Developing Computational Models for Predicting Diagnoses of Depression

Sandy H Huang[1], Paea LePendu[1], Srinivasan V Iyer[1], Anna Bauer-Mehren[1], Cliff Olson[2], and Nigam H Shah[1]

[1]Stanford University, Stanford, CA; [2]Palo Alto Medical Foundation, Palo Alto, CA

Abstract

Depression is a prevalent disorder that is difficult to diagnose. We developed and evaluated a model that uses electronic health record data for predicting the diagnosis of depression. The model was trained and tested on a set of 35,000 patients (1:6 ratio of depressed to non-depressed patients) selected from a database of 1.16 million patients from the Palo Alto Medical Foundation. Using disease and drug ingredient terms extracted from clinical notes in addition to ICD-9 codes and patient demographics as features, the model performed with good sensitivity and specificity (AUROC = 0.81 to 0.82), accurately predicting a diagnosis of depression up to 12 months beforehand. Thus, this model has the potential to both serve as a screening tool for identifying high-risk patients for follow-up as well as enable better cohort building for clinical studies.

Introduction and Background

Affecting about 13% of individuals in the United States at some point in their lives, depression is a prevalent disorder[1]. An estimated 10 to 20% of primary care visits are related to depression, making it the second most chronic disorder seen by primary care physicians[2,3]. The economic cost of depression is also staggering. In the United States, recent estimates put the direct expenses and loss of productivity resulting from depression at about $44 billion[4].

Despite the prevalence of depression, diagnosing it is a challenge. A meta-analysis performed by Mitchell et al. found primary care physicians identify only about 50% of true depression cases[2]. A variety of factors contribute to this low rate of identification. For one, patients may be reluctant to describe their emotional problems; they choose to describe somatic symptoms instead**,** such as loss of appetite or insomnia. Given its social stigma, doctors may also be reluctant to consider depression in a patient before ruling out other causes for the patient's symptoms.

The growing amount of data available from electronic health record (EHR) systems has proven useful in predicting diagnoses of other disorders. For example, Reis et al. built a model for predicting a diagnosis of domestic abuse based on medical billing records[5]. However, very few efforts utilize free-text derived features in such predictive models. In particular, to our knowledge, no such model has been built for depression.

We evaluated the effectiveness of utilizing both structured and unstructured EHR data to predict whether a given patient will be diagnosed with depression, and to determine how early we can accurately make such a prediction. Our model achieves good sensitivity at a high level of specificity and has the potential to assist in the timely diagnosis of depression. It can serve as an early level of screening for patients, and those patients classified as having a high risk of being diagnosed with depression can be examined further. In addition, our model can also be used to enable better cohort building for clinical studies on depression. Several investigators have demonstrated the utility of using EHR-derived features for automated cohort building[6-9].

Methods

Data

We use EHR data from the Palo Alto Medical Foundation (PAMF), which spans 13 years of patient data. It contains data from 1.16 million patients, 78 million visits, 622 million coded ICD-9 diagnoses, and a combination of progress notes, pathology, radiology, and transcription reports totaling over 65 million unstructured clinical texts. Each billing code and note is dated, and patients' gender, ethnicity, and year of birth are specified. The gender split is roughly 54% female. Ages range from 0 to 90, with an average age of 44. We process the clinical notes as described previously[10,11]. In brief, we use an optimized version of the NCBO Annotator with a set of 22 clinically relevant ontologies, remove ambiguous terms[12-14], and flag negated terms and terms attributed to family history to reduce the false positive rate[15]. The output of the annotation process is a mapping from each medical note to the terms in that note, including whether each term is flagged as negated and/or related to family history. In the final step, concepts are aggregated based on the hierarchies of the ontologies, and drugs are normalized to their active ingredients using RxNorm[16].

Table 1 Selected ICD-9 codes for depression

ICD-9 Code	Description
296.2[0-6]	Major depressive disorder, single episode
296.3[0-6]	Major depressive disorder, recurrent episode
296.82	Atypical depressive disorder
298.0	Depressive type psychosis
300.4	Dysthymic disorder
311	Depressive disorder, not elsewhere classified

Cohort Selection

To train and test the model for diagnosing depression, we created a cohort of depressed patients and a control cohort of non-depressed patients. The *depression patient cohort* consists of all patients who have substantial medical history and meet strict criteria that ensure they indeed have depression (Figure 1)[17]. To meet these criteria, a patient must have a depression-related ICD-9 code (Table 1)[18,19], as well as a depression term and an anti-depressive drug ingredient in the patient's clinical text. The depression terms are chosen from the terms in the lexicon, and were manually filtered to remove any terms that resulted in a significant number of false positives (i.e. patients who had the term in their clinical text but in fact did not have depression). We selected depression drug ingredients by first retrieving drugs that treat depression from the Medi-Span® (Wolters Kluwer Health, Indianapolis, IN) Drug Indications Database™, mapping them to active ingredients using RxNorm[16], and finally filtering out those active ingredients that are also present in drugs with primary indications other than depression.

