



RESEARCH ARTICLE

REVISED Knowledge discovery for Deep Phenotyping serious mental illness from Electronic Mental Health records [version 2; referees: 2 approved]

Richard Jackson ¹, Rashmi Patel ^{1,2}, Sumithra Velupillai ^{1,3}, George Gkotsis ¹, David Hoyle ⁴, Robert Stewart ^{1,2}

¹Institute of Psychiatry, Psychology and Neuroscience, King's College London, London, SE5 8AF, UK

²South London and Maudsley NHS Foundation Trust, London, SE5 8AZ, UK

³School of Computer Science and Communication, TH Royal Institute of Technology, Stockholm, SE-100 44, Sweden

⁴Independent Researcher, Manchester, UK

v2 First published: 21 Feb 2018, 7:210 (doi: [10.12688/f1000research.13830.1](https://doi.org/10.12688/f1000research.13830.1))
 Latest published: 08 May 2018, 7:210 (doi: [10.12688/f1000research.13830.2](https://doi.org/10.12688/f1000research.13830.2))

Abstract

Background: Deep Phenotyping is the precise and comprehensive analysis of phenotypic features in which the individual components of the phenotype are observed and described. In UK mental health clinical practice, most clinically relevant information is recorded as free text in the Electronic Health Record, and offers a granularity of information beyond what is expressed in most medical knowledge bases. The SNOMED CT nomenclature potentially offers the means to model such information at scale, yet given a sufficiently large body of clinical text collected over many years, it is difficult to identify the language that clinicians favour to express concepts.

Methods: By utilising a large corpus of healthcare data, we sought to make use of semantic modelling and clustering techniques to represent the relationship between the clinical vocabulary of internationally recognised SMI symptoms and the preferred language used by clinicians within a care setting. We explore how such models can be used for discovering novel vocabulary relevant to the task of phenotyping Serious Mental Illness (SMI) with only a small amount of prior knowledge.

Results: 20 403 terms were derived and curated via a two stage methodology. The list was reduced to 557 putative concepts based on eliminating redundant information content. These were then organised into 9 distinct categories pertaining to different aspects of psychiatric assessment. 235 concepts were found to be expressions of putative clinical significance. Of these, 53 were identified having novel synonymy with existing SNOMED CT concepts. 106 had no mapping to SNOMED CT.

Conclusions: We demonstrate a scalable approach to discovering new concepts of SMI symptomatology based on real-world clinical observation. Such approaches may offer the opportunity to consider broader manifestations of SMI symptomatology than is typically assessed via current diagnostic frameworks, and create the potential for enhancing nomenclatures such as SNOMED CT based on real-world expressions.

Open Peer Review

Referee Status:

	Invited Referees	
	1	2
REVISED		
version 2	report	report
published		
08 May 2018		
version 1		
published	report	report
21 Feb 2018		

1 **Julian Hong** , Duke University School of Medicine, USA

Jessica Tenenbaum , Duke University School of Medicine, USA

2 **Karin Verspoor** , The University of Melbourne, Australia
 The University of Melbourne, Australia

Discuss this article

Comments (0)

Keywords

word2vec, natural language processing, serious mental illness, electronic health records, schizophrenia

Corresponding author: Robert Stewart (robert.stewart@kcl.ac.uk)

Author roles: **Jackson R:** Conceptualization, Formal Analysis, Investigation, Methodology, Project Administration, Software, Validation, Visualization, Writing – Original Draft Preparation; **Patel R:** Data Curation, Validation, Writing – Review & Editing; **Velupillai S:** Methodology, Writing – Review & Editing; **Gkotsis G:** Methodology, Writing – Review & Editing; **Hoyle D:** Methodology, Supervision, Writing – Review & Editing; **Stewart R:** Data Curation, Funding Acquisition, Project Administration, Resources, Supervision, Validation, Writing – Review & Editing

Competing interests: RJ and RS have received research funding from Roche, Pfizer, J&J and Lundbeck.

How to cite this article: Jackson R, Patel R, Velupillai S *et al.* **Knowledge discovery for Deep Phenotyping serious mental illness from Electronic Mental Health records [version 2; referees: 2 approved]** *F1000Research* 2018, 7:210 (doi: [10.12688/f1000research.13830.2](https://doi.org/10.12688/f1000research.13830.2))

Copyright: © 2018 Jackson R *et al.* This is an open access article distributed under the terms of the [Creative Commons Attribution Licence](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. Data associated with the article are available under the terms of the [Creative Commons Zero "No rights reserved" data waiver](#) (CC0 1.0 Public domain dedication).

Grant information: This paper represents independent research funded by the National Institute for Health Research (NIHR) Biomedical Research Centre at South London and Maudsley NHS Foundation Trust and King's College London. RP has received support from a Medical Research Council (MRC) Health Data Research UK Fellowship and a Starter Grant for Clinical Lecturers (SGL015/1020) supported by the Academy of Medical Sciences, The Wellcome Trust, MRC, British Heart Foundation, Arthritis Research UK, the Royal College of Physicians and Diabetes UK. SV is supported by the Swedish Research Council (2015-00359), Marie Skłodowska Curie Actions, Cofund, Project INCA 600398 *The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.*

First published: 21 Feb 2018, 7:210 (doi: [10.12688/f1000research.13830.1](https://doi.org/10.12688/f1000research.13830.1))

REVISED Amendments from Version 1

This revision includes amendments that we hope address the issues raised by the peer review process. A response to each comment can be found in the 'response to reviewer' section that accompanies the article, but the changes can be summarised as follows:

1. Improvements to the clarity of the methods section, addressing some comprehension issues that were raised such as consistency of terminology and the description of techniques employed
2. An expanded rationale for several decisions that were made in the development of the approach, against alternatives that were available
3. The citation of additional relevant literature for this domain, such as work on automated term recognition and existing work on symptom grouping
4. Some additional results regarding the counts of unigrams, bigrams and trigrams
5. A reference to a publicly available code repository that demonstrates the approach (since sharing the underlying data is not possible)
6. Several minor grammatical errors

We offer our gratitude to both sets of reviewers for their time and valuable assistance.

See referee reports

Introduction

The dramatic decrease of genetic sequencing costs, coupled with the growth of our understanding of the molecular basis of diseases, has led to the identification of increasingly granular subsets of disease populations that were once thought of as homogenous groups. As of 2010, the molecular basis for nearly 4 000 Mendelian disorders has been discovered¹, subsequently leading to the development of around 2 000 clinical genetic tests². The resulting 'precision medicine' paradigm has been touted as the logical evolution of evidence-based medicine.

Precision medicine has arisen in response to the fact that the real-world application of many treatments have a lower efficacy and a differential safety profile compared to clinical trials, most likely due to genetic and environmental differences in the disease population. Precision medicine seeks to obtain deeper genotypic and phenotypic knowledge of the disease population, in order to offer tailored care plans with evidence-based outcomes. Amongst the challenges presented by precision medicine is the requirement to obtain highly granular phenotypic knowledge that can adequately explain the variable manifestation of disease.

To realise the ambitions of precision medicine, large amounts of phenotypic data are required to provide sufficient statistical power in tightly defined patient cohorts (so called 'Deep Phenotyping'³). Historical clinical data mined from Electronic Health Record (EHR) systems are frequently employed to meet the related use case of observational epidemiology. As such, EHRs are often posited as the means to provide extensive phenotypic information with a relatively low cost of collection^{4,5}.

In order to standardise knowledge representation of clinically relevant entities and the relationships between them, phenotyping from EHRs often employs curated terminology systems, most commonly SNOMED CT. The use of such resources creates a common domain language in the clinical setting, theoretically allowing an unambiguous interpretation of events to be shared within and between healthcare organisations. The anticipated value of such a capability has prompted the UK National Information Board to recommend the adoption of SNOMED CT across all care settings by 2020⁶. However, the task of representing the sprawling and ever-changing landscape of healthcare in such a fashion has proven complex⁷⁻¹⁰. Although a complete description of the structure and challenges of SNOMED CT are beyond the scope of this paper, we describe how aspects of these problems manifest themselves in accordance with the task of phenotyping serious mental illness (SMI) from a real-world EHR system.

Phenotyping SMI

The quest for empirically validated criteria for assessing the symptomatology of mental illness has been a long term goal of evidence-based psychiatry. SMI is a commonly used umbrella term to denote the controversial diagnoses of schizophrenia (encoded in SNOMED as SCTID: 58214004), bipolar disorder (SCTID: 13746004), and schizoaffective disorder (SCTID: 68890003). While field trials of DSM-5 have revealed promising progress in reliably delineating these three conditions in clinical assessment¹¹, such diagnostic entities continue to have low clinical utility¹²⁻¹⁴. Recent evidence from genome-wide association studies appears to suggest that such disorders share common genetic loci, further countering the argument that SMI can be classified into discrete, high level diagnostic units¹⁵. In terms of clinical practice, the presenting symptomatology of SMI is usually the basis for treatment. This is often characterised by abnormalities in various mental processes, which are in turn categorised according to broad groupings of clinically observable behaviours. For instance, 'positive symptoms' refer to the presence of behaviours not seen in unaffected individuals, such as hallucinations, delusional thinking and disorganised speech. Conversely, 'negative symptoms', such as poverty of speech and social withdrawal refer to the absence of normal behaviours. Such symptomatology assessments are organised via an appropriate framework such as Positive and Negative Symptom Scale¹⁶ (PANSS) or Brief Negative Symptom Scale¹⁷. Accordingly, SNOMED CT includes coverage for many of these symptoms, generally within the 'Behaviour finding' branch (SCTID: 844005).

A qualifying factor regarding the adoption of SNOMED amongst SMI specialists might therefore require that the list of clinical 'finding' entities in SNOMED are sufficiently expansive and diverse to represent their own experiences during patient interactions. Specifically, this may manifest as two key challenges for terminology developers.

First, insight must be obtained regarding real-world language usage such that universally understood medical concepts, encompassing hypernymy, synonymy and hyponymy. Similarly, the abundant use of acronyms in the medical domain means that a

large percentage of acronyms to have two or more meanings¹⁸, creating word sense disambiguation problems. As such, significant efforts have arisen to supplement these types of knowledge bases with appropriate real-world synonym usage extracted from EHR datasets¹⁹. The problem may be considered analogous to difficulties in the recognition, classification and mapping of technical terminology variants throughout the biomedical literature, which is known to be an impediment to the construction of knowledge representation systems (see 20 for a review).

Second, if there is controversy over international consensus in a particular area of medicine, the use of ‘global’ perspectives may not be sufficient to meet local reporting/investigatory requirements. Such issues are particularly pertinent in mental health where many diseases defy precise definition and biomarker development has yielded few successes²¹. More generally, all medical knowledge bases are incomplete to one degree or another. The opportunity to utilise large amounts of EHR data to discover novel observations and relationships arising from real-world clinical practise must not be overlooked.

