Mapping Drug.com Info

All preprocessing files are stored in preprocessing/

The output\_full.csv file was deleted before moving to O2 since it was taking up too much space. Run scrape.py to re-scrape the data if needed.

All maping files are stored in output/

All missing mapping information are stored in missed mappings/

All dictionaries/mapping references are stored in ../mapping data/

# Output\_full.csv

## Information about the dataset

1. Scraped form drug.com
2. 15121 rows, 9 columns
3. 9 columns are **'disease name', 'other names', 'RX/OTC', 'drug', 'generic name', 'drug class', 'brand names', 'pregnancy label', 'CSA label'**

# Disease-drug.csv

## Information about the dataset

1. The data set is a cleaned version of output\_full.csv file.
2. 15121 rows, 9 columns.
3. 14624 rows are RX drugs.
4. 9 columns are **'disease name', 'other names', 'RX/OTC', 'drug', 'generic name', 'drug class', 'brand names', 'pregnancy label', 'CSA label'**

## Cleansing Standard:

1. All characters are lowercased
2. “|” represents “or”
3. All “and”, “ & ”, “ + ”, “, ” replaced with “/” since they indicate “or” (Manually checked).
4. All original “ \ “ and “ / “ replaced with “|” since they indicate “or” (Manually checked).
5. Parentheses are kept since they are used in UMLS (Manually checked).
6. “-” are kept since they are used in UMLS (Manually checked).
7. All other punctuations are cleaned.

# Disease-PheCode

## Exact Mapping

### Methodology

Create exact matching from disease names using the mapping data/broad\_new\_full\_dict.txt file.

### Information about the dataset

1. The data set contains drugs with only RX tags in the disease-drug.csv file
2. 14624 rows
3. 9 columns are 'disease name', 'other names', 'RX/OTC', 'drug', 'generic name', 'drug class', 'brand names', 'pregnancy label', 'CSA label'

### Generic\_name - RXNorm Mapping Rate

0.2732752846617549

### Generic\_name - RXNorm Output Data

* Column name: “exact\_pheCodes”.
* Appended to the cleaned disease-drug.csv as the last column. Stored as PheCode\_rxNorm.csv.

### Brand\_name - RXNorm Mapping Rate

0.2732752846617549

### Brand\_name - RXNorm Output Data

* Column name: “exact\_pheCodes”.
* Appended to the cleaned disease-drug.csv as the last column. Stored as PheCode\_rxNorm.csv.

## 

## Stemmed Exact Mapping

### Methodology and Info

Same as exact mapping, but all strings are stemmed.

### Mapping Rate

0.16945746818486268

Comment: This should be >= exact matching. Not sure what’s going wrong.

### Output Data

* Column name: “stemmed\_excat\_pheCodes”
* Appended to the PheCode\_rxNorm.csv as the last column. Stored as PheCode\_rxNorm.csv.

## Disease - CUI (multiple) - PheCode

### Methodology

1. Map all names of diseases to CUIs
2. Use CUIs to find corresponding PheCodes through the broad\_new\_full\_dict.txt

### Disease - CUI

#### Mapping Rate

0.7756195579370395

#### Output Data

* Column name: “cui\_from\_disease”
* Appended to the cleaned PheCode\_rxNorm.csv as the last column. Stored as PheCode\_rxNorm.csv.

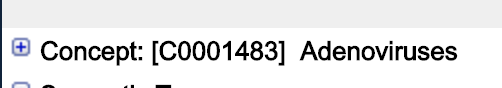
### Enhanced Disease - CUI

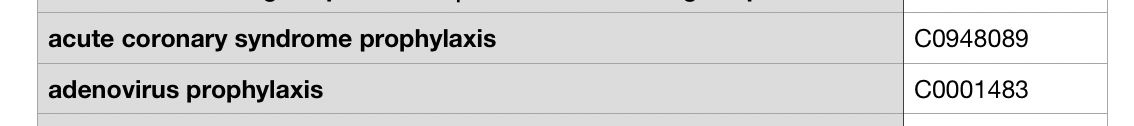
#### Evaluation 1

335 disease names are missing in the CUI mapping. They are stored in the missed CUIs from disease.csv file.

#### Enhancement 1

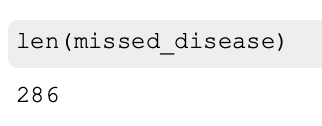
Disease names with “prophylaxis” could be mapped to CUI without “prophylaxis”.