For each depression patient, we define the *time of first diagnosis* as the first date in the patient's medical history at which both a depression-related ICD-9 code and an anti-depressive drug ingredient have occurred[17]. A patient is considered to have substantial medical history if the time span of visits before the patient's time of first diagnosis is at least a year and a half. This requirement ensures the model is supplied adequate medical information for each patient. There are 35,102 PAMF patients that meet these criteria, and are thus included in the depression cohort (Figure 1). The estimated PPV for the depression cohort is 96%, calculated from a manual review of 100 randomly sampled patients in the cohort.

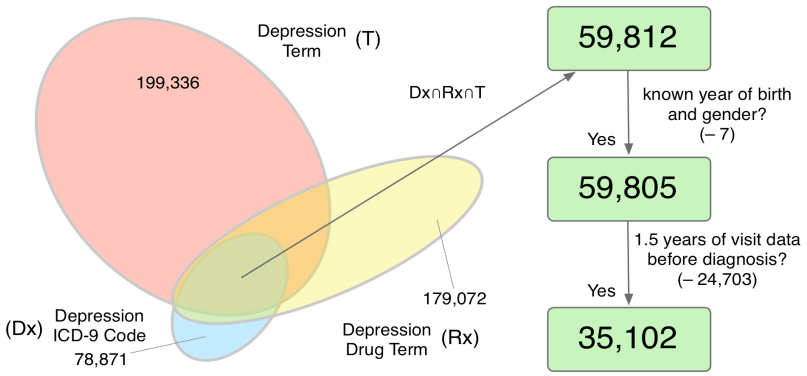

Figure 1 Selection of depression cohort from PAMF data

To create the *depression control cohort*, we matched six randomly selected non-depressed patients to each of 5,000 depressed patients randomly selected from the depression cohort. This 1:6 matching ratio mirrors the 14% prevalence of depression in the general population and is within the range seen in primary care[1,2]. Control patients are defined as patients having neither a depression-related ICD-9 code nor a depression term mentioned in their medical history. In addition, the number of days between the first and last visit of each control patient must be within six months (182 days) more than the matched depression patient's first visit to the time of first diagnosis. Matching patients based on length of visit history reduces the likelihood that the lack of a depression diagnosis in the control patient's medical record is due to inadequate medical record length, rather than the patient's actual lack of depression. Based on these requirements, all of the 5,000 depression patients were successfully matched to controls, resulting in a control cohort of 30,000 patients. The estimated PPV for the control cohort is 97%, calculated from a manual review of 100 randomly sampled patients in the cohort, 90 of which had enough clinical note data to determine whether they had depression or not.

Model Description

We base our prediction model on a Naïve Bayes classifier, a straightforward but powerful approach that has been used for similar problems with success[5,20]. In a Naïve Bayes classifier, cases are represented as feature vectors, and for each case, the model calculates the probability that the case is in a particular class, given the case's feature vector representation – under the central assumption that the features are independent. In this instance, the cases are patients, and the classes are low risk, and high risk, of an impending diagnosis of depression. The features we use are gender, age, average number of visits per year, ICD-9 codes, and disease and drug ingredient terms found in the annotated notes, excluding those codes and terms used to select the depression cohort or flagged as negated and/or related to family history. Each patient is represented as a Boolean feature vector, in which each feature corresponds to a unique dimension of the vector. A patient's vector has a value of *true* for a dimension if that patient's medical history or demographics information contains the corresponding feature, and *false* otherwise. Since age and average number of visits per year are continuous variables, they are divided into discrete bins, each of which is considered as a separate Boolean feature. Age is divided into ten bins of width ten: [0, 10), [10, 20) … [80, 90), and [90, ∞). The average number of visits is divided into ten bins of width two: [0, 2), [2, 4) … [16, 18), and [18, ∞). This type of binning is similar to equal width interval binning, which has been shown to produce results comparable to that of supervised methods, which use class labels to determine ideal bin boundaries[21,22]. Each patient's feature vector thus has a value of *true* for the bins that the patient's age and average number of visits per year fall into, and *false* for all other bins.

