

Dose Building Using Example Vanderbilt EHR Data

Introduction

We have provided the medExtractR output and gold standards for the tacrolimus and lamotrigine test sets used to develop the dose building algorithm detailed in this paper. This data comes from Vanderbilt's EHR system. In this vignette, we show how to access this data, how to implement the algorithm, and how to compare the algorithm output to the gold standard using the tacrolimus data. More details of the functions used in the algorithm can be found in our EHR vignette for Extract-Med and Pro-Med-NLP.

medExtractR Output

Several rows of the medExtractR output for tacrolimus are shown below.

```
tac_mxr_fn <- system.file("examples", "tac_mxr_out.csv", package = "EHR")
tac_mxr <- read.csv(tac_mxr_fn, na = '')
tac_mxr[c(135:139,163:167,283:289,343:346),]
```

##	filename	entity	expr	pos
## 135	X240866534_2010-01-28_4070129	DrugName	Tacrolimus	839:849
## 136	X240866534_2010-01-28_4070129	DrugName	Prograf	851:858
## 137	X240866534_2010-01-28_4070129	Strength	1 mg	860:864
## 138	X240866534_2010-01-28_4070129	DoseAmt	4	874:875
## 139	X240866534_2010-01-28_4070129	Frequency	every twelve hours	888:906
## 163	X240866534_2010-01-28_9659069	DrugName	Tacrolimus	150:160
## 164	X240866534_2010-01-28_9659069	DrugName	Prograf	162:169
## 165	X240866534_2010-01-28_9659069	Strength	1 mg	171:175
## 166	X240866534_2010-01-28_9659069	DoseAmt	4	185:186
## 167	X240866534_2010-01-28_9659069	Frequency	every twelve hours	199:217
## 283	X410930205_2006-06-20_3473651	DrugName	Prograf	870:877
## 284	X410930205_2006-06-20_3473651	Dose	3mg	878:881
## 285	X410930205_2006-06-20_3473651	Frequency	BID	882:885
## 286	X410930205_2006-06-20_3473651	DrugName	prograf	943:950
## 287	X410930205_2006-06-20_3473651	Strength	1mg	951:954
## 288	X410930205_2006-06-20_3473651	DoseAmt	3	955:956
## 289	X410930205_2006-06-20_3473651	Frequency	bid	961:964
## 343	X410930205_2006-06-20_2809083	DrugName	prograf	560:567
## 344	X410930205_2006-06-20_2809083	Strength	1mg	568:571
## 345	X410930205_2006-06-20_2809083	DoseAmt	3	572:573
## 346	X410930205_2006-06-20_2809083	Frequency	bid	578:581

Part I

The first step of Part I of our algorithm is parsing the raw NLP output. This results in a standardized form of the data that includes a row for each drug mention and columns for all entities anchored to that drug mention. Here, we use the `parseMedExtractR` function since we are using medExtractR output as an example.

```
tac_mxr_parsed <- parseMedExtractR(tac_mxr_fn)
```

Below are the rows of the parsed output corresponding to the raw NLP output from above.

```

##          filename          drugname      strength
## 1 X240866534_2010-01-28_4070129 Tacrolimus::839::849
## 2 X240866534_2010-01-28_4070129   Prograf::851::858 1 mg::860::864
## 3 X240866534_2010-01-28_9659069 Tacrolimus::150::160
## 4 X240866534_2010-01-28_9659069   Prograf::162::169 1 mg::171::175
## 5 X410930205_2006-06-20_3473651   Prograf::870::877
## 6 X410930205_2006-06-20_3473651   prograf::943::950 1mg::951::954
## 7 X410930205_2006-06-20_2809083   prograf::560::567 1mg::568::571
##
##          dose route          freq      dosestr
##          1
##          2 4::874::875      every twelve hours::888::906
##          3
##          4 4::185::186      every twelve hours::199::217
##          5                      BID::882::885 3mg::878::881
##          6 3::955::956      bid::961::964
##          7 3::572::573      bid::578::581
##
##          dosechange lastdose
##          1
##          2
##          3
##          4
##          5
##          6
##          7

