Journal of Affective Disorders 260 (2020) 617–623

Contents lists available at ScienceDirect

Journal of Affective Disorders

journal homepage: www.elsevier.com/locate/jad

Research paper

Using text mining to extract depressive symptoms and to validate the diagnosis of major depressive disorder from electronic health records

Chi-Shin Wu[a,b], Chian-Jue Kuo[c,d], Chu-Hsien Su[a], Shi-Heng Wang[e], Hong-Jie Dai[f,g,*]

[a] Department of Psychiatry, National Taiwan University Hospital, Taipei, Taiwan R.O.C
[b] College of Medicine, National Taiwan University, Taipei, Taiwan R.O.C
[c] Taipei City Psychiatric Center, Taipei City Hospital, Taipei, Taiwan R.O.C
[d] Department of Psychiatry, School of Medicine, College of Medicine, Taipei Medical University, Taiwan R.O.C
[e] Department of Public Health and Department of Occupational Safety and Health, China Medical University, Taichung, Taiwan R.O.C
[f] Department of Electrical Engineering, National Kaohsiung University of Science and Technology, Kaohsiung, Taiwan R.O.C
[g] School of Post-Baccalaureate Medicine, College of Medicine, Kaohsiung Medical University, Kaohsiung, Taiwan R.O.C

ARTICLE INFO

Keywords:
Text mining
Information extraction
Major depressive disorder

ABSTRACT

Background: Many studies have used Taiwan's National Health Insurance Research database (NHIRD) to conduct psychiatric research. However, the accuracy of the diagnostic codes for psychiatric disorders in NHIRD is not validated, and the symptom profiles are not available either. This study aimed to evaluate the accuracy of diagnostic codes and use text mining to extract symptom profile and functional impairment from electronic health records (EHRs) to overcome the above research limitations.

Methods: A total of 500 discharge notes were randomly selected from a medical center's database. Three annotators reviewed the notes to establish gold standards. The accuracy of diagnostic codes for major psychiatric illness was evaluated. Text mining approaches were applied to extract depressive symptoms and function profiles and to identify patients with major depressive disorder.

Results: The accuracy of the diagnostic code for major depressive disorder, schizophrenia, and dementia was acceptable but that of bipolar disorder and minor depression was less satisfactory. The performance of text mining approach to recognize depressive symptoms is satisfactory; however, the recall for functional impairment is lower resulting in lower F-scores of 0.774–0.753. Using the text mining approach to identify major depressive disorder, the recall was 0.85 but precision was only 0.69.

Conclusions: The accuracy of the diagnostic code for major depressive disorder in discharge notes was generally acceptable. This finding supports the utilization of psychiatric diagnoses in claims databases. The application of text mining to EHRs might help in overcoming current limitations in research using claims databases.

1. Introduction

In the era of big data, the secondary use of claims databases to conduct epidemiological studies, comparative effective research, and active pharmacovigilance has increased dramatically (Chen et al., 2010; Harpe, 2009; Schneeweiss and Avorn, 2005). However, important information, such as symptom profiles, disease severity, and laboratory or image findings, is not available in most claims databases. Electronic health records (EHRs) are the source of this abundant and important information and also contain records of the clinical judgment of physicians and treatment responses and complications. Extracting information from EHRs could overcome the limitations of claims data and boost epidemiological and medical research.

Manual chart review can obtain comprehensive information from EHRs. Unfortunately the information usually buries in the unstructured text of EHRs (Jensen et al., 2012), which bring on time-consuming and labor-intensive review process. Text mining could be an efficient alternative to automatically extract information from EHRs. Text mining is a combination of the development of natural language processing, data mining, and statistical learning (Meystre et al., 2008). Information extraction is one important field of text mining, which transforms information from unstructured text into structured data-frame (Dai et al., 2014). Approaches to information extraction can be categorized into rule-based and machine-learning approaches (Meystre et al., 2008).

* Corresponding author Department of Electrical Engineering, National Kaohsiung University of Science and Technology, No. 415, Jiangong Rd., Sanmin Dist., Kaohsiung City 80778, Taiwan.
E-mail address: hjdai@nkust.edu.tw (H.-J. Dai).

https://doi.org/10.1016/j.jad.2019.09.044
Received 10 March 2019; Received in revised form 29 July 2019; Accepted 8 September 2019
Available online 11 September 2019
0165-0327/

Rule-based approaches mainly rely on patterns or dictionaries of key words and require domain experts to manually develop the patterns and dictionaries that may not be comprehensive to cover all possible linguistic variations. In contrast, machine learning approaches are flexible and could overcome the aforementioned issues but require large amounts of annotated training data, which is also time-consuming and labor-intensive work (Dai et al., 2019b).

Utilization of text mining in EHRs in psychiatric research has increased in recent years (Abbe et al., 2015). One study used text mining from EHRs to identify symptoms of severe mental illness, including psychotic and manic symptoms but not depressive symptoms (Jackson et al., 2017). Furthermore, there are studies using EHRs to explore treatment resistant depression (Perlis et al., 2012), to improve the predictability of suicide (McCoy et al., 2016), and to identify phenotype for genomic research (Smoller, 2018). Recently, Dai and Jonnagaddala (2018) used patients' initial psychiatric evaluation records to study the application of various convolutional neural network architectures and their performance in predicting the severity of positive valence symptoms.

