Identifying Patients with Depression Using Free-text Clinical Documents

Article *in* Studies in health technology and informatics · August 2015

Li Zhou
Brigham and Women's Hospital

Amol S Navathe
University of Pennsylvania

Margarita Sordo
Massachusetts General Hospital

Maxim Topaz
Harvard Medical School

MEDINFO 2015: eHealth-enabled Health
I.N. Sarkar et al. (Eds.)

doi:10.3233/978-1-61499-564-7-629

Identifying Patients with Depression Using Free-text Clinical Documents

Li Zhou[a,b], Amy W. Baughman[b], Victor J. Lei[a], Kenneth H. Lai[b], Amol S. Navathe[c], Frank Chang[a], Margarita Sordo[a,b], Maxim Topaz[b], Feiran Zhong[b], Madhavan Murrali[d], Shamkant Navathe[d], Roberto A. Rocha[a,b]

[a] *Clinical Informatics, Partners eCare, Partners Healthcare Inc. Boston, MA, USA*
[b]*Division of General Internal Medicine and Primary Care, Brigham & Women's Hospital, Harvard Medical School, Boston, MA,USA*
[c] *Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania, USA*
[d] *School of Computer Science, College of Computing, Georgia Institute of Technology, Atlanta, GA, USA*

Abstract

About 1 in 10 adults are reported to exhibit clinical depression and the associated personal, societal, and economic costs are significant. In this study, we applied the MTERMS NLP system and machine learning classification algorithms to identify patients with depression using discharge summaries. Domain experts reviewed both the training and test cases, and classified these cases as depression with a high, intermediate, and low confidence. For depression cases with high confidence, all of the algorithms we tested performed similarly, with MTERMS' knowledge-based decision tree slightly better than the machine learning classifiers, achieving an F-measure of 89.6%. MTERMS also achieved the highest F-measure (70.6%) on intermediate confidence cases. The RIPPER rule learner was the best performing machine learning method, with an F-measure of 70.0%, and a higher precision but lower recall than MTERMS. The proposed NLP-based approach was able to identify a significant portion of the depression cases (about 20%) that were not on the coded diagnosis list.

Keywords:
Depression; Natural Language Processing; Text Classification; Machine Learning'

Introduction

Nearly 1 in 10 adults in the United States have been reported as suffering from clinical depression [1] and similar percentages are reported worldwide [2]. The associated personal, societal, and economic costs are significant [3]. In the U.S., over 15% of depressed people commit suicide, accounting for 30,000 deaths each year. Annual economic consequences are estimated at $83 billion in the U.S. due to higher healthcare utilization and decreased worker productivity [4]. People who have suffered from chronic diseases such as heart disease or diabetes are at greater risk for depression than the overall population. Importantly, those with both depression and other chronic diseases also have worse health outcomes and significantly increased costs compared to their non depressed counterparts [5]. Recently, depression has also been recognized as a major hospital readmission risk factor, and the 30-day readmission rate for mood disorders has been reported at about 15% [6]. A recent study found that 16% of hospitalized adult patients screened positive for mild depressive symptoms and another 24% for moderate or severe depression [7].

Fortunately, most patients with depression can be reliably diagnosed and successfully treated by non-specialists. Only a fraction of patients with major depression require more complex interventions by mental health specialists [8]. Research shows that about 80% of patients with depression will improve after first-line treatment [9]. However, a significant portion of patients do not receive or benefit from depression treatments; only one third of depression patients receive treatment and as many as half of depression cases will go underreported or under-diagnosed [10-13]. Because many complex medical conditions (e.g., heart failure, coronary artery disease, and dementia) are comorbid with depression, it is not surprising that many patients with depression are missed or under-treated as these other issues may take priority during a patient visit. Systematic methods to identify at-risk individuals have the potential to improve patient well-being and quality of life as well as decrease morbidity and suffering.

One strategy to improve this under-treatment is appropriate documentation of a depression diagnosis in electronic health records (EHRs) to better identify depressed patients; and provide evidence-based treatment in a timely fashion. However, emerging evidence shows that using only structured EHR data (such as diagnoses and medications), which are mostly collected for billing purposes, often identifies only a fraction of all documented depression patients. Rather, much of the depression related information is documented in clinical notes [14]. As a result, we need new strategies to more consistently identify depressed patients, particularly those with comorbid medical conditions. Modern information technology methods, such as natural language processing (NLP) can potentially improve the identification and treatment of depression, as well as enable more accurate quality metrics and support new care models targeted at high-risk patients.