```

Next, the parsed entities are paired using the `buildDose` function. This results in a dataset with a column for each entity and a row for each pairing.

```
tac_mxr_part1_out <- buildDose(tac_mxr_parsed)
```

The output is shown below.

```

##          filename  drugname strength dose route
## 51 X240866534_2010-01-28_4070129 Tacrolimus <NA> <NA> <NA>
## 52 X240866534_2010-01-28_4070129   Prograf   1 mg   4 <NA>
## 55 X240866534_2010-01-28_9659069 Tacrolimus <NA> <NA> <NA>
## 56 X240866534_2010-01-28_9659069   Prograf   1 mg   4 <NA>
## 104 X410930205_2006-06-20_2809083   prograf   1mg   3 <NA>
## 105 X410930205_2006-06-20_3473651   Prograf   <NA> <NA> <NA>
## 106 X410930205_2006-06-20_3473651   prograf   1mg   3 <NA>
##          freq dosestr dosechange lastdose drugname_start
## 51          <NA> <NA>          <NA>          <NA>          839
## 52 every twelve hours <NA>          <NA>          <NA>          851
## 55          <NA> <NA>          <NA>          <NA>          150
## 56 every twelve hours <NA>          <NA>          <NA>          162
## 104          bid <NA>          <NA>          <NA>          560
## 105          BID 3mg <NA>          <NA>          870
## 106          bid <NA>          <NA>          <NA>          943

```

Comparing to Gold Standard

We have provided the gold standard that we generated for part 1. Several rows are shown below.

```
tac_gs_part1 <- read.csv(system.file("examples", "tac_gs_part1.csv", package = "EHR"),
  stringsAsFactors = FALSE, na = '')
```

```
##          filename  drugname drugname_start strength dose route
## 51 X240866534_2010-01-28_4070129 Tacrolimus      839    <NA> <NA>   NA
## 52 X240866534_2010-01-28_4070129  Prograf      851    1 mg    4    NA
## 53 X240866534_2010-01-28_9659069 Tacrolimus      150    <NA> <NA>   NA
## 54 X240866534_2010-01-28_9659069  Prograf      162    1 mg    4    NA
## 104 X410930205_2006-06-20_3473651  Prograf      870    <NA> <NA>   NA
## 105 X410930205_2006-06-20_3473651  prograf      943     1mg    3    NA
## 107 X410930205_2006-06-20_2809083  prograf      560     1mg    3    NA
##          freq dosestr dosechange
## 51          <NA>    <NA>    <NA>
## 52 every twelve hours  <NA>    <NA>
## 53          <NA>    <NA>    <NA>
## 54 every twelve hours  <NA>    <NA>
## 104          BID      3mg    <NA>
## 105          bid    <NA>    <NA>
## 107          bid    <NA>    <NA>
```

The following code compares the gold standard to the Part I output and provides the recall and precision measures.

```
precall <- function(dat, gs) {
  tp1 <- sum(dat %in% gs)
  fp1 <- sum(!(dat %in% gs))
  fn1 <- sum(!(gs %in% dat))
  r1 <- c(tp1, tp1 + fn1)
  p1 <- c(tp1, tp1 + fp1)
  r <- rbind(r1,p1)
  dimnames(r) <- list(c('recall','prec'), c('num','den'))
  cbind(r, prop = round(r[,1] / r[,2], 2))
}

colsToCompare <- c('filename','drugname','strength','dose','route','freq',
  'dosestr','dosechange','drugname_start')
tac_mxr_part1_out <- tac_mxr_part1_out[,colsToCompare]
tac_gs_part1 <- tac_gs_part1[,colsToCompare]

tacxrrow <- do.call(paste, c(tac_mxr_part1_out, sep = '|'))
gs.tacxrrow <- do.call(paste, c(tac_gs_part1, sep = '|'))

precall(tacxrrow, gs.tacxrrow)

##          num den prop
## recall 285 285    1
## prec   285 285    1
```

Part II

In part II of the algorithm, the final datasets are formed containing dose intake and daily dose, and redundancies are removed at the note and date level for each patient.

