

# Development of a Corpus Annotated with Medications and their Attributes in Psychiatric Health Records

Jaya Chaturvedi<sup>1</sup>, Natalia Viani<sup>1</sup>, Jyoti Sanyal<sup>2</sup>, Chloe Tytherleigh<sup>1</sup>, Idil Hassan<sup>3</sup>, Kate Baird<sup>1</sup>,  
Sumithra Velupillai<sup>1</sup>, Robert Stewart<sup>1,2</sup>, Angus Roberts<sup>1</sup>

<sup>1</sup>King's College London, <sup>2</sup>South London and Maudsley NHS Foundation Trust, <sup>3</sup>University of Leicester

<sup>1</sup>De Crespigny Road, London SE5 9RS, <sup>2</sup>Camberwell, London SE5 8AZ, <sup>3</sup>University Rd, Leicester LE1 7RH

{jaya.chaturvedi, natalia.viani, chloe.tytherleigh, kate.baird, sumithra.velupillai, robert.stewart, angus.roberts}@kcl.ac.uk  
jyoti.jyoti@slam.nhs.uk, fah15@student.le.ac.uk

## Abstract

Free text fields within electronic health records (EHRs) contain valuable clinical information which is often missed when conducting research using EHR databases. One such type of information is medications which are not always available in structured fields, especially in mental health records. Most use cases that require medication information also generally require the associated temporal information (e.g. current or past) and attributes (e.g. dose, route, frequency). The purpose of this study is to develop a corpus of medication annotations in mental health records. The aim is to provide a more complete picture behind the mention of medications in the health records, by including additional contextual information around them, and to create a resource for use when developing and evaluating applications for the extraction of medications from EHR text. Thus far, an analysis of temporal information related to medications mentioned in a sample of mental health records has been conducted. The purpose of this analysis was to understand the complexity of medication mentions and their associated temporal information in the free text of EHRs, with a specific focus on the mental health domain.

**Keywords:** Medication, Drug, Natural Language Processing, Electronic Health Records, Temporality, Information Extraction

## 1. Introduction

Electronic health records (EHRs) contain combinations of structured and unstructured (free text) fields. The free text within EHRs contains a large proportion (Stewart *et al.*, 2009) and variety of information about clinical encounters that might go unnoticed if not explored. This is especially true for mental health records where many presentations, contextual factors, interventions and outcomes are not captured in structured fields (such as symptoms, behaviours, and self-reported experiences) so that the extraction of such information from free text generates a more accurate picture (Velupillai *et al.*, 2018). In this domain, it is more likely for information such as drug prescriptions and related context to be available in the text rather than in the structured fields (Kadra *et al.*, 2015). In the UK, many mental health medications are prescribed in primary rather than specialist care and therefore would not be recorded in hospital prescribing systems, even if these were in use. Consequently, hospital clinicians will routinely ask patients to list their medications and record these as free text in their clinical notes.

There is unlimited value in being able to automatically extract information on medications for a variety of research purposes, and this is the main motivation for this particular study. Medication information is essential for research investigating treatment resistance, characterising prescribing patterns and polypharmacy (prescribing more than one drug), and studying the effects as well as side-effects of particular medications on the general health of particular groups of patients. Natural Language Processing (NLP) techniques provide great value in such situations by helping improve the quality of research and in turn potentially facilitating and improving processes in healthcare (Jagannathan *et al.*, 2009). In this paper, we describe the process being undertaken to develop a manually created corpus of medication annotations, for use when developing and evaluating algorithms for the automated extraction of medication information from the text of EHRs.

