Package 'EHR'

October 8, 2020

Version 0.3-1

Date 2020-09-28

Title Electronic Health Record (EHR) Data Processing and Analysis Tool

Maintainer Leena Choi <naturechoi@gmail.com>

Description Process and analyze Electronic Health Record (EHR) data. The 'EHR' package provides modules to perform diverse medication-related studies using data from EHR databases. Especially, the package includes modules to perform pharmacokinetic/pharmacodynamic (PK/PD) analyses using EHRs, as outlined in Choi, Beck, McNeer, Weeks, Williams, James, Niu, Abou-Khalil, Birdwell, Roden, Stein, Bejan, Denny, and Van Driest (2020) <doi:10.1002/cpt.1787>. Additional modules will be added in future. In addition, this package provides various functions useful to perform Phenome Wide Association Study (PheWAS) to explore associations between drug exposure and phenotypes obtained from EHR data, as outlined in Choi, Carroll, Beck, Mosley, Roden, Denny, and Van Driest (2018) <doi:10.1093/bioinformatics/bty306>.

Depends R (>= 2.10)

License GPL (>= 3)

Imports stats, utils, data.table, methods

Suggests glmnet, logistf, medExtractR, knitr, rmarkdown, ggplot2, formula.tools

NeedsCompilation no

RoxygenNote 7.1.1

VignetteBuilder knitr

Author Leena Choi [aut, cre] (<https://orcid.org/0000-0002-2544-7090>), Cole Beck [aut] (<https://orcid.org/0000-0002-6849-6255>), Hannah Weeks [aut], Elizabeth McNeer [aut]

Repository CRAN

Date/Publication 2020-10-08 10:40:02 UTC

R topics documented:

`EHR-package` *Electronic Health Record (EHR) Data Processing and Analysis Tool*

Description

The 'EHR' package provides modules to perform diverse medication-related studies using data from EHR databases.

Details

Package functionality:

- Process and analyze Electronic Health Record (EHR) data.
- Implement modules to perform diverse medication-related studies using data from EHR databases. Especially, the package includes modules to perform pharmacokinetic/pharmacodynamic (PK/PD) analyses using EHRs, as outlined in Choi et al. (2020).
- Implement three statistical methods for Phenome Wide Association Study (PheWAS). Contingency tables for many binary outcomes (e.g., phenotypes) and a binary covariate (e.g., drug exposure) can be efficiently generated by `zeroOneTable`, and three commonly used statistical methods to analyze data for PheWAS are implement by `analysisPheWAS`.

Author(s)

Maintainer: Leena Choi <naturechoi@gmail.com> (ORCID)

Authors:

- Cole Beck <cole.beck@vumc.org> (ORCID)
- Hannah Weeks <hannah.l.weeks@vanderbilt.edu>
- Elizabeth McNeer <elizabeth.mcneer@vumc.org>

References

1. Development of a system for postmarketing population pharmacokinetic and pharmacodynamic studies using real-world data from electronic health records.
Choi L, Beck C, McNeer E, Weeks HL, Williams ML, James NT, Niu X, Abou-Khalil BW, Birdwell KA, Roden DM, Stein CM, Bejan CA, Denny JC, Van Driest SL.
Clin Pharmacol Ther. 2020 Apr;107(4):934-943. doi: 10.1002/cpt.1787.
2. Evaluating statistical approaches to leverage large clinical datasets for uncovering therapeutic and adverse medication effects.
Choi L, Carroll RJ, Beck C, Mosley JD, Roden DM, Denny JC, Van Driest SL.
Bioinformatics. 2018 Sep 1;34(17):2988-2996. doi: 10.1093/bioinformatics/bty306.

`addLastDose` *Add Lastdose Data*

Description

Add lastdose data to data set from the `buildDose` process.

Usage

```
addLastDose(buildData, lastdoseData)
```

Arguments

buildData	data.frame, output of buildDose function.
lastdoseData	data.frame with columns filename, ld_start, lastdose, raw_time, time_type

Details

Lastdose is a datetime string associated with dose data. Information on time of last dose can be extracted within the extractMed function (i.e., medExtractR) using the argument lastdose=TRUE. Raw extracted times should first be processed using the processLastDose function to convert to datetime format before providing to addLastDose. This function then combines the processed last dose times with output from the buildDose process by file name to pair last dose times with dosing regimens based on position. Alternatively, the user can provide their own table of lastdose data. In this case, with position information absent, the lastdose data should be restricted to one unique last dose time per unique patient ID-date identifier.

In the case where lastdoseData is output from processLastDose, it is possible to have more than one extracted last dose time. In this case, rules are applied to determine which time should be kept. First, we give preference to an explicit time expression (e.g., "10:30pm") over a duration expression (e.g., "14 hour level"). Then, we pair last dose times with drug regimens based on minimum distance between last dose time start position and drug name start position.