This part of the algorithm requires more detailed meta data associated with each clinical note file. This is shown below using our example tacrolimus data.

```

bmd <- function(x) {
  fns <- strsplit(x, '_')
  pid <- sapply(fns, `[`, 1)
  date <- as.Date(sapply(fns, `[`, 2), format = '%Y-%m-%d')
  note <- sapply(fns, `[`, 3)
  data.frame(filename = x, pid, date, note, stringsAsFactors = FALSE)
}
tac_metadata <- bmd(tac_mxr_part1_out[['filename']])

```

```

##                filename      pid      date      note
## 51 X240866534_2010-01-28_4070129 X240866534 2010-01-28 4070129
## 55 X240866534_2010-01-28_9659069 X240866534 2010-01-28 9659069
## 104 X410930205_2006-06-20_2809083 X410930205 2006-06-20 2809083
## 105 X410930205_2006-06-20_3473651 X410930205 2006-06-20 3473651

```

Below, a few rows of the note level and date level collapsing are shown for our example tacrolimus data.

```

tac_part2 <- collapseDose(tac_mxr_part1_out, tac_metadata, naFreq='most')

```

Note level:

```

##                filename drugname strength dose  route freq dosestr
## 40 X240866534_2010-01-28_4070129 Prograf      1 mg    4 orally bid    <NA>
## 42 X240866534_2010-01-28_9659069 Prograf      1 mg    4 orally bid    <NA>
## 68 X410930205_2006-06-20_2809083 prograf      1mg     3 orally bid    <NA>
## 69 X410930205_2006-06-20_3473651 Prograf      <NA> <NA> orally bid    3mg
##
##                dosechange drugname_start dosestr.num strength.num doseamt.num
## 40                <NA>           851           NA           1           4
## 42                <NA>           162           NA           1           4
## 68                <NA>           560           NA           1           3
## 69                <NA>           870           3           NA           NA
##
##                freq.num dose.intake intaketime dose.seq dose.daily
## 40                2           4           <NA>           NA           8
## 42                2           4           <NA>           NA           8
## 68                2           3           <NA>           NA           6
## 69                2           3           <NA>           NA           6

```

Date level:

```

##                filename drugname strength dose  route freq dosestr
## 29 X240866534_2010-01-28_4070129 Prograf      1 mg    4 orally bid    <NA>
## 42 X410930205_2006-06-20_2809083 prograf      1mg     3 orally bid    <NA>
##
##                dosechange drugname_start dosestr.num strength.num doseamt.num
## 29                <NA>           851           NA           1           4
## 42                <NA>           560           NA           1           3
##
##                freq.num dose.intake intaketime dose.seq dose.daily
## 29                2           4           <NA>           NA           8
## 42                2           3           <NA>           NA           6

```

Comparing to Gold Standard

We have provided the gold standards that we generated for part 2.

Note level:

```
tac_gs_part2_note <- read.csv(  
  system.file("examples", "tac_gs_part2_note.csv", package = "EHR"),  
  stringsAsFactors = FALSE, na = ''  
)
```

```
##           filename drugname drugname_start strength dose route  
## 40 X240866534_2010-01-28_4070129 Prograf          851      1  4  NA  
## 41 X240866534_2010-01-28_9659069 Prograf          162      1  4  NA  
## 68 X410930205_2006-06-20_3473651 Prograf          870     NA  NA  NA  
## 70 X410930205_2006-06-20_2809083 prograf          560      1  3  NA  
##   freq intaketime dosestr dosechange doseintake daily  
## 40    2      <NA>      NA      <NA>          4    8  
## 41    2      <NA>      NA      <NA>          4    8  
## 68    2      <NA>      3      <NA>          3    6  
## 70    2      <NA>      NA      <NA>          3    6
```

