Knowledge discovery for Deep Phenotyping serious mental illness from Electronic Mental Health records [version 1; referees: awaiting peer review]

Richard Jackson¹, Rashmi Patel¹,², Sumithra Velupillai¹,³, George Gkotsis¹, David Hoyle⁴, Robert Stewart¹,²

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

Abstract

Background: Deep Phenotyping is the precise and comprehensive analysis of phenotypic features, where 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 that 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’s difficult to identify the language that clinicians favour to express concepts.

Methods: Vector space models of language seek to represent the relationship between words in a corpus in terms of cosine distance between a series of vectors. When utilising a large corpus of healthcare data and combined with appropriate clustering techniques and manual curation, we explore how such models can be used for discovering vocabulary relevant to the task of phenotyping Serious Mental Illness (SMI) with only a small amount of prior knowledge.

Results: 20 403 n-grams 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 (42%) concepts were found to be depictions of putative clinical significance. Of these, 53 (10%) were identified having novel synonymy with existing SNOMED CT concepts. 106 (19%) had no mapping to SNOMED CT.

Conclusions: We demonstrate a scalable approach to discovering new depictions 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 depictions.
Introduction
The dramatic decrease in 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 discovered, subsequently leading to the development of around 2,000 clinical genetic tests. 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’). 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.

In order to standardise knowledge representation of clinically relevant entities and the relationships between them, phenotyping from EHRs often employ 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 2020. However, the task of representing the sprawling and ever-changing landscape of healthcare in such a fashion has proven complex. 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 assessment, such diagnostic entities continue to have low clinical utility. 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 units. In terms of clinical practice, the presenting phenotype 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 Scale or Brief Negative Symptom Scale. 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 extensive 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 entities, encompassing hypernymy, synonymy and hyponymy adequately represent models of concepts. Similarly, the abundant use of acronyms in the medical domain has caused a large percentage of acronyms to have two or more meanings, 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.

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 successes. 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 authored by hundreds of clinical staff over several years, it is often difficult to predict the diversity 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 resource, based upon the contents of such frameworks. During the course of manually
reviewing clinical text, we made two subjective observations regarding the authorship of 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 prodromes\(^2\). 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 patterns of 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 depictions 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.

---

**Figure 1. Overview of project workflow.**

---

**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)\(^3\) infrastructure programme has been operating to offer a pseudonymised and de-identified 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 suite\(^24\). 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\(^25\) suite and the sampling method proposed by Mikolov \textit{et al.}\(^26\). 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 n-grams. 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 Harris\textsuperscript{27}, 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.\textsuperscript{28} has risen to prominence, owing to its ability to accurately capture semantic relationships whilst scaling to large corpora of text.\textsuperscript{29} Recently, the CBOW model has been used to identify the semantic similarities between single word entities in biomedical literature and clinical text,\textsuperscript{30} 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 conditions.\textsuperscript{32}

A full description of the CBOW architecture is discussed in 30. 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 n-grams, 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 n-grams 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 n-gram 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 n-grams, 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 n-grams (usually between 5 and 10) that move over each token in a document (potentially ignoring natural grammatical structures). For a given input n-gram, 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 n-gram 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 n-grams 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 n-grams 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 n-grams, 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 partitioning 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 framework\textsuperscript{31} 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',\textsuperscript{32} 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.
To identify a putative cluster of SMI symptomatology, we devised a simple ‘relevance’ cluster scoring approach based upon prior knowledge of common SMI symptom concepts. We selected 38 symptoms of SMI based upon their depictions in SMI frameworks and on their specificity in clinical use (Table 1). 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’. 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. We produced stems, appropriate synonyms/acronyms and/or n-grams of each concept as described in Table 1.

For each concept, we then calculated a vector of the minimum count per concept across all clusters:

$$u_i = \min_{j \in J} x_{ij}, \ i = 1, \ldots, m.$$  

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

$$v_i = \max_{j \in J} x_{ij}, \ i = 1, \ldots, m.$$  

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

$$x'_{ij} = \frac{x_{ij} - u_i}{v_j - u_i}$$

The purpose of rescaling in such a way was to prevent over-represented concepts unduly influencing the overall result (for instance, a concept 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 concepts equally, regardless of their frequency).

