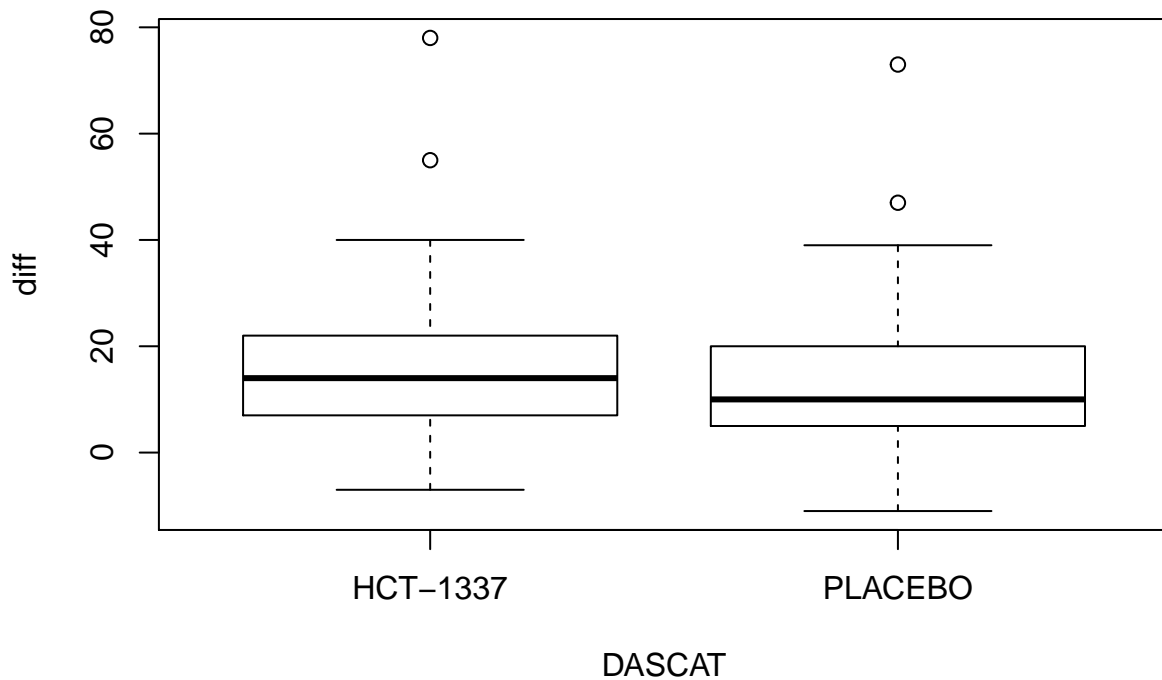
## The following objects are masked from 'package:stats':
##
## filter, lag

## The following objects are masked from 'package:base':
##
## intersect, setdiff, setequal, union

QS %>%
  # mutate() は列を加工して新たな列を追加します。QSSTRESC の数値への変換。
  mutate(qsstrescn = as.numeric(as.character(QSSTRESC))) %>%
  # select() は特定の列だけを取り出します。USUBJID と QSTESTCD と qsstrescn だけに。
  select(USUBJID, QSTESTCD, qsstrescn) %>%
  # USUBJID でグループ化
  group_by(USUBJID) %>%
  # spread() は Normalized format を、1 行 1 症例のフォーマットに変換します。
  # QSTESTCD の値をそれぞれ列にして、qsstrescn を値にします。

```

```
spread(QSTESTCD,qsstrescn) %>%
# POSTQとPRESQの差をとって、diffという列に追加します。
mutate(diff = POSTQ - PREQS) %>%
# DAドメインの表を結合
inner_join(DA,by="USUBJID") %>%
# DASCATごとにdiffの値を箱ひげ図にします
plot(diff ~ DASCAT, data=.)
```



7 R AnalyticFlow

- **R AnalyticFlow** by Ef-prime, Inc. is a data analysis software that utilizes the R environment for statistical computing.