Given a sufficiently large corpus of documents, typically written by hundreds of clinical staff over several years, it is often difficult to track the evolution of vocabulary used within the local EHR setting to describe potentially important clinical constructs. In previous work, we describe our attempts to extract fifty well known SMI symptomatology concepts from a large electronic mental health database resource²², based upon the contents of such frameworks. During the course of manually reviewing clinical text, we made two subjective observations of the documentation resulting from clinician/patient interactions:

- The tendency of clinicians to use non-technical vocabulary in describing their observations

- The occasional appearance of highly detailed, novel observations that do not readily fit into known symptomatology frameworks

Such observations may feasibly have clinical relevance, for example, as non-specific symptomatology prodromes²³. On the basis that the modelling of SMI for precision medicine approaches require the full dimensionality of the disease to be considered, we sought to explore these observations further.

In this study, we present our efforts to utilise *a priori* knowledge discovery methods to identify preferences in real-world language usage that reflect clinically relevant SMI symptomatology within the context of a large mental healthcare provider. We contrast and compare these patterns with a modern version of the UK SNOMED CT (v1.33.2), and suggest how such approaches may offer novel and/or more granular symptom expressions from patient/clinician interactions when used to supplement resources such as SNOMED CT, potentially offering alternatives to classify psychiatric disorders with finer resolution and greater real-world validity.

Methods

Our general approach for SMI knowledge discovery is composed of several discrete steps. An overview of the workflow is given in Figure 1.

Corpus creation from the Clinical Record Interactive Search

The South London and Maudsley NHS Foundation Trust (SLaM) provides mental health services to 1.2 million residents over four south London boroughs (Lambeth, Southwark, Lewisham and Croydon). Since 2007, the Clinical Record Interactive Search (CRIS)²⁴ infrastructure programme has been operating to offer a pseudonymised and de-identified

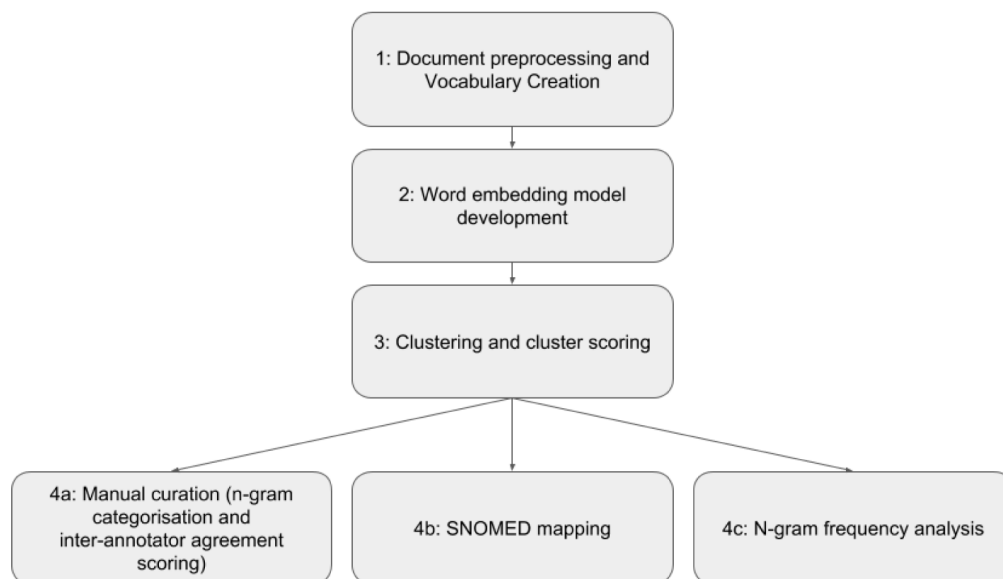


Figure 1. Overview of project workflow.

research database of SLaM's EHR system. As the CRIS resource received ethical approval as a pseudonymised and de-identified data source by Oxford Research Ethics Committee (reference 08/H0606/71+5), patient consent was not required for this study.

11 745 094 clinical documents were collected from the [CRIS database](#) from the period 01/01/2007 - 27/10/2016 on the basis that the 20 472 associated patients were assigned an SMI ICD10 code of F20, F25, F30 or F31 at some point during their care, in accordance with current clinical practice.

Pre-processing and vocabulary creation

Sentences and tokens were extracted from each document using the English Punkt tokeniser from the NLTK 3.0 suite²⁵. Each token was converted to lower case. A vocabulary was then constructed of all 1-gram types in the corpus, supplemented with frequently occurring bi-grams and tri-grams using the Gensim²⁶ suite and the sampling method proposed by Mikolov *et al.*²⁷. Bi-grams and tri-grams with a minimum frequency of 10 occurrences in the entire corpus were retained, to give a total vocabulary size of 896 195 terms (617 095 unigrams, 277 490 bigrams, 303 trigrams and 1 307 non-word entities). No further assumptions about the structure of the data, such as the need for stemming/lemmatisation, were made.

Building a word embedding model

The distributional hypothesis was first explored by Harris²⁸, which proposed that, given a sufficiently large body of text, linguistic units that co-occur in the same context are likely to have a semantically related meaning. Modelling the distribution of such units may therefore have value for a wide range of natural language processing applications. Models of distributional semantics, including word embeddings, are techniques that aim to derive models of semantically similar units in a corpus of text by co-locating them in vector space. In recent years, the use of the Continuous Bag-of-Words (CBOW) model proposed by Mikolov *et al.*²⁹ has risen to prominence, owing to its ability to accurately capture semantic relationships whilst scaling to large corpora of text²⁷. Recently, the CBOW model has been used to identify the semantic similarities between single word entities in biomedical literature and clinical text³⁰, suggesting that biomedical literature may serve as a useful proxy for clinical text, for tasks such as synonym identification and word sense disambiguation tasks under limited conditions³⁰.

A full description of the CBOW architecture is discussed in 31. For brevity, we describe only the key features used in our work here. The purpose of the architecture is to 'learn' in an unsupervised manner, a representation of the semantics of different terms, given an input set of documents. CBOW might be described as a simple feed forward neural network consisting of three layers. An input layer X composed of o nodes (where o is the number of unique terms in a corpus produced from our above described pre-processing), a hidden layer H of a user defined size n (usually between 100 and 300), and an output

layer Y that is also composed of o nodes. Every node in X is connected to every node in H , and every node in H is connected to every node in Y . Between each of the layers is a matrix of weight values; for the X and H layer, an 'input' matrix of dimensions $o \times n$ (hereafter denoted W); and between the H and the Y layer, an 'output' matrix of dimensions $n \times o$ (denoted W'). The output of training the neural network is to produce weights in each of these matrices. The weights learnt in the W matrix might be intuitively described as the semantic relationships between each term in the vocabulary as represented in vector space, with semantically similar words located in closer proximity to each other. Weights in the W' matrix represent the predictive model from the H to the Y layer. A training instance is composed of a group of terms, known as a context. A context can be composed of natural language structures, such as sentences in a document, or more complex arrangements, such as a sliding window of terms (usually between 5 and 10) that move over each token in a document (potentially ignoring natural grammatical structures). For a given input term, the input into the nodes on the hidden layer is the product of each vector index in matrix W corresponding to each context word and the average vector. From the H to the Y layer, it is then possible to score each term using the W' matrix, from which a posterior probability is obtained for each word in the vocabulary using the softmax function. The weights in each matrix are then updated using computationally efficient hierarchical softmax or negative sampling approaches. Once training is complete, the semantic similarity of terms is often measured via their cosine distance between vectors in the W matrix.

Using the Gensim implementation of CBOW and our previously constructed vocabulary, we trained a word embedding model of $n = 100$ over our SMI corpus to produce a vector space representation of our clinical vocabulary. Due to patient confidentiality, offline access to records was not feasible and so only a limited number of epochs of training could be performed. However, due to the relatively narrow/controlled vocabulary employed in clinical records (compared to normal speech/text) the range of possible input vectors was narrower than might otherwise be expected, and even a single epoch of training appeared to yield meaningful clusters that could be identified with SMI. As we were primarily intending to identify initial clusters for validation by clinical experts it was felt that single epoch of training, over the 20M clinical records available, was sufficient.

Vocabulary clustering and cluster scoring

The task of clustering seeks to group similar dataset objects together in meaningful ways. In unsupervised clustering, the definition of 'meaningfulness' is often subjectively defined by the human observers. In our task, we sought to identify clusters of terms derived from our word embedding model that represent semantically linked components of our clinical vocabulary, based on the theory that our word embedding model would cause related symptom concepts to appear close to each other within the vector space.

A particular challenge in the development of clustering algorithms is achieving scalability to large datasets. Since many clustering algorithms make use of the pairwise distance between n samples (or terms, in our case), the memory requirements of such algorithms tend to run in the order of n^2 . One such algorithm that does not suffer from this limitation is k -means clustering. k -means clustering is a partitional clustering algorithm that seeks to assign n samples into a user defined k clusters by minimising the squared error between each centroid of a cluster and its surrounding points. A global (although not necessarily optimal) solution is derived when the algorithm has minimised the sum of squared errors across all k clusters, subject to some improvement threshold or other stopping criteria. For all experiments, we used the k -means++ implementation from the Scikit-Learn framework³² with 8 runs each time, to control against centroids emerging in local minima.

The key parameter for k -means clustering is the selection of k . While techniques exist for estimating an appropriate value, such as silhouette analysis and the ‘elbow method’³³, these utilise pairwise distances between samples, creating substantial technical limitations for large matrices in terms of memory usage. To overcome this, we opted for a memory efficient version of the elbow method, involving plotting the minimum centroid distance for different values of k . The intuition behind this approach is that every increase in k is likely to result in a smaller minimum centroid distance in vector space (subject to a random seed for the algorithm). As k increases, genuine clusters should be separated by a steady decline in minimum centroid distance. However, when the slope of the decline flattens out (i.e. the ‘elbow’ of the curve), assignment of samples to new clusters is likely to be random).

With the data clustered, we sought to identify one or more clusters of interest for further examination. To this end, we devised a simple ‘relevance’ cluster scoring approach based upon prior knowledge of common SMI symptom concepts. The intuition behind our approach is that the training of the Word2Vec model will cause terms that represent ‘known’ concepts of SMI symptomatology to collocate in close proximity to each other in vector space, and the clustering approach will place them in the same cluster, along with other terms that theoretically relate to these SMI symptomatology concepts. The additional contents of this cluster may therefore hold terms that represent concepts of SMI symptomatology undefined by our team, but in natural use by the wider clinical staff of the SLaM Trust during the course of their duties. By identifying the richest cluster(s) in terms of the known SMI symptomatology lexicon, we sought to drastically reduce the search space of terms in the corpus to carry forward for human assessment.

We selected 38 internationally recognised symptom concepts of SMI based upon their expression in SMI frameworks and on their specificity in clinical use (Table 1), to form the basis of our scoring algorithm. For instance, we did not select ‘loosening of associations’, due to the different word sense that the word ‘associations’ appears in, such as ‘housing associations’, and organisational references such as ‘Stroke Association’.

Table 1. Known symptomatology concepts and Prior Concept vocabulary matching sequences used for cluster scoring. An underscore represents a bigram match.