Value

a data.frame with the 'lastdose' column added.

Examples

```
# Get build data
data(tac_mxr_parsed)
# don't combine lastdose at this stage
tac_build <- buildDose(tac_mxr_parsed, preserve = 'lastdose')
# Get processed last dose data
tac_mxr <- read.csv(system.file("examples", "tac_mxr.csv", package = "EHR"))
data(tac_metadata)
data(tac_lab)
ld_data <- processLastDose(tac_mxr, tac_metadata, tac_lab)

addLastDose(tac_build, ld_data)
```

---

analysisPheWAS *Statistical Analysis for PheWAS*

Description

Implement three commonly used statistical methods to analyze data for Phenome Wide Association Study (PheWAS)

Usage

```
analysisPheWAS(
  method = c("firth", "glm", "lr"),
  adjust = c("PS", "demo", "PS.demo", "none"),
  Exposure,
  PS,
  demographics,
  phenotypes,
  data
)
```

Arguments

method	define the statistical analysis method from 'firth', 'glm', and 'lr'. 'firth': Firth's penalized-likelihood logistic regression; 'glm': logistic regression with Wald test, 'lr': logistic regression with likelihood ratio test.
adjust	define the adjustment method from 'PS','demo','PS.demo', and 'none'. 'PS': adjustment of PS only; 'demo': adjustment of demographics only; 'PS.demo': adjustment of PS and demographics; 'none': no adjustment.
Exposure	define the variable name of exposure variable.
PS	define the variable name of propensity score.
demographics	define the list of demographic variables.
phenotypes	define the list of phenotypes that need to be analyzed.
data	define the data.

Details

Implements three commonly used statistical methods to analyze the associations between exposure (e.g., drug exposure, genotypes) and various phenotypes in PheWAS. Firth's penalized-likelihood logistic regression is the default method to avoid the problem of separation in logistic regression, which is often a problem when analyzing sparse binary outcomes and exposure. Logistic regression with likelihood ratio test and conventional logistic regression with Wald test can be also performed.

Value

estimate	the estimate of log odds ratio.
stdError	the standard error.
statistic	the test statistic.
pvalue	the p-value.

Author(s)

Leena Choi <naturechoi@gmail.com> and Cole Beck <cole.beck@vumc.org>

Examples

```
## use small datasets to run this example
data(dataPheWASsmall)
## make dd.base with subset of covariates from baseline data (dd.baseline.small)
## or select covariates with upper code as shown below
upper.code.list <- unique(sub("[.][^.]*(.).*", "", colnames(dd.baseline.small)) )
upper.code.list <- intersect(upper.code.list, colnames(dd.baseline.small))
dd.base <- dd.baseline.small[, upper.code.list]
## perform regularized logistic regression to obtain propensity score (PS)
## to adjust for potential confounders at baseline
phenos <- setdiff(colnames(dd.base), c('id', 'exposure'))
data.x <- as.matrix(dd.base[, phenos])
glmnet.fit <- glmnet::cv.glmnet(x=data.x, y=dd.base[,'exposure'],
                                family="binomial", standardize=TRUE,
                                alpha=0.1)
dd.base$PS <- c(predict(glmnet.fit, data.x, s='lambda.min'))
data.ps <- dd.base[,c('id', 'PS')]
dd.all.ps <- merge(data.ps, dd.small, by='id')
demographics <- c('age', 'race', 'gender')
phenotypeList <- setdiff(colnames(dd.small), c('id','exposure','age','race','gender'))
## run with a subset of phenotypeList to get quicker results
phenotypeList.sub <- sample(phenotypeList, 5)
results.sub <- analysisPheWAS(method='firth', adjust='PS', Exposure='exposure',
                              PS='PS', demographics=demographics,
                              phenotypes=phenotypeList.sub, data=dd.all.ps)
## run with the full list of phenotype outcomes (i.e., phenotypeList)

        results <- analysisPheWAS(method='firth', adjust='PS',Exposure='exposure',
                          PS='PS', demographics=demographics,
                          phenotypes=phenotypeList, data=dd.all.ps)
```

`buildDose` *Combine Dose Data*

Description

Output from parse process is taken and converted into a wide format, grouping drug entity information together based on various steps and rules.