Finally, we summed all rescaled concept counts per cluster, and divided by the total cluster size to provide a score per cluster $z$ representing known symptomatology contents:

$$z_j = \frac{\sum_{i=1}^{m} x'_{ij}}{s_j}$$

where $s$ is a vector of the total count of n-grams 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 to identify outliers (the distribution of our cluster scores was non-normal). 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.

### Table 1. Known symptomatology concepts and vocabulary matching criteria. An underscore represents a bigram match.

<table>
<thead>
<tr>
<th>core concept</th>
<th>concept match</th>
</tr>
</thead>
<tbody>
<tr>
<td>aggression</td>
<td>aggress</td>
</tr>
<tr>
<td>agitation</td>
<td>agitat</td>
</tr>
<tr>
<td>anhedonia</td>
<td>anhedon</td>
</tr>
<tr>
<td>apathy</td>
<td>apath</td>
</tr>
<tr>
<td>affect</td>
<td>affect</td>
</tr>
<tr>
<td>catalepsy</td>
<td>catalep</td>
</tr>
<tr>
<td>catatonic</td>
<td>cataton</td>
</tr>
<tr>
<td>circumstantial</td>
<td>circumstant</td>
</tr>
<tr>
<td>concrete</td>
<td>concrete</td>
</tr>
<tr>
<td>delusional</td>
<td>delusion</td>
</tr>
<tr>
<td>derailment</td>
<td>derail</td>
</tr>
<tr>
<td>eye contact</td>
<td>eye_contact</td>
</tr>
<tr>
<td>echolalia</td>
<td>echola</td>
</tr>
<tr>
<td>echopraxia</td>
<td>echopra</td>
</tr>
<tr>
<td>elation</td>
<td>elat</td>
</tr>
<tr>
<td>euphoria</td>
<td>euphor</td>
</tr>
<tr>
<td>flight of ideas</td>
<td>foi</td>
</tr>
<tr>
<td>thought disorder</td>
<td>thought_disorder</td>
</tr>
<tr>
<td>grandiosity</td>
<td>grandios</td>
</tr>
<tr>
<td>hallucinations</td>
<td>hallucinat</td>
</tr>
<tr>
<td>hostility</td>
<td>hostil</td>
</tr>
<tr>
<td>immobility</td>
<td>immobil</td>
</tr>
<tr>
<td>insomnia</td>
<td>insomn</td>
</tr>
<tr>
<td>irritability</td>
<td>irritab</td>
</tr>
<tr>
<td>coherence</td>
<td>coheren</td>
</tr>
<tr>
<td>mannerisms</td>
<td>mannerism</td>
</tr>
<tr>
<td>mutism</td>
<td>mute</td>
</tr>
<tr>
<td>paranoia</td>
<td>paranoi</td>
</tr>
<tr>
<td>persecution</td>
<td>persecut</td>
</tr>
<tr>
<td>motivation</td>
<td>motivat</td>
</tr>
<tr>
<td>rapport</td>
<td>rapport</td>
</tr>
<tr>
<td>posturing</td>
<td>postur</td>
</tr>
<tr>
<td>rigidity</td>
<td>rigid</td>
</tr>
<tr>
<td>stereotypy</td>
<td>stereotyp</td>
</tr>
<tr>
<td>stupor</td>
<td>stupor</td>
</tr>
<tr>
<td>tangential</td>
<td>tangenti</td>
</tr>
<tr>
<td>thought block</td>
<td>thought_block</td>
</tr>
<tr>
<td>waxy</td>
<td>waxy</td>
</tr>
</tbody>
</table>
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 a simple manual data cleaning process to reduce the total number of concepts via grouping misspellings, tokenisation failures and other constructs that did not yield additional information content.

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.

