# Example study

## Introduction

This is an example of how datasets can be created using the `amp.dm` package. It mainly shows how quarto (or rmarkdown) can be used to directly create documentation. Also it demonstrate how meta data is handled and used in the documentation and how the analysis functions can help in the creation of a (NONMEM) dataset. To fully follow the workflow, the code blocks are shown. These will typically be hidden for the final documentation. For more information also check the underlying quarto file of this vignette.

## Version history

- v1: First version

## Study Description

This is an adaptation of the original Theophylline dataset where additional subjects, dose arms and covariates are added. With this data, source data is created that can be used to demonstrate how the package works.

## Dataset instructions

This section would normally contain important information about requirements for the NONMEM dataset, such as necessary NONMEM parameters for compartments and dose records.

## Other

Arbitrary sections can be added here to provide additional information. For instance items like assumptions, special attention and data excluded (although this can be added through `cmnt` function as well; see below for examples)

# Data management

## Import source data

For this example all source data is created and saved as SAS export files. Note the `read_data` function and how it logs information (see tables at the end).

```
dm <- read_data(system.file("example/SOURCE/DM.xpt",package="amp.dm"),
                comment = "demographic data")
ex <- read_data(system.file("example/SOURCE/EX.xpt",package="amp.dm"),
                comment = "dosing data")
pc <- read_data(system.file("example/SOURCE/PC.xpt",package="amp.dm"),
                comment = "pk data")
vs <- read_data(system.file("example/SOURCE/VS.xpt",package="amp.dm"),
                comment = "vital signs data")
```

## Demographic data

The demographics age, sex and race are available in the DM domain. Information regarding height, weight and BMI was available in the VS domain. Note functions `filterr`, `left_joinr` and `srce` and how information is logged (see tables at the end)

```
cmnt("There are duplicate subjects, these are excluded in DM domain")
dm1 <- filterr(dm, !duplicated(USUBJID), comment="Dropped duplicate subjects") |>
  mutate(SEX    = ifelse(SEX=='M', 0, 1),
         TRT    = as.numeric(as.factor(ARM)),
         CNTRY = as.numeric(as.factor(COUNTRY)))

vs1  <- tidyr::pivot_wider(vs,names_from = VSTESTCD, values_from = VSSTRESN) |>
  select(-STUDYID)
subj <- left_joinr(dm1,vs1,by='USUBJID',comment = "Combine covariates")
subj <- select(subj, STUDYID, USUBJID, TRT, CNTRY, SEX, AGE, WEIGHT, HEIGHT, BMI)

srce(CNTRY,dm.COUNTRY)
srce(BMI,c(vs.WEIGHT,vs.HEIGHT),'d')
```

Assumptions and special attention:

- There are duplicate subjects, these are excluded in DM domain

## PK observation data

PK observations were taken from the PC domain and adapted for the NONMEM analysis

```
pk <- pc |>
  mutate(variable = "PKSample",
         STIME    = PCTPTNUM,
         dattim   = as.POSIXct(PCDTC, format="%Y-%m-%dT%H:%M:%S"),
         FLAGPK   = case_when(is.na(PCSTRESN) ~ 1, PCSTRESN==0 ~ 2, .default = 3)) |>
  rename(Stime = PCTPTNUM, DV = PCSTRESN) %>% select(-c(PCTESTCD, STUDYID, PCDTC))
```

## Dosing data

The dose data were provided in the EX domain.

```
dose <- ex|>
  mutate(variable = "Dose",
         STIME    = 0,
         dattim   = as.POSIXct(EXSTDTC, format = "%Y-%m-%dT%H:%M:%S")) %>%
  rename(AMT = EXDOSE) %>% select(-c(EXSTDTC, STUDYID))
```