Early evidence suggests that NLP methods may be very effective. Few previous studies have focused on the application of NLP technologies to extract psychiatric and mood disorder information from clinician notes [14-16]. A recent study created and validated an NLP application to extract symptomatic remission and treatment resistance information from outpatient psychiatric provider notes for patients with major depressive disorder [15]. Another pilot study has used Twitter data to identify mental health symptoms for several conditions, including depression [16]. Fischer et al [14] identified depression among diabetes patients using an NLP-based approach by recognizing depression terms and negations from office notes. The authors compared these extracted cases with administrative database codes and found that almost a third more depression cases were identified through NLP only.

The goal of this study is to identify patients with depression using discharge summaries by applying a general NLP system and machine learning classification algorithms. This work is a part of two greater efforts: one study is using NLP to improve patient problem lists by identifying depression diagnoses from free-text notes; and the other utilizes unstructured data from clinical narratives to identify patients at high risk for hospital readmission, augmenting existing methods based on structured administrative data. Depression is of particular interest because it is a risk factor that can enhance the predictive power of existing models of risk-stratification.

Methods

Our methods are summarized in Figure 1. We first manually reviewed a random sample of discharge summaries, identified depression cases, and created a gold standard with the help of domain experts. We then trained and tested our NLP system (known as MTERMS [17]) and classification algorithms. Finally, we assessed our automated approach by comparing system-generated classification against the gold standard. We also compared depression cases extracted from free-text data versus coded diagnoses.

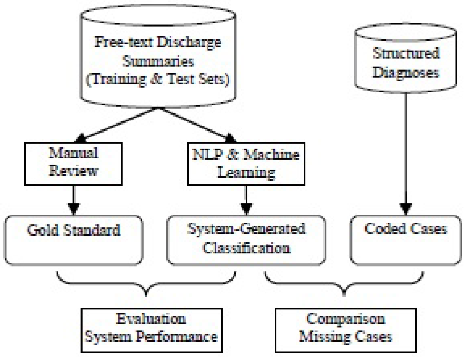

Free-text Discharge Summaries (Training & Test Sets)
Structured Diagnoses
Manual Review
NLP & Machine Learning
Gold Standard
System-Generated Classification
Coded Cases
Evaluation System Performance
Comparison Missing Cases

Figure 1-Methods Overview

Data Collection

We used a retrospective cohort of patients from our readmission study. Patients in this cohort had a history of ischemic heart disease. They were hospitalized between 01/01/2011 and 12/31/2013 at different hospitals in Partners HealthCare, a large integrated healthcare network in Boston, Massachusetts, US. We randomly selected 1,200 patients, each with one discharge summary. We then randomly selected 600 discharge summaries for training and used the remaining 600 for testing. At these institutions, discharge summaries include a complete copy of admission note, as well as standard discharge summary information including detailed hospital course, discharge medications, and plan of care.

Identifying Depressed Patients

A pharmacy doctoral student in consultation with an internal medicine physician reviewed both training and testing cases. They classified these cases as depression with high, intermediate and low confidence, based on information in the discharge summary. Twenty randomly selected patient medical records in each classification category were reviewed by one of the clinician authors. This was to validate each category and confirm whether there was enough clinical evidence in the record to support a diagnosis of depression. These findings were then used by the research group to modify our classification algorithms introduced below as needed.

- **High confidence** cases were asserted when depression diagnosis terms were present in notes (e.g., depressive disorder or depression was listed in Past Medical History), indicating a history of depression.
- **Intermediate confidence** cases were identified when combinations of antidepressant treatment (e.g., medications for depression), psychiatry consultation (e.g., referrals to a mental health specialist), or depressive symptomatology were documented. Combinations of above situations can provide clinicians with more information than singular instances when reviewing notes. More specifically, intermediate confidence cases were identified when one of the following scenarios were mentioned in discharge summaries: 1) at least one depressive symptom (e.g., suicidal ideation) and a psychiatric consult was involved (e.g., psychiatry was consulted); 2) at least one symptom of depression was present and an antidepressant medication was prescribed (e.g., citalopram) that did not have an indication in instructions for another problem (e.g., for insomnia); 3) an antidepressant medication was present that did not have an indication in instructions for another problem and a psychiatric consult was involved.
- **Low confidence** cases were asserted when depression diagnosis terms or synonyms were absent and none of the criteria for intermediate confidence case were satisfied. The lack of documentation related to depression cannot exclude a history of depression, but in the scope of discharge notes these were considered as negative or unknown cases.