- With this software, the procedure of data cleaning can be composed by connecting some of icons which contains a set of frequently used R commands.
- Once an experienced data manager prepares a set of commands, even operators who do not have any experience of R programming can just use power of R.
- The software is licensed under GNU LGPL license. We can use the software free for any use.

8 Import dataset by R AnalyticFlow

- Each connected icon has a function. Parameters can be set by user.
- Each icon has corresponding R code.

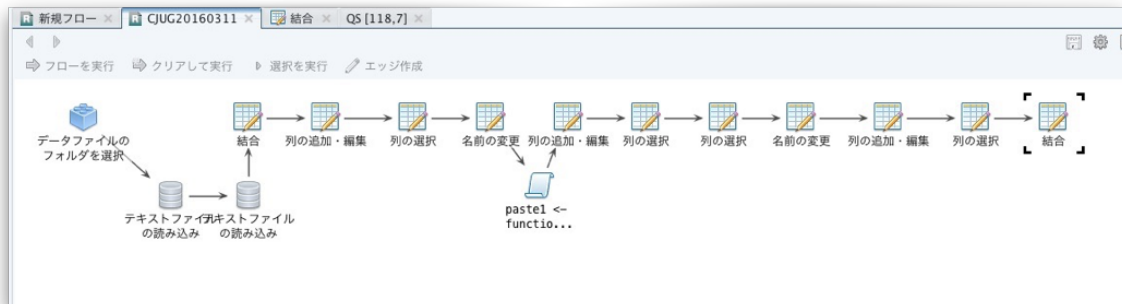


図 1 R AnalyticFlow Screenshot

9 Merge tables by R AnalyticFlow

- Merge two tables

10 Define custom function, Edit variable

- If you have experiences of R programming, you can define new “icon” with your R code.
- With calculations by R built-in functions or your custom functions, you can make a new variable.

