Comparison of Drug Resources to Build a Dictionary of Medications

Jaya Chaturvedi, Angus Roberts
King's College London, United Kingdom

Introduction

This paper focuses on the creation of a gazetteer using several pre-existing lists in order to create a unified dictionary of medication names that incorporates all the nuances of these different pre-existing lists. We describe the initial steps in the development of a gazetteer for medication names from a number of openly available medication resources. The gazetteer is compared to an existing medication gazetteer, and to a corpus of EHR documents manually labelled for medications.

Results

As noticed by comparing Table 2 and Table 3, there is a substantial increase in the percentage of matches amongst the different sources when they are compared post-tokenisation, based on the first token of the medication name only. Whilst eliminating part of the medication name might increase the number of matches, these cases require further investigation and consultation from a clinician to understand the impact of this elimination, and how using such a partial name in a gazetteer, might introduce ambiguity into the meaning of the term, and potentially change the meaning of the medication mention captured. All sources appear to match the most with the GP+SLaM gazetteer in the post-tokenisation comparison, followed by the CRIS medications gazetteer, followed by dm+d and DrugBank, with dm+d showing higher match percentages amongst the two. The highest match percentage in the post-tokenisation comparison is seen between PCA and GP+SLaM at 100%, followed by PCA and dm+d at 92%.

Table 2 . Summary of comparison between different sources (pre-tokenisation)

	CRIS gaz.	Annotations	PCA	DrugBank	SLaM	NHS_GP	GP + SLaM	dm+d
CRIS gaz.		5%	19%	21%	23%	24%	35%	30%
Annotations	64%		36%	46%	50%	32%	56%	55%
PCA	57%	8%		48%	38%	24%	44%	86%
DrugBank	15%	3%	11%		11%	5%	13%	28%
SLaM	78%	17%	54%	63%		34%	100%	80%
NHS_GP	54%	6%	18%	15%	18%		100%	23%
GP + SLaM	56%	7%	23%	26%	39%	74%		37%
dm+d	26%	4%	37%	39%	20%	20%	22%	

Table 3 . Summary of comparison between different sources (post-tokenisation)

	CRIS gaz.	Annotations	PCA	DrugBank	SLaM	NHS_GP	GP + SLaM	dm+d
CRIS gaz.		7%	30%	28%	16%	37%	52%	31%
Annotations	79%		62%	60%	33%	60%	72%	63%
PCA	83%	14%		81%	37%	44%	100%	92%
DrugBank	18%	3%	19%		9%	10%	20%	30%
SLaM	83%	16%	76%	74%		42%	100%	83%
NHS_GP	88%	13%	39%	38%	19%		100%	39%
GP + SLaM	84%	14%	49%	47%	41%	84%		70%
dm+d	45%	7%	49%	69%	21%	24%	54%	

Methods

There are various medication knowledge resources, which have been constructed and made available for a variety of reasons. In order to create a comprehensive medications gazetteer, we decided to incorporate openly available sources of medication names both from primary as well as secondary care, as the texts in secondary care EHRs often refer to medications prescribed by primary care physicians. We describe the datasets that were incorporated below.

Table 1 Sources of medication name data (no duplicates)

Source	Total number of medications (pre-tokenisation)	Total number of medications (post-tokenisation)
NHS Digital - GP prescribing data	1955	1528
Prescription Cost Analysis	1442	1202
SLaM pharmacy	679	559
DrugBank	6175	5466
Manually extracted medication names	315	264
CRIS medications gazetteer	4172	3717
dm+d	3534	2298

Discussion

The description of patient medications and prescribing presents a problem familiar to NLP: the available structured resources that describe the terminology and conceptual knowledge of the domain do not provide a ready-to-use set of the terms used in the natural language resources of the domain. Use of these structured resources to construct NER gazetteers therefore requires careful analysis and processing We are undertaking such an analysis of drug resources for the construction of a comprehensive medication gazetteer. The process undertaken so far highlights the complex nature of creating such a gazetteer.

Further work is being carried out on how normalisation of the medication names affects the quality of matching, and whether important information is being missed or misclassified by such normalisation. We are also considering how such a gazetteer can be kept up to date. Once complete, this gazetteer will be made openly available for other researchers to use.

Corresponding author: Jaya Chaturvedi,
Department of Biostatistics & Health Informatics,
King's College London
Email: jaya.chaturvedi@kcl.ac.uk

JC is funded by the Medical Research Council (MRC) King's Mental Health Data Pathfinder (grant number MC/PC/17214) project grant. AR are supported by the National Institute of Health Research Biomedical Research Centre at the South London and Maudsley NHS Foundation Trust and King's College London. AR is also supported by the Maudsley Charity and by Health Data Research UK, an initiative funded by UK Research and Innovation, Department of Health and Social Care (England) and the devolved administrations, and leading medical research charities. The views expressed are those of the authors and not necessarily those of the NHS, the NIHR or the Department of Health. The above funding had no role in the study design; in the collection, analysis and interpretation of the data; in the writing of the report; and in the decision to submit the article for publication.