Source Augmentation Script ReadMe

Cynthia Yeh

April 12 2021

Contents

Before getting started, the essential files are:

1. In r folder

augment.source 03032021_v3.R

source.functions.R

1. In csv folder

source_data

data_augmented_xl

Please make sure that there are no blanks in source.numbers column and the numbers are matched between Outcomes DB and Source DB. :)

note: The below example is from AMD source DB.

DEFAULTS

Dependencies

Please install dplyr, purrr, tidyverse, lubridate before running the script.

```r
#install.packages("dplyr")
#install.packages("purrr")
#install.packages("tidyverse")
#install.packages("lubridate")
library(dplyr)
library(purrr)
library(tidyverse)
library(lubridate)
```

Directories

Set up directories and make sure it is in the current database r folder

```r
getwd()
```

```
## [1] "C:/Users/cyeh/OneDrive - Certara/Desktop/Augmentation Practice/r"
```

```r
dirSource <- '../csv'
dirResults <- '../csv'
sourcedb <- 'source_data.csv'
augmentdb<-'data_augmented_xl.csv'
```

DATA IMPORT

Import the data from csv folder and cleaning the Outcomes and Source DB.

Check the original data dimensions

```r
dim(OutcomesDB)
dim(SourceDB)
```

```
## [1] 9595  300
## [1] 375  25
```

DATA PREP

Create SourceDB dataset

Convert date columns to desired format using lubridate package

```r
datSource$date.transform <- mdy(datSource$search.date)
```

Remove internal columns

For example, the last 6 columns for the database are listed below: comment, e.copy, database, entry, qc, date.transform. Here, are internal columns that need to be removed.

Please change colnames in internalCols to remove internal columns (ignore date.transform column).

```
internalCols <- c("entry", "qc")
```

The updated columns are listed here: control, n.patients, n.arm, comment, e.copy, database. In most cases, **database** should be the last column in the source DB.

Subset Outcomes DB

```
OutcomesDB <- OutcomesDB %>%
  select(url:publication.type,
         study,
         registry.study.id,
         n.study,
         arm,
         control.arm,
         randomized.drug)
```

DATA FORMAT CHECKS

In this section, we will check if there are any missing columns, unallowed non-numeric columns, and unify the columns format.

Example output as below (table 1):

```
checkReqSourceCols(datSource)
```

```
## The following 1 columns are missing and required:
```

```
## [1] "n.study"
```

From the message above, please change the n.patients columns to n.study.

```
checkNumSourceCols(datSource)
```

```
## There are no columns with unallowed non-numerics.
```

Uncomment the following 2 lines in the R-script to correct unallowed non-numerics.

```
#x <- c("source.number", "publication.year")
#datSource[x] <- sapply(datSource[x], as.numeric)
```

Table 1: Column Names

c1	c2
search	reason
search.date	reason.descr
url	study
source.number	registry.study.id
authors	treatment
publication.year	control
title	n.study
journal	n.arm
volume	comment
pages	e.copy
abstract	database
include	

Table 2: Named Ranges

reasonType	searchType
blanks	pubmed
indication	clinicaltrials.gov
design	fda
treatment	ema
comparison	conference
endpoints	clinical study report
population	company website
sample size	cross-reference
availability	web search
reliability	
secondary reference	
other	
study duration	

NAMED RANGES

Check if the named ranges are matching (table 2). Please manually change them if there are unallowed names.

Example: This chunk of codes will list out unallowed named ranges. Run it line by line.

```r
unique(datSource$search[!datSource$search %in% searchType])
unique(datSource$include[!datSource$include %in% yesNoUnclear])
unique(datSource$reason[!datSource$reason %in% reasonType])
unique(datSource$e.copy[!datSource$e.copy %in% yesNo])
unique(datSource$database[!datSource$database %in% yesNo])
```

URL HARMONIZATION

It makes the pubmed/clinical.trial urls consistent.

REFERENCES CHECKS

Two messages to read in this section.

1st message: Whether there are references marked as database = yes in Source DB, but they are actually not found in Outcomes DB.

2nd message: Check if all included references are curated or not.

```
checkSourceYes(wDB.no)
```

```
## Perfect! Let's go.
```

```
checkSourceNo(sDB.include.yes)
```

```
## Great!!! All included references are curated.
```

CREATE DATASETS

This section will reshape the dataset, check if we have **any references that have more than 1 study**, and will also combine the information for multiple studies in single references.

Identify the source numbers for multiple studies references

```
checkMultiStudy(datUniRef)
```

```
## The following source numbers have more than 1 study

##  [1]       14       15       17       64 22555112 23084240 28779006 30768224
##  [9] 30986442 31791663 32574761
```

After data wrangling, the first dimension of **datUniRef** and **datUniTrt** datasets should be the same.

```
dim(datUniRef)
```

```
## [1] 91 16
```

```
dim(datUniCtrl)
```

```
## [1] 91  2
```

```
dim(datUniTrt)
```

```
## [1] 91  2
```

FILL IN MISSING INFORMATION

This part, we will overwrite Source database by gathering information from augmented Outcomes database. After this step, data be lined up between databases.

Some databases, oncology database especially, there is no control arm due to single arm study design. Here, we will fill in control column of the Source DB as **no control**.

```
datUniRef.join$control.join[is.na(datUniRef.join$control.join)] <- "no control"
```

CHECK INFORMATION FOR REFERENCES MORE THAN 1 STUDY

A glimpse of combined information for multiple studies references.

DATA FINAL CLEANING

Finally, we clean the data again and write the output to **source_augmented.csv** file.

```
# Unify study name format (remove quotation marks)
datSource$study <- gsub("=|\"", "", datSource$study)

datSource <- select(datSource, !ends_with("join"))
write.csv(datSource, paste(dirResults, "source_augmented.csv", sep = "/"), na = "",
    row.names = FALSE)
```

MESSAGES

Just a second. we still have a few more friendly reminders here. In this section, we can check if there are any more filling needs to be done manually.

```
## Please manually fill back info for following 2 references: 71, 70
```

```
## No missing URLs. You are good to go!
```

The augmentation is complete. Thanks! :)