Introduction to drugprepCPRD

Belay B. Yimer, David Sleby, Meghna Jani, Goran Nenadic,Mark Lunt, William G. Dixon

Jun 2021

Motivation

The Clinical Practice Research Datalink (CPRD), a UK database of anonymised primary care electronic health records (EHRs), is one of the widely used EHR to study the effectiveness and safety of medications. However, prescription data from CPRD are often messy, with inherent issues such as missing information on stop date, quantity, and unstructured free-text instructions. Data preparation steps of exposure information are rarely fully reported in pharmacoepidemiology studies, yet assumptions made during this stage can have considerable implications for risk attribution of possible adverse events (AEs). We have previously developed a framework for dealing with missing information on stop dates and other issues such as overlapping prescriptions. The framework was implemented using `STATA` software. Beside being only available for `STATA` users, the earlier algorithm did not deal with the free text prescriptions. The current r-package `drugprepCPRD`, build up on the earlier algorithm, and aims to make the framework available to wider audience through the implementation in a free open-source software.

This vignette describes how to use the `drugprepCPRD` package to transform CPRD drug data contained in the 'therapy.txt' file into information on individuals' drug use over time. We will walk through all the steps needed to perform the transformation. We assume the user are familiar with CPRD data and have basic knowledge of R-software.

The CPRD Data

The CPRD Gold data follows the CPRD Gold specification (https://cprdcw.cprd.com/_docs/CPRD_GOLD_Full_Data_Specification_v2.0.pdf). It is made-up of several tables containing information related to the patients. One of the CPRD tables called 'Therapy' contain the prescription information for a patient. This table can be linked to the 'commondoage' lookup table to get the text instruction for the prescribed product. Table 1 below presents a hypothetical prescription data for two individuals.

```
#>    patid pracid event_date prodcode dossageid                      text qty
#> 1   2156    156 2011-06-16        1         1    TAKE 1 OR 2 4 TIMES/DAY  40
#> 2   2156    156 2011-06-24        1         1    TAKE 1 OR 2 4 TIMES/DAY  40
#> 3   2156    156 2011-07-10        2         2 TAKE 1-2 THREE TIMES A DAY  24
#> 4   2156    156 2011-07-17        2         2 TAKE 1-2 THREE TIMES A DAY  24
#> 5   2156    156 2011-07-17        2         2 TAKE 1-2 THREE TIMES A DAY  24
#> 6   2156    156 2011-08-16        1         1    TAKE 1 OR 2 4 TIMES/DAY  NA
#> 7   2156    156 2011-08-24        1        NA                             40
#> 8   2156    156 2011-09-10        2         2 TAKE 1-2 THREE TIMES A DAY  NA
#> 9   2156    156 2011-09-17        2        NA                             24
#> 10  2256    160 2011-06-16        1         1    TAKE 1 OR 2 4 TIMES/DAY  40
#> 11  2256    160 2011-06-24        1         1    TAKE 1 OR 2 4 TIMES/DAY  40
#> 12  2256    160 2011-07-10        2         2 TAKE 1-2 THREE TIMES A DAY  24
#> 13  2256    160 2011-07-17        2         2 TAKE 1-2 THREE TIMES A DAY  24
#> 14  2256    160 2011-07-17        2         2 TAKE 1-2 THREE TIMES A DAY  24
```

```
#> 15  2256    160 2011-08-16        1         1    TAKE 1 OR 2 4 TIMES/DAY  NA
#> 16  2256    160 2011-08-24        1        NA                             40
#> 17  2256    160 2011-09-10        2         2 TAKE 1-2 THREE TIMES A DAY  NA
#> 18  2256    160 2011-09-17        2        NA                             24
#>    numdays dose_duration
#> 1        7             6
#> 2        7             6
#> 3        6             5
#> 4        6             5
#> 5        6             5
#> 6        7             6
#> 7        7             6
#> 8        6             5
#> 9        6             5
#> 10       7             6
#> 11       7             6
#> 12       6             5
#> 13       6             5
#> 14       6             5
#> 15       7             6
#> 16       7             6
#> 17       6             5
#> 18       6             5
```

The event_date in the above table is often used as the start date of exposure but the stop date of the exposure is not available as per se and we have to compute it from the available information. CPRD provides two options namely, numdays and dose_duration which can be used to define the stop date. However, this values are often missing and does not give flexibility for the researcher in defining the stop date in the case of prescription with a variable dose frequency (number of times the prescription to be taken per day) and dose number (e.g., number of tablets to take at a time).