Usage

```
buildDose(
  dat,
  dn = NULL,
  preserve = NULL,
  dist_method,
  na_penalty,
  neg_penalty,
```

```
  greedy_threshold,
  checkForRare = FALSE
)
```

Arguments

dat	data.table object from the output of parseMedExtractR, parseMedXN, parseMedEx, or parseCLAMP
dn	Regular expression specifying drug name(s) of interest.
preserve	Column names to include in output, whose values should not be combined with other rows. If present, dosechange is always preserved.
dist_method	Distance method to use for calculating distance of various paths. Alternatively set the 'ehr.dist_method' option, which defaults to 'minEntEnd'.
na_penalty	Penalty for matching extracted entities with NA. Alternatively set the 'ehr.na_penalty' option, which defaults to 32.
neg_penalty	Penalty for negative distances between frequency/intake time and dose amounts. Alternatively set the 'ehr.neg_penalty' option, which defaults to 0.5.
greedy_threshold	Threshold to use greedy matching; increasing this value too high could lead to the algorithm taking a long time to finish. Alternatively set the 'ehr.greedy_threshold' option, which defaults to 1e8.
checkForRare	Indicate if rare values for each entity should be found and displayed.

Details

The buildDose function takes as its main input (dat), a data.table object that is the output of a parse process function (parseMedExtractR, parseMedXN, parseMedEx, or parseCLAMP). Broadly, the parsed extractions are grouped together to form wide, more complete drug regimen information. This reformatting facilitates calculation of dose given intake and daily dose in the collapseDose process.

The process of creating this output is broken down into multiple steps:

1. Removing rows for any drugs not of interest. Drugs of interest are specified with the dn argument.
2. Determining whether extractions are "simple" (only one drug mention and at most one extraction per entity) or complex. Complex cases can be more straightforward if they contain at most one extraction per entity, or require a pairing algorithm to determine the best pairing if there are multiple extractions for one or more entities.
3. Drug entities are anchored by drug name mention within the parse process. For complex cases, drug entities are further grouped together anchored at each strength (and dose with medExtractR) extraction.
4. For strength groups with multiple extractions for at least one entity, these groups go through a path searching algorithm, which computes the cost for each path (based on a chosen distance method) and chooses the path with the lowest cost.
5. The chosen paths for each strength group are returned as the final pairings. If route is unique within a strength group, it is standardized and added to all entries for that strength group.

The user can specify additional arguments including:

- `dist_method`: The distance method is the metric used to determine which entity path is the most likely to be correct based on minimum cost.
- `na_penalty`: NA penalties are incurred when extractions are paired with nothing (i.e., an NA), requiring that entities be sufficiently far apart from one another before being left unpaired.
- `neg_penalty`: When working with dose amount (DA) and frequency/intake time (FIT), it is much more common for the ordering to be DA followed by FIT. Thus, when we observe FIT followed by DA, we apply a negative penalty to make such pairings less likely.
- `greedy threshold`: When there are many extractions from a clinical note, the number of possible combinations for paths can get exponentially large, particularly when the medication extraction natural language processing system is incorrect. The greedy threshold puts an upper bound on the number of entity pairings to prevent the function from stalling in such cases.

If none of the optional arguments are specified, then the `buildDose` process uses the default option values specified in the EHR package documentation.

For additional details, see McNeer, et al. 2020.

Value

A data.frame object that contains columns for filename (of the clinical note, inherited from the parse output object `dat`), drugname, strength, dose, route, freq, duration, and drugname_start

Examples

```
data(lam_mxr_parsed)

buildDose(lam_mxr_parsed)
```

`collapseDose` *Collapse Dose Data*

Description

Splits drug data and calls `makeDose` to collapse at the note and date level.

Usage

```
collapseDose(x, noteMetaData, naFreq = "most", ...)
```

Arguments

`x`	data.frame containing the output of `buildDose`, or the output of `addLastDose` if last dose information is being incorporated.
`noteMetaData`	data.frame containing identifying meta data for each note, including patient ID, date of the note, and note ID. Column names should be set to 'filename', 'pid', 'date', 'note'. Date should have format YYYY-MM-DD.

naFreq	Expression used to replace missing frequencies with, or by default use the most common.
...	drug formulations to split by

Details

If different formulations of the drug (e.g., extended release) exist, they can be separated using a regular expression (e.g., 'xr|er'). This function will call makeDose on parsed and paired medication data to calculate dose intake and daily dose and remove redundancies at the note and date level.

Value

A list containing two dataframes, one with the note level and one with the date level collapsed data.

Examples

```
data(lam_mxr_parsed)
data(lam_metadata)

lam_build_out <- buildDose(lam_mxr_parsed)

lam_collapsed <- collapseDose(lam_build_out, lam_metadata, naFreq = 'most', 'xr|er')
lam_collapsed$note # Note level collapsing
lam_collapsed$date # Date level collapsing
```

---

dataTransformation *Data Transformation*

Description

Convenience function for making small modifications to a data.frame.