NLP and Classification Algorithms

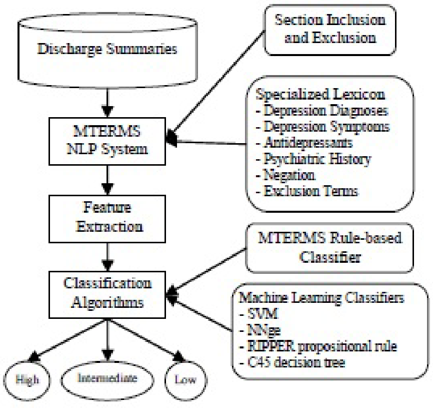


Figure 2 - Overview of NLP and Classification Algorithms

As shown in Figure 2, we first used the MTERMS NLP system [17] to identify relevant sections of the discharge summaries. We included sections that contained medical history, diagnoses, and treatments (e.g., "History of Present Illness", "Social History", "Assessment" and "Medication"). We excluded sections such as "Discharge Instructions", which may mention depression-related terms but do not indicate that the patient actually has depression (e.g., "Call your doctor if you have...suicidal feelings...").

We then created a specialized lexicon containing terms related to depression. For depression diagnoses (e.g., "depression", "depressive"), we used terms that SNOMED-CT (September 2014 Release) classified as preferred terms and synonyms of depression [18]. We identified a list of symptoms (e.g., "insomnia", "suicidal ideation") from criteria listed in Diagnostic and Statistical Manual of Mental Disorders, 4th Edition (DSM-IV) [19] and phrases derived from the Patient Health Questionnaire – 9 (PHQ-9) [20], as well as from our previous clinical experience. We used a list of antidepressants (e.g., "amitriptyline", "citalopram") provided by Lexicomp [21], augmented with brand names from the National Library of Medicine's DailyMed website [22]. We also included terms related to a patient's psychiatric history (e.g., "counselor", "psychotherapist"), that we found previously. In addition, we identified terms related to medical electrocardiogram (EKG) findings (e.g., "ST depression") that we found in the training data, since they often contain the word "depression" but are not related to mood disorders. We used MTERMS to extract terms in our specialized lexicon as well as negation terms (e.g., "denies", "no", and "absent").

We then used these terms as features for our classification algorithms, experimenting with both rule-based (within MTERMS) and machine learning classifiers. For machine learning classifiers, we used Weka open-source toolkit [23]. We used 10-fold cross-validation on the training set to select the four best-performing machine learning algorithms to compare on the test set: a support vector machine (SVM) using sequential minimal optimization algorithm [24], a generalized nearest neighbor (NNge) classifier [25], a Repeated Incremental Pruning to Produce Error Reduction (RIPPER) propositional rule learner [26], and a C4.5 decision tree learner [27]. We also manually created a decision tree using Weka's interactive UserClassifier (see Figure 3). At each vertex of the tree, we chose the feature that splits training data most accurately, as long as it was consistent with our clinical knowledge. For example, rules such as "if a symptom is negated, then classify as depression with high confidence" were considered idiosyncrasies of training data, and were therefore not included. We subsequently incorporated our decision tree into MTERMS' classification logic.

System Performance Evaluation

We used standard metrics to evaluate system performance which included precision (p), recall (r), and the F-measure (f). For each category (high and intermediate confidence cases), let TP, TN, FP, and FN be number of true positives, true negatives, false positives, and false negatives, respectively. Then p=TP/(TP+FP), r=TP/(TP+FN), and f=2pr/(p+r). The F-measure is harmonic mean of precision and recall represents their combined quality.