11 Write reports in R Markdown format

- **R Markdown** is a simple text file format, but R code can be embedded in the file.
- With **knitr** R package or **Rstudio** development software, we can convert R Markdown documents into HTML, Word, PDF, or Presentation like this.
- When converting, the embedded R codes are interpreted, then the results of calculations are also embedded into the final document.
- Now, we do not need to make any ‘temporary files’. Just write a R code to read the data from the original file, and codes to make tables and graphics inside the report.
- By using R markdown, there is no more problems like:
- Original data changed, but the reports are not updated.
- No one remembers how to read original data and update reports in detail.
- Use filenames like ‘Analysis_2017Feb22.RData’ to control version of temporary analysis files.

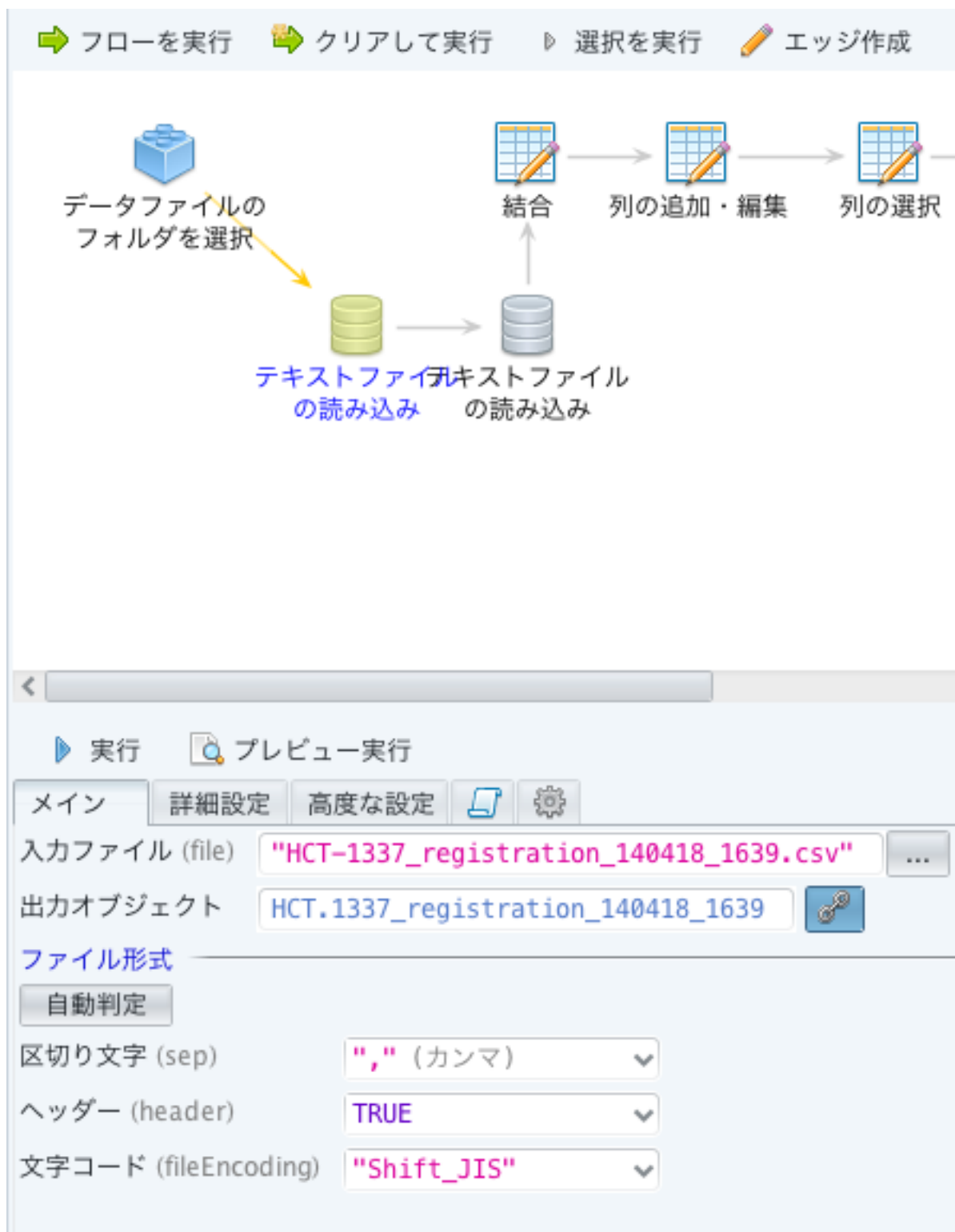