Usage

```
dataTransformation(x, select, rename, modify)
```

Arguments

x	a data.frame
select	columns to select
rename	character vector with names for all columns
modify	list of expressions used to transform data set

Value

The modified data.frame

dd *dd*

Description

Simulated outcome data example from Phenome Wide Association Study (PheWAS) that examines associations between drug exposure and various phenotypes at follow-up after the drug exposure. The dataset includes 1505 variables: subject identification number ('id'), drug exposure ('exposure'), 3 demographic variables ('age', 'race', 'gender'), and 1500 phenotypes.

Usage

```
data(dataPheWAS, package = 'EHR')
```

Format

A data frame with 10000 observations on 1505 variables.

Examples

```
data(dataPheWAS)
```

dd.baseline *dd.baseline*

Description

Simulated baseline data example from a Phenome Wide Association Study (PheWAS) obtained at baseline before drug exposure. The dataset includes 1505 variables: subject identification number ('id'), drug exposure ('exposure'), 3 demographic variables ('age', 'race', 'gender'), and 1500 phenotypes.

Usage

```
data(dataPheWAS, package = 'EHR')
```

Format

A data frame with 10000 observations on 1505 variables.

Examples

```
data(dataPheWAS)
```

dd.baseline.small *dd.baseline.small*

Description

A smaller subset of baseline data example, dd.baseline. The dataset includes 55 variables: subject identification number ('id'), drug exposure ('exposure'), 3 demographic variables ('age', 'race', 'gender'), and 50 phenotypes.

Usage

```
data(dataPheWASsmall, package = 'EHR')
```

Format

A data frame with 2000 observations on 55 variables.

Examples

```
data(dataPheWASsmall)
```

dd.small *dd.small*

Description

A smaller subset of outcome data example, 'dd'. The dataset includes 55 variables: subject identification number ('id'), drug exposure ('exposure'), 3 demographic variables ('age', 'race', 'gender'), and 50 phenotypes.

Usage

```
data(dataPheWASsmall, package = 'EHR')
```

Format

A data frame with 2000 observations on 55 variables.

Examples

```
data(dataPheWASsmall)
```

`extractMed` *Extract medication information from clinical notes*

Description

This function is an interface to the `medExtractR` function within the **medExtractR** package, and allows drug dosing information to be extracted from free-text sources, e.g., clinical notes.

Usage

```
extractMed(note_fn, drugnames, drgunit, windowlength, max_edit_dist = 0, ...)
```

Arguments

`note_fn`	File name(s) for the text file(s) containing the clinical notes. Can be a character string for an individual note, or a vector or list of file names for multiple notes.
`drugnames`	Vector of drug names for which dosing information should be extracted. Can include various forms (e.g., generic, brand name) as well as abbreviations.
`drgunit`	Unit of the drug being extracted, e.g., 'mg'
`windowlength`	Length of the search window (in characters) around the drug name in which to search for dosing entities
`max_edit_dist`	Maximum edit distance allowed when attempting to extract `drugnames`. Allows for capturing misspelled drug name information.
`...`	Additional arguments to `medExtractR`, for example `lastdose=TRUE` to extract time of last dose (see **medExtractR** package documentation for details)

Details

Medication information, including dosing data, is often stored in free-text sources such as clinical notes. The `extractMed` function serves as a convenient wrapper for the **medExtractR** package, a natural language processing system written in R for extracting medication data. Within `extractMed`, the `medExtractR` function identifies dosing data for drug(s) of interest, specified by the `drugnames` argument, using rule-based and dictionary-based approaches. Relevant dosing entities include medication strength (identified using the `unit` argument), dose amount, dose given intake, intake time or frequency of dose, dose change keywords (e.g., 'increase' or 'decrease'), and time of last dose. For more details, see Weeks, et al. 2020. After applying `medExtractR` to extract drug dosing information, `extractMed` appends the file name to results to ensure they are appropriately labeled.

Value

A data.frame with the extracted dosing information, labeled with file name as an identifier
Sample output:

filename	entity	expr	pos

note_file1.txt	DoseChange	decrease	66:74
note_file1.txt	DrugName	Prograf	78:85
note_file1.txt	Strength	2 mg	86:90
note_file1.txt	DoseAmt	1	91:92
note_file1.txt	Frequency	bid	101:104
note_file1.txt	LastDose	2100	121:125

Examples

```
tac_fn <- list(system.file("examples", "tacpid1_2008-06-26_note1_1.txt", package = "EHR"),
               system.file("examples", "tacpid1_2008-06-26_note2_1.txt", package = "EHR"),
               system.file("examples", "tacpid1_2008-12-16_note3_1.txt", package = "EHR"))

extractMed(tac_fn,
           drugnames = c("tacrolimus", "prograf", "tac", "tacro", "fk", "fk506"),
           drgunit = "mg",
           windowlength = 60,
           max_edit_dist = 2,
           lastdose=TRUE)
```

---

freqNum *Convert Character Frequency to Numeric*

Description

This function converts the frequency entity to numeric.

