EHR Vignette for *Build-PK-Oral*

Introduction

The `EHR` package provides several modules to perform diverse medication-related studies using data from electronic health record (EHR) databases. Especially, the package includes modules to perform pharmacokinetic/pharmacodynamic (PK/PD) analyses using EHRs, as outlined in Choi *et al.*[1], and additional modules will be added in the future. This vignette describes one of the PK data building modules in the system for medications that are typically orally administrated. One of the key data for this module is drug dose data that can be provided by users or generated from unstructured clinical notes using extracted dosing information with the *Extract-Med* module and processed with the *Pro-Med-NLP* module in the system.

```
library(EHR)
```

Build-PK-Oral

Population PK datasets require a certain format in order to be analyzed by software systems specialized for PK analysis such as NONMEM. The Build-PK-Oral module requires dose data and concentration data. Demographic data (provided by users or the *Pro-Demographic* module) and laboratory data (provided by users or the *Pro-Laboratory* module) may optionally be included. Building PK datasets from EHR-extracted information may require some assumptions regarding dosing. The major functions performed in this module by `run_Build_PK_Oral()` are:

- Build PK data for orally administered medications using assumed dosing schedule when last-dose times are not provided.
- Build PK data for orally administered medications using last-dose times if they are provided or extracted from clinical notes in EHRs.

An example data pre-processed from EHR-extracted data follows:

```
# Data generating function for examples
mkdat <- function() {
  npat=3
  visits <- floor(runif(npat, min=2, max=6))
  id <- rep(1:npat, visits)
  dt <- as.POSIXct(paste(as.Date(sort(sample(700, sum(visits))),
                                 origin = '2019-01-01'), '10:00:00'), tz = 'UTC')
  + rnorm(sum(visits), 0, 1*60*60)
  dose_morn <- sample(c(2.5,5,7.5,10), sum(visits), replace = TRUE)
  conc <- round(rnorm(sum(visits), 1.5*dose_morn, 1),1)
  ld <- dt - sample(10:16, sum(visits), replace = TRUE) * 3600
  ld[rnorm(sum(visits)) < .3] <- NA
  age <- rep(sample(40:75, npat), visits)
  weight <- rep(round(rnorm(npat, 180, 20)),visits)
  hgb <- round(rep(rnorm(npat, 10, 2), visits),1)
  data.frame(id, dt, dose_morn, conc, age, weight, hgb, ld)
}
```

```r
# Make example data
set.seed(30)
dat <- mkdat()
dat
```

```
##   id                  dt dose_morn conc age weight  hgb                  ld
## 1  1 2019-01-30 10:00:00       2.5  1.9  50    197 10.6                <NA>
## 2  1 2019-07-08 10:00:00      10.0 14.8  50    197 10.6                <NA>
## 3  2 2019-09-27 10:00:00       2.5  4.5  56    154  7.8                <NA>
## 4  2 2019-11-01 10:00:00      10.0 14.1  56    154  7.8 2019-10-31 22:00:00
## 5  2 2019-11-12 10:00:00       2.5  4.5  56    154  7.8                <NA>
## 6  3 2020-03-11 10:00:00       2.5  5.2  40    147 10.8 2020-03-11 00:00:00
## 7  3 2020-04-05 10:00:00       5.0  6.4  40    147 10.8                <NA>
## 8  3 2020-06-06 10:00:00       7.5 10.7  40    147 10.8 2020-06-05 19:00:00
```

There are 3 individuals in the dataset, each has a set of EHR-extracted dose and blood concentrations data along with demographic data and information commonly found with laboratory data:

- Subject identification number (`id`)
- Time of concentration measurement (`dt`)
- Dose (`dose_morn`)
- drug concentration (`conc`)
- Age (`age`)
- Weight (`weight`)
- Hemoglobin (`hgb`)
- Last-dose time (`ld`)

All concentrations are being taken in the morning. Given that this is a drug which should be taken orally every 12 hours, we can construct a reasonable dosing schedule which details the amount and timing of each dose.

Assume that individuals take their morning dose of medicine 30 minutes after their blood is drawn for labs to check trough concentrations at clinic visit. All doses following that first one will then occur every 12 hours until the next measured concentration is within 6 to 18 hours (when no extracted last-dose times are provided). The timing of the subsequent concentration will then dictate the next sequence of doses in the same way, and so on until the final extracted concentration.