図2 Read CSV by R AnayticFlow

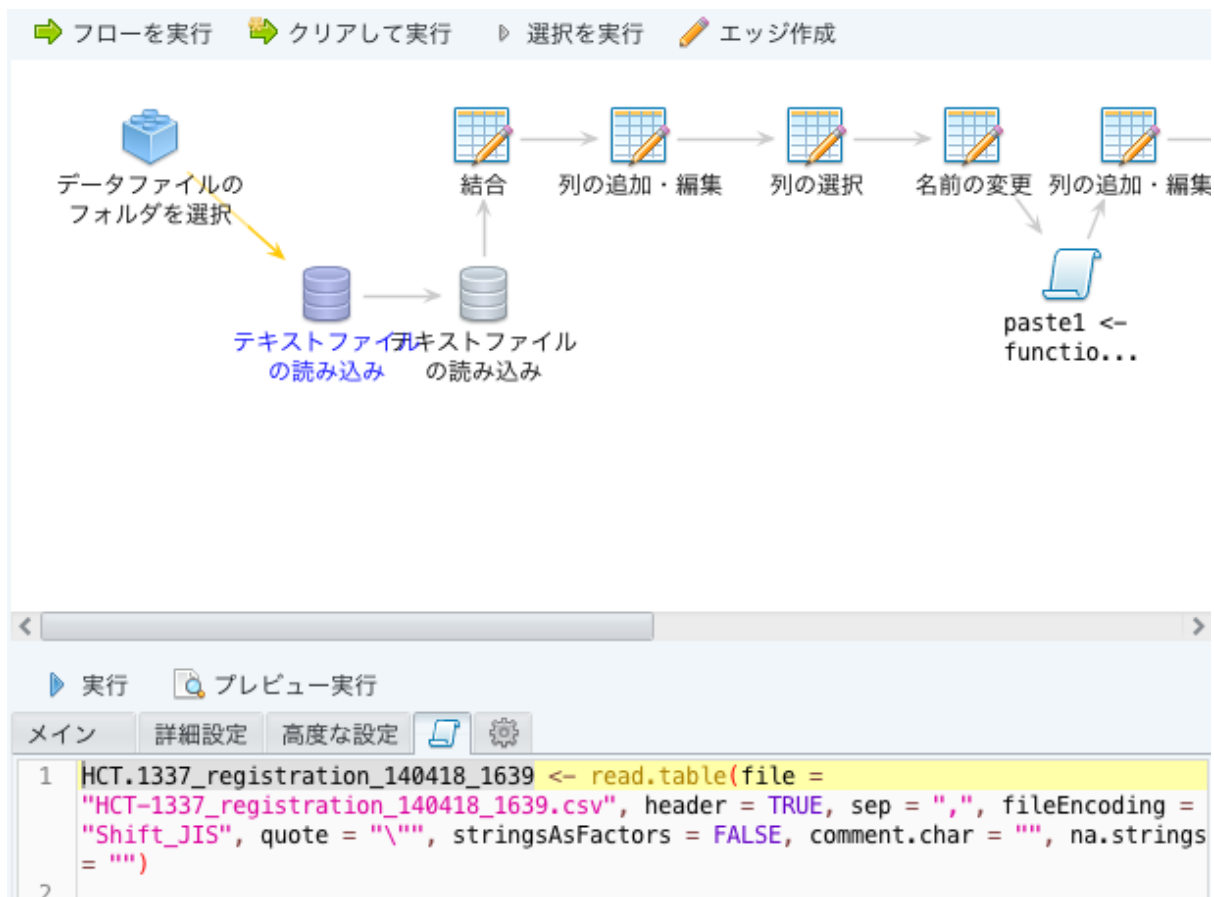


図3 Read CSV command by R AnayticFlow

12 Embedding R code in your report document

Like that:

The code inside backquotes are interpreted when converting document into HTML/Word/PDF, and the result(string representation of CM variable) will be displayed in the final document even there is no actual data in the R markdown document.

From the Rmd document below:

```
Currently, there are `r length(unique(DM$USUBJID))` subjects. Number of non-treated arm is `r length`
```

You will get this Word file:

```
Currently, there are 32 subjects. Number of non-treated arm is 8.
```

This technique will improve **reproducibility** of the document, because the actual data is not written in document. Every time we need the values, it will be re-calculated from the original dataset. This concept is known as 'Reproducible Research'.