Usage

```
freqNum(x)
```

Arguments

x	character vector of extracted frequency values

Value

numeric vector

Examples

```
f <- stdzFreq(c('in the morning', 'four times a day', 'with meals'))
freqNum(f)
```

lam_metadata *Example of Metadata for Lamotrigine Data*

Description

An example of the metadata needed for the `processLastDose`, `makeDose`, and `collapseDose` functions.

Usage

```
data(lam_metadata, package = 'EHR')
```

Format

A data frame with 5 observations on the following variables.

filename A character vector, filename for the clinical note

pid A character vector, patient ID associated with the filename

date A character vector, date associated with the filename

note A character vector, note ID associated with the filename

Examples

```
data(lam_metadata)
```

lam_mxr_parsed *Example of Lamotrigine Output from 'parseMedExtractR'*

Description

The output after running `parseMedExtractR` on 4 example clinical notes.

Usage

```
data(lam_mxr_parsed, package = 'EHR')
```

Format

A data frame with 10 observations on the following variables.

filename A character vector, filename for the clinical note

drugname A character vector, drug name extracted from the clinical note along with start and stop positions

strength A character vector, strengths extracted from the clinical note along with start and stop positions

dose A character vector, dose amounts extracted from the clinical note along with start and stop positions

route A character vector, routes extracted from the clinical note along with start and stop positions

freq A character vector, frequencies extracted from the clinical note along with start and stop positions

dosestr A character vector, dose intakes extracted from the clinical note along with start and stop positions

dosechange A character vector, dose change keywords extracted from the clinical note along with start and stop positions

lastdose A character vector, last dose times extracted from the clinical note along with start and stop positions

Examples

```
data(lam_mxr_parsed)
```

`Logistf` *Firth's penalized-likelihood logistic regression with more decimal places of p-value than* `logistf` *function in the R package 'logistf'*

Description

Adapted from `logistf` in the R package 'logistf', this is the same as `logistf` except that it provides more decimal places of p-value that would be useful for Genome-Wide Association Study (GWAS) or Phenome Wide Association Study (PheWAS).

Usage

```
Logistf(
  formula,
  data,
  pl = TRUE,
  alpha = 0.05,
  control,
  plcontrol,
  firth = TRUE,
  init,
  weights,
  plconf = NULL,
  flic = FALSE,
  model = TRUE,
  ...
)
```

Arguments

formula	a formula object, with the response on the left of the operator, and the model terms on the right. The response must be a vector with 0 and 1 or FALSE and TRUE for the outcome, where the higher value (1 or TRUE) is modeled. It is possible to include contrasts, interactions, nested effects, cubic or polynomial splines and all S features as well, e.g. `Y ~ X1*X2 + ns(X3,df=4)`. From version 1.10, you may also include offset() terms.
data	a data.frame where the variables named in the formula can be found, i. e. the variables containing the binary response and the covariates.
pl	specifies if confidence intervals and tests should be based on the profile penalized log likelihood (`pl=TRUE`, the default) or on the Wald method (`pl=FALSE`).
alpha	the significance level (1-α the confidence level, 0.05 as default).
control	Controls Newton-Raphson iteration. Default is `control=logistf.control(maxstep,maxit,maxhs,lconv,gconv,xconv)`
plcontrol	Controls Newton-Raphson iteration for the estimation of the profile likelihood confidence intervals. Default is `plcontrol=logistpl.control(maxstep,maxit, maxhs,lconv,xconv,ortho,pr)`
firth	use of Firth's penalized maximum likelihood (`firth=TRUE`, default) or the standard maximum likelihood method (`firth=FALSE`) for the logistic regression. Note that by specifying `pl=TRUE` and `firth=FALSE` (and probably a lower number of iterations) one obtains profile likelihood confidence intervals for maximum likelihood logistic regression parameters.
init	specifies the initial values of the coefficients for the fitting algorithm.
weights	specifies case weights. Each line of the input data set is multiplied by the corresponding element of `weights`.
plconf	specifies the variables (as vector of their indices) for which profile likelihood confidence intervals should be computed. Default is to compute for all variables.
flic	If TRUE, intercept is altered such that the predicted probabilities become unbiased while keeping all other coefficients constant
model	If TRUE the corresponding components of the fit are returned.
...	Further arguments to be passed to logistf.

Value

same as `logistf` except for providing more decimal places of p-value.

