Validating a strategy for psychosocial phenotyping using a large corpus of clinical text

Adi V Gundlapalli,[1,2] Andrew Redd,[1,2] Marjorie Carter,[1,2] Guy Divita,[1,2] Shuying Shen,[1,2] Miland Palmer,[1] Matthew H Samore[1,2]

▸ Additional material is published online only. To view please visit the journal online (http://dx.doi.org/10.1136/amiajnl-2013-001946).

[1]IDEAS Center, VA Salt Lake City Health Care System, Salt Lake City, Utah, USA
[2]Departments of Internal Medicine and Biomedical Informatics, University of Utah School of Medicine, Salt Lake City, Utah, USA

Correspondence to
Dr Adi V Gundlapalli, Mail Code 182, VA Salt Lake City Health Care System, 500 Foothill Drive, Salt Lake City, UT 84148, USA; adi.gundlapalli@hsc.utah.edu

Received 24 April 2013
Revised 2 October 2013
Accepted 8 October 2013
Published Online First
29 October 2013

To cite: Gundlapalli AV, Redd A, Carter M, *et al*. *J Am Med Inform Assoc* 2013;**20**:e355–e364.

ABSTRACT

Objective To develop algorithms to improve efficiency of patient phenotyping using natural language processing (NLP) on text data. Of a large number of note titles available in our database, we sought to determine those with highest yield and precision for psychosocial concepts.

Materials and methods From a database of over 1 billion documents from US Department of Veterans Affairs medical facilities, a random sample of 1500 documents from each of 218 enterprise note titles were chosen. Psychosocial concepts were extracted using a UIMA-AS-based NLP pipeline (v3NLP), using a lexicon of relevant concepts with negation and template format annotators. Human reviewers evaluated a subset of documents for false positives and sensitivity. High-yield documents were identified by hit rate and precision. Reasons for false positivity were characterized.

Results A total of 58 707 psychosocial concepts were identified from 316 355 documents for an overall hit rate of 0.2 concepts per document (median 0.1, range 1.6–0). Of 6031 concepts reviewed from a high-yield set of note titles, the overall precision for all concept categories was 80%, with variability among note titles and concept categories. Reasons for false positivity included templating, negation, context, and alternate meaning of words. The sensitivity of the NLP system was noted to be 49% (95% CI 43% to 55%).

Conclusions Phenotyping using NLP need not involve the entire document corpus. Our methods offer a generalizable strategy for scaling NLP pipelines to large free text corpora with complex linguistic annotations in attempts to identify patients of a certain phenotype.

INTRODUCTION

The psychosocial phenotype of a patient generally refers to their psychological and social characteristics. These include evidence of mental health disorders (such as depression), personality traits (such as violence and anger), life stressors (such as unemployment and incarceration), and social isolation (such as living alone, being divorced, etc). An understanding of this phenotype has been shown to be important in identifying risk factors for homelessness,[1] forecasting outcomes of mental illness,[2] and predicting risk of heart disease[3] and readmission rates.[4] Apart from this, associations of these phenotypes with genotypes are waiting to be discovered.[5] For these reasons, it is important for researchers to improve the psychosocial phenotyping of patients using all available data in the electronic medical record.

While structured data such as demographic data and International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM) codes are ‘low hanging fruit’ with respect to availability and ease of access, there are limitations to using these data in deciphering the true phenotype of the patient.[5] Not all symptoms or diagnoses are coded and they therefore may offer only a limited glimpse into the characteristics of the patient depending on the phenotype sought. The free text of the clinical note is a rich resource where providers record the history as told to them by their patients and their interpretation of the history. A review of these notes would provide the best snapshot of the patient's symptoms and history; however, the high cost of manual review has limited the use of this resource for large-scale projects.

The advent of natural language processing (NLP) has allowed healthcare systems to tap into the free text of electronic health records for clinical care, quality improvement, public health surveillance, and research.[6–9] Identifying phenotypic characteristics of patients has been an important next step in using NLP. The free text has been used to supplement structured data in identifying specific phenotypes of patients for clinical operations and recruitment of patients for clinical trials and research.[10–13]

Working with the free text of medical records poses challenges in the form of the availability of NLP tools and the expertise to use them.[9] The availability of ‘big data’ poses another challenge in terms of computing resources, timeliness, and completeness of trying to process millions or billions of records. While all free text notes may be available in large healthcare systems and academic medical centers, depending on the question being asked, it may not be necessary to process each and every document. There are limited studies on strategies designed to increase the efficiency of processing document corpora by identifying documents with the highest yield.[14]

We sought to address the challenges of phenotyping a large population based on information extracted from text notes. We investigated a strategy that increases the efficiency and usefulness of an NLP pipeline by targeting those notes that have the highest yield of the concepts of interest. This project addresses the following two key research questions. (1) Of all the different note titles used in a large healthcare system such as the US Department of Veterans Affairs (VA) medical facilities, which titles have the highest yield in terms of returning the most extracted psychosocial concepts per document processed using an NLP pipeline (‘hit rate’ for a given set of psychosocial concepts)? (2) Of those note titles with the highest hit rate for concepts, which are high yield in terms of precision (or positive predictive value) of the various concepts identified by NLP? The goal is to develop evidence-based algorithms to improve efficiency and high-throughput phenotyping using NLP on the free text of clinical documents that can supplement the phenotyping of patients from administrative data.

METHODS

Setting and note corpus

The VA is one of the largest providers of healthcare in the USA. Approximately 15 million Veterans of the US military are enrolled in the VA system, and, each year, nearly 3 million Veterans are provided with care in nearly 150 facilities spread across the country. Research using VA data is conducted using various national databases available within the Veterans Informatics, Information, and Computing Infrastructure (VINCI), which provides centralized access to VA data resources and creates an integrated suite of databases in a computing environment to provide secure access to comprehensive VA healthcare data, including the full spectrum of text notes and administrative data.[15] Of the nearly 1.6 billion free text documents available to researchers, a corpus analysis revealed a set of 220 enterprise standard note titles used throughout the VA system ranging from 'acupuncture' to 'wound care'. Approximately 650 000 documents are added to the warehouse each day from VA clinical operations.