## 2. Literature Review

A review of the literature was conducted to understand how other researchers have extracted medication information from EHRs, and what features have been the focus of previous work. A study by Chhieng *et al.* (2007) focused on the extraction of the names of the medications using a software application called NegEx (for identifying negated findings) and a UMLS-based drug lexicon (Chhieng *et al.*, 2007) while a method described by Gold *et al.* (2008) aimed to extract drug names appearing with and without dosage information, misspelled drug names, and contextual information, such as dose, route and frequency, using parsing rules as a set of regular expressions (Gold *et al.*, 2008). In these studies, gazetteers and rules were developed for identifying the information of interest. A paper by Xu *et al.* (2010), describes the creation of a medication information extraction system called MedEx. MedEx is a medication parser which was built using a semantic-based approach where the semantic types and patterns were used at a fine level of granularity. An integration of the semantic tagger and a chart parser was expected to capture the medication names along with the major categories of attributes, such as dose, route, and frequency. This system performed well (F score of over 0.90) in identifying drug names, strength, frequency, and route in discharge summaries (Xu *et al.*, 2010). In 2009, the i2b2 medication extraction challenge was organised to support the development of NLP systems for medication extraction (Uzuner, Solti and Cadag, 2010). In this challenge, most developed NLP solutions relied on rule-based approaches, with some systems also exploiting machine learning classifiers. A medication extraction system called medExtractR was recently developed by Weeks *et al.* (2019) using the R programming language to extract dose and timing information associated with medication mentions (Weeks *et al.*, 2019). MedExtractR uses a combination of lexicons and regular expression patterns to identify relevant medication information such as drug name, strength, dose, intake time, frequency, and last

dose (Weeks *et al.*, 2019). A recent paper by Aberdeen *et al.* (2019) used a corpus of ambulatory prescriptions to create a modelling schema for prescription regimens. Most of the attributes were extracted from the structured fields using a novel set of semantic tags. The main information extracted from free text was that of “directions” (such as “take 2 tablets as needed”). Their automated annotation used Conditional Random Fields to train a model, and various other methods, based on the manual annotations (Aberdeen *et al.*, 2019). These studies all reinforce the importance of NLP methods for extraction of medication information from text, and also provide some insight on the schemas used to describe medications, and the different techniques that can potentially be employed.

### 3. Methods

#### 3.1 Setting

In this study, we considered mental health records from the South London and Maudsley NHS Foundation Trust (SLaM), one of the largest mental health care providers in Western Europe. It serves around 1.36 million residents of four south London boroughs (Lambeth, Southwark, Lewisham and Croydon) and provides comprehensive secondary, and a range of tertiary, mental health care services. All clinical records in SLaM services have been electronic since 2006 (including imported legacy data from earlier years for some services) and have been made available for research since 2008 following the establishment of the Clinical Record Interactive Search (CRIS) platform. CRIS was set up and subsequently supported by the NIHR Biomedical Research Centre at SLaM and King’s College London. CRIS has been described in detail (Perera *et al.*, 2016). It enables researcher access to full but de-identified data from the electronic mental health record within a robust, patient-led governance model, and with ethical approval for secondary analysis (Oxford C Research Ethics Committee, reference 18/SC/0372). For the de-identification of free text within CRIS, masking strings have been used to replace names. Patient names are replaced with ZZZZZ and names of relatives and close contacts, with QQQQQ (Fernandes *et al.*, 2013). A large amount of valuable information is contained in the free text of the CRIS database, which includes progress notes, written assessments and correspondence documents (Stewart *et al.*, 2009). CRIS currently accesses about 30 million case notes and correspondence, with an average of 90 documents per patient (Velupillai *et al.*, 2018).

Over 60 NLP algorithms are currently deployed on the CRIS database (with at least another 50 currently in development), including those that extract information about cognitive function, smoking status, diagnostic statements and pharmacotherapy/medications (Perera *et al.*, 2016). The medications application has been described in a paper by Kadra *et al.* (2015) and highlights the development and validation of this application (Kadra *et al.*, 2015). This application was built using GATE (General Architecture for Text Engineering) by developers with language engineering skills, using annotation definitions agreed with clinicians and epidemiologists (Cunningham *et*

*al.*, 2013). Medication information extracted by this application is made available to researchers, together with any structured medication information already available in CRIS (Perera *et al.*, 2016).

The current CRIS medications application is rules-based and had been initially validated with a single drug (clozapine) only. Its precision was originally tested against a manual search of 279 documents, and recall was ascertained on a random set of 200 documents containing the word clozapine (Perera *et al.*, 2016). Further validation was recently conducted considering a wider range of medications, and recommendations were made on how this can be further improved. The F-scores ranged from 0.70 to 0.85 for a number of other antipsychotic drugs (Perera *et al.*, 2016). The paper by Perera *et al.* (2016) recommends further bespoke validation of the application prior to its use with other drugs and use cases. The application has a large and complex rule base, which is difficult to maintain. Anecdotally, it performs poorly on assigning temporal status to medications (e.g. current or past medications). Finally, it does not make use of recent advances in NLP. For these reasons, it was decided to develop an extended NLP algorithm to be validated on additional pharmacotherapy types as well as to capture additional contextual attributes (e.g. temporality, modality and the subject of the medication mention – for example, the patient or a relative). This application will provide the initial use cases for the corpus of medication mentions reported here. To develop the corpus, we initially selected a cohort of patients, and are creating reference standard annotations on documents associated with these patients. Whilst these annotations will be used to develop and evaluate a new CRIS medications application, they have potential for other uses.