Comparison of Coded vs. Free-text Identification of Depression Patients

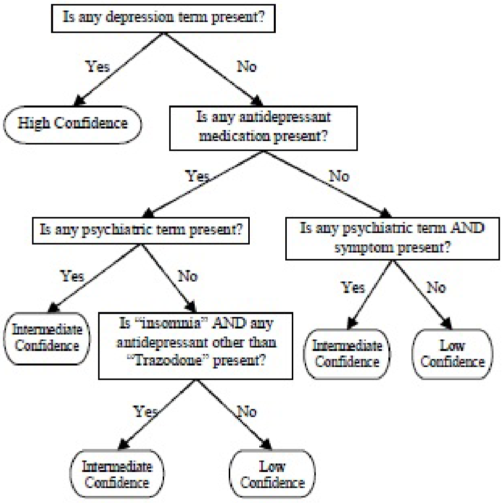

Is any depression term present?
Yes
No
High Confidence
Is any antidepressant medication present?
Yes
No
Is any psychiatric term present?
Is any psychiatric term AND symptom present?
Yes
No
Intermediate Confidence
Is "insomnia" AND any antidepressant other than "Trazodone" present?
Yes
No
Intermediate Confidence
Low Confidence
Yes
No
Intermediate Confidence
Low Confidence

Figure 3- A Knowledge-based Decision Tree

In order to measure how the NLP-based approach in addition to using structured data can enhance the detection of depressed patients, we calculated the percentage of cases that were additionally identified by our NLP system. Coded identification was done with structured diagnosis data which was retrieved from inpatient discharge diagnoses as well as patient problem lists from our outpatient EHR system, when available. We also compared the discrepancies between these data sources.

Results

System Performance

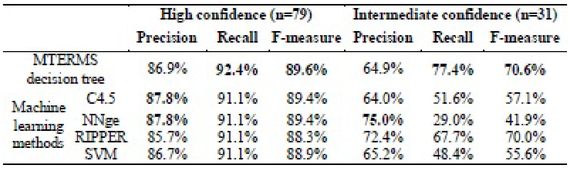
The gold standard established by manual review indicated that out of 600 training cases, 89 depression cases were identified with high confidence and 22 with intermediate confidence; out of 600 test cases, 79 were identified with high confidence and 31 with intermediate confidence. On high confidence cases, all of the algorithms we tested performed similarly; with MTERMS' knowledge-based decision tree slightly better than the machine learning classifiers, achieving an F-measure of 89.6%. MTERMS achieved highest F-measure on intermediate confidence cases, 70.6%. The RIPPER rule learner was the best performing machine learning method, with an F-measure of 70.0%, and a higher precision but lower recall than MTERMS. Full results on our test set are shown in Table 1. The confusion matrix for our knowledge-based decision tree is shown in Table 2.

Comparison of Coded vs. Free-text Data

Using both coded diagnoses and discharge summaries, 140 (23.3%) of the 600 patients in the training set were identified with depression with high confidence, while depression rate in test set was 20.8%, as shown in Table 3.

Even though about 80% of diagnoses were documented in structured data, a significant portion of diagnoses (approximately 20%) were only mentioned in discharge summaries that were identified by our NLP system. Additionally, for cases identified with intermediate confidence from discharge summaries, more than a quarter (6 of 22 cases in training set and 8 of 31 in test set) were documented in coded diagnosis list, indicating that those intermediate cases that were not on the diagnosis lists may help capture more depressed patients and serve as a good screening candidate for depression.

Table 1- System Performance on Test Set



		High confidence (n=79)			Intermediate confidence (n=31)		
		Precision	Recall	F-measure	Precision	Recall	F-measure
MTERMS decision tree		86.9%	92.4%	89.6%	64.9%	77.4%	70.6%
Machine learning methods	C4.5	87.8%	91.1%	89.4%	64.0%	51.6%	57.1%
	NNge	87.8%	91.1%	89.4%	75.0%	29.0%	41.9%
	RIPPER	85.7%	91.1%	88.3%	72.4%	67.7%	70.0%
	SVM	86.7%	91.1%	88.9%	65.2%	48.4%	55.6%

Table 2-Confusion Matrix for our Knowledge-based Decision Tree

Gold Standard	Knowledge-based Deision Tree		
	High confidence	Intermediate confidence	Low confidence
High confidence	73	2	4
Intermediate confidence	1	24	6
Low confidence	10	11	469

Table 3- Depression Cases Identified using both Coded Diagnoses and NLP

	Coded Diagnoses n (%)	Discharge Summaries & NLP* n (%)	Total Patients with Depression n (%)
Training Set	114 (81.4)	26 (18.6)	140 (23.3)
Test Set	99 (79.2)	26 (20.8)	125 (20.8)

**cases only mentioned in discharge summaries and identified by NLP*

Discussion

This study applied NLP and machine learning classification algorithms to identify clinical depression cases based on discharge summaries. We found a high (over 20%) prevalence of depression among hospitalized patients with a history of ischemic heart disease, consistent with clinical observations that acute myocardial infarction is associated with high risk for on-going depression. Compared to the structured problem list, our automated approach identified about 20% additional depression cases. In Fischer's study [14], 10% of diabetes patients were identified with depression using both administrative data and an NLP method, and an additional 5.3% were identified through NLP alone. While other studies [28] used structured EHR fields (such as billing diagnoses, problem list and medication list) to identify a diagnosis of depression by the primary care physician, our approach can serve as a complementary means to recognize additional cases, particularly in high-risk individuals such as those with ischemic heart disease.