As described below, the `drugprepCPRD` package will extract the dose frequency and dose number from the text instruction and provides a series of data processing steps with multiple options to define start and stop dates of a given prescription. The package works at the prodcode level to give much granularity.

The algorithm

The `drugprepCPRD` is made up of the following function that must be executed sequentially.

- **Compute ndd**: Extract dose frequency (number of times the prescription to be taken per day) and dose number (e.g., number of tablets to take at a time) from the free text and compute the number of daily dose.
- **Define implusible values**: Given "Plausible" values (e.g. based on prescribing guidelines and clinical experience), this stage identifies those values outside the plausible values range.
- **Dec1-Handle implausible qty**: Values outside the "Plausible" range may be [1a] ignored, [1b] set to missing, or imputed. Imputation options include: [1c1] set to the mean value for that patient for that product code, [1c2] set to the mean value for that practice for that product code, [1c3] set to the mean value for the whole cohort for that product code, etc,. See the package manual for all possible options.
- **Dec2-Handle missing qty**: – Options for missing qty are: [2a] leave as missing, [2b1] set to the mean value for that patient for that product code, [2b2] set to the mean value for that practice for that product code, [2c] set to the mean value for the whole cohort for that product code, etc,.
- **Dec3-Handle implausible ndd**: Options for implausible ndd are the same as for decision 1.

- **Dec4-Handle missing ndd**: Options for missing ndd are the same as for decision 2.
- **Dec5-Clean duration**: cleans implausibly high values for each of the three available duration variables (numdays, dose_duration, and qty/ndd). Options for cleaning each duration variable are: [5a] make no changes, [5b(X)] set to missing if duration is greater than X months, or [5c(X)] set to X if duration is greater than X months. X is 6, 12, or 24.
- **Dec6-Select stop date**: defines a stop date for each prescription. calculate stop date as prescription start date + one of the following duration definitions: [6a] numdays, [6b] dose_duration, [6c] qty/ndd, or [6d(X)]
- **Dec7-Handle missing stop date**: if stop date is missing: [7a] keep as missing, [7b] set to the mean value for that product code for that patient, [7c] set to the mean value for that product code for the whole cohort, [7d] set to the mean value for that product code for that patient, otherwise set to the mean value for that product code for the whole cohort.
- **Dec8-Handle multiple prescriptions**: for multiple prescriptions for the same product code on the same day, but with different stop dates, options are: [8a] do nothing, ; [8b] calculate the mean duration of prescriptions and drop redundant records; [8c] keep the record with the shortest duration; [8d] keep the record with with the longest duration; [8e] sum the durations and drop redundant records.
- **Dec9-Handle overlaping prescriptions**: for consecutive records with overlapping start and stop dates, options are [9a] to ignore the overlap but sum ndds; [9b] move the overlapping time.
- **Dec10-Handle short gaps between prescriptions**: handles small gaps between consecutive prescriptions by either allowing these gaps to remain classified as unexposed or reclassifying the gaps as exposed when the gap is less than a specified number of days. options include [10a] do nothing – the gap remains classified as unexposed; [10b(X)] move the stop date of the preceding prescription to "fill in" the gap, reclassifying the time as exposed, if the gap between consecutive prescriptions is less than X days. X is 15, 30, or 60

R package drugprepCPRD

You can install the latest development version from `GitHub` with these `R` commands:

```
install.packages("devtools")
devtools::install_github("belayb/drugprepCPRD")
```

Once the package is installed, we have to load the `drugprepCPRD` package in the R-environment as follow.

```
library(drugprepCPRD)
```