#### 3.2 Data Cohort

In order to decide which diagnosis groups should be included in the cohort, an SQL query was run on the CRIS database to identify the number of patients for each primary diagnosis ICD-10 code<sup>1</sup>. No filter was applied to the date of diagnosis. The top four diagnosis groups which accounted for the greatest number of patients were: depression (ICD10: F32x, F33x) accounting for 10% of all primary diagnoses within CRIS, schizophrenia (F20x) at 4% of all primary diagnoses, dementia (F00x, F02x, F01x) at 5% of all primary diagnoses, and stress and anxiety (F43x, F41x) at 7% of all primary diagnoses within CRIS. These were the cohorts that were used in the data extraction. While these diagnosis codes don’t make up the majority of the records within CRIS, these are the most frequently coded primary diagnoses and were expected to provide a good reflection of how medications are expressed within the free text.

##### 3.2.1 Text Sources

The CRIS database consists of 23 different sources of text. A medications corpus created from these was thus expected to reflect the relevant text sources. In order to determine which text sources should be used, an SQL query was run to extract the top 1000 records of each text source, and an

---

<sup>1</sup> <https://icd.who.int/browse10/2010/en>

analysis was carried out to determine what percent of the records had text, the average number of documents per patient for each source, as well as the minimum, maximum, and average length (in characters) of the texts in these documents. Out of the 23 different text sources, 13 were chosen based on how much text they contained. Over 80% of the records within these 13 text sources contained free text, while the other text sources were less populated with text. Following this, another SQL query was run to check the frequency of mentions of four common psychotropic drugs (clozapine, risperidone, sertraline, mirtazapine) and four common general use drugs (paracetamol, atorvastatin, insulin, tamsulosin). These particular drugs were chosen based on those which were most frequently found when the existing medications application was run on a similar set of documents. This narrowed down the text sources to two main ones – Attachments and Events. These were the tables that were included for the final extraction of documents for the annotation corpus.

The attachments table consists of a variety of different documents and letters. A majority of these are letters to primary care physicians, and referral forms. Some of the other types of documents include discharge letters, assessment documents, review forms, appointment letters, etc. The events table consists of case notes made by the clinicians/nurses at face to face encounters with the patient. These tend to be shorter in length but greater in number than the attachments.

### 3.2.2 Data Extraction

To make a manual annotation task feasible, it was decided that documents for 50 patients would be extracted for each diagnosis group and for each text source. Each patient had 1-2 documents each. Table 1 gives a brief description of the average number of tokens and characters within these text sources to provide a better understanding of the length of documents contained in each.

	Attachments	Events
Avg. number of tokens	2,458	785
Avg. length in characters	18,139	6,000

Table 1: Average tokens and characters

Only documents that had at least 500 tokens in events and at least 1000 tokens in attachments were extracted in order to give substantial information per document. Due to the generally shorter length of events, a minimum of 500 tokens were considered for these documents, rather than 1000. A summary of the extracted documents can be seen in Table 2 below.

Cohort	Number of documents (for 50 patients each)	
	Attachments	Events
Depression	64	54
Schizophrenia	68	56
Dementia	56	58
Stress/Anxiety	54	55
Total	242	223

Table 2: Summary of documents extracted

### 3.3 Reference Standard

The creation of the reference standard annotations was carried out by three medical students. A set of annotation guidelines were created by three NLP researchers to facilitate them in this task. The annotation guidelines described the task of annotating medication names and the different attributes that required filling in. They also covered nuances in annotations, such as what date formats were to be used and how to deal with specific situations such as dose changes in medications within the text. Examples were provided for each attribute as well. The annotation guidelines were developed on a small set of documents over multiple iterations, to check their applicability.