In order to simulate the early prediction of a diagnosis of depression, the EHR data of depressed patients is truncated at three different cutoffs: the time of first diagnosis, six months before the first diagnosis, and one year before the first diagnosis. We train the model using EHR data until the first cutoff for depressed patients, and using all available EHR data for controls. Then, we test the model on three different test sets: one for each of the three cutoffs defined previously. For all three test sets, we use all available EHR data for control patients. Thus, an accurate prediction of a depression diagnosis by the classifier, using EHR data up to the six months point, would be half a year earlier than the doctor's diagnosis of depression in the patient. Similarly, an accurate prediction using EHR data restricted to the time of first diagnosis would be analogous to the doctor's diagnosis.

Model Validation

We trained the model on a randomly selected 70% of the 35,000 total depressed and non-depressed patients and tested it on the remaining 30%. There were about 10,600 features total in the training set. To improve the model's performance, the feature set used by the model is limited to the top features based on information gain, which counteracts the tendency of Naïve Bayes classifiers to be over-sensitive to irrelevant and redundant features[23]. Information gain is a commonly used and effective method for ranking features in feature selection[24]. Through parameter selection, we decided on using the top 50 features more likely in depression patients and the top 25 more likely in non-depressed patients.

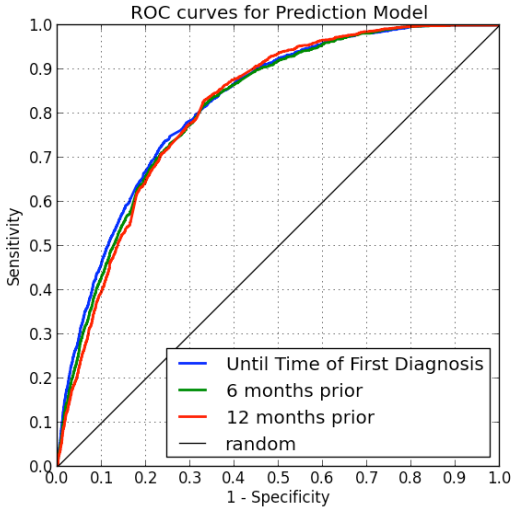

Figure 2 ROC curves for the model's performance on test data restricted to three cutoff points

Results

Model Performance

With data for depressed patients up until the time of first diagnosis, the model achieved an area under the receiver operating characteristic (ROC) curve of 0.823. This cutoff corresponds with a timely prediction – one that is not early. The area under the ROC curve decreased to 0.815 for the six-month cutoff and 0.814 for the earliest, one-year cutoff. Figure 2 shows the ROC curves for these three cutoffs, which display the full range of sensitivity and specificity values achieved by the model.

Top Features

Table 2 lists the most predictive features, based on information gain. Out of the top 75 features (the ones used by the model), just over half were ICD-9 codes, about a fourth were drug ingredient terms, and 5% were disease terms. The age buckets [0, 10) and [80, 90), as well as the visit buckets [14, 16) and [18, ∞), were also within the top features, and gender was the 26th most predictive feature.

Discussion

The accuracy of our predictive model, even when restricting data from depressed patients to one year prior to the time of diagnosis, rivals that of primary care physicians, whom Mitchell et al. found had a sensitivity of about 50% and a specificity of just over 80%[2]. In comparison, our model has a sensitivity of around 65% at a specificity of 80%, for all three cutoff points.

The top features show that a high visit density indicated a patient was more likely to be depressed, and young children were unlikely to be depressed while older patients were more likely to be depressed. Anxiety was also a predictive feature of a diagnosis of depression, which concurs with previous studies linking depression to anxiety[25,26].

Limitations

One limitation of our work is the accuracy of our annotation workflow in extracting disease and drug ingredient terms from clinical notes. Annotation errors impair the performance of our model. Such annotation errors include inaccurate mappings of notes to terms and erroneous flagging of terms as negated or related to family history.

Our work is also limited by the quality of our gold standard: the depression and control cohorts. The labels (depressed vs. non-depressed) we assign to patients may not be entirely correct, introducing slight inaccuracies and potential bias in both training and testing our model. As mentioned, the estimated PPVs of our depression and control cohorts are 96% and 97%, respectively. It is worth noting that the estimated PPV of our depression cohort is on par or better than that of previous studies using EHR data to select cohorts[27,28]. The actual PPVs may be different, and not just due to annotation errors. For instance, certain patients in the control cohort may have depression but are never diagnosed with it, or some patients in the depression cohort may be misdiagnosed with depression. A significant number of patients may fall in these categories, due to the difficulty of diagnosing depression[2].