An NLP pipeline

An NLP tool to extract psychosocial concepts was developed using annotators from v3NLP, a UIMA-AS-based framework developed by VINCI. v3NLP is closest in terms of architecture and underlying techniques to the clinical Text Analysis and Knowledge Extraction System (cTAKES) and uses several morpho/syntactic components, including its part of speech tagger and phrase chunker.[16] Savova *et al*[16] benchmarked cTAKES' components and overall performance on five external tasks, with reports of precision of 0.89 and recall of 0.77. v3NLP inherits its phrase-to-concept mapping techniques from MetaMap, using MetaMap directly to retrieve Unified Medical Language System (UMLS) concepts and using an evolved lexical lookup technique (from MetaMap's java implementation: MMTx) when looking for local terminology. MetaMap has been benchmarked in several clinical and biomedical literature domains including emergency department reports (precision 0.56, recall 0.72), radiology reports (precision 0.63, recall 0.80, F score 0.70), history and physical notes (precision 0.65, recall 0.92, F score 0.76), and for discharge summaries (precision 0.82, recall 0.88, F score 0.85).[17–19] v3NLP inherited its sectionizer and implementation of negation[20] from Hitex. Hitex has been benchmarked on external tasks such as the i2b2 smoking cessation challenge (overall precision reported as 0.82).[21] Annotator modules have also been used to identify slot:value and check box type of statements in the free text.

Development of lexicon for psychosocial concepts related to homelessness and NLP pipeline

We first developed a set of lexical terms for all psychosocial concepts related to homelessness based on the published literature, review of medical records of patients experiencing homelessness, and domain expertise. The 300+ individual terms in the lexicon were mapped to UMLS using a concept extraction tool similar to MetaMap (in-house tool called Sophia) using the UMLS 2011AA Level 0+Level 9 terminologies version of the Metathesaurus. We found that approximately 28% of the lexicon terms mapped to UMLS concepts; these same terms also mapped to SNOMED-CT IDs. Human review was performed with the aid of National Library of Medicine's (NLM) Semantic Navigator tool, which visually displays the semantic quality of query terms. The tool allows one to review and traverse the hierarchical relationships between terms within the UMLS. Synonymy was considered along with other lexical variation to the extent that the synonymy was generated via NLM's Lexical Variant Generation tool and to the extent that NLM's Semantic Navigator Tool utilizes the synonymy inherent within the construction of UMLS concepts.

The individual psychosocial concepts were binned into 16 concept categories. Some of these, such as alcohol abuse, substance abuse, and psychiatric disorders, are based on existing hierarchical and ontology-based semantic categories from standardized terminologies including SNOMED-CT. Several categories were developed on the basis of clinical experience (eg, direct evidence of homelessness, social stressors, incarceration-related); these are patient-oriented and informative (self-explanatory) when reporting the psychosocial phenotype of patients and have, in general, poor mapping to standardized terminologies. The use of a local concept annotator module was determined to be more efficacious than using a generic concept extraction annotator such as MetaMap to identify and label clinical statements with UMLS concept attributes[22] or using SNOMED-CT. This also resulted in greater processing speed when scaling to large corpora of complex clinical text.

Selecting a random corpus for high-yield document analysis and determining 'hit rate'

We define a hit as a concept extracted from the document through NLP. Hit rates from NLP of free text documents are considered in the first step without regard to whether those hit rates represent true positives or false positives at the concept level.

To address the first research question (which of the 220 enterprise standard note titles have the best yield in terms of hit rate of finding concepts), the metrics calculated through NLP of the document corpus were: (1) raw mean number of concepts per all documents; (2) percentage of documents containing a concept hit; (3) mean number of concepts per document, given that at least one concept was found in a document (containing document). Although the metrics are quite similar, they give three different perspectives on the data. Metric 1 was used as the primary outcome of interest; however, the more conservative estimates from metric 2 were used to calculate the sample sizes. Sample sizes were determined to give a 95% CI width of less than 5% for any proportion with 80% power on the proportion in metric 2. This yielded a sample size of 1500 documents on each of the 220 enterprise note titles.

Documents were chosen as a simple random sample within each enterprise note title from the corpus of available documents in VINCI. Time and patient dependence were not considered, as documents were chosen from a large corpus of 1.6 billion documents from at least 9 million Veterans over a 20-year time period. The probability that two or more documents were from the same individual Veteran was considered to be small.

Determining precision by human review of extracted concepts

To address the second question (which of the note titles with the highest yield for concepts (hit rate) provided the best resource in terms of precision for the concepts identified by the

NLP pipeline), we performed a manual review of a subset of the hits. The criteria for choosing documents and concepts for review were based on: (1) highest raw hit rate; (2) highest percentage of documents containing concepts; (3) rate of concepts per containing documents to account for documents containing clusters of various concepts.

We reviewed the top 25 document titles plus another 10 note titles that were determined to be high yield by overlapping definitions of hit rates described above. We reviewed a subset of documents to estimate the false positive rate, which was then used to determine the precision of the NLP pipeline with respect to the various concept categories. On the basis of the need to obtain estimated CIs that are less than 15% total width and assuming a working value of 80% for precision (based on the preliminary evaluation of the NLP pipeline), we determined that we would need to review at least 185 extracted concepts per high yield title.

Documents were randomly sampled from each title to obtain at least 185 concepts per note title for human review. Documents containing concepts were displayed using an annotation tool developed for ease of annotating false positives and the reasons for the false positivity.[23] The file format for the documents was based on Knowtator, a widely used annotation tool.[24] Two experienced reviewers performed the false positive error analysis, with one checking the work of the other. Discrepancies between the reviewers were adjudicated by discussion between the two reviewers. The metrics reported were precision rate per note title, the reasons for the false positivity, and a ranking of the note titles by precision of concept categories to determine the best documents for psychosocial phenotyping using an NLP pipeline.

Sensitivity (or recall) analyses by human review of documents

Two trained human reviewers manually evaluated a random set of documents drawn from the top 35 high-yield (and other) note titles as described above to annotate psychosocial concepts based on the lexicon. Documents were selected at random from the entire set of titles without considering note title. This was to estimate the marginal sensitivity over the high-yield documents. Marginal performance was chosen to be estimated over title-specific performances because of the prohibitively large sample sizes that would have been required to estimate title-specific sensitivity. The performance of v3NLP on the set of documents was compared with the human review (considered to be the reference standard for this task). We report the traditional sensitivity in terms of annotated concepts. We also report how v3NLP fared on those documents that were found to have no concepts based on human review.

Data analyses were performed using the open source R statistical package.[25] This study was reviewed and approved by the Research and Development Committee of the VA Salt Lake City Health Care System and the Institutional Review Board of the University of Utah.