Additionally, it is reasonable to assume that the individual has been taking the drug of interest prior to their first measured concentration. For this reason assume that there is regular dosing leading up to the first extracted concentration, the duration of which is set by the argument `first_interval_hours`; this duration should be long enough for trough concentrations to reach a steady state.

- `run_Build_PK_Oral` builds a PK dataset by following this logic, given specification of appropriate columns. The `run_Build_PK_Oral` function arguments are as follows:
  - `idCol`: subject identification number
  - `dtCol`: time of concentration measurement
  - `doseCol`: dose
  - `concCol`: drug concentration
  - `ldCol`: last-dose time; the default is `NULL` to ignore
  - `first_interval_hours`: Hours of regular dosing leading up to the first concentration; the default is 336 hours = 14 days
  - `imputeClosest`: Vector of columns for imputation of missing data using last observation carried forward or, if unavailable, next observation propagated backward; the default is `NULL` to ignore

```r
dat2 <- dat[,-8]
# Build PK data without last-dose times
run_Build_PK_Oral(x = dat2,
```

```
                  idCol = "id",
                  dtCol = "dt",
                  doseCol = "dose_morn",
                  concCol = "conc",
                  ldCol = NULL,
                  first_interval_hours = 336,
                  imputeClosest = NULL)
```

```
##    id   time  amt   dv mdv evid addl II               date age weight  hgb
## 1   1    0.0  2.5   NA   1    1   27 12 2019-01-16 10:00:00  50    197 10.6
## 2   1  336.0   NA  1.9   0    0   NA NA 2019-01-30 10:00:00  50    197 10.6
## 3   1  336.5  2.5   NA   1    1  317 12 2019-01-30 10:30:00  50    197 10.6
## 4   1 4151.0   NA 14.8   0    0   NA NA 2019-07-08 10:00:00  50    197 10.6
## 5   2    0.0  2.5   NA   1    1   27 12 2019-09-13 10:00:00  56    154  7.8
## 6   2  336.0   NA  4.5   0    0   NA NA 2019-09-27 10:00:00  56    154  7.8
## 7   2  336.5  2.5   NA   1    1   69 12 2019-09-27 10:30:00  56    154  7.8
## 8   2 1176.0   NA 14.1   0    0   NA NA 2019-11-01 10:00:00  56    154  7.8
## 9   2 1176.5 10.0   NA   1    1   21 12 2019-11-01 10:30:00  56    154  7.8
## 10  2 1441.0   NA  4.5   0    0   NA NA 2019-11-12 10:00:00  56    154  7.8
## 11  3    0.0  2.5   NA   1    1   27 12 2020-02-26 09:00:00  40    147 10.8
## 12  3  336.0   NA  5.2   0    0   NA NA 2020-03-11 10:00:00  40    147 10.8
## 13  3  336.5  2.5   NA   1    1   49 12 2020-03-11 10:30:00  40    147 10.8
## 14  3  936.0   NA  6.4   0    0   NA NA 2020-04-05 10:00:00  40    147 10.8
## 15  3  936.5  5.0   NA   1    1  123 12 2020-04-05 10:30:00  40    147 10.8
## 16  3 2424.0   NA 10.7   0    0   NA NA 2020-06-06 10:00:00  40    147 10.8
```