SMI symptom	Prior Concept matching character sequence
aggression	aggress
agitation	agitat
anhedonia	anhedon
apathy	apath
affect	affect
cataplexy	catalep
catatonic	cataton
circumstantial	circumstant
concrete	concrete
delusional	delusion
derailment	derail
eye contact	eye_contact
echolalia	echola
echopraxia	echopra
elation	elat
euphoria	euphor
flight of ideas	foi
thought disorder	thought_disorder
grandiosity	grandios
hallucinations	hallucinat
hostility	hostil
immobility	immobil
insomnia	insomn
irritability	irritab
coherence	coheren
mannerisms	mannerism
mutism	mute
paranoia	paranoi
persecution	persecut
motivation	motivat
rapport	rapport
posturing	postur
rigidity	rigid
stereotypy	stereotyp
stupor	stupor
tangential	tangenti
thought block	thought_block
waxy	waxy

Rather, we chose symptoms such as ‘aggression’, ‘apathy’ and ‘agitation’, which are less likely to have different word sense interpretations in the context of SMI clinical documents.

For each of the 38 concepts, we produced a set of terms constituting stems and appropriate synonyms/acronyms as described

in Table 1, in order to produce a set of character sequences representing existing domain knowledge, or ‘prior concepts’ (hereafter, termed PCs) that could be matched against each term in each cluster via regular expressions. With this matching criterion, we scored each cluster based on the number of hits to derive a cluster/PC count matrix x where $x_{i,j}$ represents the count of the i th PC in the j th cluster. For example, a cluster containing the 1-gram ‘insomnia’ and ‘insomniac’ would receive a count of two for the ‘insomni’ PC. For each PC, we then calculated a vector of the minimum count per concept across all clusters:

$$u_i = \min_{j \in J} x_{ij}, \quad i = 1, \dots, m. \quad (1)$$

where m is 38 (denoting the number of PCs we describe in Table 1). Similarly, we generated a vector of maximum count per PC across all clusters:

$$v_i = \max_{j \in J} x_{ij}, \quad i = 1, \dots, m. \quad (2)$$

to enable us to rescale the value of each PC/cluster count to between 0 and 1 into a matrix x' :

$$x'_{ij} = \frac{x_{i,j} - u_i}{v_i - u_i} \quad (3)$$

The purpose of rescaling in such a way was to prevent overrepresented PCs unduly influencing the overall result (for instance, a PC with many hits in a cluster would unduly bias the score towards that concept, whereas we sought a scoring mechanism that would weigh all input PCs equally, regardless of their frequency).

Finally, we summed all rescaled PC counts per cluster, and divided by the total cluster size to provide a score per cluster z representing the value of the:

$$z_j = \frac{\sum_{i=1}^m x'_{ij}}{S_j} \quad (4)$$

where s is a vector of the total count of terms in each cluster. The purpose of dividing by cluster size was to prevent the tendency of larger clusters to score higher on account of their size.

To select clusters for further investigation, the robust median absolute deviation (MAD) statistic was chosen (the distribution of our cluster scores was non-normal). This precipitated clusters that were the most valuable, in terms of the breadth of PC concept hits they contain. We adopted a conservative approach to cluster selection by choosing clusters that scored at least six MAD above the median score for further processing, which is approximately equivalent to four standard deviations for a normally distributed dataset.

We provide a worked example of this technique in the code repository that accompanies this paper, using publically available data.

Expert curation of symptom concepts, frequency analysis and SNOMED CT mapping

The contents of the top scoring clusters underwent a two stage curation process. The first stage was performed by an informatician, and involved several simple string processing tasks to filter out uninteresting terms. Such processes included removal of terms that contained tokenisation failures (for example, single character non-word tokens such as ‘y’, ‘p’) and other constructs that had low information content, such as terms composed of stop words. A final manual check followed to reduce the annotator burden required by the clinical team.

The second, more important, stage was composed of independent annotation of the curated concept list by two psychiatrists, to identify likely synonyms and new symptomatology based on their clinical experience. Each concept was assigned to one of the below 8 ‘substantive’ categories, or a 9th ‘other’ category. The categories were derived from 34, and the experience of the team Clinical Psychiatrists.

Appearance/Behaviour Implying a real-time description of the way a patient appears or behaves (including their interaction with the recording clinician)

Speech Anything implying a description of any vocalisation (i.e. theoretically a subset of behaviour but restricted to vocalisations)

Affect/Mood Implying clinician-observed mood/emotional state (i.e. theoretically a subset of appearance but restricted to observed emotion), or implying self-reported mood/emotional state (i.e. has to imply a description that a patient would make of their own mood; theoretically a subset of thought)

Thought Implying any other thought content

Perception Implying any described perception

Cognition Implying anything relating to the patient’s cognitive function

Insight Implying anything relating to insight (awareness of health state)

Personality Anything implying a personality trait or attitude (i.e. something more long-standing than an observed behaviour at interview)

Other A mixed bag of definable terms that do not fit into the above. Common examples included anything implying information that will have been collected as part of a patient’s history, often of behaviours that would have to have been reported as occurring in the past and cannot have been observed at interview, but also which cannot be termed a personality trait. Alternatively, anything where insufficient context was available to make a decision

Inter annotator agreement (IAA) was measured with the Cohen's Kappa agreement statistic³⁵.

To explore the frequency of both our prior symptomatology concepts and the newly curated ones in our symptom clusters, we counted the number of unique patient records and the number of unique documents in which the stems of each term appeared. To protect patient anonymity, we discarded any concept that appeared in ten or fewer unique patient records. Finally, we mapped the remaining concepts to SNOMED CT, UK version v1.33.2, using the following method. First, the root morpheme of each concept was matched to a relevant finding, observable entity or disorder type in SNOMED CT. If a match could not be found, SNOMED CT was explored for potential synonymy, or other partial match. If a clear synonym could not be found, we classified the concept as novel.

Results

Word embedding model training

Processing the corpus of SMI clinical documents took approximately 100 hours on an 8-core commodity hardware server. Documents were fed sequentially from an SQL Server 2008 database operating as a shared resource, with an additional overhead likely resulting from network latency.

Parameter selection for *k*-means clustering

Figure 2 shows a scatterplot of variable values of *k* and the resulting minimum centroid distance. This suggests a *k* value of around 50–75 may be optimal for our data. On this basis, we chose a *k* value of 75.

Cluster scoring

The application of our relevancy scoring algorithm to the 75 derived clusters resulted in a median score was 0.000229 and a MAD of 0.000277, and is visualised in Figure 3.

Three clusters emerged with a score at least six MADs outside of the median cluster score: No. 52 (score: 0.002883), containing 6 665 terms, No. 69, containing 9 314 (score: 0.002282) terms and No. 49 (score: 0.001940), containing 4 424 terms. Taken together, these three clusters contained a total of 20 403 terms.

Expert curation of symptom concepts, frequency analysis and SNOMED CT mapping

The combined 20 403 terms were taken forward for curation as described above. The first phase of curation reduced the list to 519 putative concepts. The majority of eliminated terms were morphological variations, misspellings and tokenisation anomalies of singular concepts. For instance, 84 variations were detected for the stem 'irrit*' (as in 'irritable'). Other terms were removed because insufficient context was available for a reasonable clinical interpretation, such as 'fundamentally unchanged', 'amusing' and 'formally tested'. Finally, terms that appeared to have no relevance to symptomatology at all were removed, such as dates and clinician names.

Expert curation by two psychiatrists of the 557 concepts (519 discovered concepts and 38 prior concepts) produced a Cohen's Kappa agreement score of 0.45, where 337 concepts were assigned to one of our 9 categories independently by expert psychiatric curation. Of the 337 concepts, 235 were assigned to a substantive category (i.e. not the indeterminate 'other' group). Table 2 shows the number of terms per category where agreement was reached.

Supplementary File 1 is a CSV table of all 557 terms. In addition to the term itself, the table contains the following information; the counts of the unique patient records of our 20 472 patient SMI cohort in which the term was detected; the counts of the unique documents of the 11 745 094 clinical

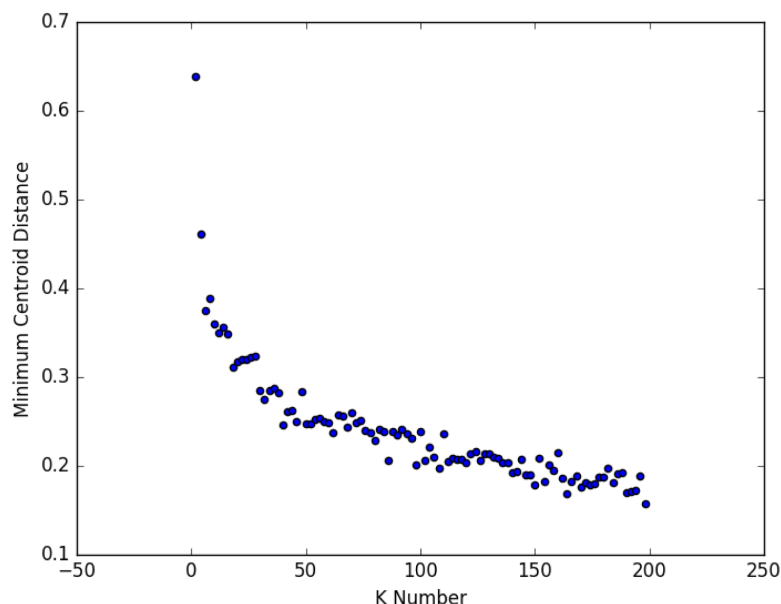


Figure 2. Selecting *K* for *K*-means++.

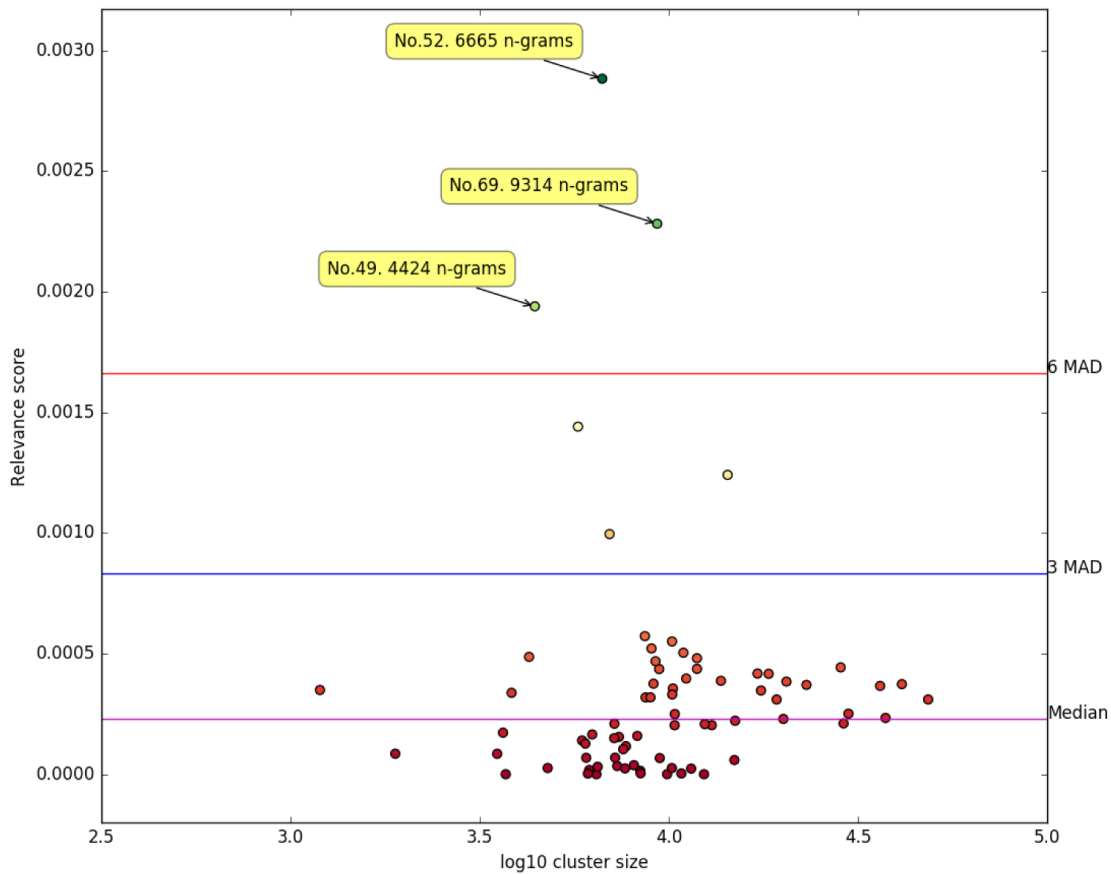


Figure 3. Scoring of clusters according to known symptomatology content. Each dot represents a unique cluster. The unique cluster IDs of the most relevant clusters according to our scoring algorithm are labelled.