Our study used discharge summaries which are primarily focused on acute medical problems related to the reason for admission. While depression was often clearly documented as a medical problem, we could not always confirm this diagnosis using more comprehensive criteria such as DSM-IV diagnostic criteria or positive results on the PHQ-9 due to limitations in documentation, even when accessing the full medical record. A few previous studies have used psychiatric notes for identifying depression patients. However, most studies do not have access to psychiatric notes, which are often blocked to other providers or are in a separate medical system because of their sensitive content; an algorithm based on this kind of note has limited generalizablity. Our study used a much more pragmatic approach by utilizing a common and readily available type of clinical note; therefore, our methods can be easily replicated within hospitals or primary care practices. By focusing on discharge notes, we also identify patients in a particularly vulnerable time window – after an acute event requiring hospitalization and within the post-discharge period.

Our system achieved an F-measure of 89.6% in identifying high confidence cases and 70.6% for intermediate confidence cases. MTERMS' performance was slightly better than machine learning classifiers. Recall was higher than precision (92.4% vs. 86.9% for high confidence cases and 77.4% vs. 64.9% for intermediate confidence cases), which indicates that such a system will be useful for retrieving relevant instances. Identified cases can then be reviewed by clinicians and researchers. By examining the classification errors made by the system, we found that there were several cases in which our knowledge-based system classified a patient who did not have depression as depressed with either high or intermediate confidence. In some of these false positive cases, a depression-related term was negated, but the negation was outside our NLP algorithm's scope. For example, in the sentence "Negative for fevers / chills / sweats / ... (many other symptoms) ... / depression / ...", the word "depression" was extracted because the negation was too far away. In other cases, depression-related words were used to describe non-mood disorders (e.g., "mild depression of the lateral tibial plateau", "moderate depression of systolic function"). Our lexicon was able to correctly discover some but not all of these phrases. In addition, there were a few cases in which a patient possibly having depression with intermediate confidence was not identified by our system. These false negative cases occurred mainly because the system did not identify symptoms outside our lexicon (e.g., "becomes angry / altered mood"). Our study has a few limitations. First, we only used clinical narrative reports from a single institution's EHR system, thus our results may not be generalizable to other healthcare institutions. We may also have missed terms that physicians use in other institutions. Use of unsupervised algorithms to identify these other terms may be helpful in the future.

Second, we only used discharge summaries. Future work will include applying our system to other types of clinical notes, such as clinic visit notes in outpatient settings. Lastly, patients included in this study had ischemic heart disease and hence results may vary for a general patient population.

Acknowledgments

This study was funded by the Partners Siemens Research Council.

References

[1] Center for Disease Control and Prevention. Current Depression Among Adults - United States, 2006 and 2008. Morbidity and Mortality Weekly Report. Center for Disease Control and Prevention, 2010; 59(38):1229-1235.

[2] Andrade L, Caraveo-anduaga JJ, Berglund P, Bijl RV, Graaf RD, Vollebergh W, Dragomirecka E, Kohn R, Keller M, Kessler RC, Kawakami N, Kiliç C, Offord D, Ustun TB, Wittchen HU. The epidemiology of major depressive episodes: results from the International Consortium of Psychiatric Epidemiology (ICPE) Surveys. International Journal of Methods in Psychiatric Research. 2003;12(1): 3-21.

[3] National Alliance on Mental Illness. The Impact and Cost of Mental Illness: The Case of Depression. National Alliance on Mental Illness (NAMI). Retrieved from: http://www.nami.org/Template.cfm?Section=Policymakers_Toolkit&Template=/ContentManagement/ContentDisplay.cfm&ContentID=19043.

[4] Donohue JM, Pincus HA. Reducing the societal burden of depression. Pharmacoeconomics. 2007;25(1):7-24.