In addition, the time of first diagnosis calculated for the depressed patients is an approximation to the actual date at which the patient was first diagnosed with depression. The PAMF dataset we used also excludes notes from mental health professionals, due to privacy regulations. Inclusion of these notes may have improved the accuracy of our model for patients with these notes in their medical record.

Future research

We plan to develop and compare other types of models for predicting diagnoses of depression, in particular models such as the Cox proportional hazards model that take the sequence of events into account. In addition, we will evaluate the performance of our prediction model when applied to a specific demographic of patients, for instance patients of a particular gender or age range.

Table 2: Top five features, based on information gain, when training the model. The "% Dep" column is the percentage of depressed patients with this feature, and the "% Control" column is that for controls

Feature	Description	IG	% Dep	% Control
V20.2	(ICD-9) Routine infant or child health check	0.025	5.1	30.0
Drug ingredient	Fluorides	0.024	0.5	19.3
V06.1	Vaccination of DTP-DTaP	0.019	6.5	28.4
Age: [0, 10)	Age between 0 and 10	0.019	0.0	13.8
300.00	(ICD-9) Anxiety state, unspecified	0.016	20.1	4.9

We are also evaluating the portability of this model by testing it on an external dataset composed of approximately 17,000 patients from Group Health Research Institute (GHRI) in Seattle who have been treated for depression and scored using the Patient Health Questionnaire (PHQ-9). The dataset provides a standard for assessing severity and is intended to be used to predict treatment outcomes using scored data on patients before and after treatment. This dataset has the advantage of having verified depression patients. By adding randomly selected patients as controls, we should also be able to evaluate the performance of our model for cohort selection. This cohort selection model would be identical to our prediction model, except it would not restrict depression patients' EHR data and would also include depression-related ICD-9 features, disease terms, and drug terms. (These depression-related features were excluded from the prediction model because they were used to select the depression and control cohorts.) With these highly relevant additional features, we expect to improve upon the results of our prediction model, and our model's performance should be on par with that of existing cohort selection models for other disorders[8,29]. While predicting the diagnosis of depression is an important goal in itself, we plan to extend our methods toward gaining a better understanding of treatment outcomes given a new patient's medical history – for example, whether talk therapy would work better than medication for a particular patient.

Conclusion

We developed and assessed a model that uses EHRs for predicting the diagnosis of depression. The model uses ICD-9 codes, disease and drug ingredient terms extracted from clinical notes, and patient demographics as features to achieve an AUROC of 0.81 to 0.82 for predicting a diagnosis of depression in patients, up to 12 months before the first diagnosis of depression. Our results suggest the use of EHR data can improve the timely diagnosis of depression, a disorder that primary care physicians often miss. Even up to a year before their diagnosis of depression, patients show patterns in their medical history that our model can detect. We believe that this model has the potential to both serve as a screening tool in identifying high-risk patients for closer examination as well as enable better cohort building for clinical studies on depression.

Acknowledgements: The authors acknowledge support from the NIH grant U54 HG004028 for the National Center for Biomedical Ontology. We also acknowledge Rave Harpaz, Kenneth Jung, and Tyler Cole for useful discussions.

References

1. Hasin DS, Goodwin RD, Stinson FS, Grant BF. Epidemiology of major depressive disorder: results from the National Epidemiologic Survey on Alcoholism and Related Conditions. Arch Gen Psychiatry 2005;62:1097-106.
2. Mitchell AJ, Vaze A, Rao S. Clinical diagnosis of depression in primary care: a meta-analysis. Lancet 2009;374:609-19.
3. Sharp LK, Lipsky MS. Screening for depression across the lifespan: a review of measures for use in primary care settings. Am Fam Physician 2002;66:1001-8.
4. Cain RA. Navigating the Sequenced Treatment Alternatives to Relieve Depression (STAR*D) study: practical outcomes and implications for depression treatment in primary care. Prim Care 2007;34:505-19, vi.
5. Reis BY, Kohane IS, Mandl KD. Longitudinal histories as predictors of future diagnoses of domestic abuse: modelling study. BMJ 2009;339:b3677.