Table 2. Counts of terms where annotators independently agreed by category.

Category	Count
Affect/Mood	6
Appearance/Behaviour	78
Cognition	6
Insight	2
Mood/Anxiety/Affect	26
Other	102
Perception	9
Personality	23
Speech	63
Thought	22

document corpus wherein the term was detected; the category assigned to the term by each of our clinical annotators, and the SNOMED CT ID code for each term, where mapping was possible.

The most frequently detected concept mentions include ‘affect’ (detected in 91% of patients), ‘eye contact’ (85%), ‘hallucinations’ (85%), ‘delusions’ (83%) and ‘rapport’ (81%). Other concepts follow a long tailed distribution, with mentions of the top 407 concepts found in at least 100 unique patient records.

Regarding SNOMED CT mapping, it was possible to suggest direct mappings for 177 concepts and to suggest synonymy or partial mappings for another 53 concepts. This left a remaining 327 concepts that did not appear to be referenced in SNOMED CT, of which 106 were classified as belonging to a substantive symptom category by independent curation.

Figure 4 visualises the top 20% most frequent terms by appearance in unique patient records, where annotators agreed and were not classified as our ‘other’ grouping.

Owing to the difficulty of the IAA and categorisation task, an extended analysis of the top 40% most frequent terms by appearance in unique patient records, irrespective of IAA and categorisation is provided in Supplementary Figure 1.

In this project, we sought to explore SMI symptomatology and other language constructs as expressed by clinicians in their

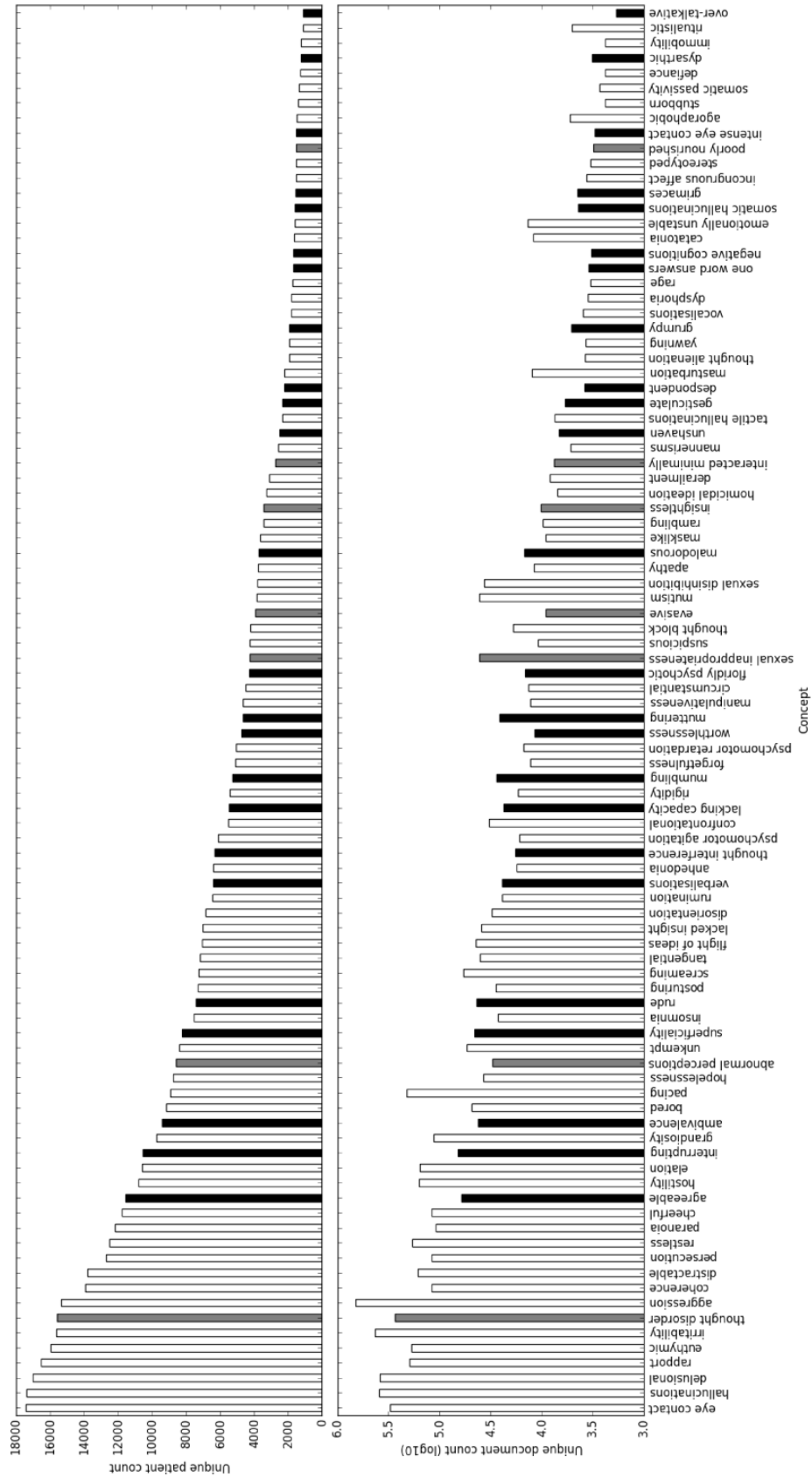


Figure 4. Frequency of terms across all SMI documents in CRIS. White bars represent concepts that were found to exist in SNOMED CT. Grey bars represent partial/uncertain matches, or novel synonyms of existing SNOMED CT concepts. Black bars represent concepts with no SNOMED mapping.

own words, using more than ten years of observations made during real-world clinician/patient interactions from more than 20 000 unique SMI cases. Within the context of a large mental healthcare provider, the results of our vocabulary curation efforts suggest that psychiatrists make use of a wide range of vocabulary to describe detailed symptomatic observations.

Many of the curated entities where both annotators agreed upon a substantive category map directly to preferred terms or synonyms of well known symptomatology constructs as described in SNOMED CT. Reassuringly, many of most frequently encountered entities as represented by unique patient count are represented in SNOMED CT, suggesting that SNOMED CT offers a reasonable coverage of what clinicians deem to be the most salient features of a psychiatric examination.

Nevertheless, our work produces evidence to suggest that many suitable synonyms are currently missing from SNOMED CT symptom entities. For instance, ‘aggression’ is commonly observed in SMI patients. Our results indicate that this construct might also be referred to by adjectives and phrases such as ‘combative’ [*sic*], ‘assaultative’ [*sic*], ‘truculent’, ‘stared intimidatngly’ and ‘stared menacingly’, amongst others. Similarly, direct synonyms of ‘paranoia’ might include ‘suspiciousness’, ‘mistrustful’ and ‘conspirational’ [*sic*].

In addition, many of the curated constructs appear to reflect more granular observations of known symptomatology. For example, the PANSS utilises a 30-point scale of different symptomatology constructs. Specifically regarding abnormal speech, the PANSS provide guidance amounting to the high level clinical scrutiny of ‘lack of spontaneity & flow of conversation’. However, clinical expressions of speech within our dataset suggest around 68 distinct states, including ‘making animal noises’, ‘staccato quality’, ‘easily interruptible’, ‘prosody’ and ‘silently mouthing’.

We note the occurrence of several constructs that defy classification under existing schemas of SMI symptomatology, such as behaviours of ‘over politeness’, ‘over complimentary’, ‘spending recklessly’ and ‘shadow boxing’. The clinical interpretation of such entities is a non-trivial exercise, and is out of scope for this piece. Nevertheless, word embedding models may offer the potential to gain insight into potentially novel symptomatology constructs observed from real-world clinician/patient interactions. Future work might explore the context for such constructs in more detail.

The emergence of such diverse language in turn has implications for how SNOMED CT might be implemented within an SMI context, raising the question of whether such gaps represent significant barriers to the use of SNOMED CT as a phenotyping resource. The issue of SNOMED CT’s sufficiency in this context has previously been raised for other areas, such as rare disease³⁶, psychological assessment instruments³⁷ and histopathology findings³⁸. However, in fairness, SNOMED CT is not a static resource, but an international effort dependent on the contributions of researchers. Perhaps a more pertinent question for the future development of SNOMED CT concerns balancing its objective to be a comprehensive terminology of

clinical language (capable of facilitating interoperability and modelling deep phenotypes within disparate healthcare organisations across the globe) and the overwhelming complexity it would need to encompass in order to not constrain its users. Certainly, at more than 300 000 entities in its current incarnation, its size already presents problems in biomedical applications³⁹.

Limitations and future work

On the basis that manifestations of symptoms are the result of abnormal mental processes, novel symptom entities possibly represent observations of clinical significance. However, one particular complication in validating the clinical utility of novel symptomatology constructs with historic routinely recorded notes arises from systemic biases in EHR data. Specifically, the breadth and depth of symptomatic reporting is likely to be highly variable for a number of reasons. For instance, established symptoms as defined by current diagnostic frameworks are likely to be preferentially recorded, as clinicians are mandated to capture such entities in their assessments. On the other hand, constructs that fall outside of such frameworks may only be recorded as tangential observations made during patient/clinician interactions. Regardless of whether they are observed or not, without an established precedent of their clinical utility, they may be subject to random variation as to whether they are documented in a patient’s notes. This is borne out by the tendency of SNOMED CT-ratified concepts to appear more frequently in unique documents compared to our derived expressions. The validation of new symptoms from historic data is therefore something of a ‘chicken and egg’ situation, a widely-discussed limitation of the reuse of EHR data^{40,41}. Nevertheless, our frequency analysis of our discovered constructs suggests that there is evidence that many are observed often enough to warrant their consideration within an expanded framework. Similarly, older frameworks with a limited scope of symptomatic expression were likely designed with pragmatic constraints around speed and reproducibility of assessment in mind. However, modern technology allows for a far greater scope of data capture and validation going forward, creating opportunities to develop new frameworks that maximise the value of psychiatric assessment. Future work in this domain might seek statistical validation via randomised experimental design, as opposed to observational study.