The annotation tool used was eHOST: The Extensible Human Oracle Suite of Tools (South *et al.*, 2012). eHOST is a Java-based prototype annotation system which provides an open-source and standalone client for manual annotations.

The annotation schema included a number of attributes, some which were already used by the existing medications application as well as other studies done on the topic, and some of which were newly proposed for this study. Table 3 lists the attributes. The most commonly used attributes among previous studies reviewed are drug name alone (Levin *et al.*, 2007), or in combination with dose, frequency and route (Jagannathan *et al.*, 2009; Doan *et al.*, 2010; Xu *et al.*, 2010; Sohn *et al.*, 2014), with one study also looking at modality (Iglesias *et al.*, 2009). We believe that all other attributes are novel to the study reported here.

Attribute	Explanation	Example
Drug Name	Name of the drug	“Clozapine”
Dose value	The numeric value of the dose	“200”
Dose unit	The unit of the dose	“mg”, “ml”, etc.
Frequency	The number of times the drug has to be taken	“2” in twice every 4 days
Interval	At what intervals the drug needs to be taken	“4” in twice every 4 days
Route	The route of intake of the drug	“Oral”, “IV”, etc.
When	What time of day the drug should be taken	“night”
Initiation	When the drug was started	Current/past/future
Initiation time	The exact/vague date for the start of the drug	“March 2012”, “03/12/1998”, etc.
Cessation	When the drug was stopped	Current/past/future
Cessation time	The exact/vague date the drug was stopped	“March 2012”, “03/12/1998”, etc.
Subject	The person the drug is	Patient/Other (“Other” could be a relative of the

	mentioned in relation to	patient). The default entry for this field is “Patient”
Modality	Whether the drug mentioned was taken by the patient	Taken by patient (including overdose and side effects) / Not taken by patient (mentions of allergy or hypothetical discussions). The default entry for this field is “Taken by patient” except when the subject is entered as “Other” in which case the modality will default to “Not taken by the patient”

Table 3: Attributes in the annotation schema

The documents from Attachments were annotated using the first version of the guidelines. Upon analysis of the first set of annotations, certain shortcomings in the guidelines became apparent, such as lack of a standardised format for dates. The annotation guidelines were subsequently updated as part of the iterative process with any clarifications on conflicts noticed between annotations. The corpus thus far only consists of attachments, with annotations that were created using the first version of the annotation guidelines. Annotation of documents from events using the updated version of annotation guidelines is currently ongoing. However, the differences between the annotations made using the first version of guidelines and the second shouldn’t be too inconsistent since the updates to the guidelines were small differences such as date formats, which were accounted for in the adjudication process for the annotations made using the first version. The results presented in this paper are from the annotations conducted on the attachments only. Each document was double annotated, and adjudication was carried out by one NLP researcher following a set of adjudication guidelines, for the creation of the final reference standard. The adjudication process ensured consistent resolution of any disagreements in annotations, as well as any corrections to the annotations where required.

## 4. Results

### 4.1 Initial Analysis

The annotations for the documents from the attachments have been completed, and some preliminary analysis has been carried out on these. The macro average of the number of annotations between the two annotators for each cohort is shown in Table 4. Stress/anxiety had the least number of annotations while schizophrenia had the most.

Cohort	No. of annotations
Depression	577
Schizophrenia	945
Dementia	464
Stress/Anxiety	186

Table 4: Number of annotations

When looking at the availability of information on the different attributes within the text, with regards to “modality” it was found that about 95% of the time the medication was taken by the patient, while 5% of the time it was not. This 5% includes situations such as hypothetical mentions (e.g. “..Ivabradine is an option for the patient if Bisoprolol is ineffective..”), allergy information, or mentions related to blood tests given to test for the presence of or effect of a drug. With regards to temporality, the “initiation” (past/current/future) for the medication was identifiable in 93% of the documents, with a majority of this indicating initiation in the “past”. However, “cessation” (past/current/future) was not identifiable for 80% of the documents, with an exception of the schizophrenia cohort where the “cessation” was identified as “past” in 76% of the documents. However, this attribute only indicates whether the initiation and cessation were in the past, current or future. The attribute for exact initiation and cessation times, such as full dates or years, were not identifiable in about 85% of the documents. The “initiation time” attribute was an exact date in 8% of the mentions (such as “12/09/2010”) and a vague date in 10% of the mentions (such as “November 2012”, or “in 2014”). The “cessation time” attribute was an exact date in only 4% of the mentions, and a vague date in 4% of the mentions too. A visual representation of some of these results is shown in the figures below. Figure 1 shows what percent of documents indicated initiation within the four different cohorts. Null indicates that the information was missing.