([https://uts.nlm.nih.gov/metathesaurus.html;jsessionid=864F36E3F49E3F93BA04D82CD91A6C93#C0001483;0;1;CUI;2020AA;EXACT\_MATCH;CUI;\*;](https://uts.nlm.nih.gov/metathesaurus.html;jsessionid=864F36E3F49E3F93BA04D82CD91A6C93#C0001483;0;1;CUI;2020AA;EXACT_MATCH;CUI;*;))

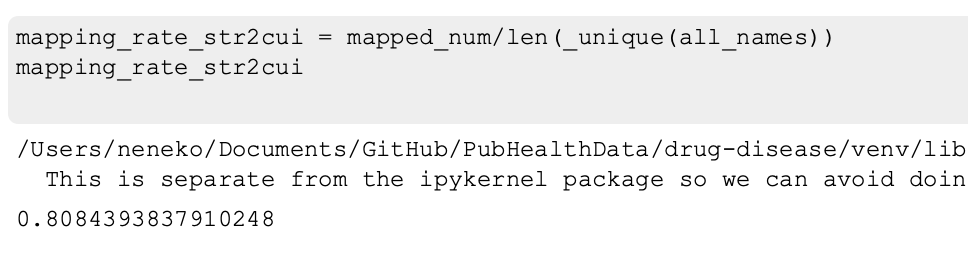


#### Evaluation 2

The missing diseases are reduced to 287.



Mapping rate increase to 0.8084393837910248



#### Enhancement 2

1. Get rid of “,” in the broad.txt file.

#### 



But the broad.txt dictionary does not contain a similar string.

1. Disease names with “human” could be mapped to CUI without “prophylaxis”.



([https://uts.nlm.nih.gov/metathesaurus.html;jsessionid=864F36E3F49E3F93BA04D82CD91A6C93#C1720505;0;1;CUI;2020AA;EXACT\_MATCH;CUI;\*;](https://uts.nlm.nih.gov/metathesaurus.html;jsessionid=864F36E3F49E3F93BA04D82CD91A6C93#C1720505;0;1;CUI;2020AA;EXACT_MATCH;CUI;*;))



1. States are described in UMLS in the following way. But this term cannot be found in the broad\_txt we are using,





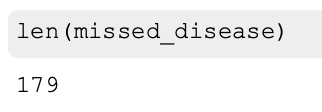
1. Manually filled out the CUI mapping using UMLS. The mapping is stored in the file mapping data/missed\_CUIs\_from\_disease.csv. The terms are missing because of misspellings, different sequences of word orderings and etc.

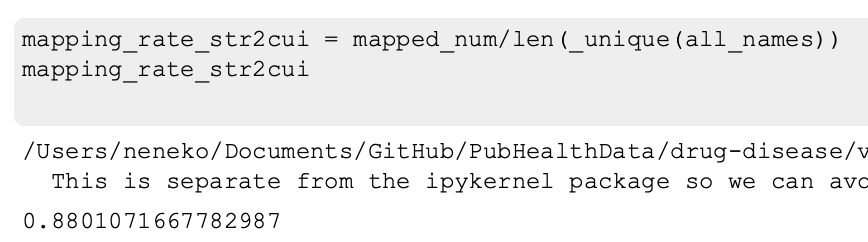
#### Mapping Rate

0.8801071667782987

#### Output Data

* Column name: “cui\_from\_disease+”
* Appended to the cleaned PheCode\_rxNorm.csv as the last column. Stored as PheCode\_rxNorm.csv.

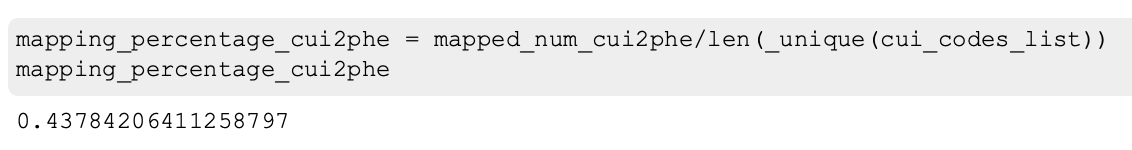




### CUI - PheCode

#### Mapping Rate

0.43784206411258797



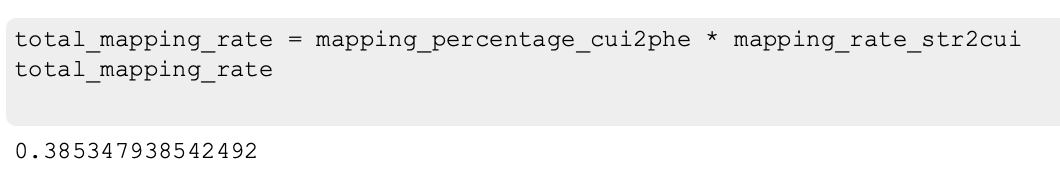
#### Output Data

* Column name: “phe\_from\_CUI”
* Appended to the cleaned PheCode\_rxNorm.csv as the last column. Stored as PheCode\_rxNorm.csv.