## Combine data

In this part of the dataset the pk and dose records were combined, and the subject covariates included, to create a final NONMEM data set. Note the `time_calc` function here that directly calculates TIME, TAFD and TALD based on date/time.

```
cmnt(paste("For combining PK/observations with demographics,",
           "records in **first** are used (`all.x=TRUE`)"))

nm <- bind_rows(dose,pk) |> left_join(subj, by="USUBJID") |>
  rename(ID=USUBJID) |>
  time_calc(datetime = "dattim") |>
  mutate(STUDYID = as.numeric(as.factor(STUDYID)),
         CMT     = ifelse(variable=="Dose", 1, 2),
         EVID    = ifelse(CMT==1, 1, 0),
         MDV     = ifelse(CMT==2 & FLAGPK==3, 0, 1)) %>%
  select(STUDYID, ID, TRT, CMT, AMT, STIME, TIME, TAFD, TALD, DV, EVID, MDV,
         CNTRY, SEX, AGE, WEIGHT, HEIGHT, BMI, FLAGPK)
```

Assumptions and special attention:

- For combining PK/observations with demographics, records in **first** are used (`all.x=TRUE`)

Note the `attr_xls` function that obtains meta data from an excel file, this is then added to the data using`attr_add`. The `output_data` function can create csv and xpt file output and performs various checks on the data.

```
attr <- attr_xls(system.file("example/Attr.Template.xlsx",package = "amp.dm"))
nmf  <- attr_add(nm, attr)

# Write csv and/or xpt file (notice file is named same as script)
output_data(nmf, csv = paste0(tempdir(),"/",get_script(), ".csv"),
            xpt = paste0(tempdir(),"/",get_script(), ".xpt"),
            attr = paste0(tempdir(),"/",get_script(), ".rds"),
            readonly = TRUE)

# Save current workspace
# lognfo <- get_log()
# save.image(paste0(get_script(),".RData"))
```

## Dataset overview

The tables in this section are useful for reviewing and documenting the data management process. The following functions are used:

- `define_tbl`; Uses the meta data or attributes to create a define table that can also be used for eSubmission
- `stats_df`; Provide simple statistics of available data to spot for instance outliers and missing data
- `counts_df`; Provide counts for number of samples and subjects, stratified by variable(s)
- `log_df`; Provide results from functions that log information such as `read_data`, `filterr` or `left_joinr`
- `check_nmdata`; Provide checks for NONMEM data either essential or triggering further investigation
- `session_tbl`; Provide information on the session that was used to run the code

### Dataset define

```
define_tbl(attr, ret="tbl")
```

Table 1: Dataset define form

| Data.Item | Description | Unit | Remark |
|---|---|---|---|
| STUDYID | Study identifier | - | - |
| ID | Unique subject identifier | - | - |
| TRT | Treatment arm | - | 1 = 300 mg theoph form 1, 2 = 300 mg theoph form 2 |
| CMT | Compartment | - | 1 = Dosing compartment, 2 = Central compartment |
| AMT | Amount administered | mg | Original dose units set to mg |
| STIME | Scheduled time | h | - |
| TIME | Time | h | - |
| TAFD | Time after first dose | h | - |
| TALD | Time after last dose | h | - |
| DV | Concentration Theoph | ng/mL | - |
| EVID | Event ID | - | 0 = Observations, 1 = Dosing event |
| MDV | missing DV | - | 0 = Other, 1 = Dose records and missing observations |
| CNTRY | Country | - | 1 = BEL, 2 = FRA, 3 = GER, 4 = NED [source: dm.COUNTRY (copied)] |
| SEX | Gender | - | 0 = Male, 1 = Female |
| AGE | Age | y | - |
| WEIGHT | Weight | kg | - |
| HEIGHT | Height | m | - |
| BMI | Body mass index | kg/m2 | - [source: vs.WEIGHT, vs.HEIGHT (derived)] |

Table 1: Dataset define form

| Data.Item | Description | Unit | Remark |
|---|---|---|---|
| FLAGPK | Flag for type of PK record | - | 1 = Missing PK, 2 = PK below LOQ, 3 = Valid PK sample |

```r
# We could directly output a define.pdf
# define_tbl(attr, outnm= paste0("define.",get_script(), ".tex"),
#             show=FALSE, title="Data define overview")
```