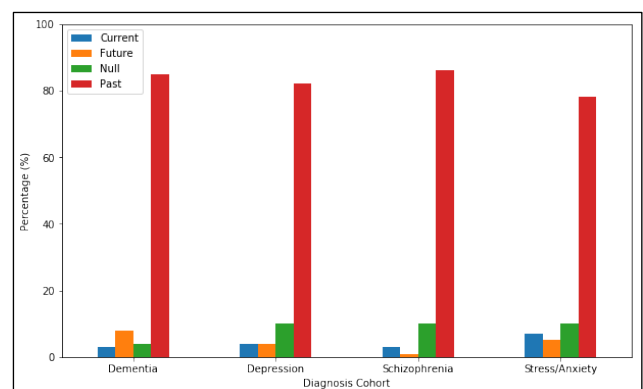


Figure 1: Initiation attribute

As seen in Figure 1, all four cohorts have a majority of the initiation attribute recorded as “past”. The second most frequent attribute is “null”, followed by “future” for the dementia and the stress/anxiety cohort, and “current” for the schizophrenia cohort. The occurrence of “current” and “future” seem to be almost equal for the depression cohort. Figure 2 shows what percent of documents indicated cessation within the four cohorts. As seen in the figure, a



majority of the cessation attribute was recorded as “null” which indicates that this information was not available in the free text of the records. The second most common attribute for this field is “past”, followed by “future”, and the least frequent one being “current”.

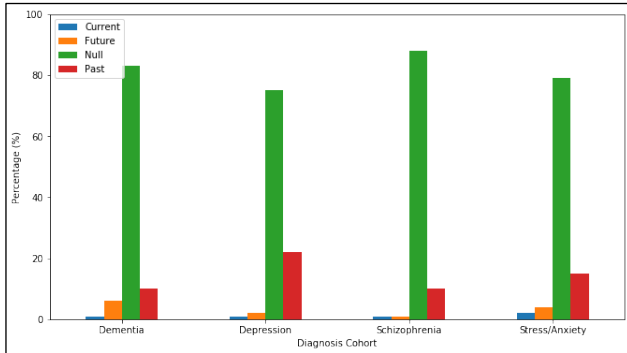


Figure 2: Cessation attribute

Some of the other medications mentions were categorised into four different groups based on their characteristics – ‘Generic mention’ which referred to the mention of a drug class such as antipsychotics or antidepressants (e.g. “patient has been on antidepressants in the past..”), ‘Dose change’ which referred to annotations where the dose was being altered in any way for a particular drug (such as “...clozapine 250mg to be increased to 400mg..”), ‘Blood test mentions’ which refers to mentions associated with monitoring blood tests (such as “clozapine blood test”), and ‘Overdose’ which is associated with mentions of drugs in the context of overdose by the patient. These groups were identified and formed during the analysis of the annotations rather than as part of the annotation task. Figure 3 shows the frequency of mentions from these groups in the 4 cohorts. ‘Overdose’ appears to be more frequent in the depression and stress/anxiety cohort, while ‘dose change’ is most frequent in the dementia cohort, and ‘generic mention’ in the schizophrenia cohort.

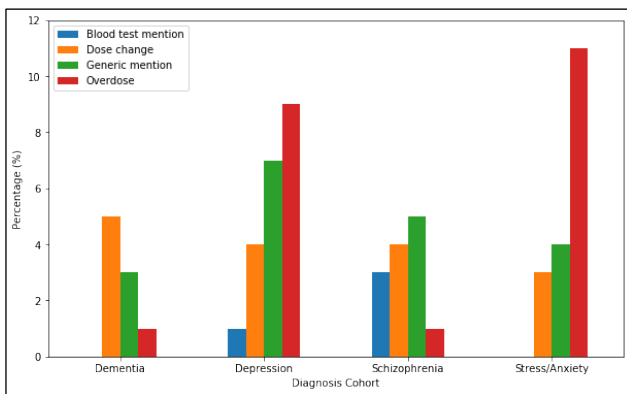


Figure 3: Other categories of mentions

Dose information (“Dose unit” and “Dose value”) associated with the medication mention was available in 47% of the documents, while route information (oral, IV etc.) was available only in 17% of the documents. Another attribute was “subject” which was to indicate if the mention of a medication was in relation to the patient or someone

else (such as a family member). However, within these 4 cohorts, the subject was the patient 100% of the times.