**Appearance/Behaviour** Imposing 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** Imposing 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** Imposing any other thought content

**Perception** Imposing any described perception

**Cognition** Imposing anything relating to the patient’s cognitive function

**Insight** Imposing 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 statistic3.

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 that the stems of each n-gram appeared in. To protect patient anonymity, we discarded any concept that appeared in ten or less 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 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 n-grams.

**Term curation, frequency analysis and SNOMED CT mapping**

The combined 20 403 n-grams were taken forward for curation as described above. The first phase of curation reduced the list to 519 putative concepts. The majority of eliminated concepts were morphological variations, misspellings and tokenisation anomalies of singular concepts. For instance, 84 variations were detected for the stem ‘irrit*’ (as in ‘irritable’). Other n-grams were removed because insufficient context was available for a reasonable clinical interpretation, such as ‘fundamentally unchanged’, ‘amusing’ and ‘formally tested’. Finally, n-grams 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 terms (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 n-grams per category where agreement was reached.

Supplementary File 1 is a CSV table of all 557 n-grams. In addition to the n-gram itself, the table contains the following information: the counts of the unique patient records of our 20 472 patient SMI cohort in which the n-gram was detected; the counts of the unique documents of the 11 745 094 clinical
**Figure 2.** Selecting K for K-means++: Elbow analysis to suggest optimal value for K based on minimum centroid distance. The elbow of the curve indicates the point at which clearly defined clusters no longer emerge with higher values of K.

**Figure 3.** Scoring of clusters according to known symptomatology content. Each dot represents a unique cluster. The unique cluster ID’s of the most relevant clusters according to our scoring algorithm are labelled.
document corpus wherein the n-gram was detected; the category assigned to the n-gram by each of our clinical annotators, and the SNOMED CT ID code for each n-gram, 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.

Table 2. Counts of n-grams where annotators independently agreed by category.

<table>
<thead>
<tr>
<th>Category</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Affect/Mood</td>
<td>6</td>
</tr>
<tr>
<td>Appearance/Behaviour</td>
<td>78</td>
</tr>
<tr>
<td>Cognition</td>
<td>6</td>
</tr>
<tr>
<td>Insight</td>
<td>2</td>
</tr>
<tr>
<td>Mood/Anxiety/Affect</td>
<td>26</td>
</tr>
<tr>
<td>Other</td>
<td>102</td>
</tr>
<tr>
<td>Perception</td>
<td>9</td>
</tr>
<tr>
<td>Personality</td>
<td>23</td>
</tr>
<tr>
<td>Speech</td>
<td>63</td>
</tr>
<tr>
<td>Thought</td>
<td>22</td>
</tr>
</tbody>
</table>

Regarding SNOMED CT mapping, it was possible to suggest direct mappings for 177 concepts and to suggest synonymy or partial mapping 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 n-grams 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 n-grams 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 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], ‘assaultivative’ [sic], ‘truculent’, ‘stared intimidatingly’ 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 Positive and Negative Symptom Scale (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 depictions 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 disease34, psychological assessment instruments35 and histopathology findings36. 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 over 300 000 entities in its current incarnation, its size already presents problems in biomedical applications37.

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 the egg’ situation, a widely discussed limitation of the reuse of EHR data38,39. 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 n-gram, and that the nature of SMI symptom observation is somewhat subjective, perhaps it is to be expected that agreement was not higher. Nevertheless, future work might seek more formal definitions of such constructs via expert panel discussion and engagement with international collaborative efforts in SMI research.

Our method for vocabulary building produced nearly 1 million n-grams. 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 n-grams taken forward for curation, we employed a word embedding model with
a clustering algorithm. With our cluster scoring methodology, we were able to successfully produce meaningful clusters of n-grams 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 n-grams 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 non-symptomatological assessment contexts.

After clustering, a two stage manual curation of more than 20 000 n-grams 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.

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 South London and Maudsley 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. 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

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 non-substantive ‘other’ classification restrictions lifted.

Click here to access the data.
References


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