RESULTS

Relative proportion of note titles in the VA database

As shown in figure 1, the 220 note titles varied widely in their relative proportion in the entire document corpus. Nursing notes, addendum, primary care, mental health, psychiatry, and social work were among the most prevalent. A large proportion of notes that did not contain clinical information were labeled 'null-undefined-empty'. As an example, history and physical notes (H&P) and homeless program notes were relatively underrepresented in the document corpus.

Hit rate of concepts on random corpus of 316 455 documents representing all note titles

The random corpus comprised 316 455 documents from 218 enterprise standard note titles (as opposed to an expected set of 220×1500=330 000), as 34 note titles had between 1490 and 1499 documents and nine sparingly used note titles had between 1 and 807 documents. The 'null-undefined-empty' note title, which represents 6.3% of the entire database corpus, was not sampled, and another note title was not sampled because of a technical issue (0.3% of the entire database corpus).

The NLP pipeline returned a total of 58 707 concepts from this large corpus in 16 concept categories for an overall hit rate of 0.2 concepts per document (median 0.1, range 1.6–0). The concept categories used for this analysis were granular in nature and provided the basis for psychosocial phenotyping based on

Figure 1 Relative proportions of 220 US Department of Veterans Affairs (VA) enterprise note titles in document corpus.

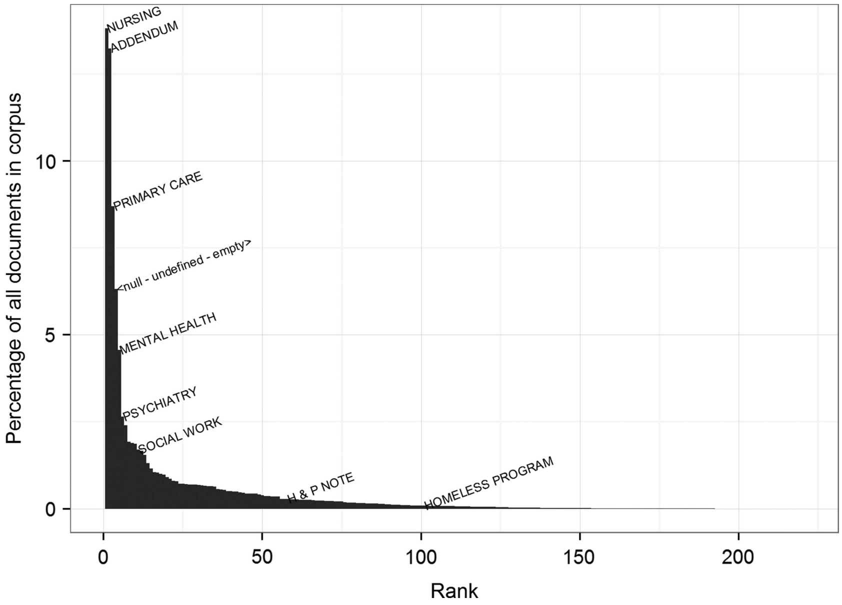


Table 1 Concept categories and individual concept terms developed as part of a lexicon on psychosocial concepts related to homelessness and numbers of concepts extracted from a large document corpus of 316 455 documents representing 218 note titles from the US Department of Veterans Affairs (VA) electronic medical record database

Concept category	Examples of concepts	Raw number of concepts extracted from 316 355-document corpus (% of 58 707 total concepts)
Homelessness related		
Direct evidence of homelessness	Sleeping in park, living on the street, living in the shelter, Domiciliary Care for Homeless Veterans, VA Supportive Housing (VASH), Healthcare for Homeless Veterans	4860 (8.3)
Doubling up (at risk of homelessness)	Doubled up, couch surfing, crashing in friend's house	69 (0.1)
Homelessness-related needs	Needs socks, needs sleeping bag	6 (0)
Mental health		
Psychiatric disorders	Severe and persistent mental illness, schizophrenia, psychosis, depression, bipolar disorder, antisocial personality disorder	14 831 (25.3)
Suicide related	High risk for suicide	102 (0.2)
Behavioral health related		
Alcohol related	Alcohol or ethanol abuse/dependence, binge drinking, Alcoholics Anonymous groups	8093 (13.8)
Substance abuse related	Substance abuse treatment program, substance abuse groups, drug rehab, detoxification	12 552 (21.4)
Alcohol and substance abuse related	Alcohol/drug abuse	3932 (6.7)
Social stressors	Limited social support, unemployed, divorced, separated	6210 (10.6)
Legal and incarceration related	On parole, on probation, history of incarceration, legal problems, Healthcare for Re-Entry Veterans	1225 (2.1)
Pathological gambling	Pathological gambling	18 (0)
Utilization and non-compliance	Non-adherence, non-compliant, missed appointments, emergency medical services, EMS	577 (1)
Trauma related		
Military sexual trauma	Military sexual trauma, MST	1339 (2.3)
Other sexual trauma	Sexual abuse, sexual trauma	2990 (5.1)
Other risks		
Ethnicity	Native American ethnicity	305 (0.5)
Comorbidities: HIV and hepatitis C	HIV, AIDS, hepatitis C	1598 (2.7)

the presence of individual concept terms (table 1). The concept categories comprising the greatest number of hits were mental health (25.3% of total), followed by substance abuse related (21.4%), alcohol related (13.8%), social stressors (10.6%), and direct evidence of homelessness (8.3%). The full set of enterprise standard note titles available in the VA database is presented in online supplementary table S1 with their psychosocial concept hit rates.

High-yield note titles based on hit rate

With regard to hit rate for psychosocial concepts, there was large variation among the 218 note titles, with 'blood banking transfusion' and 'immunization note' yielding a hit rate of nearly 0 and the 'homeless program' note yielding a hit rate of 1.6 (figure 2). The average number of concepts per note title was 276 (median 161, range 3–2396).

In terms of the number of concepts by note title (hit rate), the top 25 titles are shown in table 2. Notes associated with services provided to homeless Veterans were among the high-yield set for hit rate (homeless program note, addiction severity index note, addiction psychiatry, substance abuse treatment program). The notes with the highest hit rate for psychosocial concepts were greatly under-represented in the main corpus, with only four note titles (substance abuse treatment program, psychiatry, mental health, and social work) crossing the 1% mark. The top 25 note titles account for just under 12% of all documents in the corpus, yet contain 44.4% of all the concepts extracted by the NLP pipeline (26 087 of 58 707). Figure 3 shows the prevalence of concepts by category among the top 25 note titles compared with the total concepts extracted. Other than 'homelessness-related needs' and 'ethnicity', other concept categories are well represented in the high hit rate set.