The inter-annotator agreement (IAA) was calculated on the spans of the annotations on all four cohorts and averaged at 0.90. A breakdown of the IAA scores for each cohort is shown in Table 5.

Cohort	IAA of annotation span
Depression	0.88
Schizophrenia	0.86
Dementia	0.89
Stress/Anxiety	0.96

Table 5: IAA scores for annotation spans

The agreement between the different attributes is shown in Table 6. The IAA scores are a macro average of all four diagnosis groups.

Attribute	IAA
Drug name	0.81
Dose value	0.84
Dose unit	0.94
Frequency	0.86
Interval	0.90
Route	0.54
When	0.84
Initiation	0.89
Initiation time	0.79
Cessation	0.85
Cessation time	0.88
Subject	0.99
Modality	0.98

Table 6: IAA summary for attributes

As seen in Table 5, most attributes have an IAA score of over 0.80, with some over 0.90. The IAA score for ‘drug name’ can be further improved with some post-processing of the annotations, since most disagreements appeared to be due to entry of incomplete drug names (such as “levothyroxine” vs. “levothyroxine sodium”) and spelling mistakes. The attributes that were available in drop-down format (such as initiation, cessation, subject, and modality) scored better than the others which were free form and therefore affected by how the attributes were entered. The attributes were normalised for case (all made lower case) and the ones with dates were normalised for date formats (DD/MM/YYYY, or Mon YYYY, or just YYYY) before the calculation of the IAA scores. The least scoring attribute is “route”. One of the reasons for this is the inconsistency in entering this information, such as oral vs. tablets, injection vs. depot. Some of it was normalised, but there were too many variations in the way this field was entered to capture all of them. Apart from the inconsistency in this attribute, this information was also missed from some annotations which also contributed to the lower score.

## 5. Discussion

Preliminary analysis of the manual annotation of documents indicates that a good variety of information is

available within the corpus, and the agreements between annotators along with adjudication indicates that the reference standard is of good quality. During the i2b2 medication challenge in 2009, a community annotation experiment was conducted for ground truth generation, where manual annotations were conducted and IAA scores were calculated (Uzuner *et al.*, 2010). The results of this annotation challenge were quite similar to the results obtained in this study. The IAA scores for drug names were higher in the challenge compared to this study (0.86 vs. 0.81) but the scores here have potential to improve with some further normalisation of the values entered, while the scores for dose and frequency are almost equal in both cases (Uzuner *et al.*, 2010).

One limitation that has become apparent from the initial analysis is the lack of sufficient information on exact initiation and cessation times for the medications mentioned in the source text. One alternative to overcome this limitation might be to use the document date as the date associated with the commencement of the medication, which is what is currently done. However, this approach cannot be used for cessation times as it could lead to erroneous dates. It might be best to leave it as 'null' and let the use cases decide. The 'subject' attribute that was included in the annotation task yielded results of the subject always being the patient. This might, however, not hold true on a different cohort such as child and adolescent records, where there might be more mentions of medications in relation to family members of the patient. A preliminary analysis can potentially be run on this group of patients to gain a better understanding on how frequently the subject might or might not be the patient, and the results can be used to make decisions on how this could be incorporated into future algorithms, if required. Given the large amounts of free text data available in CRIS, it will be interesting to see how future applications built with this reference standard perform when run over all fields. This might further help refine the design of applications.

It is quite challenging to overcome the ambiguity that resides in the clinical notes, specifically when there are mentions of medications in a hypothetical context. The proportion of annotations where the medication is actually taken by the patient versus other mentions of medications in relation to the patient is quite imbalanced and might not make for a good training set. A potential way to overcome this or at least be more certain that the mentions of medications are in relation to actual prescriptions, might be to link mentions of medications with any sections or paragraphs within documents that evidence these particular mentions as being related to the patient. Such sections include those with sub-headings of 'Medication Plan' or 'Prescription'.