[5] Moussavi S, Chatterji S, Verdes E, Tandon A, Patel V, Ustun B. Depression, chronic diseases, and decrements in health: results from the World Health Surveys. The Lancet. 2007;370(9590):851-858.

[6] Elixhauser A, Steiner C. Readmissions to U.S. Hospitals by Diagnosis, 2010: Statistical Brief #153. Healtcare Cost and Utilization Project. Retrieved from: http://wwwhcup-usahrqgov/reports/statbriefs/sb153pdf.

[7] Cancino RS, Culpepper L, Sadikova E, Martin J, Jack BW, Mitchell SE. Dose-response relationship between depressive symptoms and hospital readmission. Journal of Hospital Medicine. 2014;9(6):358-364.

[8] Gilbody S, Bower P, Fletcher J, Richards D, Sutton AJ. Collaborative care for depression: a cumulative meta-analysis and review of longer-term outcomes. Archives of Internal Medicine. 2006;166(21):2314-2321.

[9] Campbell KP, Lanza A, Dixon R, Chattopadhyay S, Molinari N, Finch RA, editors. A purchaser's guide to clinical preventive services: moving science into coverage. Washington, DC: National Business Group on Health; 2006.

[10] Harman JS, Edlund MJ, Fortney JC. Disparities in the adequacy of depression treatment in the United States. Psychiatric Services. 2004;55(12):1379-1385.

[11] Simpson SM, Krishnan LL, Kunik ME, Ruiz P. Racial disparities in diagnosis and treatment of depression: a literature review. Psychiatric Quarterly. 2007;78(1):3-14.

[12] Mitchell AJ, Vaze A, Rao S. Clinical diagnosis of depression in primary care: a meta-analysis. The Lancet. 2009;374(9690):609-619.

[13] Edlund MJ, Unutzer J, Wells KB. Clinician screening and treatment of alcohol, drug, and mental problems in primary care: results from healthcare for communities. Medical Care. 2004;42(12):1158-1166.

[14] Fischer LR, Rush WA, Kluznik JC, O'Connor PJ, Hanson AM. Abstract C-C1-06: identifying depression among diabetes patients using natural language processing of office notes. Clinical Medicine & Research. 2008;6(3-4):125-126.

[15] Perlis R, Iosifescu D, Castro V, Murphy S, Gainer V, Minnier J, Fava M, Weilburg JB, Churchill SE, Kohane IS, Smoller JW. Using electronic medical records to enable large-scale studies in psychiatry: treatment resistant depression as a model. Psychological Medicine. 2012;42(01):41-50.

[16] Harman GCMDC. Quantifying mental health signals in Twitter. ACL 2014, 51.

[17] Zhou L, Plasek JM, Mahoney LM, Karipineni N, Chang F, Yan X, Rocha RA. Using Medical Text Extraction, Reasoning and Mapping System (MTERMS) to process medication information in outpatient clinical notes. AMIA Annual Symposium proceedings. 2011; 1639-1648.

[18] SNOMED CT. Retrieved from: http://www.ihtsdo.org/.

[19] Diagnostic and statistical manual of mental disorders, 4th ed. American Psychiatric Association, 2000.

[20] Kroenke K, Spitzer RL, Williams JB. The Phq‐9. Journal of General Internal Medicine. 2001;16(9):606-613.

[21] Lexicomp. Retrieved from: http://www.lexi.com/.

[22] DailyMed. Retrieved from: http://dailymed.nlm.nih.gov/dailymed/index.cfm.

[23] Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH. The WEKA data mining software: an update. SIGKDD Explorations. 2009;11(1):10-18.

[24] Smola AJ, Schölkopf B. A tutorial on support vector regression. Statistics and Computing.2004;14(3):199-222.

[25] Martin B. Instance-based learning: nearest neighbour with generalisation. University of Waikato, 1995.

[26] Cohen, WW. Fast effective rule induction. In Machine Learning: Proceedings of the Twelfth International Conference, Lake Tahoe, California, 1995.

[27] Quinlan, JR. C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers, 1993.

[28] Trinh N-HT, Youn SJ, Sousa J, Regan S, Bedoya CA, Chang TE, Fava M, Yeung A. Using electronic medical records to determine the diagnosis of clinical depression. International Journal of Medical Informatics. 2011;80(7):533-540.

Address for correspondence

Li Zhou, MD, PhD

Clinical Informatics, Partners eCare, Partners HealthCare

93 Worcester Street, Wellesley, MA 02481

lzhou2@partners.org