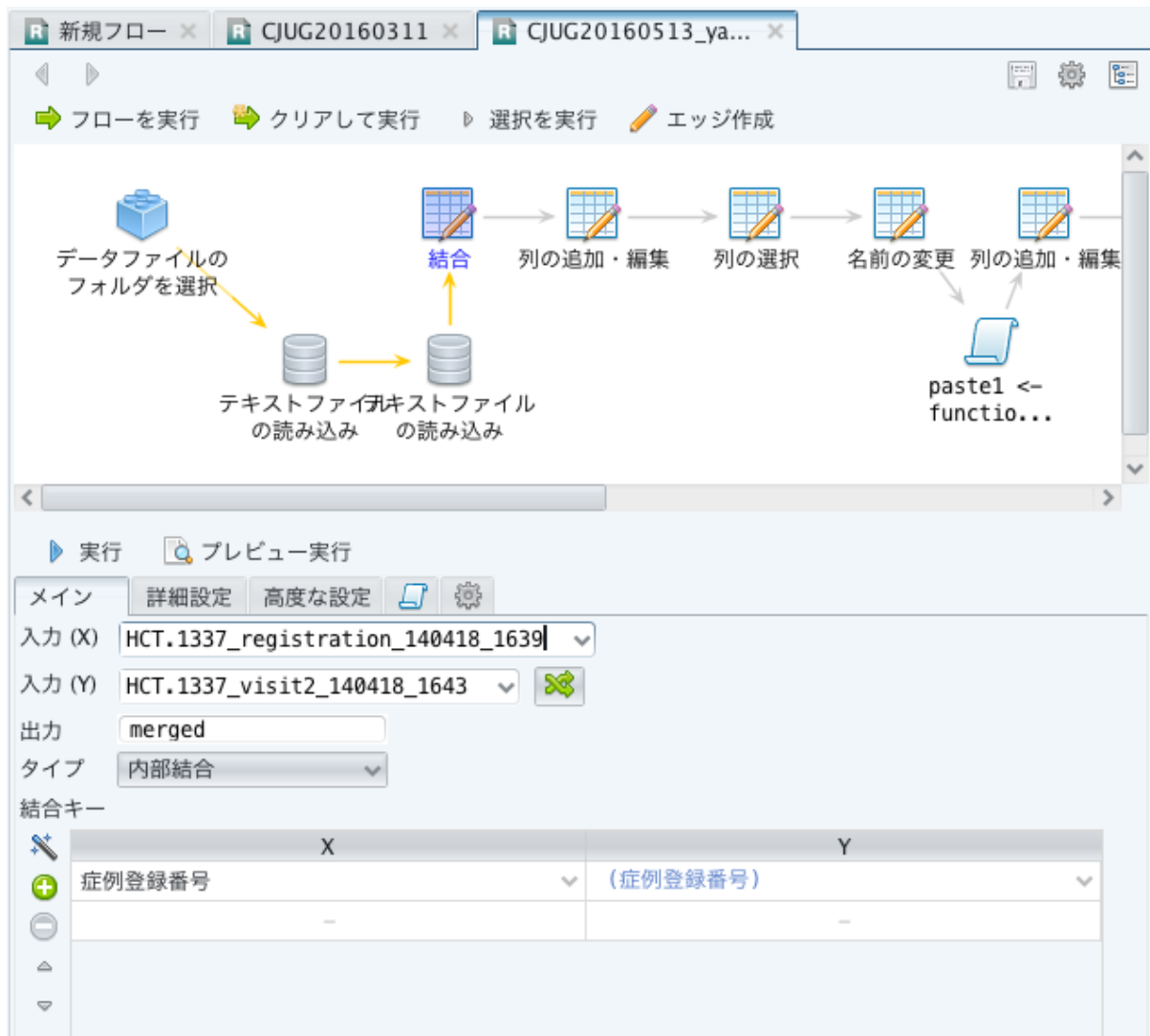


図 4 Merge tables by R AnalyticFlow

13 Version control of R script and R markdown document

- All activities in R Team are stored in the GitHub repository. <http://github.com/mokjpn/cjugar/>
- Git is a version control system, mainly for text files. R source code, R markdown documents, or any XML documents can be well managed by git.
- Git focuses on collaboration of programmers. All members of a project can copy('clone') the current version, then make some changes, and merge('commit') the changes. If there is any conflict of changes, members can discuss which commit is acceptable.
- All change logs are preserved, so we can back into any old revision at any time.
- GitHub is a web-based hosting service of Git Repository, with many original features that help collaboration.