Date level:

```
tac_gs_part2_date <- read.csv(  
  system.file("examples", "tac_gs_part2_date.csv", package = "EHR"),  
  stringsAsFactors = FALSE, na = ''  
)
```

```
##           filename drugname drugname_start strength dose route  
## 29 X240866534_2010-01-28_4070129 Prograf          851      1  4  NA  
## 42 X410930205_2006-06-20_3473651 Prograf          870     NA  NA  NA  
##   freq intaketime dosestr dosechange doseintake daily  
## 29    2      <NA>      NA      <NA>          4    8  
## 42    2      <NA>      3      <NA>          3    6
```

The following code compares the gold standard to the Part II output and provides the recall and precision measures for note level and date level collapsing for dose intake and daily dose. In order to replicate the results from this paper, we use the Part I gold standard as the input to `collapseDose`.

```
precall <- function(dat, gs) {  
  tp1 <- sum(dat %in% gs)  
  fp1 <- sum(!(dat %in% gs))  
  fn1 <- sum(!(gs %in% dat))  
  r1 <- c(tp1, tp1 + fn1)  
  p1 <- c(tp1, tp1 + fp1)  
  r <- rbind(r1,p1)  
  dimnames(r) <- list(c('recall','prec'), c('num','den'))  
  cbind(r, prop = round(r[,1] / r[,2], 2))  
}  
  
metaData <- bmd(unique(tac_gs_part1$filename))  
tacxr <- collapseDose(tac_gs_part1, metaData, 'bid')  
tacxr.note <- tacxr[['note']]  
tacxr.date <- tacxr[['date']]  
  
tacxr.note$pid <- sub("_.*", "", tacxr.note$filename)  
tacxr.date$pid <- sub("_.*", "", tacxr.date$filename)  
tac_gs_part2_note$pid <- sub("_.*", "", tac_gs_part2_note$filename)  
tac_gs_part2_date$pid <- sub("_.*", "", tac_gs_part2_date$filename)
```

```

tacxrrow.note.intake <- do.call(paste, c(tacxr.note[,c('pid','dose.intake',
                                                'dosechange']], sep = '|'))
tacxrrow.note.daily <- do.call(paste, c(tacxr.note[,c('pid','intaketime','dose.daily',
                                                'dosechange']], sep = '|'))
tacxrrow.date.intake <- do.call(paste, c(tacxr.date[,c('pid','dose.intake',
                                                'dosechange']], sep = '|'))
tacxrrow.date.daily <- do.call(paste, c(tacxr.date[,c('pid','intaketime','dose.daily',
                                                'dosechange']], sep = '|'))

gs.tacxrrow.note.intake <- do.call(paste, c(tac_gs_part2_note[,c('pid','doseintake',
                                                                'dosechange']], sep = '|'))
gs.tacxrrow.note.daily <- do.call(paste, c(tac_gs_part2_note[,c('pid','intaketime','daily',
                                                                'dosechange']], sep = '|'))
gs.tacxrrow.date.intake <- do.call(paste, c(tac_gs_part2_date[,c('pid','doseintake',
                                                                'dosechange']], sep = '|'))
gs.tacxrrow.date.daily <- do.call(paste, c(tac_gs_part2_date[,c('pid','intaketime','daily',
                                                                'dosechange']], sep = '|'))

precall(tacxrrow.note.intake, gs.tacxrrow.note.intake)
precall(tacxrrow.note.daily, gs.tacxrrow.note.daily)
precall(tacxrrow.date.intake, gs.tacxrrow.date.intake)
precall(tacxrrow.date.daily, gs.tacxrrow.date.daily)

```

```

##          num den prop
## recall  205 205    1
## prec    205 205    1

##          num den prop
## recall  205 206    1
## prec    205 205    1

##          num den prop
## recall  116 116    1
## prec    116 116    1

##          num den prop
## recall  116 117 0.99
## prec    116 116 1.00

```