Join SDTM with its SUPP dataset (function)

Noory Kim

started 2025-05-19, updated 2025-05-22

```r
library(safetyData) # a package with sample CDISC data sets
library(tidyverse)
```

NOTES:

- with tidyverse, column names are case sensitive
- tidyr
  - pivot_wider() : transpose from long to wide
  - mutate() : change column from character to numeric
- dplyr
  - select()
    - keep/drop columns
      - select(-any_of(vars)) : drop columns
    - reorder columns
      - everything() : columns not explicitly named
  - slice_head(n=) : keep the first n rows
- References:
  - select(-any_of(vars)): https://tidyselect.r-lib.org/reference/all_of.html

UPDATES

- 2025-05-22: Added parameters ds_name and supp_name to the function

```r
## Get list of datasets
dataset_list <- data(package = "safetyData")$results[ , "Item"]
dataset_list
```

```
##  [1] "adam_adae"     "adam_adlbc"    "adam_adlbh"    "adam_adlbhy"
##  [5] "adam_adqsadas" "adam_adqscibc" "adam_adqsnpix" "adam_adsl"
##  [9] "adam_adtte"    "adam_advs"     "sdtm_ae"       "sdtm_cm"
## [13] "sdtm_dm"       "sdtm_ds"       "sdtm_ex"       "sdtm_lb"
## [17] "sdtm_mh"       "sdtm_qs"       "sdtm_relrec"   "sdtm_sc"
## [21] "sdtm_se"       "sdtm_suppae"   "sdtm_suppdm"   "sdtm_suppds"
## [25] "sdtm_supplb"   "sdtm_sv"       "sdtm_ta"       "sdtm_te"
## [29] "sdtm_ti"       "sdtm_ts"       "sdtm_tv"       "sdtm_vs"
```

The function

Parameters

- domain: SDTM domain abbreviation

- ds_name: name of main dataset
- supp_name (default: NULL): name of existing SUPP– dataset

```
join_domain_with_supp_by_seq <- function(domain, ds_name, supp_name=NULL){

## concatenate string to get name of datasets to join and of the SEQ variable
  name_main <- paste0(ds_name)

  if(!is.null(supp_name)){
    name_supp <- paste0(supp_name)
    name_seq  <- paste0(toupper(domain), "SEQ")
  }

## get main dataset
  if(is.null(supp_name)){
    output <- get(name_main) %>%
      as_tibble()
  }
  else if(!is.null(supp_name)){
    main <- get(name_main) %>%
      as_tibble() %>%
      rename("SEQ"=name_seq) ## rename --SEQ as SEQ, to simplify join statement below

    ## Columns in SUPP-- not needed for ADaMs or TLFs
    cols_to_drop <- c("STUDYID", "RDOMAIN", "IDVAR", "QLABEL", "QORIG", "QEVAL")

    ## get SUPP-- dataset and transpose
    supp_t <- get(name_supp) %>%
      as_tibble() %>%
      pivot_wider(names_from=QNAM, values_from=QVAL) %>%
      mutate(SEQ = as.numeric(IDVARVAL)) %>%
      select(-any_of(cols_to_drop)) %>%
      select(SEQ, everything())

    ## join datasets
    output <- left_join(main, supp_t, by=c("USUBJID"="USUBJID", "SEQ"="SEQ")) %>%
      rename(!!name_seq := "SEQ") ## rename SEQ back to --SEQ, now that the join is done
  }

  ## output result
  return(output)
}
```

Function calls

DS domain

```
ds <- join_domain_with_supp_by_seq(domain = "ds", ds_name = "sdtm_ds", supp_name = "sdtm_suppds")
```

```
## Warning: Using an external vector in selections was deprecated in tidyselect 1.1.0.
## i Please use `all_of()` or `any_of()` instead.
##   # Was:
##   data %>% select(name_seq)
```

```
##
##   # Now:
##   data %>% select(all_of(name_seq))
##
## See <https://tidyselect.r-lib.org/reference/faq-external-vector.html>.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```

Warning appears even though any_of() is being used.

Documented workaround shows an example for all_of(): https://tidyselect.r-lib.org/reference/faq-external-vector.html

```r
# show the first few rows of a tibble
ds %>%
  slice_head(n=3)
```

```
## # A tibble: 3 x 15
##   STUDYID  DOMAIN USUBJID DSSEQ DSSPID DSTERM DSDECOD DSCAT VISITNUM VISIT DSDTC
##   <chr>    <chr>  <chr>   <dbl>  <int> <chr>  <chr>   <chr>    <dbl> <chr> <chr>
## 1 CDISCPI~ DS     01-701~     1     NA PROTO~ COMPLE~ DISP~       13 WEEK~ 2014~
## 2 CDISCPI~ DS     01-701~     2     NA FINAL~ FINAL ~ OTHE~       13 WEEK~ 2014~
## 3 CDISCPI~ DS     01-701~     1     24 ADVER~ ADVERS~ DISP~        5 WEEK~ 2012~
## # i 4 more variables: DSSTDTC <chr>, DSSTDY <int>, IDVARVAL <int>,
## #   ENTCRIT <int>
```

```r
# show as a dataframe rather than as a tibble (which gets truncated)
ds %>%
  as.data.frame() %>%
  head(3)
```