Author(s)

Leena Choi <naturechoi@gmail.com> and Cole Beck <cole.beck@vumc.org>

References

same as those provided in the R package 'logistf'.

Examples

```
data(dataPheWAS)
fit <- Logistf(X264.3 ~ exposure + age + race + gender, data=dd)
summary(fit)
```

---

makeDose *Make Dose Data*

Description

Takes parsed and paired medication data, calculates dose intake and daily dose, and removes redundant information at the note and date level.

Usage

```
makeDose(x, noteMetaData, naFreq = "most")
```

Arguments

x	data.frame containing the output of buildDose, or the output of addLastDose if last dose information is being incorporated.
noteMetaData	data.frame containing identifying meta data for each note, including patient ID, date of the note, and note ID. Column names should be set to 'filename', 'pid', 'date', 'note'. Date should have format YYYY-MM-DD.
naFreq	Replacing missing frequencies with this value, or by default the most common value across the entire set in x.

Details

This function standardizes frequency, route, and duration entities. Dose amount, strength, and frequency entities are converted to numeric. Rows with only drug name and/or route are removed. If there are drug name changes in adjacent rows (e.g., from a generic to brand name), these rows are collapsed into one row if there are no conflicts. Missing strengths, dose amounts, frequencies, and routes are borrowed or imputed using various rules (see McNeer et al., 2020 for details). Dose given intake and daily dose are calculated. Redundancies are removed at the date and note level. If time of last dose is being used and it is unique within the level of collapsing, it is borrowed across all rows.

Value

A list containing two dataframes, one with the note level and one with the date level collapsed data.

Examples

```
data(lam_mxr_parsed)
data(lam_metadata)

lam_build_out <- buildDose(lam_mxr_parsed)

lam_collapsed <- makeDose(lam_build_out, lam_metadata)
lam_collapsed[[1]] # Note level collapsing
lam_collapsed[[2]] # Date level collapsing
```

---

parseCLAMP *Parse CLAMP NLP Output*

Description

Takes files with the raw medication extraction output generated by the CLAMP natural language processing system and converts it into a standardized format.

Usage

```
parseCLAMP(filename)
```

Arguments

filename	File name for a single file containing CLAMP output.

Details

Output from different medication extraction systems is formatted in different ways. In order to be able to process the extracted information, we first need to convert the output from different systems into a standardized format. Extracted expressions for various drug entities (e.g., drug name, strength, frequency, etc.) each receive their own column formatted as "extracted expression::start position::stop position". If multiple expressions are extracted for the same entity, they will be separated by backticks.

CLAMP output files anchor extractions to a specific drug name extraction through semantic relations.

Value

A data.table object with columns for filename, drugname, strength, dose, route, and freq. The filename contains the file name corresponding to the clinical note. Each of the entity columns are of the format "extracted expression::start position::stop position".

`parseMedEx` *Parse MedEx NLP Output*

Description

Takes files with the raw medication extraction output generated by the MedEx natural language processing system and converts it into a standardized format.

Usage

```
parseMedEx(filename)
```

Arguments

`filename`	File name for a single file containing MedEx output.

Details

Output from different medication extraction systems is formatted in different ways. In order to be able to process the extracted information, we first need to convert the output from different systems into a standardized format. Extracted expressions for various drug entities (e.g., drug name, strength, frequency, etc.) each receive their own column formatted as "extracted expression::start position::stop position". If multiple expressions are extracted for the same entity, they will be separated by backticks.

MedEx output files anchor extractions to a specific drug name extraction.

Value

A data.table object with columns for filename, drugname, strength, dose, route, and freq. The filename contains the file name corresponding to the clinical note. Each of the entity columns are of the format "extracted expression::start position::stop position".

`parseMedExtractR` *Parse medExtractR NLP Output*

Description

Takes files with the raw medication extraction output generated by the medExtractR natural language processing system and converts it into a standardized format.

Usage

```
parseMedExtractR(filename)
```

Arguments

filename	File name for a single file containing medExtractR output.

Details

Output from different medication extraction systems is formatted in different ways. In order to be able to process the extracted information, we first need to convert the output from different systems into a standardized format. Extracted expressions for various drug entities (e.g., drug name, strength, frequency, etc.) each receive their own column formatted as "extracted expression::start position::stop position". If multiple expressions are extracted for the same entity, they will be separated by backticks.

The medExtractR system returns extractions in a long table format, indicating the entity, extracted expression, and start:stop position of the extraction. To perform this initial parsing, entities are paired with the closest preceding drug name. The one exception to this is the dose change entity, which can occur before the drug name (see Weeks, et al. 2020 for details).