### 

### Net Mapping Rate

0.385347938542492



# Drug - RX\_Norm

## Exact Mapping

### Methodology

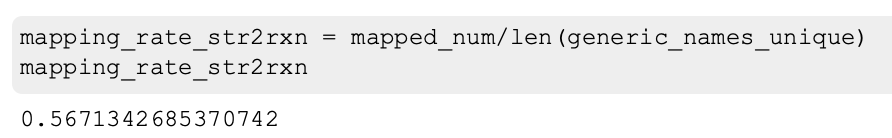
1. Create exact matching from generic names using the mapping data/RXNORM-ingredient-base.csv file.
2. Create exact matching from brand names using the mapping data/RXNORM-ingredient-base.csv file.

### Information about the dataset

1. The data set contains drugs with only RX tags in the disease-drug.csv file
2. 14624 rows
3. 9 columns are 'disease name', 'other names', 'RX/OTC', 'drug', 'generic name', 'drug class', 'brand names', 'pregnancy label', 'CSA label'

### Generic Mapping Rate

0.5671342685370742

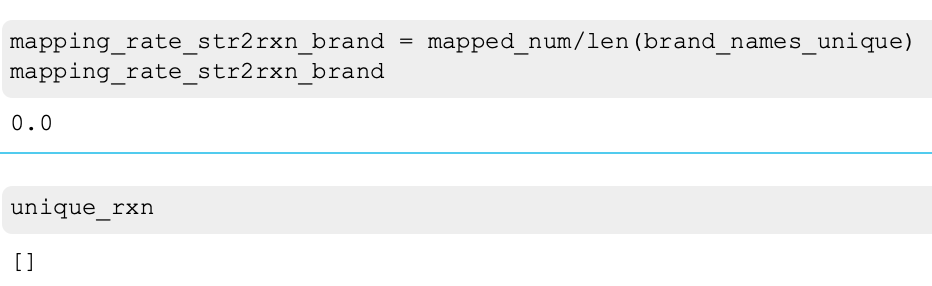


### Generic Output Data

* Column name: “exact\_rxNorm\_generic”.
* Appended to the cleaned PheCode\_rxNorm.csv as the last column. Stored as PheCode\_rxNorm.csv.

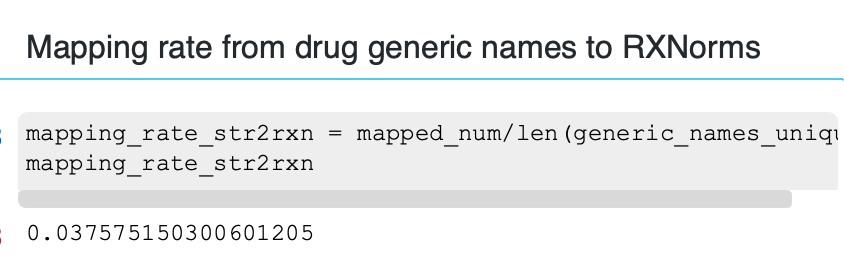
### Brand Mapping Rate

Nothing is mapped.



#### Evaluation

I doubt that the base\_str is not for brand names. By looking at the data file, it should be a more detailed version of the generic name column. I thus tried to map generic names using the base\_str entry, and ~0.038 of the terms are matched. This result is not written to the output file since it’s a relatively small enhancement.



## Generic/Brand - CUI (multiple) - RXNorms

### Generic/Brand - CUI (multiple)

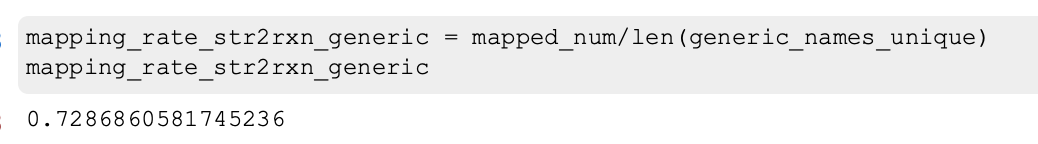
#### Methodology

Map all names of drugs to CUIs

#### Generic Names - CUI

#### Mapping Rate

0.7286860581745236



#### Output Data

* Column name: “cui\_from\_generic”
* Appended to the cleaned PheCode\_rxNorm.csv as the last column. Stored as PheCode\_rxNorm.csv.