```
##        STUDYID DOMAIN     USUBJID DSSEQ DSSPID             DSTERM
## 1 CDISCPILOT01     DS 01-701-1015     1     NA PROTOCOL COMPLETED
## 2 CDISCPILOT01     DS 01-701-1015     2     NA    FINAL LAB VISIT
## 3 CDISCPILOT01     DS 01-701-1023     1     24      ADVERSE EVENT
##           DSDECOD             DSCAT VISITNUM   VISIT            DSDTC
## 1       COMPLETED DISPOSITION EVENT       13 WEEK 26       2014-07-02
## 2 FINAL LAB VISIT       OTHER EVENT       13 WEEK 26 2014-07-02T11:45
## 3   ADVERSE EVENT DISPOSITION EVENT        5  WEEK 4       2012-09-02
##      DSSTDTC DSSTDY IDVARVAL ENTCRIT
## 1 2014-07-02    182       NA      NA
## 2 2014-07-02    182       NA      NA
## 3 2012-09-02     29       NA      NA
```

AE domain, with SUPPAE

```r
ae <- join_domain_with_supp_by_seq(domain = "ae", ds_name = "sdtm_ae", supp_name = "sdtm_suppae")

# show as a dataframe rather than as a tibble (which gets truncated)
ae %>%
  as.data.frame() %>%
  head(3)
```

```
##        STUDYID DOMAIN     USUBJID AESEQ AESPID                    AETERM
## 1 CDISCPILOT01     AE 01-701-1015     1    E07 APPLICATION SITE ERYTHEMA
## 2 CDISCPILOT01     AE 01-701-1015     2    E08 APPLICATION SITE PRURITUS
## 3 CDISCPILOT01     AE 01-701-1015     3    E06                 DIARRHOEA
##                       AELLT AELLTCD                   AEDECOD AEPTCD    AEHLT
## 1 APPLICATION SITE REDNESS      NA APPLICATION SITE ERYTHEMA     NA HLT_0617
## 2 APPLICATION SITE ITCHING      NA APPLICATION SITE PRURITUS     NA HLT_0317
## 3                 DIARRHEA      NA                 DIARRHOEA     NA HLT_0148
##   AEHLTCD    AEHLGT AEHLGTCD
## 1      NA HLGT_0152       NA
## 2      NA HLGT_0338       NA
## 3      NA HLGT_0588       NA
##                                               AEBODSYS AEBDSYCD
## 1 GENERAL DISORDERS AND ADMINISTRATION SITE CONDITIONS       NA
## 2 GENERAL DISORDERS AND ADMINISTRATION SITE CONDITIONS       NA
## 3                           GASTROINTESTINAL DISORDERS       NA
##                                                  AESOC AESOCCD AESEV AESER
## 1 GENERAL DISORDERS AND ADMINISTRATION SITE CONDITIONS      NA  MILD     N
## 2 GENERAL DISORDERS AND ADMINISTRATION SITE CONDITIONS      NA  MILD     N
## 3                           GASTROINTESTINAL DISORDERS      NA  MILD     N
##   AEACN    AEREL                      AEOUT AESCAN AESCONG AESDISAB AESDTH
## 1    NA PROBABLE NOT RECOVERED/NOT RESOLVED      N       N        N      N
## 2    NA PROBABLE NOT RECOVERED/NOT RESOLVED      N       N        N      N
## 3    NA   REMOTE         RECOVERED/RESOLVED      N       N        N      N
##   AESHOSP AESLIFE AESOD      AEDTC    AESTDTC    AEENDTC AESTDY AEENDY IDVARVAL
## 1       N       N     N 2014-01-16 2014-01-03       <NA>      2     NA        1
## 2       N       N     N 2014-01-16 2014-01-03       <NA>      2     NA        2
## 3       N       N     N 2014-01-16 2014-01-09 2014-01-11      8     10        3
##   AETRTEM
## 1       Y
## 2       Y
## 3       Y
```

QS domain, without SUPP–

```r
qs <- join_domain_with_supp_by_seq(domain = "qs", ds_name = "sdtm_qs")

# show as a dataframe rather than as a tibble (which gets truncated)
qs %>%
  as.data.frame() %>%
  head(3)
```

```
##        STUDYID DOMAIN     USUBJID QSSEQ QSTESTCD           QSTEST
## 1 CDISCPILOT01     QS 01-701-1015  5001  ACITM01 WORD RECALL TASK
## 2 CDISCPILOT01     QS 01-701-1015  5016  ACITM01 WORD RECALL TASK
## 3 CDISCPILOT01     QS 01-701-1015  5031  ACITM01 WORD RECALL TASK
##                                  QSCAT QSSCAT QSORRES QSORRESU QSSTRESC
## 1 ALZHEIMER'S DISEASE ASSESSMENT SCALE   <NA>       3     <NA>        3
## 2 ALZHEIMER'S DISEASE ASSESSMENT SCALE   <NA>       2     <NA>        2
## 3 ALZHEIMER'S DISEASE ASSESSMENT SCALE   <NA>       4     <NA>        4
##   QSSTRESN QSSTRESU QSBLFL QSDRVFL VISITNUM    VISIT VISITDY      QSDTC QSDY
## 1        3     <NA>      Y    <NA>        3 BASELINE       1 2014-01-02    1
## 2        2     <NA>   <NA>    <NA>        8   WEEK 8      56 2014-03-05   63
## 3        4     <NA>   <NA>    <NA>       10  WEEK 16     112 2014-05-07  126
```