Value

A data.table object with columns for filename, drugname, strength, dose, route, freq, dosestr, dosechange and lastdose. The filename contains the file name corresponding to the clinical note. Each of the entity columns are of the format "extracted expression::start position::stop position".

Examples

```
mxr_output <- system.file("examples", "lam_mxr.csv", package = "EHR")
mxr_parsed <- parseMedExtractR(mxr_output)
mxr_parsed
```

---

parseMedXN *Parse MedXN NLP Output*

Description

Takes files with the raw medication extraction output generated by the MedXN natural language processing system and converts it into a standardized format.

Usage

```
parseMedXN(filename, begText = "^[R0-9]+_[0-9-]+_[0-9]+_")
```

Arguments

filename	File name for single file containing MedXN output.
begText	A regular expression that would indicate the beginning of a new observation (i.e., extracted clinical note).

Details

Output from different medication extraction systems is formatted in different ways. In order to be able to process the extracted information, we first need to convert the output from different systems into a standardized format. Extracted expressions for various drug entities (e.g., drug name, strength, frequency, etc.) each receive their own column formatted as "extracted expression::start position::stop position". If multiple expressions are extracted for the same entity, they will be separated by backticks.

MedXN output files anchor extractions to a specific drug name extraction.

In MedXN output files, the results from multiple clinical notes can be combined into a single output file. The beginning of some lines of the output file can indicate when output for a new observation (or new clinical note) begins. The user should specify the argument `begText` to be a regular expression used to identify the lines where output for a new clinical note begins.

Value

A data.table object with columns for filename, drugname, strength, dose, route, freq, and duration. The filename contains the file name corresponding to the clinical note. Each of the entity columns are of the format "extracted expression::start position::stop position".

Examples

```
mxn_output <- system.file("examples", "lam_medxn.csv", package = "EHR")
mxn_parsed <- parseMedXN(mxn_output, begText = "^ID[0-9]+_[0-9-]+_")
mxn_parsed
```

---

processLastDose *Process and standardize extracted last dose times*

Description

This function takes last dose times extracted using the **medExtractR** system and processes the times into standardized datetime objects using recorded lab data where necessary. The raw output from `extractMed` is filtered to just the LastDose extractions. Time expressions are standardized into HH:MM:SS format based on what category they fall into (e.g., a time represented with AM/PM, 24-hour military time, etc.). When the last dose time is after 12pm, it is assumed to have been taken one day previous to the note's date. For any duration extractions (e.g. "14 hour level"), the last dose time is calculated from the labtime by extracting the appropriate number of hours. The final dataset is returned with last dose time formatted into a POSIXct variable.

Usage

```
processLastDose(mxrData, noteMetaData, labData)
```

Arguments

mxrData	data.frame containing output from the medExtractR system
noteMetaData	data.frame with meta data (pid (patient ID) and date) for the file names contained within mxrData
labData	data.frame that contains lab dates and times associated with the file names within mxrData. Must contain columns pid and date, as well as labtime. The date column must be in the same format as date in noteMetaData, and labtime must be a POSIXct

Value

data.frame with identifying information (e.g., filename, etc) as well as processed and standardized last dose times as a POSIXct column

Examples

```
tac_mxr <- read.csv(system.file("examples", "tac_mxr.csv", package = "EHR"))
data(tac_metadata)
data(tac_lab)

processLastDose(mxrData = tac_mxr, noteMetaData = tac_metadata, labData = tac_lab)
```

---

readTransform *Read and Transform*

Description

Convenience function for reading in a CSV file, and making small modifications to a data.frame.

Usage

```
readTransform(file, ...)
```

Arguments

file	filename of a CSV file
...	additional information passed to dataTransformation

Details

If read.csv needs additional arguments (or the file is in a different format), the user should load the data first, then directly call dataTransformation.

Value

The modified data.frame

stdzDose *Standardize Dose Entity*

Description

This function standardizes the dose entity.

Usage

```
stdzDose(x)
```

Arguments

x	character vector of extracted dose values

Details

Some dose strings may include multiple values and additional interpretation may be needed. For example '2-1' likely indicates a dose of 2 followed by a dose of 1. Currently it would be converted to the average of 1.5.

Value

numeric vector

Examples

```
stdzDose(c('one tablet', '1/2 pill', '1-3 tabs'))
```

stdzDuration *Standardize Duration Entity*

Description

This function standardizes the duration entity.

Usage

```
stdzDuration(x)
```

Arguments

x	character vector of extracted duration values

Value

character vector

Examples

```
stdzDuration(c('1 month', 'three days', 'two-weeks'))
```

`stdzFreq` *Standardize Frequency Entity*

Description

This function standardizes the frequency entity.