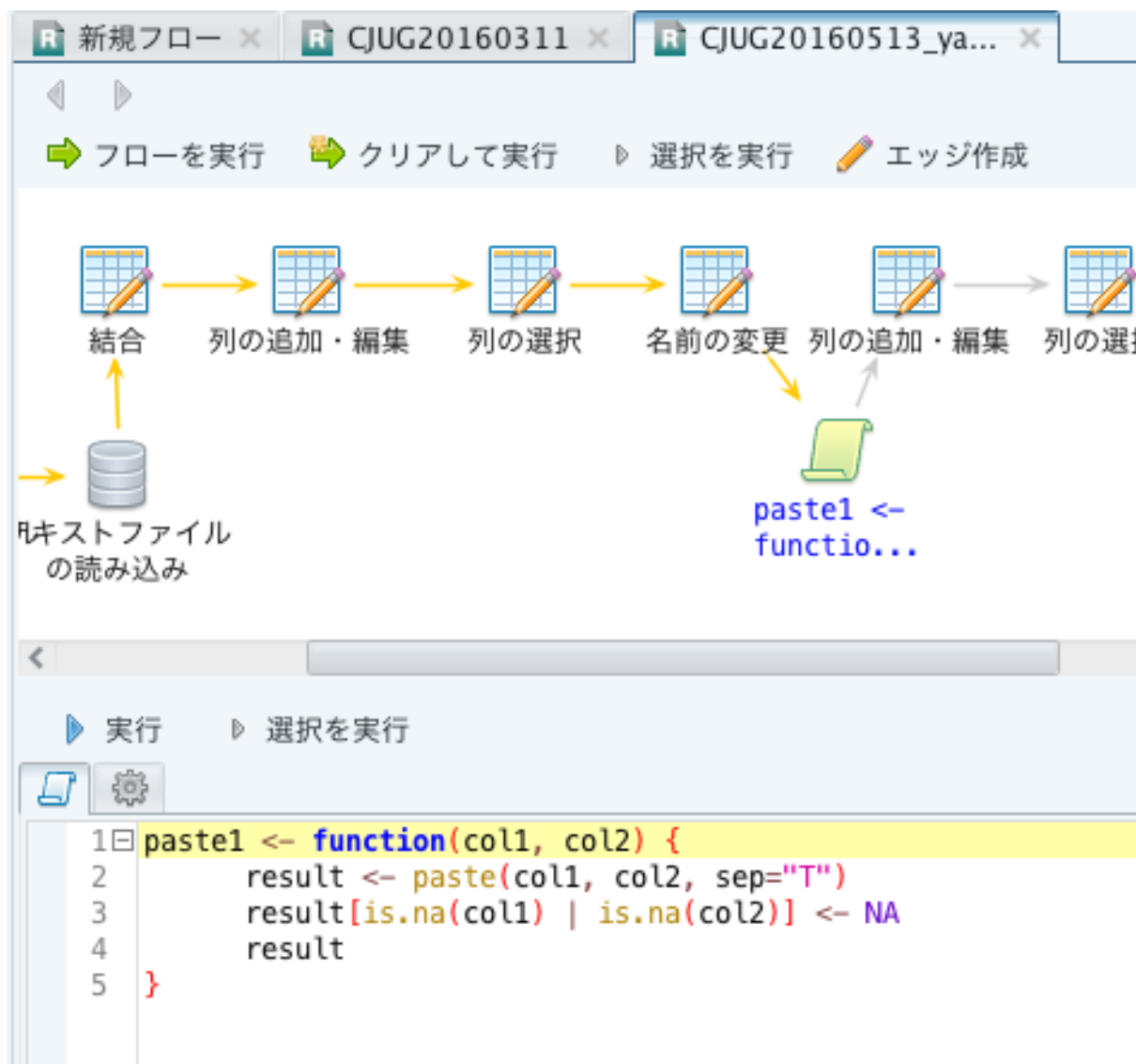


図 5 Custom Function in R AnalyticFlow

14 Summary

- CJUG SDTM+R Team members have learnt R language environment and related software, and discussed how to make data managers' work efficient with R.
- We recommend "R AnalyticFlow" tool as an efficient tool to compose everyday workflow with the power of R language.
- For the purpose of authoring data reports, we recommend using "R Markdown" format, to keep authoring work simple and improve reproducibility of documents.
- Also we recommend git version control system to manage revisions of R script source and R markdown document, that enables collaboration of statistical programmers and skillful data managers.