Precision of NLP-extracted concepts as determined by human review

The sample set for human review consisted of 3526 documents from 35 note titles with a total of 6031 concepts in all 16 concept categories. Evaluation of this sample of concepts by two trained reviewers (with an inter-rater agreement of 95%) yielded an overall precision of 80% with 1223 false positive concepts. There was a large variation in the precision among the different concept categories and note titles. Table 3 shows the precision of various concept categories by grouped note titles. For example, concepts related to alcohol and substance abuse (typical terms listed in table 1) were found in 33 of the 35 note titles except for 'donor note' and 'military sexual trauma' note. Precision ranged from a low of 33% in 'medical specialty' notes (95% CI 1% to 43%) to 83% in 'mental health' notes (95% CI 37% to 97%). Direct evidence of homelessness, mental health, and social stressor concepts were found in nearly all 35 note titles reviewed and had an overall high precision of 83–100%, with the exception of primary care notes. Social stressors were found in all note titles, and the extraction performed well, with the precision being greater than 80% (95%

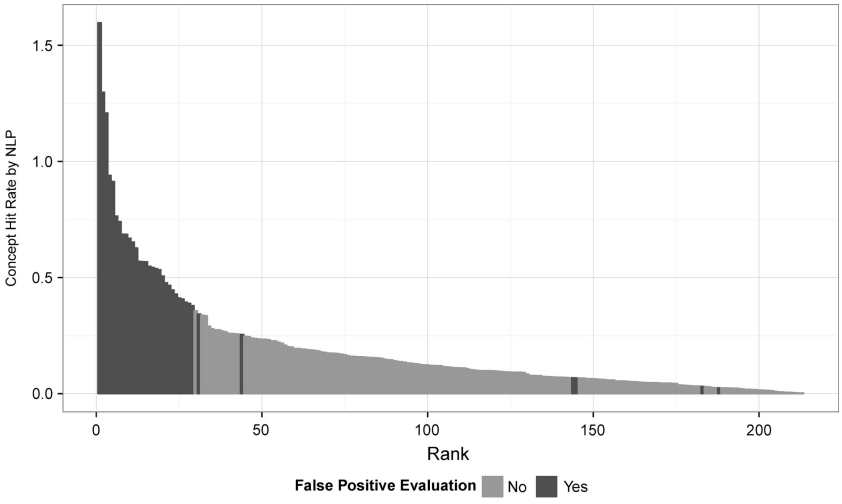


Figure 2 Histogram of rank of the concept hit rate of psychosocial concepts extracted by the natural language processing (NLP) pipeline per note title of 218 note titles from the US Department of Veterans Affairs (VA) database. Concepts from the dark shaded note titles were evaluated by human reviewers for the false positive analysis as described in the methods.

CI 62% to 97%) in most notes. Concepts related to utilization of and non-compliance with care were among the poorest performers, with an overall precision of 59% (95% CI 13% to 93%), with several note titles exhibiting 100% false positivity. With the inherent limitations of comparing our work with prior trials in the field with regard to concepts of interest, our overall precision of 80% (95% CI 42% to 93%) compares favorably with those reported from emergency department reports (precision 0.72), radiology reports (0.63), history and physical notes (0.65), discharge summaries (0.82), and smoking status identification (0.82).[18 19 21] The full table of the precision of concept categories by individual note titles (for the 35 note titles reviewed) is presented as online supplementary table S2A. Online supplementary table S2B provides the upper and lower bounds for the precision by concept category and by document title.

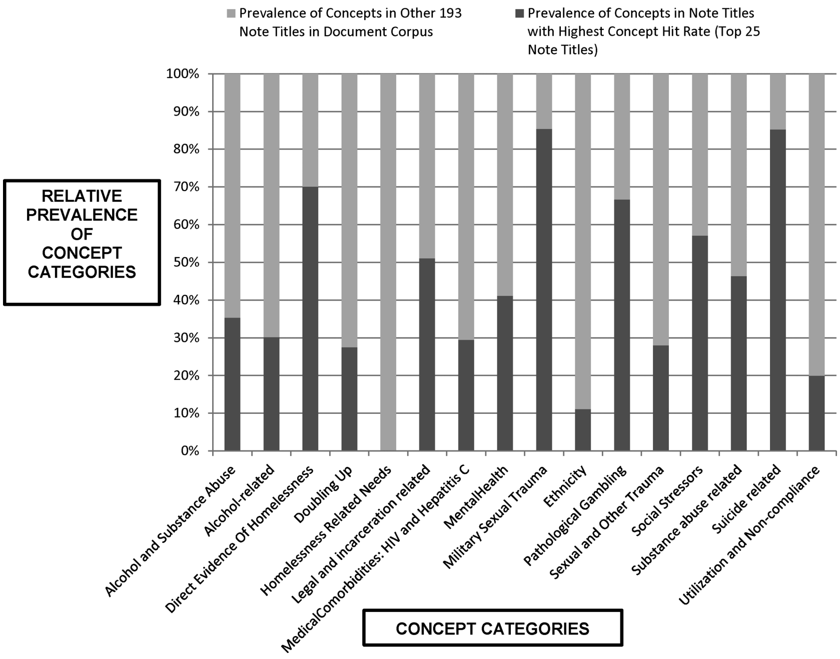
Reasons for false positivity

The reasons for false positivity were binned into five categories by human review: (1) context of the word or alternate meaning

Table 2 High-yield note titles for psychosocial concepts related to homelessness based on hits using natural language processing of free text of documents from the US Department of Veterans Affairs (VA) database