Computation of ndd from free text

This step uses another package that we developed for extraction of dose frequency and dose number, r-package `doseminer`, and compute the number of daily dose(ndd) using the following formula

$$ndd = DF * DN/DI$$

where DF is the dose frequency, DN the dose number, and DI is the dose interval. The user must define here what DF or DN value to use. Possible values are min, max, and mean. In the case of regular prescriptions (i.e., prescriptions with fixed instructions such as take 2 tablet 4 times a day), the min, max, and mean value will be the same. Computation of ndd can be done using the `compute_ndd` function as follow.

```
dataset1<-compute_ndd(dataset1, "min_min")
as.data.frame(dataset1)
#>    patid pracid event_date prodcode dossageid                      text qty
#> 1   2156    156 2011-06-16        1         1    TAKE 1 OR 2 4 TIMES/DAY  40
```

```
#> 2   2156    156 2011-06-24        1         1    TAKE 1 OR 2 4 TIMES/DAY  40
#> 3   2156    156 2011-07-10        2         2 TAKE 1-2 THREE TIMES A DAY  24
#> 4   2156    156 2011-07-17        2         2 TAKE 1-2 THREE TIMES A DAY  24
#> 5   2156    156 2011-07-17        2         2 TAKE 1-2 THREE TIMES A DAY  24
#> 6   2156    156 2011-08-16        1         1    TAKE 1 OR 2 4 TIMES/DAY  NA
#> 7   2156    156 2011-08-24        1        NA                             40
#> 8   2156    156 2011-09-10        2         2 TAKE 1-2 THREE TIMES A DAY  NA
#> 9   2156    156 2011-09-17        2        NA                             24
#> 10  2256    160 2011-06-16        1         1    TAKE 1 OR 2 4 TIMES/DAY  40
#> 11  2256    160 2011-06-24        1         1    TAKE 1 OR 2 4 TIMES/DAY  40
#> 12  2256    160 2011-07-10        2         2 TAKE 1-2 THREE TIMES A DAY  24
#> 13  2256    160 2011-07-17        2         2 TAKE 1-2 THREE TIMES A DAY  24
#> 14  2256    160 2011-07-17        2         2 TAKE 1-2 THREE TIMES A DAY  24
#> 15  2256    160 2011-08-16        1         1    TAKE 1 OR 2 4 TIMES/DAY  NA
#> 16  2256    160 2011-08-24        1        NA                             40
#> 17  2256    160 2011-09-10        2         2 TAKE 1-2 THREE TIMES A DAY  NA
#> 18  2256    160 2011-09-17        2        NA                             24
#>    numdays dose_duration optional ndd
#> 1        7             6        0   4
#> 2        7             6        0   4
#> 3        6             5        0   3
#> 4        6             5        0   3
#> 5        6             5        0   3
#> 6        7             6        0   4
#> 7        7             6        0  NA
#> 8        6             5        0   3
#> 9        6             5        0  NA
#> 10       7             6        0   4
#> 11       7             6        0   4
#> 12       6             5        0   3
#> 13       6             5        0   3
#> 14       6             5        0   3
#> 15       7             6        0   4
#> 16       7             6        0  NA
#> 17       6             5        0   3
#> 18       6             5        0  NA
```

Here, we specified to use the minimum values for both the DF and DN in the computation of `ndd`. Running `compute_ndd()` creates an additional column names `ndd`.

Defining implausible values

The next stage is to define the cut-off for plausible values of prescription quantity and number of daily dose for each product. The information has to be provided by the users in a table format. The table should have column names: prodcode, max_qty, min_qty, max_rec_ndd, min_rec_ndd. Such information might be obtained from British National Formulary (BNF) - NICE (https://bnf.nice.org.uk/). For our hypothetical case, we defined the min_max data as follow.

```
#>   prodcode qty_max qty_min max_rec_ndd min_rec_ndd
#> 1        1     100       1          10           1
#> 2        2      50       1          10           1
```