Our work suggests an approximate correlation between patient and document count, such that intra and inter patient symptomatological clinical language usage varies relatively consistently. However, some notable exceptions to this correlation (i.e. with a higher document level frequency to patient record level frequency) include ‘aggression’, ‘pacing’, ‘sexual inappropriateness’, ‘sexual disinhibition’ and ‘mutism’. Further work might seek to study these effects in greater detail, to uncover whether they represent a systemic bias in how such concepts are represented in the EHR.

The results of our IAA exercise between two experienced psychiatrists suggested a moderate level of agreement in categorising the newly identified constructs. Given that this annotation exercise did not provide any context beyond the term, and that the nature of SMI symptom observation is somewhat subjective, perhaps it is to be expected that agreement was not higher.

As suggested during peer review, providing a concordance of some of the instances of each term, along with expert panel discussion and engagement with international collaborative efforts in SMI research may prove valuable in seeking more formal definitions of the identified concepts.

Our method for vocabulary building produced nearly 1 million terms. A manual annotation of this list may have resulted in further discoveries, although would have been intractable in practical terms. To reduce the volume of terms taken forward for curation, we employed a word embedding model with a clustering algorithm. With our cluster scoring methodology that makes use of existing domain knowledge, we were able to successfully produce meaningful clusters of terms reflecting the semantics of SMI symptomatology. However, as with many unsupervised tasks, it is difficult to determine whether an optimal solution has been achieved. In particular, the emergence of three ‘symptom’ clusters instead of one indicates sub-optimal localisation of symptom constructs in vector space. Addressing such a problem is multifaceted. For technical reasons, only a single epoch of training was possible in this exercise. Additional epochs would likely contribute to better cluster definition, in turn allowing us to reduce the value of our k parameter. In addition, spell checking and collapsing terms into their root forms may also have assisted. However, the latter may have also created new word sense disambiguation problems if common, symptom-like morphemes also appear in nonsymptomatological assessment contexts.

After clustering, a two stage manual curation of more than 20 000 terms was necessary. Methods that produce a smaller vocabulary might conceivably reduce annotator burden. This might include the use of spell checkers and stemming/lemmatisation to correct and normalise tokens, at the risk of introducing new issues associated with morphological forms in word embedding model building. For this attempt, we took the conscious decision to make as few assumptions about the underlying structure of the data as possible.

During peer review, it was suggested that recent advancements in topic modelling approaches may be relevant to our work. Many groups have sought to combine the popular technique of Latent Dirichlet Allocation (LDA)⁴² with word embedding models to derive appropriate terminology for a given topic^{43–45}. For instance, Nguyen *et al.*⁴⁶ propose an extension of LDA that makes use of a word embedding model trained on a very large corpus of text to improve the performance of topic coherence modelling on several datasets. Future work might seek to explore such techniques, and (assuming regulatory barriers can be overcome), the potential of creating word embedding models from very large clinical text corpora by combining data with other care organisations.

Conclusions

Evidence-based mental health has long sought to produce disease model definitions that are both valid, in the sense they

represent useful clinical representations that can inform treatment, and reliable, in that they can be consistently applied by different clinicians to achieve the same outcomes. In practice this has proven difficult, due to the often subjective nature of psychiatric examination/phenotyping and insufficient knowledge about the underlying mechanisms of disorders such as SMI. Here, we demonstrate that clinical staff make use of a diverse vocabulary in the course of their interactions with patients. This vocabulary often references findings that are not represented in SNOMED CT, raising questions about whether clinicians should observe the constraints of SNOMED CT or whether SNOMED CT should incorporate greater flexibility to reflect the nature of mental health. It is outside the scope of this work to explore how the granularity of symptom-based phenotyping affects patient outcomes, although the possibility of offering a fully realised picture of symptom manifestation may prove valuable in future endeavours of precision medicine.

Data availability

The CRIS dataset is a pseudonymised and de-identified case registrar of electronic health records of the SLaM NHS Trust. It operates under a security model that does not allow for open publication of raw data. However, access can be granted for research use cases under a patient-led security model. For further information and details on the application process, please contact cris.administrator@kcl.ac.uk or visit the website: <https://www.maudsleybrc.nihr.ac.uk/facilities/clinical-record-interactive-search-cris/>. Alternatively, you may write to the CRIS team at:

PO Box 92 Institute of Psychiatry, Psychology & Neuroscience at King’s College London 16 De Crespigny Park London SE5 8AF

Example code used in this analysis is available at: https://github.com/RichJackson/clustering_w2v

Competing interests

RJ and RS have received research funding from Roche, Pfizer, J&J and Lundbeck.

Grant information

This paper represents independent research funded by the National Institute for Health Research (NIHR) Biomedical Research Centre at South London and Maudsley NHS Foundation Trust and King’s College London. RP has received support from a Medical Research Council (MRC) Health Data Research UK Fellowship and a Starter Grant for Clinical Lecturers (SGL015/1020) supported by the Academy of Medical Sciences, The Wellcome Trust, MRC, British Heart Foundation, Arthritis Research UK, the Royal College of Physicians and Diabetes UK. SV is supported by the Swedish Research Council (2015-00359), Marie Skłodowska Curie Actions, Cofund, Project INCA 600398.

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Supplementary material

Supplementary File 1: This file contains all of the 557 terms taken forward for expert annotation. It includes SNOMED mappings where possible, unique document and patient counts within the corpus, and the annotations provided by RP and RS.

[Click here to access the data.](#)

Supplementary Figure 1: This file is an expanded visualisation of the frequency analysis figure contained in the main manuscript, with the agreement and nonsubstantive ‘other’ classification restrictions lifted.

[Click here to access the data.](#)

References

- Amberger J, Bocchini C, Hamosh A: **A new face and new challenges for Online Mendelian Inheritance in Man (OMIM®).** *Hum Mutat.* 2011; **32**(5): 564–567. ISSN 10597794.
[PubMed Abstract](#) | [Publisher Full Text](#)
- Mirnezami R, Nicholson J, Darzi A: **Preparing for precision medicine.** *N Engl J Med.* 2012; **366**(6): 489–491. ISSN 0028-4793, 1533-4406.
[PubMed Abstract](#) | [Publisher Full Text](#)
- Robinson PN: **Deep phenotyping for precision medicine.** *Hum Mutat.* 2012; **33**(5): 777–780. ISSN 10597794.
[PubMed Abstract](#) | [Publisher Full Text](#)
- Pathak J, Kho AN, Denny JC: **Electronic health records-driven phenotyping: challenges, recent advances, and perspectives.** *J Am Med Inform Assoc.* 2013; **20**(e2): e206–11.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Castro VM, Minnier J, Murphy SN, *et al.*: **Validation of electronic health record phenotyping of bipolar disorder cases and controls.** *Am J Psychiatry.* 2015; **172**(4): 363–372. ISSN 0002-953X, 1535-7228.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- NATIONAL INFORMATION BOARD: **Personalised Health and Care 2020.** 2014. [Reference Source](#)
- Lee D, Cornet R, Lau F, *et al.*: **A survey of SNOMED CT implementations.** *J Biomed Inform.* 2013; **46**(1): 87–96. ISSN 15320464.
[PubMed Abstract](#) | [Publisher Full Text](#)
- Barnes M: **Lessons learned from the implementation of clinical messaging systems.** *AMIA Annu Symp Proc.* 2007; 36–40. ISSN 1942-597X.
[PubMed Abstract](#) | [Free Full Text](#)
- The future of healthcare informatics: it is not what you think.** *Glob Adv Health Med.* 2012; **1**(4): 5–6. ISSN 2164-957X, 2164-9561.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Gordon D: **Merging multiple institutions: Information architecture problems and solutions.** *Proc AMIA Symp.* 1999; 785–789. ISSN 1531-605X.
[PubMed Abstract](#) | [Free Full Text](#)
- Freedman R, Lewis DA, Michels R, *et al.*: **The initial field trials of DSM-5: new blooms and old thorns.** *Am J Psychiatry.* 2013; **170**(1): 1–5. ISSN 0002-953X, 1535-7228.
[PubMed Abstract](#) | [Publisher Full Text](#)
- Kendell R, Jablensky A: **Distinguishing between the validity and utility of psychiatric diagnoses.** *Am J Psychiatry.* 2003; **160**(1): 4–12. ISSN 0002-953X, 1535-7228.
[PubMed Abstract](#) | [Publisher Full Text](#)
- Chmielewski M, Bagby RM, Markon K, *et al.*: **Openness to experience, intellect, schizotypal personality disorder, and psychoticism: resolving the controversy.** *J Pers Disord.* 2014; **28**(4): 483–99. ISSN 1943-2763.
[PubMed Abstract](#) | [Publisher Full Text](#)
- Adam D: **Mental health: On the spectrum.** *Nature.* 2013; **496**(7446): 416–418. ISSN 0028-0836, 1476-4687.
[PubMed Abstract](#) | [Publisher Full Text](#)
- Cross-Disorder Group of the Psychiatric Genomics Consortium: **Identification of risk loci with shared effects on five major psychiatric disorders: a genome-wide analysis.** *Lancet.* 2013; **381**(9875): 1371–1379. ISSN 01406736.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Kay SR, Fiszbein A, Opler LA: **The positive and negative syndrome scale (PANSS) for schizophrenia.** *Schizophr Bull.* 1987; **13**(2): 261–76. ISSN 0586-7614. Cited by 8221.
[PubMed Abstract](#) | [Publisher Full Text](#)
- Kirkpatrick B, Strauss GP, Nguyen L, *et al.*: **The brief negative symptom scale: psychometric properties.** *Schizophr Bull.* 2010; **37**(2): 300–305. ISSN 0586-7614, 1745-1701. Cited by 0000.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Liu H, Aronson AR, Friedman C: **A study of abbreviations in MEDLINE abstracts.** *Proc AMIA Symp.* 2002; 464–468. ISSN 1531-605X.
[PubMed Abstract](#) | [Free Full Text](#)
- Henriksson A, Conway M, Duneld M, *et al.*: **Identifying synonymy between SNOMED clinical terms of varying length using distributional analysis of electronic health records.** *AMIA Annu Symp Proc.* 2013; **2013**: 600–609. ISSN 1942-597X.
[PubMed Abstract](#) | [Free Full Text](#)
- Krauthammer M, Nenadic G: **Term identification in the biomedical literature.** *J Biomed Inform.* 2004; **37**(6): 512–526. ISSN 15320464.
[PubMed Abstract](#) | [Publisher Full Text](#)
- Boksa P: **A way forward for research on biomarkers for psychiatric disorders.** *J Psychiatry Neurosci.* 2013; **38**(2): 75–85. ISSN 11804882.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Jackson RG, Patel R, Jayatilake N, *et al.*: **Natural language processing to extract symptoms of severe mental illness from clinical text: The Clinical Record Interactive Search Comprehensive Data Extraction (CRIS-CODE) project.** *BMJ Open.* 2017; **7**(1): e012012. ISSN 2044-6055, 2044-6055.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- McGorry PD: **The next stage for diagnosis: Validity through utility.** *World Psychiatry.* 2013; **12**(3): 213–215. ISSN 17238617.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Perera G, Broadbent M, Callard F, *et al.*: **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.* 2016; **6**(3): e008721. ISSN 2044-6055, 2044-6055.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Bird S, Klein E, Loper E: **Natural Language Processing with Python.** O'Reilly, Beijing; Cambridge [Mass.], 1st ed edition, 2009. ISBN 978-0-596-51649-9. OCLC: ocn301885973.
[Reference Source](#)
- Řehůřek R, Sojka P: **Software Framework for Topic Modelling with Large Corpora.** In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks.* 2010; 45–50. Valletta, Malta, ELRA.
[Publisher Full Text](#)
- Mikolov T, Sutskever I, Chen K, *et al.*: **Distributed representations of words and phrases and their compositionality.** *Adv Neural Inf Process Syst.* 2013; 3111–3119.
[Reference Source](#)
- Harris ZS: **Distributional Structure.** *WORD.* 1954; **10**(2–3): 146–162. ISSN 0043-7956, 2373-5112.
[Publisher Full Text](#)
- Mikolov T, Chen K, Corrado G, *et al.*: **Efficient estimation of word representations in vector space.** *arXiv preprint arXiv: 1301.3781.* 2013.
[Reference Source](#)
- Pakhomov SV, Finley G, McEwan R, *et al.*: **Corpus domain effects on distributional semantic modeling of medical terms.** *Bioinformatics.* 2016; **32**(23): 3635–3644. ISSN 1367-4803, 1460-2059.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Rong X: **Word2vec parameter learning explained.** *arXiv preprint arXiv: 1411.2738.* 2014.
[Reference Source](#)
- Pedregosa F, Varoquaux G, Gramfort A, *et al.*: **Scikit-learn: Machine Learning in**