## 6. Conclusion and future work

This study describes the development of a corpus of medication annotations from the free text of mental health records. This includes the processes for the selection of documents and creation of reference standards, along with an overview of the features that were found in the annotated documents through an initial analysis. The documents that have been manually annotated are a valuable resource. A

comparison will be run between these manually annotated documents and the same set of documents run on the existing GATE based medications application to better identify where the existing application is failing and what other factors need to be considered in the construction of potential new applications. Based on the findings from the manual annotations, and the discovery of new categories of mentions such as blood test mentions, generic mentions, and overdose, there is potential to build some add-on components to future applications that could be employed when required, such as when research studies are interested in incidents of overdose in a particular cohort, or prescription of generic classes of drugs, and also in studies looking at patients who might be on blood test monitoring, especially with antipsychotics such as clozapine.

## 7. Availability

The corpus and annotations are part of the CRIS case register at the South London and Maudsley NHS Foundation Trust. CRIS access is controlled by an information governance framework that includes both project approval and researcher approval<sup>2</sup>. This corpus of medication annotations is available to anyone who has obtained the required approvals for access to CRIS. The latest version of the annotation guidelines document is open and available for reference and use<sup>3</sup>.

## 8. Acknowledgements

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article. This work was supported by the Clinical Records Interactive Search (CRIS) system funded and developed by the NIHR Biomedical Research Centre at South London and Maudsley NHS Foundation Trust and King's College London. JC is funded by the Medical Research Council (MRC) King's Mental Health Data Pathfinder (grant number MC/PC/17214) project grant. RS and NV are part-supported, SV and 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. RS, SV and NV declare previous research funding received from Janssen. RS and AR declare current research funding received from Takeda. 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. The remaining authors have not received funding.

## 9. Bibliographical References

Aberdeen, J., Bayer, S., Clark, C., Keybl, M. and Tresner-Kirsch, D. (2019) 'An annotation and modeling schema for prescription regimens', *Journal of Biomedical Semantics*,

<sup>2</sup> See <https://www.slam.nhs.uk/research/cris> for details

<sup>3</sup>[https://github.com/jayachaturvedi/medications\\_application](https://github.com/jayachaturvedi/medications_application)