Note that `addl` and `II` dictate an every-twelve-hour dosing schedule which leads up to the proceeding concentration. Covariates are preserved and a `time` variable which represents hours since first dose is generated. This data is now in an appropriate format for PK analysis but makes no use of the last-dose times although they are extracted along with some (but not all) concentrations. When last-dose times are present in the input data and they are specified in the argument `ldCol`, the sequence of doses leading up to the extracted dose is reduced and a new row is inserted which accurately describes the timing of the dose which precedes the relevant concentration.

```
# Build PK data with last-dose times
run_Build_PK_Oral(x = dat,
                  idCol = "id",
                  dtCol = "dt",
                  doseCol = "dose_morn",
                  concCol = "conc",
                  ldCol = "ld",
                  first_interval_hours = 336,
                  imputeClosest = NULL)
```

```
##    id   time  amt   dv mdv evid addl II               date age weight  hgb
## 1   1    0.0  2.5   NA   1    1   27 12 2019-01-16 10:00:00  50    197 10.6
## 2   1  336.0   NA  1.9   0    0   NA NA 2019-01-30 10:00:00  50    197 10.6
## 3   1  336.5  2.5   NA   1    1  317 12 2019-01-30 10:30:00  50    197 10.6
## 4   1 4151.0   NA 14.8   0    0   NA NA 2019-07-08 10:00:00  50    197 10.6
## 5   2    0.0  2.5   NA   1    1   27 12 2019-09-13 10:00:00  56    154  7.8
## 6   2  336.0   NA  4.5   0    0   NA NA 2019-09-27 10:00:00  56    154  7.8
## 7   2  336.5  2.5   NA   1    1   68 12 2019-09-27 10:30:00  56    154  7.8
## 8   2 1164.0  2.5   NA   1    1    0 NA 2019-10-31 22:00:00  56    154  7.8
## 9   2 1176.0   NA 14.1   0    0   NA NA 2019-11-01 10:00:00  56    154  7.8
## 10  2 1176.5 10.0   NA   1    1   21 12 2019-11-01 10:30:00  56    154  7.8
```

```
## 11  2 1441.0   NA  4.5   0    0   NA NA 2019-11-12 10:00:00  56    154  7.8
## 12  3    0.0  2.5   NA   1    1   26 12 2020-02-26 09:00:00  40    147 10.8
## 13  3  326.0  2.5   NA   1    1    0 NA 2020-03-11 00:00:00  40    147 10.8
## 14  3  336.0   NA  5.2   0    0   NA NA 2020-03-11 10:00:00  40    147 10.8
## 15  3  336.5  2.5   NA   1    1   49 12 2020-03-11 10:30:00  40    147 10.8
## 16  3  936.0   NA  6.4   0    0   NA NA 2020-04-05 10:00:00  40    147 10.8
## 17  3  936.5  5.0   NA   1    1  122 12 2020-04-05 10:30:00  40    147 10.8
## 18  3 2409.0  5.0   NA   1    1    0 NA 2020-06-05 19:00:00  40    147 10.8
## 19  3 2424.0   NA 10.7   0    0   NA NA 2020-06-06 10:00:00  40    147 10.8
```

Individual 1 has no extracted last-dose times so their data is unchanged from before. Compare, however, rows 7-9 to rows 7-8 of the previous dataset constructed without last-dose times. The measured concentration of 14.1 on `date` 2019-11-01 is associated with a last-dose time. `addl` drops from 69 to 68 and the extracted last-dose is added in row 8 with additional `date` 2019-10-31 20:58:36, the last-dose time extracted from clinical notes. Notice that the number of doses leading up to the concentration is unchanged and the timing of the final dose has been adjusted to reflect information in the EHR (i.e., the calculated time of 1162.70 for `time`). This dataset still relies on assumptions about dosing, but should reflect the actual dosing schedule better by incorporating last-dose times from the EHR.

- The final PK data includes ID and standard NONMEM formatted variables:
  - `time`: time of either dosing or concentration measurement.
  - `amt`: dose amount; `NA` for concentration events.
  - `dv`: drug blood concentration value, which is DV (dependent variable) as NONMEM data item; `NA` for a dose event.
  - `mdv`: missing dependent variable; 1 for indicating that there is no dependent variable (in this case, blood concentration), 0 for dependent variable.
  - `evid`: event ID; 1 for indicating dose event (amt, II, and addl for this record will be used for the drug dose information if evid = 1), 0 for observation (or dependent variable if mdv = 0).
  - `addl`: additional doses; the number of times for additional oral dose to be repeated, which is 1 less than total number of repeated (identical) doses.
  - `II`: interdose interval, the amount of time between each additional dose.
  - `date`: date and time for concentration measurement or assumed dosing
  - `age`: an example covariate from demographic data
  - `weight`: an example covariate from demographic data
  - `hgb`: an example covariate from lab data

References

1. Choi L, Beck C, McNeer E, Weeks HL, Williams ML, James NT, Niu X, Abou-Khalil BW, Birdwell KA, Roden DM, Stein CM. Development of a System for Post-marketing Population Pharmacokinetic and Pharmacodynamic Studies using Real-World Data from Electronic Health Records. Clinical Pharmacology & Therapeutics. 2020 Apr; 107(4): 934-943.