- Python.** *J Mach Learn Res.* 2011; **12**: 2825–2830.
[Reference Source](#)
33. Kodinariya TM, Makwana PR: **Review on determining number of Cluster in K-Means Clustering.** *Int J.* 2013; **1**(6): 90–95.
[Reference Source](#)
34. Harrison PJ, Cowen P, Burns T, *et al.*: **Shorter Oxford book of psych.** In *Shorter Oxford Textbook of Psychiatry.* Oxford University Press, Oxford, seventh edition edition, 2018; 44. ISBN 978-0-19-874743-7.
35. Cohen J: **A Coefficient of Agreement for Nominal Scales.** *Educ Psychol Meas.* 1960; **20**(1): 37–46. ISSN 0013-1644, 1552-3888.
[Publisher Full Text](#)
36. Sollie A, Sijmons RH, Lindhout D, *et al.*: **A new coding system for metabolic disorders demonstrates gaps in the international disease classifications ICD-10 and SNOMED-CT, which can be barriers to genotype-phenotype data sharing.** *Hum Mutat.* 2013; **34**(7): 967–973. ISSN 10597794.
[PubMed Abstract](#) | [Publisher Full Text](#)
37. Ranallo PA, Adam TJ, Nelson KJ, *et al.*: **Psychological assessment instruments: a coverage analysis using SNOMED CT, LOINC and QS terminology.** *AMIA Annu Symp Proc.* 2013; **2013**: 1333–1340. ISSN 1942-597X.
[PubMed Abstract](#) | [Free Full Text](#)
38. Campbell WS, Campbell JR, West WW, *et al.*: **Semantic analysis of SNOMED CT for a post-coordinated database of histopathology findings.** *J Am Med Inform Assoc.* 2014; **21**(5): 885–892. ISSN 1067-5027, 1527-974X.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
39. López-García P, Schulz S: **Can SNOMED CT be squeezed without losing its shape?** *J Biomed Semantics.* 2016; **7**(1): 56. ISSN 2041-1480.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
40. Weiskopf NG, Weng C: **Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research.** *J Am Med Inform Assoc.* 2013; **20**(1): 144–151. ISSN 1067-5027, 1527-974X.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
41. Chan KS, Fowles JB, Weiner JP: **Review: electronic health records and the reliability and validity of quality measures: a review of the literature.** *Med Care Res Rev.* 2010; **67**(5): 503–527. ISSN 1077-5587, 1552-6801.
[PubMed Abstract](#) | [Publisher Full Text](#)
42. Blei DM, Ng AY, Jordan MI: **Latent dirichlet allocation.** *J Mach Learn Res.* 2003; **3**: 993–1022.
[Reference Source](#)
43. Cao Z, Li S, Liu Y, *et al.*: **A Novel Neural Topic Model and Its Supervised Extension.** In *AAAI.* 2015; 2210–2216.
[Reference Source](#)
44. Hinton GE, Salakhutdinov RR: **Replicated softmax: An undirected topic model.** In *Adv Neural Inf Process Syst.* 2009; 1607–1614.
[Reference Source](#)
45. Srivastava N, Salakhutdinov RR, Hinton GE: **Modeling documents with deep boltzmann machines.** *arXiv preprint arXiv:1309.6865,* 2013.
[Reference Source](#)
46. Nguyen DQ, Billingsley R, Du L, *et al.*: **Improving topic models with latent feature word representations.** *Trans Assoc Comput Linguist.* 2015; **3**: 399–313.
[Reference Source](#)

Open Peer Review

Current Referee Status:  

Version 2

Referee Report 24 May 2018

doi:10.5256/f1000research.16078.r33789

 **Karin Verspoor**  ^{1,2}

¹ School of Computing and Information Systems, The University of Melbourne, Melbourne, VIC, Australia

² Health and Biomedical Informatics Centre, The University of Melbourne, Melbourne, VIC, Australia

The authors have done a rigorous job of addressing the comments of the initial round of reviews. Thank you for that effort.

To clarify my comments on word embeddings in the context of topic modelling, I wasn't actually suggesting to use external resources for word embeddings; rather, since you are already building word embeddings, to use an alternative -- potentially more effective -- strategy for producing the clusters from those word embeddings. But your modifications to address this point are fine.




Competing Interests: No competing interests were disclosed.

Referee Expertise: biomedical natural language processing

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Referee Report 16 May 2018

doi:10.5256/f1000research.16078.r33788

 **Julian Hong**  ¹, **Jessica Tenenbaum**  ²

¹ Department of Radiation Oncology, Duke University School of Medicine, Durham, NC, USA

² Department of Biostatistics and Bioinformatics, Duke University School of Medicine, Durham, NC, USA

The authors have addressed our concerns. Thank you for clarifying the method, and thanks also for the WordToVec reference!

Competing Interests: No competing interests were disclosed.

Referee Expertise: Patient stratification in mental health using EHR data, structured and free text

We have read this submission. We believe that we have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Version 1

Referee Report 22 March 2018

doi:10.5256/f1000research.15033.r31100

**Karin Verspoor**  1,2¹ School of Computing and Information Systems, The University of Melbourne, Melbourne, VIC, Australia² Health and Biomedical Informatics Centre, The University of Melbourne, Melbourne, VIC, Australia

The paper introduces a strategy for unsupervised analysis of a corpus of documents in order to identify terminology related to Serious Mental Illness (SMI). It applies a process of (1) identification of frequent terms (n-grams), (2) clustering of terms based of word embedding vector similarity using k-means, (3) scoring of clusters using mappings to known SMI concepts, (4) manual annotation of concepts/terms (which?) to categories. The authors then provide a detailed analysis supported by manual review by two psychiatrists. A substantial number of new relevant symptomology terms are identified through this process.

The authors apply standard approaches/tools for doing the text analysis, which are generally well explained and easy to follow. The authors do not directly justify the frequency floor of 10, or provide details of how many of each type of n-gram (uni/bi/tri-grams) were identified in the data. There appear to be very few trigrams, for instance. Using only frequency, how do you prevent uninteresting patterns such as "of the"?

The manual data cleaning process could have been performed semi-automatically with simple string processing tools; was this considered? (Why else would an informatician specifically need to do it?)

The intuitions underlying the cluster scoring functions are not clearly stated; we are told it is related to prior knowledge of concepts but it is unclear what the objectives of the specific formulas presented/used are. Why are outliers of particular interest?

For future work, the authors may be interested in experimenting with topic modeling rather than k-means clustering; see ¹ for an approach which couples word embeddings with topic modeling. This could be more effective than k-means clustering, in particular due to the challenge of having to determine a good value for "k".

The notion of "n-gram" is not used entirely consistently with its broader usage in the literature; usually that refers specifically to a term of a given length (e.g. 1-grams/unigrams, 2-grams/bigrams) while different length terms are mixed here. The authors might consider using the word "term", or they have referred to "concepts" which seem to be equivalent to terms.

Regarding the limitation that no context was provided to the annotators; would it make sense to provide a concordance of some of the instances of the terms to the annotators in future efforts?

Also, the IAA is tied to the 9 categories defined on page 7; where do these categories come from? Are they related to standard or validated frameworks for symptoms in psychiatric assessment? If not, why were those categories chosen?

Did you perform any error analysis to explore the IAA further, e.g. a confusion matrix between categories? Is it possible that rather than considering these categories to be independent (the typical assumption for Cohen's Kappa) that some overlap between the categories might be expected?

The data is protected by patient privacy constraints and hence cannot be made openly available (indeed the authors could only work on the data in a restricted "offline" setting). Given the nature of the data, this is understandable. OTOH, given that the analysis largely makes use of existing code plus extensions for scoring functions, it would make sense to share the *methods* in an open repository.

The authors should include a suitable reference for PANSS. There is also a substantial literature on terminology induction (e.g. ²) which would be appropriate to reference.

The writing in the manuscript is generally clear, although I identified a few things that could be rephrased or clarified:

- The word "depiction" seems to mean "usage" or "phrase" or "expression" or similar; "depiction" is typically used in the context of art or illustration and I found it strange in a language-expression related context.
- Are phenotypes and symptomatology always the same thing? Is a phenotype a *set* of behaviours/symptoms?
- The abstract is not as clear as it could be. The final sentence of the Background paragraph should use "it is" rather than "it's" but more importantly it implies that the objective is to assess clinician *preferences* as opposed to actual *usage*. Are these the same? Also, n-grams, vector space models, concepts, vocabulary and depictions are all introduced; it is a bit confusing without having read the full paper. I wonder if it could be simplified somewhat?
- As a nitpick, in the Introduction the authors refer to "predicting the diversity of vocabulary"; the work does not address *prediction* of vocabulary or its diversity but rather involves *analysis* of that vocabulary.
- Another nitpick in the Introduction is the use of the term "authorship"; I suppose the authors mean "writing" or "description" or "summary" or similar.

References

1. Nguyen DQ, Billingsley R, Du L, Johnson M: Improving Topic Models with Latent Feature Word Representations. *Transactions of the Association for Computational Linguistics*. 2015; **3**: 299-313
2. Frantzi K, Ananiadou S, Mima H: Automatic recognition of multi-word terms: the C-value/NC-value method. *International Journal on Digital Libraries*. 2000; **3** (2): 115-130 [Publisher Full Text](#)

Is the work clearly and accurately presented and does it cite the current literature?