6. Wei WQ, Leibson CL, Ransom JE, et al. Impact of data fragmentation across healthcare centers on the accuracy of a high-throughput clinical phenotyping algorithm for specifying subjects with type 2 diabetes mellitus. J Am Med Inform Assoc 2012;19:219-24.
7. Kho AN, Hayes MG, Rasmussen-Torvik L, et al. Use of diverse electronic medical record systems to identify genetic risk for type 2 diabetes within a genome-wide association study. J Am Med Inform Assoc 2012;19:212-8.
8. Peissig PL, Rasmussen LV, Berg RL, et al. Importance of multi-modal approaches to effectively identify cataract cases from electronic health records. J Am Med Inform Assoc 2012;19:225-34.
9. Conway M, Berg RL, Carrell D, et al. Analyzing the heterogeneity and complexity of Electronic Health Record oriented phenotyping algorithms. AMIA Annu Symp Proc 2011;2011:274-83.
10. Shah NH, Bhatia N, Jonquet C, Rubin D, Chiang AP, Musen MA. Comparison of concept recognizers for building the Open Biomedical Annotator. BMC Bioinformatics 2009;10 Suppl 9:S14.
11. Lependu P, Iyer SV, Fairon C, Shah NH. Annotation Analysis for Testing Drug Safety Signals using Unstructured Clinical Notes. J Biomed Semantics 2012;3 Suppl 1:S5.
12. Bodenreider O, McCray AT. Exploring semantic groups through visual approaches. J Biomed Inform 2003;36:414-32.
13. Xu R, Musen MA, Shah NH. A Comprehensive Analysis of Five Million UMLS Metathesaurus Terms Using Eighteen Million MEDLINE Citations. AMIA Annu Symp Proc 2010;2010:907-11.
14. 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:e149-e56.
15. Harkema H, Dowling JN, Thornblade T, Chapman WW. ConText: an algorithm for determining negation, experiencer, and temporal status from clinical reports. J Biomed Inform 2009;42:839-51.
16. RxNorm. U.S. National Library of Medicine. (Accessed 2012 Sep 08, at http://www.nlm.nih.gov/research/umls/rxnorm/.)
17. Townsend L, Walkup JT, Crystal S, Olfson M. Mini-sentinel systematic evaluation of health outcome of interest definitions for studies using administrative data. In: U.S. Food and Drug Administration; 2010.
18. Loftus EV, Jr., Guerin A, Yu AP, et al. Increased risks of developing anxiety and depression in young patients with Crohn's disease. Am J Gastroenterol 2011;106:1670-7.
19. Valuck RJ, Orton HD, Libby AM. Antidepressant discontinuation and risk of suicide attempt: a retrospective, nested case-control study. J Clin Psychiatry 2009;70:1069-77.
20. Kononenko I. Machine learning for medical diagnosis: history, state of the art and perspective. Artif Intell Med 2001;23:89-109.
21. Dougherty J, Kohavi R, Sahami M. Supervised and Unsupervised Discretization of Continuous Features. In: Prieditis A, Russell S, editors. Proceedings of the Twelfth International Conference on Machine Learning; 1995; Tahoe City, California, USA; 1995. p. 194-202.
22. Hsu C-N, Huang H-J, Wong T-T. Implications of the dirichlet assumption for discretization of continuous variables in naive Bayesian classifiers. Machine Learning 2003;53:235-63.
23. Ratanamahatana CA, Gunopulos D. Scaling up the Naive Bayesian classifier: Using decision trees for feature selection. In: IEEE International Conference on Data Mining. Maebashi, Japan; 2002.
24. Yang Y, Pedersen JO. A Comparative Study on Feature Selection in Text Categorization. In: Proceedings of the Fourteenth International Conference on Machine Learning: Morgan Kaufmann Publishers Inc.; 1997:412-20.
25. Sartorius N, Ustun TB, Lecrubier Y, Wittchen HU. Depression comorbid with anxiety: results from the WHO study on psychological disorders in primary health care. Br J Psychiatry Suppl 1996:38-43.
26. Hirschfeld RM. The Comorbidity of Major Depression and Anxiety Disorders: Recognition and Management in Primary Care. Prim Care Companion J Clin Psychiatry 2001;3:244-54.
27. Liao KP, Cai T, Gainer V, et al. Electronic medical records for discovery research in rheumatoid arthritis. Arthritis Care Res (Hoboken) 2010;62:1120-7.
28. West SL, Richter A, Melfi CA, McNutt M, Nennstiel ME, Mauskopf JA. Assessing the Saskatchewan database for outcomes research studies of depression and its treatment. J Clin Epidemiol 2000;53:823-31.
29. Carroll RJ, Eyler AE, Denny JC. Naive Electronic Health Record phenotype identification for Rheumatoid arthritis. AMIA Annu Symp Proc 2011;2011:189-96.