### Brand Names - CUI

#### Mapping Rate

0.35707121364092276

#### 

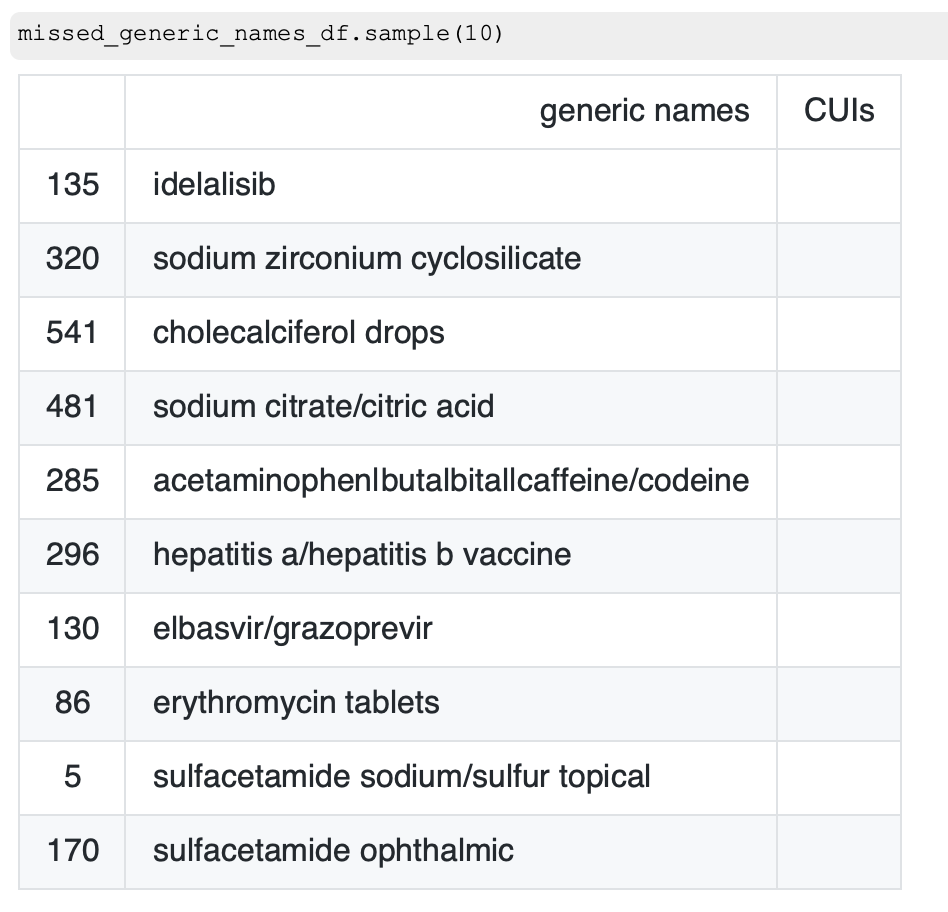
#### Output Data

* Column name: “cui\_from\_brand”
* Appended to the cleaned PheCode\_rxNorm.csv as the last column. Stored as PheCode\_rxNorm.csv.