Partly

Is the study design appropriate and is the work technically sound?

Yes

Are sufficient details of methods and analysis provided to allow replication by others?

Yes

If applicable, is the statistical analysis and its interpretation appropriate?

Yes

Are all the source data underlying the results available to ensure full reproducibility?

No

Are the conclusions drawn adequately supported by the results?

Yes

Competing Interests: No competing interests were disclosed.

Referee Expertise: biomedical natural language processing

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response 21 Apr 2018

Richard Jackson,

Thank you very much for your insightful comments.

Please see our responses below:

The authors apply standard approaches/tools for doing the text analysis, which are generally well explained and easy to follow. The authors do not directly justify the frequency floor of 10, or provide details of how many of each type of n-gram (uni/bi/tri-grams) were identified in the data. There appear to be very few trigrams, for instance. Using only frequency, how do you prevent uninteresting patterns such as "of the"?

We started this project with the explicit intention of making as few assumptions as possible about semantic relationships contained within the 20 million documents in the CRIS corpus. To this end, we kept pre-processing very light, and no attempt was made to eliminate very common n-grams. Our chief assumption in planning the methodology was that uninteresting, high frequency n-grams that appear in many contexts would occupy locations in the vector space a substantial distance away from the n-grams we were interested in (symptomatology). We therefore sought to maximise the performance of our clustering method and scoring algorithm, which we hoped would filter off the n-grams that carry little information. In addition, any uninteresting patterns that did survive the filter, we removed via some simple string processing tools (see response to your additional question below on this). Regarding the counts of different n-grams, we have added the following to the vocabulary creation subsection:

“

Sentences and tokens were extracted from each document using the English Punkt tokeniser from the NLTK 3.0 suite\cite{bird_natural_2009}. Each token was converted to lower case. A vocabulary was then constructed of all 1-gram types in the corpus, supplemented with frequently occurring bi-grams and tri-grams using the Gensim\cite{rehurek_lrec} suite and the sampling method proposed by Mikolov \textit{et al}\cite{mikolov2013distributed}. Bi-grams and tri-grams with a minimum frequency of 10 occurrences in the entire corpus were retained, to give a total vocabulary size of 896 195 terms (617 095 unigrams, 277 490 bigrams, 303 trigrams and 1307 non-word entities). No further assumptions about the structure of the data, such as the need for

stemming/lemmatisation, were made.

”

The manual data cleaning process could have been performed semi-automatically with simple string processing tools; was this considered? (Why else would an informatician specifically need to do it?)

Indeed it was, although this was not made clear in the manuscript. We have added a short piece of text (in the “expert curation” subsection) to explain this:

“

The contents of the top scoring clusters underwent a two stage curation process. The first stage was performed by an informatician, and involved several simple string processing tasks to filter out uninteresting n-grams. Such processes included removal of n-grams that contained tokenisation failures (for example, single character non-word tokens such as ‘y’, ‘p’) and other constructs that had low information content, such as n-grams composed of stop words. A final manual check followed to reduce the amount of annotator burden required by the clinical team.

”

The intuitions underlying the cluster scoring functions are not clearly stated; we are told it is related to prior knowledge of concepts but it is unclear what the objectives of the specific formulas presented/used are. Why are outliers of particular interest?

We accept this was not clearly stated and have adjusted the manuscript accordingly, in the “Vocabulary clustering and cluster scoring” section:

“

With the data clustered, we sought to identify one or more clusters of interest for further examination. To this end, we devised a simple ‘relevance’ cluster scoring approach based upon prior knowledge of common SMI symptom concepts. The intuition behind our approach is that the training of the Word2Vec model will cause n-grams that represent ‘known’ concepts of SMI symptomatology to co-locate in close proximity to each other in vector space, and the clustering approach will place them in the same cluster, along with other n-grams that theoretically relate to these SMI symptomatology concepts. The additional contents of this cluster may therefore hold n-grams that represent concepts of SMI symptomatology undefined by our team, but in natural use by the wider clinical staff of the SLAM Trust during the course of their duties. By identifying the richest cluster(s) in terms of the known SMI symptomatology lexicon, we sought to drastically reduce the search space of n-grams in the corpus to carry forward for human assessment.

”

For future work, the authors may be interested in experimenting with topic modeling rather than k-means clustering; see 1 for an approach which couples word embeddings with topic modeling. This could be more effective than k-means clustering, in particular due to the challenge of having to determine a good value for “k”.

We agree that recent advancements in topic modelling approaches are relevant to our work here. Regarding the specific case of using external word embedding models, we suspect that our target domain, (UK clinical text), is a sub-language, and the use of external word embeddings (even from

very large corpora, such as Google News in the model proposed in the suggested citation) will have limited value for discovery on our data. The concepts of interest in our work are technical in nature, and seem likely to be specific to heavily regulated documents and therefore unlikely to exist in publically available datasets. On the other hand, there is no requirement to build the word embedding model from external datasets. Ultimately, there's clearly a range of additional techniques in the literature that would be worthwhile experimenting with. We have updated our discussion as follows:

“

During peer review, it was suggested that recent advancements in topic modelling approaches may be relevant to our work. Many groups have sought to combine the popular technique of Latent Dirichlet Allocation (LDA)\cite{blei2003latent} with word embedding models to derive appropriate terminology for a given topic\cite{cao2015novel,hinton2009replicated,srivastava2013modeling}. For instance, Nguyen et al\cite{nguyen2015improving} propose an extension of LDA that makes use of a word embedding model trained on a very large corpus of text to improve the performance of topic coherence modelling on several datasets. Future work might seek to explore such techniques, and (assuming regulatory barriers can be overcome), the potential of creating word embedding models from very large clinical text corpora by combining data with other care organisations.

“

The notion of "n-gram" is not used entirely consistently with its broader usage in the literature; usually that refers specifically to a term of a given length (e.g. 1-grams/unigrams, 2-grams/bigrams) while different length terms are mixed here. The authors might consider using the word "term", or they have referred to "concepts" which seem to be equivalent to terms.

For technical clarity, we've removed references to 'n-gram' and replaced them with 'term' where appropriate. Our usage of concept refers to medical concepts (via our putative discovery process or otherwise). This is now consistent in our amendments.

Regarding the limitation that no context was provided to the annotators; would it make sense to provide a concordance of some of the instances of the terms to the annotators in future efforts? We agree this would be a useful method to assist in the decision making process for manual curation, and have adjusted the text:

“

The results of our IAA exercise between two experienced psychiatrists suggested a moderate level of agreement in categorising the newly identified constructs. Given that this annotation exercise did not provide any context beyond the n-gram, and that the nature of SMI symptom observation is somewhat subjective, perhaps it is to be expected that agreement was not higher. As suggested during peer review, providing a concordance of some of the instances of each n-gram, along with expert panel discussion and engagement with international collaborative efforts in SMI research may prove valuable in seeking more formal definitions of the identified constructs.

”

Also, the IAA is tied to the 9 categories defined on page 7; where do these categories come from? Are they related to standard or validated frameworks for symptoms in psychiatric assessment? If not, why were those categories chosen?

The categories were derived from the Shorter Oxford Textbook of Psychiatry (chapter 3, page 44), and the experience of the teams Clinical Psychiatrists. We have updated the text as follows:

“

The second, more important stage was composed of independent annotation of the curated concept list by two psychiatrists, to identify likely synonyms and new symptomatology based on their clinical experience. Each concept was assigned to one of the below 8 'substantive' categories, or a 9th 'other' category. The categories were derived from\cite{harrison_shorter_2018}, and the experience of the team Clinical Psychiatrists.

”

Did you perform any error analysis to explore the IAA further, e.g. a confusion matrix between categories? Is it possible that rather than considering these categories to be independent (the typical assumption for Cohen's Kappa) that some overlap between the categories might be expected?

No further attempts to explore the errors in IAA were made in this analysis. Given the high level of cross-sectional and longitudinal overlap between mental disorder diagnoses classified as 'SMI' and the subjectivity involved in observation, it's reasonable to think that there would be a tendency for errors to overlap in certain categories (for instance 'insight' and 'cognition'). However, we think that this is outweighed by the far more complex issue of the clinical validation of the concepts we identified (which the scope of this study did not allow for).

The data is protected by patient privacy constraints and hence cannot be made openly available (indeed the authors could only work on the data in a restricted "offline" setting). Given the nature of the data, this is understandable. OTOH, given that the analysis largely makes use of existing code plus extensions for scoring functions, it would make sense to share the methods in an open repository.

We agree this would be useful, and now provide a link to a repository in the paper:

“

Example code used in this analysis is available at: https://github.com/RichJackson/clustering_w2v

”

The authors should include a suitable reference for PANSS. There is also a substantial literature on terminology induction (e.g. 2) which would be appropriate to reference.

We think that the reviewer might have missed our original reference to the PANSS on page 3. However we mistakenly repeated the full acronym on page 10. This is now corrected. Regarding terminology induction, we have modified the following text in the introduction as follows. This now includes a reference to this article which we feel is particularly topical, given our domain:

“

First, insight must be obtained regarding real-world language usage such that universally understood medical entities, encompassing hypernymy, synonymy and hyponymy adequately

represent models of concepts. Similarly, because of the abundant use of acronyms in the medical domain, a large percentage have two or more meanings\cite{liu_study_2002}, creating word sense disambiguation problems. As such, significant efforts have arisen to supplement these types of knowledge bases with appropriate real world synonym usage extracted from EHR datasets\cite{henriksson_identifying_2013}. The problem may be considered analogous to difficulties in the recognition, classification and mapping of technical terminology variants throughout the biomedical literature, which is known to be an impediment to the construction of knowledge representation systems (see \cite{krauthammer_term_2004} for a review). Krauthammer, Michael, and Goran Nenadic. "Term identification in the biomedical literature." Journal of biomedical informatics 37.6 (2004): 512-526.

The writing in the manuscript is generally clear, although I identified a few things that could be rephrased or clarified:

The word "depiction" seems to mean "usage" or "phrase" or "expression" or similar; "depiction" is typically used in the context of art or illustration and I found it strange in a language-expression related context.

We have rephrased this language throughout

Are phenotypes and symptomatology always the same thing? Is a phenotype a set of behaviours/symptoms?

Yes - the definition of a phenotype is the set of observable characteristics of an organism resulting from its genotype and interaction with its environment. We believe that symptom/behaviour profiles can be reasonably viewed in this way (and these are commonly referred to in phenotypic terms in mental health research).

The abstract is not as clear as it could be. The final sentence of the Background paragraph should use "it is" rather than "it's" but more importantly it implies that the objective is to assess clinician preferences as opposed to actual usage. Are these the same?

You are correct to say that 'actual usage' and 'clinical preference' are different concepts here. Our work aims to capture 'preference' in clinical language constructs that do not reflect matches to industry knowledge base projects (regardless of the reason). We have made several small changes in the text to reflect this.

Also, n-grams, vector space models, concepts, vocabulary and depictions are all introduced; it is a bit confusing without having read the full paper. I wonder if it could be simplified somewhat?