Once we prepare our min_max data in a table format, we can use the function `implausible_values` to flag those prescriptions beyond the plausible values as follow

```
dataset1<-Implausible_values(dataset1, min_max_dat = min_max_dat)
as.data.frame(dataset1)
#>    patid pracid event_date prodcode dossageid                      text qty
#> 1   2156    156 2011-06-16        1         1    TAKE 1 OR 2 4 TIMES/DAY  40
#> 2   2156    156 2011-06-24        1         1    TAKE 1 OR 2 4 TIMES/DAY  40
#> 3   2156    156 2011-07-10        2         2 TAKE 1-2 THREE TIMES A DAY  24
#> 4   2156    156 2011-07-17        2         2 TAKE 1-2 THREE TIMES A DAY  24
#> 5   2156    156 2011-07-17        2         2 TAKE 1-2 THREE TIMES A DAY  24
#> 6   2156    156 2011-08-16        1         1    TAKE 1 OR 2 4 TIMES/DAY  NA
#> 7   2156    156 2011-08-24        1        NA                             40
#> 8   2156    156 2011-09-10        2         2 TAKE 1-2 THREE TIMES A DAY  NA
#> 9   2156    156 2011-09-17        2        NA                             24
#> 10  2256    160 2011-06-16        1         1    TAKE 1 OR 2 4 TIMES/DAY  40
#> 11  2256    160 2011-06-24        1         1    TAKE 1 OR 2 4 TIMES/DAY  40
#> 12  2256    160 2011-07-10        2         2 TAKE 1-2 THREE TIMES A DAY  24
#> 13  2256    160 2011-07-17        2         2 TAKE 1-2 THREE TIMES A DAY  24
#> 14  2256    160 2011-07-17        2         2 TAKE 1-2 THREE TIMES A DAY  24
#> 15  2256    160 2011-08-16        1         1    TAKE 1 OR 2 4 TIMES/DAY  NA
#> 16  2256    160 2011-08-24        1        NA                             40
#> 17  2256    160 2011-09-10        2         2 TAKE 1-2 THREE TIMES A DAY  NA
#> 18  2256    160 2011-09-17        2        NA                             24
#>    numdays dose_duration optional ndd implausible_qty implausible_ndd
#> 1        7             6        0   4           FALSE           FALSE
#> 2        7             6        0   4           FALSE           FALSE
#> 3        6             5        0   3           FALSE           FALSE
#> 4        6             5        0   3           FALSE           FALSE
#> 5        6             5        0   3           FALSE           FALSE
#> 6        7             6        0   4           FALSE           FALSE
#> 7        7             6        0  NA           FALSE           FALSE
#> 8        6             5        0   3           FALSE           FALSE
#> 9        6             5        0  NA           FALSE           FALSE
#> 10       7             6        0   4           FALSE           FALSE
#> 11       7             6        0   4           FALSE           FALSE
#> 12       6             5        0   3           FALSE           FALSE
#> 13       6             5        0   3           FALSE           FALSE
#> 14       6             5        0   3           FALSE           FALSE
#> 15       7             6        0   4           FALSE           FALSE
#> 16       7             6        0  NA           FALSE           FALSE
#> 17       6             5        0   3           FALSE           FALSE
#> 18       6             5        0  NA           FALSE           FALSE
```

Running the function `Implausible_values()` creates an additional columns named `implausible_qty` and `implausible_ndd` with values eqauls to `TRURE` if the given quantity or ndd is outside the plausible range.