## Dataset statistics

```r
stats_df(nmf,size="\\footnotesize")
```

Table 2: Statistics data frame

| Variable | Min | Max | Categories | Nna | Nmiss | Max-Char | Type |
|---|---|---|---|---|---|---|---|
| STUDYID | 1 | 1 | 1 | 0 [0%] | 0 [0%] | 1 | numeric |
| ID | 1 | 12 | More than 10 cats (12) | 0 [0%] | 0 [0%] | 2 | numeric |
| TRT | 1 | 2 | 2 / 1 | 0 [0%] | 0 [0%] | 1 | numeric |
| CMT | 1 | 2 | 2 / 1 | 0 [0%] | 0 [0%] | 1 | numeric |
| AMT | 300 | 300 | NA / 300 | 132 [91.7%] | 132 [91.7%] | 3 | numeric |
| STIME | 0 | 24 | More than 10 cats (11) | 0 [0%] | 0 [0%] | 4 | numeric |
| TIME | 0 | 25.1 | More than 10 cats (79) | 0 [0%] | 0 [0%] | 5 | numeric |
| TAFD | -0.5 | 24.6 | More than 10 cats (79) | 12 [8.33%] | 12 [8.33%] | 5 | numeric |
| TALD | -0.5 | 24.6 | More than 10 cats (79) | 12 [8.33%] | 12 [8.33%] | 5 | numeric |
| DV | 0 | 11.4 | More than 10 cats (111) | 14 [9.72%] | 14 [9.72%] | 5 | numeric |
| EVID | 0 | 1 | 0 / 1 | 0 [0%] | 0 [0%] | 1 | numeric |
| MDV | 0 | 1 | 1 / 0 | 0 [0%] | 0 [0%] | 1 | numeric |
| CNTRY | 1 | 4 | 3 / 1 / 4 / 2 | 0 [0%] | 0 [0%] | 1 | numeric |
| SEX | 0 | 1 | 1 / 0 | 0 [0%] | 0 [0%] | 1 | numeric |
| AGE | 24 | 61 | 60 / 47 / 53 / 26 / 59 / 27 / 45 / 61 / 57 / 24 | 0 [0%] | 0 [0%] | 2 | numeric |
| WEIGHT | 54.6 | 86.4 | More than 10 cats (11) | 0 [0%] | 0 [0%] | 4 | numeric |
| HEIGHT | 1.68 | 1.91 | 1.88 / 1.87 / 1.91 / 1.71 / 1.68 / 1.7 / 1.85 / 1.79 | 0 [0%] | 0 [0%] | 4 | numeric |
| BMI | 17 | 29.9 | More than 10 cats (12) | 0 [0%] | 0 [0%] | 5 | numeric |
| FLAGPK | 1 | 3 | 2 / NA / 3 / 1 | 12 [8.33%] | 12 [8.33%] | 1 | numeric |

```r
# Example for a counts table
nmf2 <- attr_factor(nmf)
counts_df(nmf2, by=c("STUDYID","TRT"), id="ID", capt="Number of records by study")
```

Table 3: Number of records by study

| STUDYID | TRT | Nobs | PERCobs | Nid | PERCid |
|---|---|---|---|---|---|
| 1 | 300 mg theoph form 1 | 72 | 50.00 | 6 | 50.00 |
| 1 | 300 mg theoph form 2 | 72 | 50.00 | 6 | 50.00 |
| Total | Total | 144 | 100.00 | 12 | 100.00 |

## Overview of data read in, excluded or merged

```
all_log   <- get_log()

log_df(get_log(), "read_nfo", ret="tbl", capt="Overview of data read-in")
```

Table 4: Overview of data read-in

| Data in | Num rows Data in | Num cols Data in | Comment |
|---|---|---|---|
| `C:/Rlibs/amp.dm/example/SOURCE/DM.xpt` | 13 | 6 | demographic data |
| `C:/Rlibs/amp.dm/example/SOURCE/EX.xpt` | 12 | 4 | dosing data |
| `C:/Rlibs/amp.dm/example/SOURCE/PC.xpt` | 132 | 6 | pk data |
| `C:/Rlibs/amp.dm/example/SOURCE/VS.xpt` | 39 | 4 | vital signs data |

```
log_df(get_log(), "filterr_nfo", ret="tbl", capt="Overview of data excluded")
```