This now reads:

"

By utilising a large corpus of healthcare data, we sought to make use of semantic modelling and clustering techniques to represent the relationship between the clinical vocabulary of internationally recognised SMI symptoms and the preferred language used by clinicians within a care setting. We explore how such models can be used for discovering novel vocabulary relevant to the task of phenotyping Serious Mental Illness (SMI) with only a small amount of prior knowledge.

"

As a nitpick, in the Introduction the authors refer to "predicting the diversity of vocabulary"; the work

does not address prediction of vocabulary or its diversity but rather involves analysis of that vocabulary.

This sentence now reads:

“

Given a sufficiently large corpus of documents, typically authored by hundreds of clinical staff over several years, it is often difficult to track the evolution of vocabulary used within the local EHR setting to describe potentially important clinical constructs.

”

Another nitpick in the Introduction is the use of the term "authorship"; I suppose the authors mean "writing" or "description" or "summary" or similar.

We've also addressed this.

Thanks once again for your valuable insights

Competing Interests: No competing interests were disclosed.

Referee Report 21 March 2018

doi:10.5256/f1000research.15033.r31101



Julian Hong ¹, **Jessica Tenenbaum** ²

¹ Department of Radiation Oncology, Duke University School of Medicine, Durham, NC, USA

² Department of Biostatistics and Bioinformatics, Duke University School of Medicine, Durham, NC, USA

The authors present a method to extract from a clinical corpus novel terms used to describe serious mental illness (SMI). They use vector space models to represent the relationship between words in the corpus and combine this approach with clustering techniques and manual curation to identify relevant n-grams (1, 2, or 3-word concepts). 106 concepts had no mapping to current SNOMED terms indicating that they have indeed discovered new knowledge, i.e. terms used by clinicians to describe patients that are not already included in SNOMED CT.

1. The introduction was unusually well written, if a little longer than strictly necessary, and provided excellent motivation for the work at hand. It has been shown that SNOMED coverage of mental health terms is sub-optimal and this is a clever approach to learning new relevant terms in a semi-automated manner.
2. For the rest of the paper, each individual part was well written, but I had a hard time seeing how they flowed together.
3. Figure 1 was a helpful overview, but I still found it difficult to follow how the sub-steps tied in together and in some cases why they were important. e.g.
 1. How did creation of the putative cluster of 38 terms help? I think it was to facilitate the scoring method, but I wasn't completely clear how.
 2. Why 38? Particularly when the clusters they later looked at were so much larger, not clear why that number was chosen.
4. I found the math/logic challenging to follow. (Admittedly, I am not a statistician, and was not previously familiar with CBOW.)

1. It was helpful that the authors included examples in some places, but they could have gone even further to make the approach concrete. Toy examples of u_1 and v_1 would help.
2. On first and second read, I was having a hard time with intuition for what a high-scoring cluster means. I now realize (I think?) it meant the cluster was particularly enriched for mental health terms. It might be helpful to state that- for some reason I was thinking it meant that the concepts were relatively similar/cohesive?
3. I had trouble wrapping my head around the sentence "we scored each cluster based on the number of per concept hits to derive a cluster/concept count matrix x where $x_{i,j}$ represents the count of the i th concept in the j th cluster." I think it means 38 rows, 1 for each concept and 3 columns, 1 for each cluster, and the value of the cell is the number of times that concept was encountered in some form in the cluster?
4. Equations could also be numbered for reference.
5. The authors report choosing not to perform stemming/lemmatization in order not to make assumptions about the structure of the data, but this decision is not very well explained or justified. Indeed they call it out as a potential limitation in Discussion. It would be useful/interesting to try the approach both ways and see if the results were different.
6. How were the 8 "substantive categories" chosen?
7. Why does inter-rater agreement matter in mapping the concepts to those categories? Was it only that the ability to map them to a single category makes it more likely the concept is semantically interesting and reliable?
8. The authors mention that the semantic similarity of n-grams is often measured via their cosine distance between vectors in the W matrix. Just out of curiosity, could distance in the W' matrix be used as well/instead?
9. My "partly" answer to "Are sufficient details of methods and analysis provided to allow replication by others?" reflects the fact that the authors very reasonably cannot publish the raw data, but they do address how to obtain the data through a formal application process. (Ergo the "Yes" to whether source data are available, even if not readily...) It would be helpful if code were to be made available.

Minor:

1. Page 3, paragraph 4 should be employS curated terminology
2. Page 6 line 5 should have) after "Table 1"

Is the work clearly and accurately presented and does it cite the current literature?

Partly

Is the study design appropriate and is the work technically sound?

Yes

Are sufficient details of methods and analysis provided to allow replication by others?

Partly

If applicable, is the statistical analysis and its interpretation appropriate?

I cannot comment. A qualified statistician is required.

Are all the source data underlying the results available to ensure full reproducibility?

Yes

Are the conclusions drawn adequately supported by the results?

Yes

Competing Interests: No competing interests were disclosed.

Referee Expertise: Patient stratification in mental health using EHR data, structured and free text.

We have read this submission. We believe that we have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however we have significant reservations, as outlined above.

Author Response 21 Apr 2018

Richard Jackson,

Thank you for your valuable comments. Please see our responses below

For the rest of the paper, each individual part was well written, but I had a hard time seeing how they flowed together.

Figure 1 was a helpful overview, but I still found it difficult to follow how the sub-steps tied in together and in some cases why they were important. e.g.

How did creation of the putative cluster of 38 terms help? I think it was to facilitate the scoring method, but I wasn't completely clear how.

Why 38? Particularly when the clusters they later looked at were so much larger, not clear why that number was chosen.

We think there's some misunderstanding of our methodology and apologise if it wasn't clear in the manuscript. We didn't create any clusters in this work by hand. All of the 896 195 terms generated from the corpus were assigned to one of 75 distinct clusters via the K-means algorithm. We then needed to identify which of the 75 clusters were worth looking at. The 38 concepts constitute the prior knowledge about SMI symptomatology that our clinical team fed into the scoring algorithm we describe in the manuscript. This revealed three clusters that we took forward for further analysis.

We've re-written the text concerning this, and introduced something we call 'Prior Concepts' to differentiate between the domain knowledge we use in cluster scoring and the clusters themselves. Please do let us know if you feel that this hasn't improved the manuscript clarity.

I found the math/logic challenging to follow. (Admittedly, I am not a statistician, and was not previously familiar with CBOW.)

It was helpful that the authors included examples in some places, but they could have gone even further to make the approach concrete. Toy examples of u_1 and v_1 would help.

We've added an example of the analysis pipeline to the accompanying code repository, and added the line:

“

We provide a worked example of this technique in the code repository that accompanies this paper, using publically available data.

“

On first and second read, I was having a hard time with intuition for what a high-scoring cluster means. I now realize (I think?) it meant the cluster was particularly enriched for mental health terms. It might be helpful to state that- for some reason I was thinking it meant that the concepts were relatively similar/cohesive?

Actually you are correct on both counts. Training the Word2Vec model causes n-grams with semantically similar meanings to co-locate near to each other in the vector space model. The application of the clustering algorithm groups semantically similar n-grams together. Three clusters scored highly in our relevancy scoring algorithm, signifying that they were enriched for our existing knowledge of SMI symptomatology.

I had trouble wrapping my head around the sentence "we scored each cluster based on the number of per concept hits to derive a cluster/concept count matrix x where $x_{i,j}$ represents the count of the i th concept in the j th cluster." I think it means 38 rows, 1 for each concept and 3 columns, 1 for each cluster, and the value of the cell is the number of times that concept was encountered in some form in the cluster?

This is almost correct, although j represents the total number of clusters (75). The result is a score per cluster that is plotted in figure 3. We have rewritten this section of text accordingly, as per the previous comment on this issue

Equations could also be numbered for reference.

This is now done

The authors report choosing not to perform stemming/lemmatization in order not to make assumptions about the structure of the data, but this decision is not very well explained or justified. Indeed they call it out as a potential limitation in Discussion. It would be useful/interesting to try the approach both ways and see if the results were different.

This is a potential limitation of our work, in that using un-stemmed tokens will have led to vastly more n-grams than we might have otherwise had to deal with, and that stemming might have led to the identification of additional n-grams of interest. However, we feel our decision not to make assumptions about the value of stemming in this context was appropriate for two reasons:

1. In the context of a mental health assessment, stemming may cause important information loss. For instance, the term 'insomnia' shares the same stem as 'insomniac'. However, short term 'insomnia' is a relatively common symptom amongst the general population for a large variety of conditions. 'Insomniac' in the context of mental illness, on the other hand, might imply a chronic condition.
2. Our IAA task would have been substantially more complex if we were to offer stemmed n-grams for human evaluation, rather than complete words.

How were the 8 "substantive categories" chosen?

The categories were derived from the Shorter Oxford Textbook of Psychiatry (chapter 3, page 44), and the experience of the teams Clinical Psychiatrists. We have updated the text as follows:

“

The second, more important stage was composed of independent annotation of the curated concept list by two psychiatrists, to identify likely synonyms and new symptomatology based on their clinical experience. Each concept was assigned to one of the below 8 'substantive' categories, or a 9th 'other' category. The categories were derived from \cite{harrison_shorter_2018}, and the experience of the team Clinical Psychiatrists.

Why does inter-rater agreement matter in mapping the concepts to those categories? Was it only that the ability to map them to a single category makes it more likely the concept is semantically interesting and reliable?

Yes, we felt that offering a binary choice per n-gram (i.e. potentially relevant/irrelevant) was likely to heavily bias our results in favour of high agreement. Rather, we thought that agreement on the mapping of the identified n-grams to defined groups of symptomatology would suggest a greater degree of robustness.

The authors mention that the semantic similarity of n-grams is often measured via their cosine distance between vectors in the W matrix. Just out of curiosity, could distance in the W' matrix be used as well/instead?

We don't believe this is possible, as the W matrix corresponds to weights between the input layer and the hidden layer of the neural network, where each row represents a single n-gram. The W' corresponds to the weights between the hidden layer and the output layer, which has different dimensions. I recommend [this](#) reference for a detailed description of the Word2Vec methodology.

My "partly" answer to "Are sufficient details of methods and analysis provided to allow replication by others?" reflects the fact that the authors very reasonably cannot publish the raw data, but they do address how to obtain the data through a formal application process. (Ergo the "Yes" to whether source data are available, even if not readily...) It would be helpful if code were to be made available.

We agree this would be useful, and now provide a link to a repository in the paper:

“

Example code used in this analysis is available at: https://github.com/RichJackson/clustering_w2v

”

Minor:

Page 3, paragraph 4 should be employS curated terminology

Page 6 line 5 should have) after "Table 1

This is now corrected.

Thanks once again for your time and efforts with our paper.

Best wishes

Richard

Competing Interests: No competing interests were disclosed.

The benefits of publishing with F1000Research:

- Your article is published within days, with no editorial bias
- You can publish traditional articles, null/negative results, case reports, data notes and more
- The peer review process is transparent and collaborative
- Your article is indexed in PubMed after passing peer review
- Dedicated customer support at every stage

For pre-submission enquiries, contact research@f1000.com

F1000Research