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.

```r
tac_mxr_fn <- system.file("examples", "tac_mxr_out.csv", package = "EHR")
tac_mxr <- read.csv(tac_mxr_fn, na = "")
```

<table>
<thead>
<tr>
<th>filename</th>
<th>entity</th>
<th>expr</th>
<th>pos</th>
</tr>
</thead>
<tbody>
<tr>
<td>X240866534_2010-01-28_4070129</td>
<td>DrugName</td>
<td>Tacrolimus</td>
<td>839:849</td>
</tr>
<tr>
<td>X240866534_2010-01-28_4070129</td>
<td>DrugName</td>
<td>Prograf</td>
<td>851:858</td>
</tr>
<tr>
<td>X240866534_2010-01-28_4070129</td>
<td>Strength</td>
<td>1 mg</td>
<td>860:864</td>
</tr>
<tr>
<td>X240866534_2010-01-28_4070129</td>
<td>DoseAmt</td>
<td>4</td>
<td>874:875</td>
</tr>
<tr>
<td>X240866534_2010-01-28_4070129</td>
<td>Frequency</td>
<td>every twelve hours</td>
<td>888:906</td>
</tr>
<tr>
<td>X410930205_2006-06-20_3473651</td>
<td>DrugName</td>
<td>Prograf</td>
<td>560:567</td>
</tr>
<tr>
<td>X410930205_2006-06-20_3473651</td>
<td>Strength</td>
<td>1 mg</td>
<td>568:571</td>
</tr>
<tr>
<td>X410930205_2006-06-20_3473651</td>
<td>DoseAmt</td>
<td>3</td>
<td>572:573</td>
</tr>
<tr>
<td>X410930205_2006-06-20_3473651</td>
<td>Frequency</td>
<td>bid</td>
<td>578:581</td>
</tr>
</tbody>
</table>

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 every twelve hours::888::906
## 3 every twelve hours::199::217
## 5 BID::882::885 3mg::878::881
## 6 bid::961::964
## 7 bid::578::581

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_mxrParsed)

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 every twelve hours <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

2
Comparing to Gold Standard

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

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

<table>
<thead>
<tr>
<th>#</th>
<th>filename</th>
<th>drugname</th>
<th>drugname_start</th>
<th>strength</th>
<th>dose</th>
<th>route</th>
<th>freq</th>
<th>dosestr</th>
<th>dosechange</th>
</tr>
</thead>
<tbody>
<tr>
<td>51</td>
<td>X240866534_2010-01-28_4070129</td>
<td>Tacrolimus</td>
<td>839</td>
<td>&lt;NA&gt;</td>
<td>&lt;NA&gt;</td>
<td>NA</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>52</td>
<td>X240866534_2010-01-28_4070129</td>
<td>Prograf</td>
<td>851</td>
<td>1 mg</td>
<td>4</td>
<td>NA</td>
<td>every twelve hours</td>
<td>&lt;NA&gt;</td>
<td></td>
</tr>
<tr>
<td>53</td>
<td>X240866534_2010-01-28_9659069</td>
<td>Tacrolimus</td>
<td>150</td>
<td>&lt;NA&gt;</td>
<td>&lt;NA&gt;</td>
<td>NA</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>54</td>
<td>X240866534_2010-01-28_9659069</td>
<td>Prograf</td>
<td>162</td>
<td>1 mg</td>
<td>4</td>
<td>NA</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>104</td>
<td>X410930205_2006-06-20_3473651</td>
<td>Prograf</td>
<td>870</td>
<td>&lt;NA&gt;</td>
<td>&lt;NA&gt;</td>
<td>NA</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>105</td>
<td>X410930205_2006-06-20_3473651</td>
<td>Prograf</td>
<td>943</td>
<td>1 mg</td>
<td>3</td>
<td>NA</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>107</td>
<td>X410930205_2006-06-20_2809083</td>
<td>Prograf</td>
<td>560</td>
<td>1 mg</td>
<td>3</td>
<td>NA</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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

```r
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', 'drugname_start', 'strength', 'dose', 'route', 'freq', 'dosestr', 'dosechange', 'drugname_start')
tac_mxr_part1_out <- tac_mxr_part1_out[,colsToCompare]
tac_gs_part1 <- tac_gs_part1[,colsToCompare]

taxrrow <- do.call(paste, c(tac_mxr_part1_out, sep = '\n'))
gs.taxrrow <- do.call(paste, c(tac_gs_part1, sep = '\n'))

precall(taxrrow, gs.taxrrow)
```

```r
# 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.

```r
bmd <- function(x) {
  fns <- strsplit(x, '\t')
  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']])
```

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

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

**Note level:**

```r
# 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
```

```r
# 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
```

```r
# 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:**

```r
# 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>
```

```r
# dosechange drugname_start dosestr.num strength.num doseamt.num
# 29 <NA> 851 NA 1 4
# 42 <NA> 560 NA 1 3
```

```r
# freq.num dose.intake intaketime dose.seq dose.daily
```
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 = ""
)

Date level:

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

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.

```r
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