Table 5: Overview of data excluded

| Data in | Num rows Data in | Num rows Data out | Num rows dropped | Reason for filter |
|---|---|---|---|---|
| dm | 13 | 12 | 1 | Dropped duplicate subjects |

```
log_df(get_log(), "joinr_nfo", ret="tbl", capt="Overview of joined data")
```

Table 6: Overview of joined data

| Data in L | Data in R | Num rows Data in L | Num rows Data in R | Rows only in L | Rows Data out | Reason for join |
|---|---|---|---|---|---|---|
| dm1 | vs1 | 12 | 13 | 0 | 12 | Combine covariates |

## Check for common errors/mistakes

```
check_nmdata(paste0(tempdir(),"/",get_script(), ".csv"), ret="tbl",type=1)
```

Table 7: Result of essential data checks

| Check | result |
|---|---|
| CSV is readable | Yes |
| More than 1 line of data | Yes |
| More than 1 data item | Yes |
| First name set (row.names set to FALSE) | Yes |
| Variables ID, TIME and DV present in data | Yes |
| Data ordered on ID and TIME | Yes |
| Quotes not present around data items | Yes |

```
check_nmdata(paste0(tempdir(),"/",get_script(), ".csv"), ret="tbl",type=2)
```

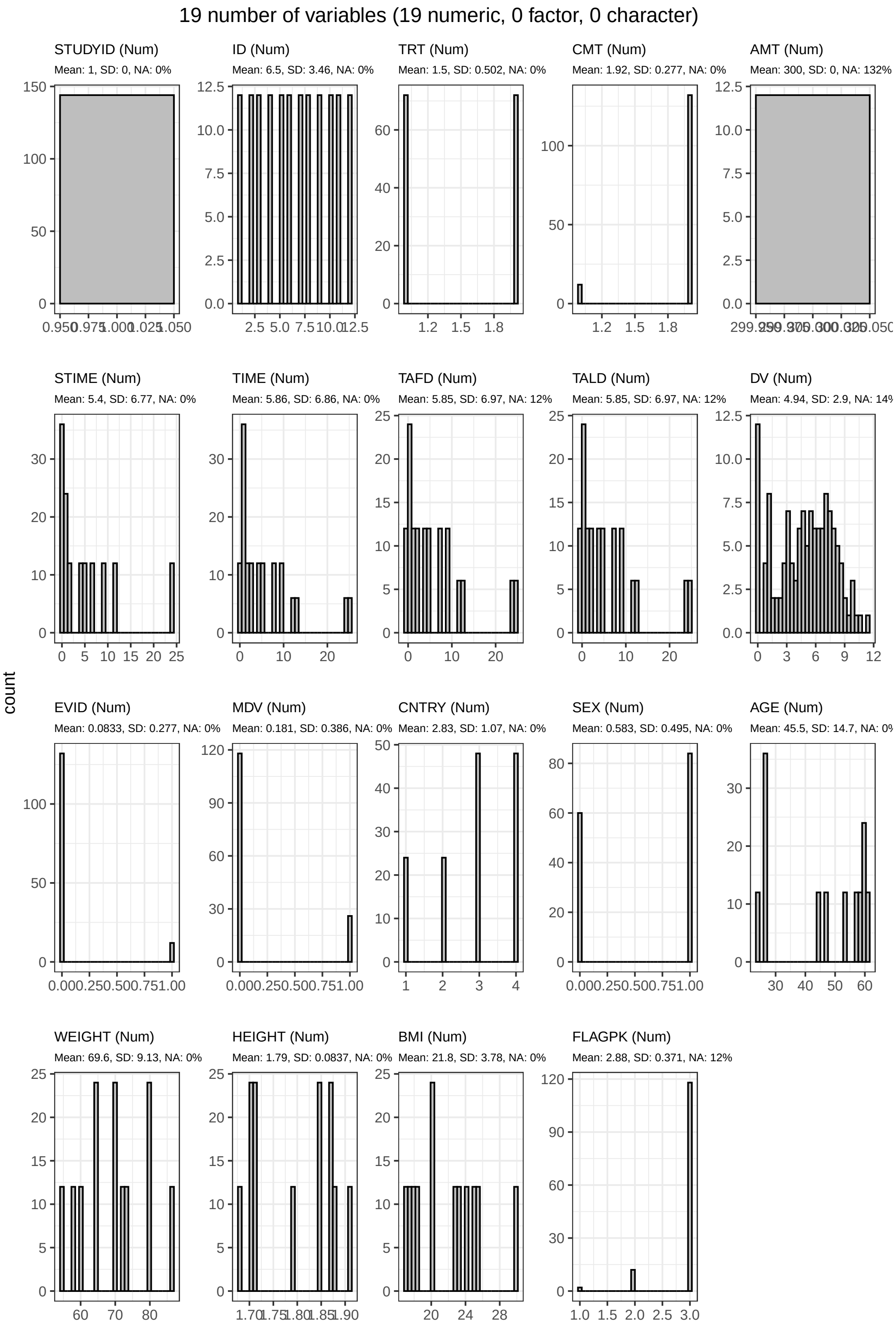
Table 8: Result of non essential data checks