Usage

```
stdzFreq(x)
```

Arguments

`x`	character vector of extracted frequency values

Value

character vector

Examples

```
stdzFreq(c('in the morning', 'four times a day', 'with meals'))
```

`stdzRoute` *Standardize Route Entity*

Description

This function standardizes the route entity.

Usage

```
stdzRoute(x)
```

Arguments

`x`	character vector of extracted route values

Value

character vector

Examples

```
stdzRoute(c('oral', 'po', 'subcut'))
```

stdzStrength *Standardize Strength Entity*

Description

This function standardizes the strength entity.

Usage

```
stdzStrength(str, freq)
```

Arguments

str	character vector of extracted strength values
freq	character vector of extracted frequency values

Details

Some strength strings may include multiple values and additional interpretation may be needed. For example '2-1' likely indicates a strength of 2 followed by a strength of 1. Thus a single element may need to be standarized into two elements. This can only happen if the frequency entity is missing or in agreement ('bid' for example). See the 'addl_data' attribute of the returned vector.

Value

numeric vector

Examples

```
stdzStrength(c('1.5', '1/2', '1/1/1'))
stdzStrength(c('1.5', '1/2', '1/1/1'), c('am', 'daily', NA))
stdzStrength(c('1.5', '1/2', '1/1/1'), FALSE)
```

tac_lab *Example of Lab Time Data for Tacrolimus*

Description

An example dataset used in processLastDose that contains lab time data. This dataset should have one row per patient ID-date pair, and contain the time a lab was performed as a datetime variable.

Usage

```
data(tac_lab, package = 'EHR')
```

Format

A data frame with 2 observations on the following variables.

pid A character vector, patient ID associated with the lab value

date A character vector, date associated with the lab value

labtime A POSIXct vector, datetime at which the lab was performed formatted as YYYY-MM-DD HH:MM:SS

Examples

```
data(tac_lab)
```

`tac_metadata` *Example of Metadata for Tacrolimus Data*

Description

An example of the metadata needed for the `processLastDose`, `makeDose`, and `collapseDose` functions.

Usage

```
data(tac_metadata, package = 'EHR')
```

Format

A data frame with 5 observations on the following variables.

filename A character vector, filename for the clinical note

pid A character vector, patient ID associated with the filename

date A character vector, date associated with the filename

note A character vector, note ID associated with the filename

Examples

```
data(tac_metadata)
```

tac_mxr_parsed *Example of Tacrolimus Output from 'parseMedExtractR'*

Description

The output after running parseMedExtractR on 3 example clinical notes.

Usage

```
data(tac_mxr_parsed, package = 'EHR')
```

Format

A data frame with 7 observations on the following variables.

filename A character vector, filename for the clinical note

drugname A character vector, drug name extracted from the clinical note along with start and stop positions

strength A character vector, strengths extracted from the clinical note along with start and stop positions

dose A character vector, dose amounts extracted from the clinical note along with start and stop positions

route A character vector, routes extracted from the clinical note along with start and stop positions

freq A character vector, frequencies extracted from the clinical note along with start and stop positions

dosestr A character vector, dose intakes extracted from the clinical note along with start and stop positions

dosechange A character vector, dose change keywords extracted from the clinical note along with start and stop positions

lastdose A character vector, last dose times extracted from the clinical note along with start and stop positions

Examples

```
data(tac_mxr_parsed)
```

zeroOneTable *Make Zero One Contingency Tables*

Description

Make contingency tables for many binary outcomes and a binary covariate

Usage

```
zeroOneTable(EXPOSURE, phenotype)
```

Arguments

EXPOSURE	binary covariate (e.g., exposure).
phenotype	binary outcome (e.g., phenotype).

Details

Generates frequency and contingency tables for many binary outcomes (e.g., large number of phenotypes) and a binary covariate (e.g., drug exposure, genotypes) more efficiently.

Value

t00	frequency for non-exposed group and non-case outcome.
t01	frequency for non-exposed group and case outcome.
t10	frequency for exposed group and non-case outcome.
t11	frequency for exposed group and case outcome.

Author(s)

Leena Choi <naturechoi@gmail.com> and Cole Beck <cole.beck@vumc.org>

Examples

```
## full example data
data(dataPheWAS)
demo.covariates <- c('id','exposure','age','race','gender')
phenotypeList <- setdiff(colnames(dd), demo.covariates)
tablePhenotype <- matrix(NA, ncol=4, nrow=length(phenotypeList),
dimnames=list(phenotypeList, c("n.nocase.nonexp", "n.case.nonexp",
"n.nocase.exp", "n.case.exp")))
for(i in seq_along(phenotypeList)) {
    tablePhenotype[i, ] <- zeroOneTable(dd[, 'exposure'], dd[, phenotypeList[i]])
}
```

Index