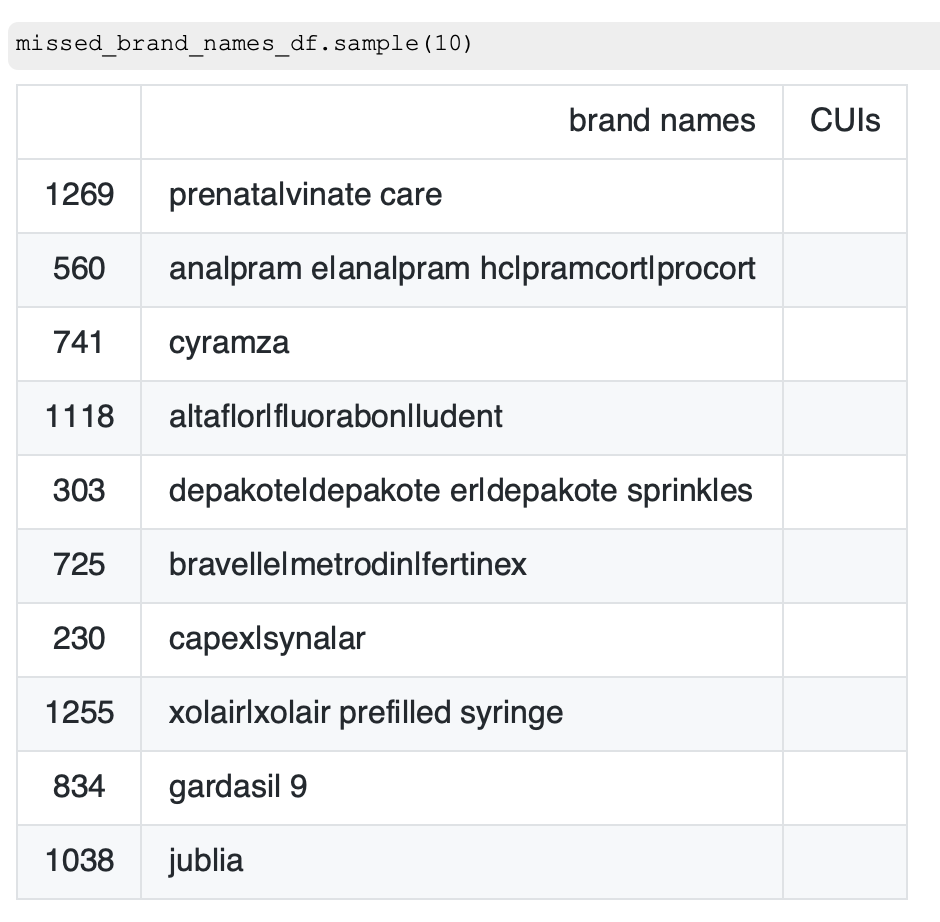
### Evaluation

* Generic names: 547 out of 1996 are missing
* Brand names: 1283 out of 1994 are missing

I randomly sampled 10 generic names in the missing\_generics file:



Except for the last one, all others could be found with an exact matching on UMLS, but not the dictionary I built from broad\_new\_dict.txt. A more comprehensive dictionary should be used, and from there, what is left could be looked up manually.

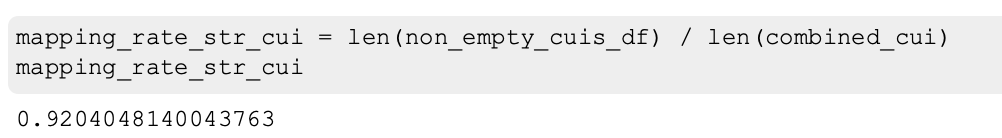


Similarly, I did a random size-10 sample for missing brand names. All of them have exact matches on UMLS.

### Combined CUI

#### Mapping Rate

Add the CUIs mapped from generic names and the CUIs mapped from brand names, we get 0.9204048140043763 of the terms mapped to at least one CUI.



#### Output Data

* Column name: “cui\_combined\_gen\_brand”
* Appended to the cleaned PheCode\_rxNorm.csv as the last column. Stored as PheCode\_rxNorm.csv.

## 

## CUI - RXNorm

### Methodology

1. Create a CUI-RXNorm dictionary from the broad\_new\_dict.txt.
2. Map the combined CUI to RXNorm using the CUI-RXNorm dictionary.

### CUI - RXNorm Dictionary

#### Mapping Rate

0.5779977764836678

#### 

#### Evaluation

Most of the missed strings should have exact mappings on UMLS. A more comprehensive dictionary from CUI to strings should be able to enhance the mapping rate greatly.

### CUI - RXNorms

#### Mapping Rate

0.6246580962800875

#### 

#### Output Data

* Column name: “rxnorms\_from\_CUI”
* Appended to the cleaned PheCode\_rxNorm.csv as the last column. Stored as PheCode\_rxNorm.csv.

#### Evaluation

The low mapping rate should be entailed by the high rate of missing links between CUIs and RXNorms. If we could fix the dictionary, then the mapping rate from CUI to RXNorm should be way higher.

### Net Mapping

#### Mapping Rate

0.5749383189230017

#### 

#### Evaluation

Slight higher (~1%) than exact mapping. But could be greatly enhanced if the broad\_new\_dictionary contains more CUI-RXNorm mappings.