| Check | result |
|---|---|
| There are no NA values in data (excluding DV column) | No |
| All non-NA DV values have MDV=0 | No |
| All observation records with DV=0 have MDV=0 | No |
| Positive DV not present at t=0 | Yes |
| There are no dose records with AMT=0 | Yes |
| Default NM variables present | MDV, CMT, EVID, AMT |
| Default NM variables not present | RECN, DOSE, RATE |
| All variable names less than 8 char | yes |
| All variable contents less than 14 char | yes |
| ID variable has less than 5 characters (consider NONMEM FORMAT option if no) | yes |

## graphical representation

```
plot_vars(nm,ppp=20)
```

```
$`1`
```


19 number of variables (19 numeric, 0 factor, 0 character)
count
STUDYID (Num)
Mean: 1, SD: 0, NA: 0%
ID (Num)
Mean: 6.5, SD: 3.46, NA: 0%
TRT (Num)
Mean: 1.5, SD: 0.502, NA: 0%
CMT (Num)
Mean: 1.92, SD: 0.277, NA: 0%
AMT (Num)
Mean: 300, SD: 0, NA: 132%
STIME (Num)
Mean: 5.4, SD: 6.77, NA: 0%
TIME (Num)
Mean: 5.86, SD: 6.86, NA: 0%
TAFD (Num)
Mean: 5.85, SD: 6.97, NA: 12%
TALD (Num)
Mean: 5.85, SD: 6.97, NA: 12%
DV (Num)
Mean: 4.94, SD: 2.9, NA: 14%
EVID (Num)
Mean: 0.0833, SD: 0.277, NA: 0%
MDV (Num)
Mean: 0.181, SD: 0.386, NA: 0%
CNTRY (Num)
Mean: 2.83, SD: 1.07, NA: 0%
SEX (Num)
Mean: 0.583, SD: 0.495, NA: 0%
AGE (Num)
Mean: 45.5, SD: 14.7, NA: 0%
WEIGHT (Num)
Mean: 69.6, SD: 9.13, NA: 0%
HEIGHT (Num)
Mean: 1.79, SD: 0.0837, NA: 0%
BMI (Num)
Mean: 21.8, SD: 3.78, NA: 0%
FLAGPK (Num)
Mean: 2.88, SD: 0.371, NA: 12%

## Session table

```
session_tbl()
```

Table 9: Session info

| parameter | value |
|---|---|
| R version | R version 4.5.2 (2025-10-31 ucrt) |
| System | x86_64-w64-mingw32/x64 |
| OS | Windows 11 x64 (build 26200) |
| Base packages | stats, graphics, grDevices, utils, datasets, methods, base |
| Other packages | amp.dm (0.2.0), dplyr (1.1.4) |
| Logged in User | r.hooijmaijers |
| Machine | LAPPC149 |
| Time | 2026-03-09 16:48:01.880227 |