Processing the CPRD prescription data

Once this preliminary stages are completed, we can use the main function of `drugprepCPRD`, `run.drugPREP` to implement the 10 decision nodes described above. This can be done by executing the following r command.

```
dataset1<-run.drugPREP(dataset1, decisions = c("1b","2b1","3b","4b1","5b_6","6c","7a","8d","9a","10b") )
#> Started executing dec1:implausible_qty
#> Started executing dec2:missing_qty
#> Started executing dec3:implausible_ndd
#> Started executing dec4:missing_ndd
```

```
#> Warning: Grouping rowwise data frame strips rowwise nature
#> Started executing dec5:clean duration
#> Started executing dec6:select stop date
#> Started executing dec7:dealing with missing stop date
#> Started executing dec8:idealing with multiple prescription
#> Started executing dec9:dealing with overlaping prescription
#> Started executing dec10:dealing with short gaps between presecriptions
as.data.frame(dataset1)
#>    patid prodcode      start  real_stop pracid dossageid
#> 1   2156        1 2011-06-16 2011-06-23    156         1
#> 2   2256        1 2011-06-16 2011-06-23    160         1
#> 3   2156        1 2011-06-24 2011-06-26    156         1
#> 4   2256        1 2011-06-24 2011-06-26    160         1
#> 5   2156        1 2011-06-27 2011-07-04    156         1
#> 6   2256        1 2011-06-27 2011-07-04    160         1
#> 7   2156        2 2011-07-10 2011-07-16    156         2
#> 8   2256        2 2011-07-10 2011-07-16    160         2
#> 9   2156        2 2011-07-17 2011-07-18    156         2
#> 10  2256        2 2011-07-17 2011-07-18    160         2
#> 11  2156        2 2011-07-19 2011-07-25    156         2
#> 12  2256        2 2011-07-19 2011-07-25    160         2
#> 13  2156        1 2011-08-16 2011-08-23    156         1
#> 14  2256        1 2011-08-16 2011-08-23    160         1
#> 15  2156        1 2011-08-24 2011-08-26    156         1
#> 16  2256        1 2011-08-24 2011-08-26    160         1
#> 17  2156        1 2011-08-27 2011-09-03    156        NA
#> 18  2256        1 2011-08-27 2011-09-03    160        NA
#> 19  2156        2 2011-09-10 2011-09-16    156         2
#> 20  2256        2 2011-09-10 2011-09-16    160         2
#> 21  2156        2 2011-09-17 2011-09-18    156         2
#> 22  2256        2 2011-09-17 2011-09-18    160         2
#> 23  2156        2 2011-09-19 2011-09-25    156        NA
#> 24  2256        2 2011-09-19 2011-09-25    160        NA
#>                          text qty numdays dose_duration ndd new_duration
#> 1     TAKE 1 OR 2 4 TIMES/DAY  40       7             6   4           10
#> 2     TAKE 1 OR 2 4 TIMES/DAY  40       7             6   4           10
#> 3     TAKE 1 OR 2 4 TIMES/DAY  40       7             6   8           10
#> 4     TAKE 1 OR 2 4 TIMES/DAY  40       7             6   8           10
#> 5     TAKE 1 OR 2 4 TIMES/DAY  40       7             6   4           10
#> 6     TAKE 1 OR 2 4 TIMES/DAY  40       7             6   4           10
#> 7  TAKE 1-2 THREE TIMES A DAY  24       6             5   3            8
#> 8  TAKE 1-2 THREE TIMES A DAY  24       6             5   3            8
#> 9  TAKE 1-2 THREE TIMES A DAY  24       6             5   6            8
#> 10 TAKE 1-2 THREE TIMES A DAY  24       6             5   6            8
#> 11 TAKE 1-2 THREE TIMES A DAY  24       6             5   3            8
#> 12 TAKE 1-2 THREE TIMES A DAY  24       6             5   3            8
#> 13    TAKE 1 OR 2 4 TIMES/DAY  40       7             6   4           10
#> 14    TAKE 1 OR 2 4 TIMES/DAY  40       7             6   4           10
#> 15    TAKE 1 OR 2 4 TIMES/DAY  40       7             6   8           10
#> 16    TAKE 1 OR 2 4 TIMES/DAY  40       7             6   8           10
#> 17                             40       7             6   4           10
#> 18                             40       7             6   4           10
#> 19 TAKE 1-2 THREE TIMES A DAY  24       6             5   3            8
```

```
#> 20 TAKE 1-2 THREE TIMES A DAY  24       6             5   3            8
#> 21 TAKE 1-2 THREE TIMES A DAY  24       6             5   6            8
#> 22 TAKE 1-2 THREE TIMES A DAY  24       6             5   6            8
#> 23                             24       6             5   3            8
#> 24                             24       6             5   3            8
#>    gap_to_next
#> 1            1
#> 2            1
#> 3            1
#> 4            1
#> 5           43
#> 6           43
#> 7            1
#> 8            1
#> 9            1
#> 10           1
#> 11          47
#> 12          47
#> 13           1
#> 14           1
#> 15           1
#> 16           1
#> 17      100000
#> 18      100000
#> 19           1
#> 20           1
#> 21           1
#> 22           1
#> 23      100000
#> 24      100000
```

The result of running the function `run.drugPREP()` provides the start and stop date (real_stop) for each prescription. Note that the value 10000 in the gap to next is given for computational purpose and should not be considered as a value indicating the gap between sequential prescriptions.