- Chhieng, D., Day, T., Gordon, G. and Hicks, J. (2007) 'Use of natural language programming to extract medication from unstructured electronic medical records.', *AMIA ... Annual Symposium proceedings. AMIA Symposium*, p. 908. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/18694008> (Accessed: 19 November 2019).
- Cunningham, H., Tablan, V., Roberts, A. and Bontcheva, K. (2013) 'Getting More Out of Biomedical Documents with GATE's Full Lifecycle Open Source Text Analytics', *PLoS Computational Biology*, 9(2). doi: 10.1371/journal.pcbi.1002854.
- Doan, S., Bastarache, L., Klimkowski, S., Denny, J. C. and Xu, H. (2010) 'Integrating existing natural language processing tools for medication extraction from discharge summaries', *Journal of the American Medical Informatics Association*. Narnia, 17(5), pp. 528–531. doi: 10.1136/jamia.2010.003855.
- Fernandes, A. C., Cloete, D., Broadbent, M. T., Hayes, R. D., Chang, C.-K., Jackson, R. G., Roberts, A., Tsang, J., Soncul, M., Liebscher, J., Stewart, R. and Callard, F. (2013) 'Development and evaluation of a de-identification procedure for a case register sourced from mental health electronic records', *BMC Medical Informatics and Decision Making*, 13(1), p. 71. doi: 10.1186/1472-6947-13-71.
- Gold, S., Elhadad, N., Zhu, X., Cimino, J. J. and Hripcsak, G. (2008) 'Extracting structured medication event information from discharge summaries.', *AMIA ... Annual Symposium proceedings. AMIA Symposium*. American Medical Informatics Association, 2008, pp. 237–41. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/18999147> (Accessed: 13 August 2019).
- Iglesias, J. E., Rocks, K., Jahanshad, N., Frias-Martinez, E., Andrada, L. P. and Bui, A. A. T. (2009) 'Tracking medication information across medical records.', *AMIA ... Annual Symposium proceedings. AMIA Symposium*. American Medical Informatics Association, 2009, pp. 266–70. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/20351862> (Accessed: 2 July 2019).
- Jagannathan, V., Mullett, C. J., Arbogast, J. G., Halbritter, K. A., Yellapragada, D., Regulapati, S. and Bandaru, P. (2009) 'Assessment of commercial NLP engines for medication information extraction from dictated clinical notes', *International Journal of Medical Informatics*. Elsevier, 78(4), pp. 284–291. doi: 10.1016/J.IJMEDINF.2008.08.006.
- Kadra, G., Stewart, R., Shetty, H., Jackson, R. G., Greenwood, M. A., Roberts, A., Chang, C.-K., MacCabe, J. H. and Hayes, R. D. (2015) 'Extracting antipsychotic polypharmacy data from electronic health records: developing and evaluating a novel process.', *BMC psychiatry*. BioMed Central, 15, p. 166. doi: 10.1186/s12888-015-0557-z.
- Levin, M. A., Krol, M., Doshi, A. M. and Reich, D. L. (2007) *Extraction and Mapping of Drug Names from Free Text to a Standardized Nomenclature*. Available at: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2655777/pdf/amia-0438-s2007.pdf> (Accessed: 25 March 2019).
- Perera, G., Broadbent, M., Callard, F., Chang, C.-K., Downs, J., Dutta, R., Fernandes, A., Hayes, R. D., Henderson, M., Jackson, R., Jewell, A., Kadra, G., Little, R., Pritchard, M., Shetty, H., Tulloch, A. and Stewart, R. (2016) 'Cohort profile of the South London and Maudsley NHS Foundation Trust Biomedical Research Centre (SLaM BRC) Case Register: current status and recent enhancement of an Electronic Mental Health Record-derived data resource.', *BMJ open*. British Medical Journal Publishing Group, 6(3), p. e008721. doi: 10.1136/bmjopen-2015-008721.
- Sohn, S., Clark, C., Halgrim, S. R., Murphy, S. P., Chute, C. G. and Liu, H. (2014) 'MedXN: an open source medication extraction and normalization tool for clinical text', *Journal of the American Medical Informatics Association*. Narnia, 21(5), pp. 858–865. doi: 10.1136/amiajnl-2013-002190.
- South, B. R., Shen, S., Leng, J., Forbush, T. B., Duvall, S. L. and Chapman, W. W. (2012) *A Prototype Tool Set to Support Machine-Assisted Annotation*. Available at: <http://code.google.com/p/ehost>. (Accessed: 24 November 2019).
- Stewart, R., Soremekun, M., Perera, G., Broadbent, M., Callard, F., Denis, M., Hotopf, M., Thornicroft, G. and Lovestone, S. (2009) 'The South London and Maudsley NHS Foundation Trust Biomedical Research Centre (SLAM BRC) case register: development and descriptive data.', *BMC psychiatry*. BioMed Central, 9, p. 51. doi: 10.1186/1471-244X-9-51.
- Uzuner, Ö., Solti, I. and Cadag, E. (2010) 'Extracting medication information from clinical text', *Journal of the American Medical Informatics Association*, 17(5), pp. 514–518. doi: 10.1136/jamia.2010.003947.
- Uzuner, Ö., Solti, I., Xia, F. and Cadag, E. (2010) 'Community annotation experiment for ground truth generation for the i2b2 medication challenge', *Journal of the American Medical Informatics Association*, 17(5), pp. 519–523. doi: 10.1136/jamia.2010.004200.
- Velupillai, S., Suominen, H., Liakata, M., Roberts, A., Shah, A. D., Morley, K., Osborn, D., Hayes, J., Stewart, R., Downs, J., Chapman, W. and Dutta, R. (2018) 'Using clinical Natural Language Processing for health outcomes research: Overview and actionable suggestions for future advances'. doi: 10.1016/j.jbi.2018.10.005.
- Weeks, H. L., Beck, C., Mcneer, E., Bejan, C. A., Denny, J. C. and Choi, L. (2019) 'medExtractR: A medication extraction algorithm for electronic health records using the

R programming language'. doi: 10.1101/19007286.

Xu, H., Stenner, S. P., Doan, S., Johnson, K. B., Waitman, L. R. and Denny, J. C. (2010) 'MedEx: a medication information extraction system for clinical narratives', *Journal of the American Medical Informatics Association*. Oxford University Press, 17(1), pp. 19–24. doi: 10.1197/jamia.M3378.