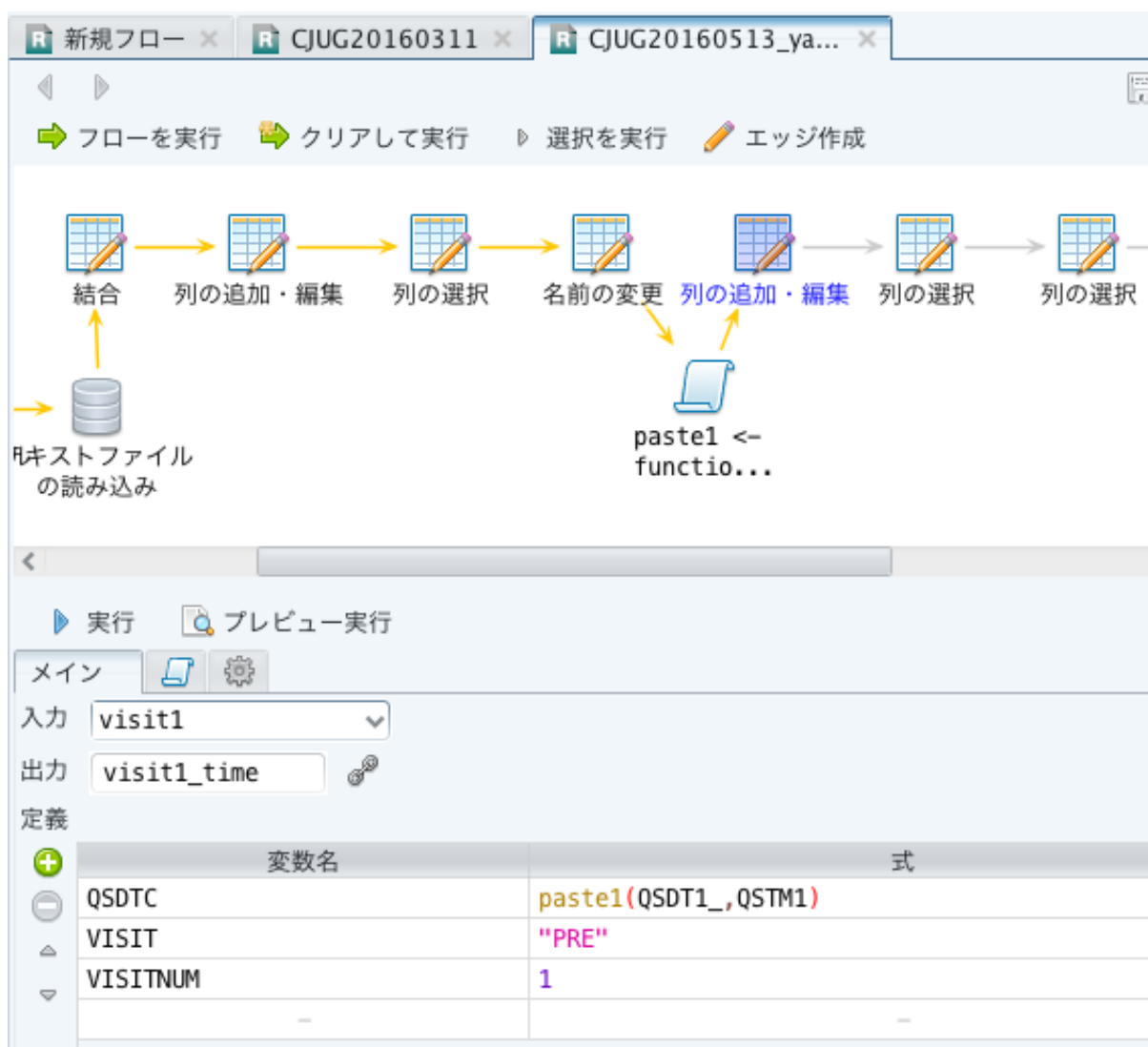


図6 Make new variables by R AnalyticFlow