Top 25 note titles based on concept hit rate	% of all documents in VA database	Total documents processed (N)	Raw number of concepts	Hit rate of concepts per N	Number of documents containing at least one concept (% of N)	Hit rate of concepts per containing documents
Homeless program	0.09	1500	2396	1.6	984 (66)	2.4
Addiction severity index note	0.02	1500	1947	1.3	1126 (75)	1.7
Military sexual trauma note	0.01	1500	1813	1.2	1174 (78)	1.5
Addiction psychiatry	0.23	1500	1410	0.9	734 (49)	1.9
Substance abuse treatment program	1.00	1500	1371	0.9	695 (46)	2.0
Psychiatry	2.64	1500	1147	0.8	616 (41)	1.9
Outreach note	0.07	1500	1111	0.7	568 (38)	2.0
Group counseling note	0.25	1500	1030	0.7	541 (36)	1.9
Crisis	0.01	1500	1029	0.7	533 (36)	1.9
Hepatology	0.04	1500	1004	0.7	604 (40)	1.7
Mental health	4.56	1500	979	0.7	573 (38)	1.7
Suicide	0.06	1500	941	0.6	599 (40)	1.6
Counseling	0.01	1499	854	0.6	481 (32)	1.8
Vocational rehabilitation	0.12	1500	853	0.6	469 (31)	1.8
Day hospitalization	0.04	1500	852	0.6	501 (33)	1.7
Certificate	0.01	1500	823	0.5	414 (28)	2.0
Domiciliary	0.36	1500	815	0.5	398 (27)	2.0
History and physical note	0.28	1500	808	0.5	455 (30)	1.8
Geriatric psychiatry	0.01	1494	798	0.5	514 (34)	1.6
Social work	1.70	1500	760	0.5	424 (28)	1.8
Traumatic brain injury note	0.01	1500	716	0.5	482 (32)	1.5
Annual evaluation note	0.02	1499	699	0.5	531 (35)	1.3
Treatment plan	0.36	1500	670	0.4	360 (24)	1.9
Conference note	0.01	1499	643	0.4	381 (25)	1.7
Neuropsychology	0.02	1500	618	0.4	402 (27)	1.5

Figure 3 Relative prevalence of psychosocial concepts extracted by natural language processing in note titles with highest concept hit rate (top 25) compared with the other 193 note titles in the document corpus.



of word or abbreviation; (2) experiencer other than Veteran; (3) formatting of text; (4) negation; (5) templating or section heading. As shown in figure 4, the note titles varied with respect to the reasons for false positivity of concepts extracted from those note titles. Notes such as 'annual evaluation note', 'donor note', 'military sexual trauma note', and 'visual rehabilitation' were found to contain mostly templated sections with question/answer format used for screening for a particular condition. These templated sections contributed to the high false positivity rates in these note titles. Alternate meanings of the word and abbreviation caused a change in contextual meaning of the concepts; this resulted in false positivity in several note titles. Similarly, mislabeling of negated concepts was a prominent reason in several note titles. In a number of situations, concepts were noted to be false positive because of the concept referring to someone other than the Veteran. The formatting of the text in some notes leading to words being split by a carriage return resulted in several false positives. Examples of phrases that resulted in false positives are presented in figure 5. Detailed tables of reasons for false positivity for all 35 note titles and by concept category are presented as online supplementary tables S3 and S4.

Sensitivity analyses based on human review of a set of documents

From a document set of 450, of the 410 documents that contained at least one concept, v3NLP detected 131 of the 266 concepts identified by human review, resulting in a sensitivity of 49% (95% CI 43% to 55%). Of the 40 documents that contained no concepts by human review, v3NLP identified 35 of these correctly.

DISCUSSION

The increase in the availability of NLP and the desire to unlock rich clinical information from the free text of electronic medical notes highlight the essential need for developing efficiencies in processing large corpora of free text. It would be ideal to process every single note available and extract all possible concepts and phenotypes for the patient from text data using one broad-based NLP pipeline (and this has been discussed in various settings). However, this would require immense resources and would probably cause bottlenecks in processing and storage of output. It is imperative that we develop efficient strategies to improve phenotyping using NLP.

We report the first large-scale psychosocial phenotyping of patients using text data. Our work demonstrates that, depending on the question being asked, it may not be necessary to process all documents using NLP. This is likely to be important for clinical research and operations-related applications of these methods.

We validate a strategy whereby high-yield document types (note titles) are identified for hit rate of psychosocial concepts and for precision at the concept level. While we acknowledge the limitations of the technique and an overall precision of 80% that matches prior work in the field, we feel these results should be of interest and use to other researchers and operational partners in the VA who have an interest in determining psychosocial phenotypes of patients. Furthermore, the strategy of scaling to large corpora is probably generalizable to other large healthcare systems and academic medical centers with 'big data'.[22]

Further validation of our strategy is the relative underrepresentation of the high-yield note titles in the general document corpus: 12% of the notes account for 44% of the concepts. Most categories of concepts are well represented in the high-yield note titles. This indicates that specific providers and services record these concepts in specific notes, thus providing a strong argument for ignoring notes with low hit rates. This is probably due to sub-language differences among clinical documents,[26] and a partial lexeme analysis can be inferred from the results presented in this study. The method presented here is an efficient way to identify notes with a high yield of true positives. Thus, extracting detailed psychosocial concepts with acceptable precision can be achieved by targeting specific note titles (table 3). This has positive implications for setting up preprocessing steps that limit the corpus of text needed to be processed through NLP. This also allows targeted education of providers and development of informatics solutions to help them improve their documentation so that the data capture at the back-end is improved.[27]

Table 3 Precision rates (as %) for various categories of psychosocial concepts extracted by v3NLP from 35 high-yield note titles from the US Department of Veterans Affairs (VA) database as determined by human review

Category	Number of documents reviewed	Number of concepts reviewed	Alcohol and substance abuse	Alcohol related	Direct evidence	Legal and incarceration related	Medical comorbidities	Mental health	Sexual and other trauma	Social stressors	Substance abuse related	Utilization and non-compliance
Administrative Certificate, conference note and donor note	240	439	53	58	97	75	51	94	50	97	53	–
Homeless program Compensated work therapy, homeless program and outreach note	279	560	78	92	97	97	100	93	83	95	74	23
Inpatient related Admission, day hospitalization, history and physical note, and discharge	425	745	78	83	90	77	61	90	48	96	70	100
Medical specialty Hepatology, medical toxicology note, neuropsychology, telemedicine, traumatic brain injury, vision rehabilitation	618	901	33	58	83	67	77	82	20	87	47	47
Mental health Addiction psychiatry, addiction severity index note, counseling, crisis, geriatric psychiatry, group counseling note, mental health, military sexual trauma, psychiatry, suicide, treatment plan	1164	2005	83	87	100	78	80	92	80	89	82	67
Primary care Annual evaluation note, initial evaluation note, telehealth	309	447	72	60	46	100	50	67	37	73	51	–
Social work Domiciliary, residential facility, social work substance abuse treatment program, vocal rehabilitation	491	935	69	95	92	86	100	100	67	86	95	40

The concept categories, homeless-related needs, military sexual trauma, doubling up, other risks, pathological gambling, and suicide-related, are not shown, as they were extracted from fewer than five note titles.
Shading indicates precision: pink, 76–100%; dark blue, 50–75%; light blue, <50%.
NLP, natural language processing.

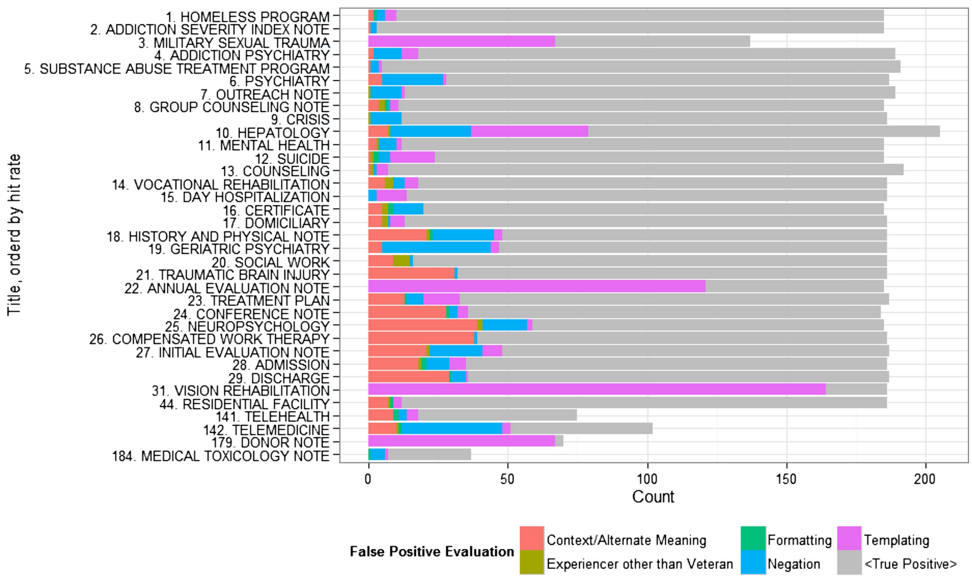
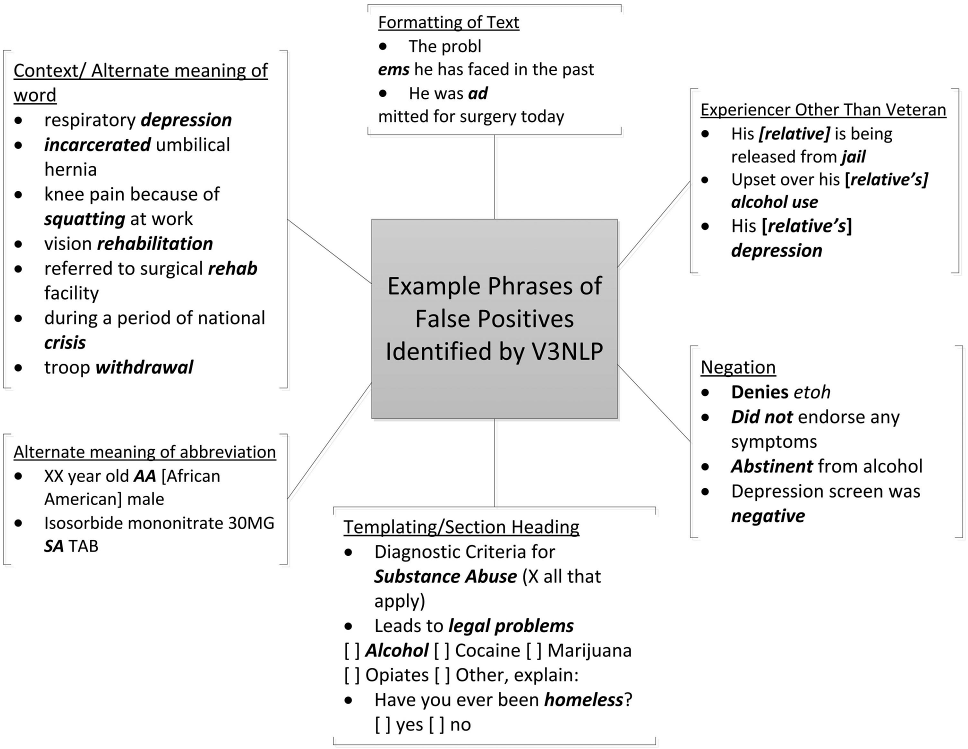


Figure 4 Bar graph representation of reasons for false positivity of 1223 concepts as determined by human review of a total of 6031 concepts extracted by natural language processing from 35 note titles; the note titles are ordered by their hit rate (highest to lowest).

Several of the reasons uncovered for false positivity have been largely described in the literature. Negation and reference to another experiencer have long been recognized as a challenge for NLP,[20 28–30] as has been disambiguation of words and abbreviations.[31 32] The templating seen in VA medical notes poses a special challenge and merits further study. Of interest, we did not see examples of anaphoric expressions as reasons for false positivity in this particular review.[33]

The usual statistical assumption is that false positivity does not correlate with baseline prevalence. We found that hit rate does somewhat crudely correlate with false positivity (figure 4), thus uncovering a situation similar to spectrum bias where the performance of a test varies when applied to different subgroups.[34] This may be due to the features noted in the VA text corpus and is likely to be seen in other large electronic medical record databases.

Some of the results are intuitive and self-evident with regard to note titles and the concepts found within them (homeless notes, addiction psychiatry, mental health, suicide, etc). In this setting, it may be argued that a time- and resource-intensive



Formatting of Text

- The probl
ems he has faced in the past
- He was *ad*
mitted for surgery today

Context/ Alternate meaning of word

- respiratory ***depression***
- ***incarcerated*** umbilical hernia
- knee pain because of ***squatting*** at work
- vision ***rehabilitation***
- referred to surgical ***rehab*** facility
- during a period of national ***crisis***
- troop ***withdrawal***

Example Phrases of False Positives Identified by V3NLP

Experiencer Other Than Veteran

- His ***[relative]*** is being released from ***jail***
- Upset over his **[*relative's*]** ***alcohol use***
- His **[*relative's*]** ***depression***

Negation

- **Denies** *etoh*
- ***Did not*** endorse any symptoms
- ***Abstinent*** from alcohol
- Depression screen was ***negative***

Alternate meaning of abbreviation

- XX year old ***AA*** [African American] male
- Isosorbide mononitrate 30MG ***SA*** TAB

Templating/Section Heading

- Diagnostic Criteria for ***Substance Abuse*** (X all that apply)
- Leads to ***legal problems***
[ ] ***Alcohol*** [ ] Cocaine [ ] Marijuana
[ ] Opiates [ ] Other, explain:
- Have you ever been ***homeless***?
[ ] yes [ ] no

Figure 5 Examples of phrases and terms leading to false positives of concepts identified by natural language processing of US Department of Veterans Affairs (VA) text documents.

NLP pipeline is unnecessary to identify these concepts. The counter argument is that the NLP pipeline offers a rich training environment for machine learning algorithms that could be used to identify specific or rare phenotypes or develop prediction models for clinical care. Furthermore, processing serial or 'lifetime' documents for patients may provide phenotyping data that are richer than what can be inferred from administrative data alone. These include severity of disease, temporality, response to treatment, epidemiologic history, and other elements not easily captured in structured format.

While this work described findings at the document level, a natural next step would be to identify phenotypes at the patient level. One concept category, or a risk factor, may be noted in a patient's chart repeatedly in different note titles by different authors. Some concept categories may also be extracted from structured data—for example, mental health disease diagnoses from ICD-9-CM CM codes. By combining NLP outputs from different notes and structured data, a more refined phenotypic inference could be made at the patient level.

The false positivity review is our starting point for the next iteration of the NLP pipeline to improve performance. Other negation annotators are now being installed in v3NLP. Separate modules are being added to address the problem of the 'experiencer' being someone other than the patient. The templating annotator modules created for this task were designed to handle the 'easy to recognize' patterns, with the expectation that a more robust machine learning technique being developed by other VA researchers would be brought on-line when available.

Developing system architectures that support the reuse of software components in a modular and interchangeable fashion, which in turn would improve efficiency and allow scaling up to medium-to-large scale corpora with complex linguistic annotations, remains a challenge for clinical and research informatics. The VA continues to develop v3NLP and other tools to address these issues.

We acknowledge several limitations of this study. Although it is known anecdotally that structured data such as ICD-9-CM codes do not capture all the diagnoses/issues that the patient is experiencing, we did not formally compare the ICD-9-CM codes associated with the visit for which the note was generated with the concepts extracted from the free text for this random document corpus. In reviewing the availability of ICD-9-CM codes for each of the concept categories, several are evident for psychiatric disorders, alcohol and substance abuse, HIV, hepatitis C, and direct evidence of homelessness. However, for many of the social stressors—legal/incarceration-related concepts, utilization/non-compliance, and sexual trauma—the coverage by ICD-9-CM codes is limited. Preliminary studies using NLP on free text of medical records have provided an adjusted prevalence of homelessness that is closer to official VA estimates and an improvement over estimates using only structured data.[35]

Any lexicon of concepts is likely to be incomplete and so it is possible that we were not able to capture all lexical and semantic variations of the psychosocial concepts found in the free text of VA documents. As can be expected of NLP surveillance techniques using keywords, terms and phrases, the sensitivity of the system is modest at best (49%). The false negativity of the NLP pipeline was not addressed in this study because of the size of the corpus and the resources that would be required to perform manual review of all notes and concepts in those notes. In this regard, the hit rate does not consider false negatives.

Further work is needed to improve performance and also to apply the NLP techniques developed to identifying patients at risk of homelessness and other issues related to psychosocial stressors such as suicide.[36] Developing a robust reference set for, and analyzing differences in number and density of concepts among, different groups of patients would be an important next step. It is possible that use of the synonymy and hierarchies in terminologies such as SNOMED-CT would be one path forward for improving the performance of the NLP technique. Similarly, a subset analysis of high- and low-yield documents might be helpful in finding areas where the performance is maximized. Applying machine learning methods to concept extraction would also be an important step forward. Acceptable performance characteristics will depend on the question being asked of the NLP system and the real world objective that is being addressed. Researchers and end-users of the system will have to strike a balance between performance, the resources and time required to further improve performance, and the clinical/operational utility of the system.

CONCLUSIONS

Our work shows that, depending on the question being posed of the NLP pipeline, high-throughput phenotyping need not necessarily involve the entire document corpus. While this is a positive result in terms of high-throughput processing of large corpora of free text documents, it is important to be aware of the limitations with regard to precision and sensitivity of identifying individual concepts from the free text. Operationally acceptable performance may be achieved by targeting specific document types from large corporate data warehouses that are known to be 'high yield' both in terms of raw hit rates and precision. With modest up-front resources and work on the domain of interest, this methodology has the potential to improve efficiency of phenotyping using NLP while scaling to large corpora. Further work is needed to maximize performance at the individual concept level and how this would relate, and add value, to identifying the psychosocial phenotype at the patient level using all available data.

Acknowledgements Resources and administrative support were provided by the VA Salt Lake City Health Care System (IDEAS Center). We thank our team members for their assistance with this project: Steve Pickard, MBA, Josh Judd, MBA, Ying Suo, MS, Tao He, MS, Yiwen Yao, MS, Warren Pettey, MPH, CPH, Rachel Peterson, MS, Tyler Forbush, Jorie Butler, PhD and Brian Sauer, PhD. We appreciate and acknowledge our colleagues at VA Informatics and Computing Infrastructure (VINCI) for their assistance with accessing VA 'big data'. The views expressed in this paper are those of the authors and do not necessarily represent the views of the Department of Veterans Affairs or the USA Government.

Contributors AVG obtained funding for the project, contributed to the conception, design, analysis and interpretation of the data as well as writing of the manuscript. AR was responsible for the design of the analysis plan, data analysis, and interpretation elements as well as revisions to the written manuscript. MC managed and contributed to the conception of the project and the interpretation of the data, as well as writing and editing the manuscript. GD developed the NLP pipeline used for concept extraction and made contributions to the writing of and revisions to the manuscript. SS developed the lexicon and made contributions to the interpretation of the data and writing/revision of the manuscript. MP was involved in the conception and design of the project and revisions to the manuscript. MHS was involved in the design of the project and critical revisions to and final approval of the submitted manuscript.

Funding Funding for this project was provided by US Department of Veterans Affairs, Veterans Health Administration, Office of Research and Development, Health Services Research and Development Project # HIR 10-002 (principal investigator: AVG).

Competing interests None.

Ethics approval University of Utah Institutional Review Board and VHA Salt Lake City R&D Committee.

Provenance and peer review Not commissioned; externally peer reviewed.

REFERENCES

1 Balshem H, Christensen V, Tuepker A, *et al*. A critical review of the literature regarding homelessness among veterans. In: US Department of Veterans Affairs, ed. *A critical review of the literature regarding homelessness among veterans*. Washington, DC: US Department of Veterans Affairs, 2011:9–43.

2 Lin A, Wood SJ, Yung AR. Measuring psychosocial outcome is good. *Curr Opin Psychiatry* 2013;26:138–43.

3 Barth J, Schneider S, von Kanel R. Lack of social support in the etiology and the prognosis of coronary heart disease: a systematic review and meta-analysis. *Psychosom Med* 2010;72:229–38.

4 Calvillo-King L, Arnold D, Eubank KJ, *et al*. Impact of social factors on risk of readmission or mortality in pneumonia and heart failure: systematic review. *J Gen Intern Med* 2013;28:269–82.

5 Denny JC, Ritchie MD, Basford MA, *et al*. PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene-disease associations. *Bioinformatics* 2010;26:1205–10.

6 Meystre SM, Savova GK, Kipper-Schuler KC, *et al*. Extracting information from textual documents in the electronic health record: a review of recent research. *Yearbook Med Inform* 2008:128–44.

7 Chapman WW. Closing the gap between NLP research and clinical practice. *Methods Inf Med* 2010;49:317–19.

8 Jha AK. The promise of electronic records: around the corner or down the road? *JAMA* 2011;306:880–1.

9 Nadkarni PM, Ohno-Machado L, Chapman WW. Natural language processing: an introduction. *J Am Med Inform Assoc* 2011;18:544–51.

10 Uzuner O. Recognizing obesity and comorbidities in sparse data. *J Am Med Inform Assoc* 2009;16:561–70.

11 Uzuner O, Goldstein I, Luo Y, *et al*. Identifying patient smoking status from medical discharge records. *J Am Med Inform Assoc* 2008;15:14–24.

12 Kho AN, Pacheco JA, Peissig PL, *et al*. Electronic medical records for genetic research: results of the eMERGE consortium. *Sci Transl Med* 2011;3:79re1.

13 Carroll RJ, Thompson WK, Eyler AE, *et al*. Portability of an algorithm to identify rheumatoid arthritis in electronic health records. *J Am Med Inform Assoc* 2012;19 (e1):e162–9.

14 South BR, Chapman WW, Delisle S, *et al*. Optimizing a syndromic surveillance text classifier for influenza-like illness: Does document source matter? *AMIA Annual Symposium proceedings/AMIA Symposium AMIA Symposium*. 2008: 692–6.

15 US Department of Veterans Affairs. VA Informatics and Computing Infrastructure (VINCI). 2013 [cited 2013. http://www.hsrd.research.va.gov/for_researchers/vinci/

16 Savova GK, Masanz JJ, Ogren PV, *et al*. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. *J Am Med Inform Assoc* 2010;17:507–13.

17 Aronson AR. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. *Proceedings/AMIA Annual Symposium AMIA Symposium*. 2001:17–21.

18 Chapman WW, Fiszman M, Dowling JN, *et al*. Identifying respiratory findings in emergency department reports for biosurveillance using MetaMap. *Stud Health Technol Inform* 2004;107(Pt 1):487–91.

19 Meystre S, Haug PJ. Evaluation of Medical Problem Extraction from Electronic Clinical Documents Using MetaMap Transfer (MMTx). *Stud Health Technol Inform* 2005;116:823–8.

20 Chapman WW, Bridewell W, Hanbury P, *et al*. A simple algorithm for identifying negated findings and diseases in discharge summaries. *J Biomed Inform* 2001;34:301–10.

21 Zeng QT, Goryachev S, Weiss S, *et al*. Extracting principal diagnosis, co-morbidity and smoking status for asthma research: evaluation of a natural language processing system. *BMC Med Inform Decis Mak* 2006;6:30.

22 Wu ST, Liu H, Li D, *et al*. Unified Medical Language System term occurrences in clinical notes: a large-scale corpus analysis. *J Am Med Inform Assoc* 2012;19(e1): e149–56.

23 South B, Shen S, Leng J, *et al*. A prototype tool set to support machine-assisted annotation. Montreal, Canada: BioNLP, 2012.

24 Ogren PV. Knowtator: a Protégé plug-in for annotated corpus construction. *Proceedings of the Human Language Technology Conference of the NAACL*. 2006; Vol 2006:273–5.

25 R Development Core Team. R: A Language and Environment for Statistical Computing. 2013.

26 Patterson O, Hurdle JF. Document clustering of clinical narratives: a systematic study of clinical sublanguages. *AMIA Annual Symposium proceedings/AMIA Symposium AMIA Symposium*. 2011;vol 2011:1099–107.

27 Hripcsak G, Albers DJ. Next-generation phenotyping of electronic health records. *J Am Med Inform Assoc* 2013;20:117–21.

28 Chapman WW, Bridewell W, Hanbury P, *et al*. Evaluation of negation phrases in narrative clinical reports. *Proceedings/AMIA Annual Symposium AMIA Symposium*. 2001:105–9.

29 Harkema H, Dowling JN, Thornblade T, *et al*. ConText: an algorithm for determining negation, experiencer, and temporal status from clinical reports. *J Biomed Inform* 2009;42:839–51.

30 South BR, Phansalkar S, Swaminathan AD, *et al*. Adaptation of the NegEx algorithm to Veterans Affairs electronic text notes for detection of influenza-like illness (ILI). *AMIA Annual Symposium proceedings/AMIA Symposium AMIA Symposium*. 2007:1118.

31 Kim Y, Hurdle J, Meystre SM. Using UMLS lexical resources to disambiguate abbreviations in clinical text. *AMIA Annual Symposium Proceedings/AMIA Symposium AMIA Symposium*. 2011;2011:715–22.

32 Xu H, Stetson PD, Friedman C. Combining corpus-derived sense profiles with estimated frequency information to disambiguate clinical abbreviations. *AMIA Annual Symposium proceedings/AMIA Symposium AMIA Symposium*. 2012;2012:1004–13.

33 Wang Y, Melton GB, Pakhomov S. It's about this and that: a description of anaphoric expressions in clinical text. *AMIA Annual Symposium Proceedings/AMIA Symposium AMIA Symposium*. 2011;vol 2011:1471–80.

34 Willis BH. Spectrum bias—why clinicians need to be cautious when applying diagnostic test studies. *Fam Pract* 2008;25:390–6.

35 Gundlapalli A, Carter M, Palmer M, *et al*. Using Natural Language Processing on the Free Text of Clinical Documents to Screen for Evidence of Homelessness Among US Veterans. AMIA Annual Symposium Proceedings, 18 Nov 2013.

36 Haerian K, Salmasian H, Friedman C. Methods for identifying suicide or suicidal ideation in EHRs. *AMIA Annual Symposium proceedings/AMIA Symposium AMIA Symposium*. 2012;2012:1244–53.