Major depressive disorder is one of the most common and severe mental disorders (Goodwin et al., 2006; Wu et al., 2017). Many studies have used Taiwan's National Health Insurance Research database (NHIRD) to explore the prevalence of mood disorder, pattern of psychopharmacological medications, and treatment of response (Chien et al., 2004; Wu et al., 2013, 2012). However, there are several limitations in these studies. The accuracy of the ICD-9 diagnostic code for major depressive disorder in NHIRD is not yet validated. In addition, the symptom profiles and disease severity of major depressive disorder are not available. Fortunately, the claims database might include EHRs in the future. Extracting information in EHRs could overcome the limitations of previous studies.

This aims of this study are threefold: 1) to validate the accuracy of diagnostic codes in the NHIRD, 2) to examine the feasibility of applying the rule-based and machine-learning approaches for detecting symptom profile and functional level of major depressive disorder in EHRs, and 3) to examine the accuracy of the diagnoses of major depressive disorders made by the developed text mining method.

2. Methods

2.1. Data source

This study utilized the Integrated Medical Database, National Taiwan University Hospital (NTUH-iMD) from January 1, 2006, to September 30, 2016. During this period, 4836 discharge notes (3087 patients) from the psychiatric unit with a principle psychiatric diagnosis (ICD-9-CM: 290–319 or ICD-10-CM: F00-F99) were included in the study. All personal information had been de-identified. The studied was approved by the IRB (Institutional Review Board) (NTUH-201610072RINA).

Among the data source there were 1546 (32.0%) notes with the principal diagnosis of schizophrenia, 989 (20.5%) notes with bipolar disorders, 883 (17.2%) notes with major depressive disorders, 181 (3.9%) notes with minor or other depressive disorder, and 152 (3.1%) notes with dementia. The discharge notes included the baseline demographic data, present illness, physical and mental examination, progress note, laboratory and image examination, prescription records, medical procedures, and discharge diagnosis with ICD-9-code, which is identical to the diagnosis in the claims database.

Fig. 1 illustrates the distributions of the original and sampled notes. The sampled dataset contains 250 notes with major depressive disorder (ICD-9: 296.2 or 296.3), 100 notes with mild depression (including dysthymia [ICD-9: 300.4], minor depressive disorder [ICD-9: 311], or adjustment disorder with depressed mood [ICD-9: 309.0 or 309.1]), 50 notes with schizophrenia (ICD-9: 295), 50 notes with bipolar disorders, and 50 notes with dementia or organic mental disorders (ICD-9: 290 or 294). For the purpose of evaluating the accuracy of the diagnostic code for major depressive disorder and the text mining performance for extracting related symptoms and function profiles, the discharge notes with the major depressive disorder as the principal diagnosis were oversampled from the NTUH-iMD so that the sampled distribution is not the same as the original distribution.

2.2. Medical chart review process and corpus generation

In order to generate gold standards for the development of the text mining system, medical chart review was conducted by two board-certified clinical psychiatrists (CSW and CJK) and one research worker (CHS) with experienced text mining studies. They reviewed 100 randomly sampled EHRs in a preliminary study and arrived at a consensus for establishing the annotation guideline. The discharge notes analyzed in this study were then annotated based on this guideline. Note that the annotators were asked to annotate all affirmed depressive symptoms and function profiles mentioned in the notes written by clinicians; the negative description of symptoms or functional impairments was not annotated. The interrater reliability is reported later in the Results section.

2.3. Validation of ICD-9-CM diagnostic codes

To assess the accuracy of ICD-9 diagnostic codes for the previously mentioned major psychiatric disorders, the clinical diagnoses were made based on the information in EHRs, including the symptom profile, functional impairment, and duration of illness. The discharge diagnosis with ICD-9 code in medical records was concealed.

The diagnoses were made by fitting the DSM-IV (Diagnostic and Statistical Manual of Mental Disorders, 4th Edition) diagnostic criteria. Two or more diagnoses in one discharge note were possible if they reach diagnostic criteria. However, the clinicians might not record comprehensively the presence or absence of symptoms or functional impairments in the discharge notes. If the information was not enough to fit the DSM-IV criteria, the diagnoses were made by following the below-mentioned principals. If the symptom profile was mainly psychotic symptoms and there was no overt cognitive impairment or mood symptoms, the diagnosis was determined to be schizophrenia. If the patient's age was older and symptom profiles included dominantly cognitive impairment, it would be dementia. If there were definite manic symptoms, which could not be attributed to schizophrenia or dementia, the diagnosis was classified as bipolar-spectrum disorder. Finally, the differential diagnosis between major depressive disorder and minor depression was based on the number of depressive symptoms, duration of illness, and functional impairment. However, in our preliminary medical chart review, we found that a significant proportion of discharge notes did not include the level of function. In this situation, if the disease course was complicated, such as repeated hospital admissions or refractory treatment response, it would be classified as major depression.

2.4. Extracting symptom profiles and functional impairment

The symptom profile of major depressive disorder was classified based on the DSM-IV diagnostic criteria, including depressive mood, loss of interest, fatigue, sleep disturbance, appetite change, psychomotor retardation or agitation, suicidality (ideation/plan/attempt), poor concentration/indecisiveness, and negative thought. There was no clear definition for functional impairment in DSM-IV diagnostic criteria; therefore, we defined the dimensions of functions based on the Sheen Disability Scale (Sheehan et al., 1996), including the impairment in occupational and academic, social and interpersonal, and family functions. In addition, general or unspecified dysfunction was added because we found some medical charts did not specific which functional domains were impaired.

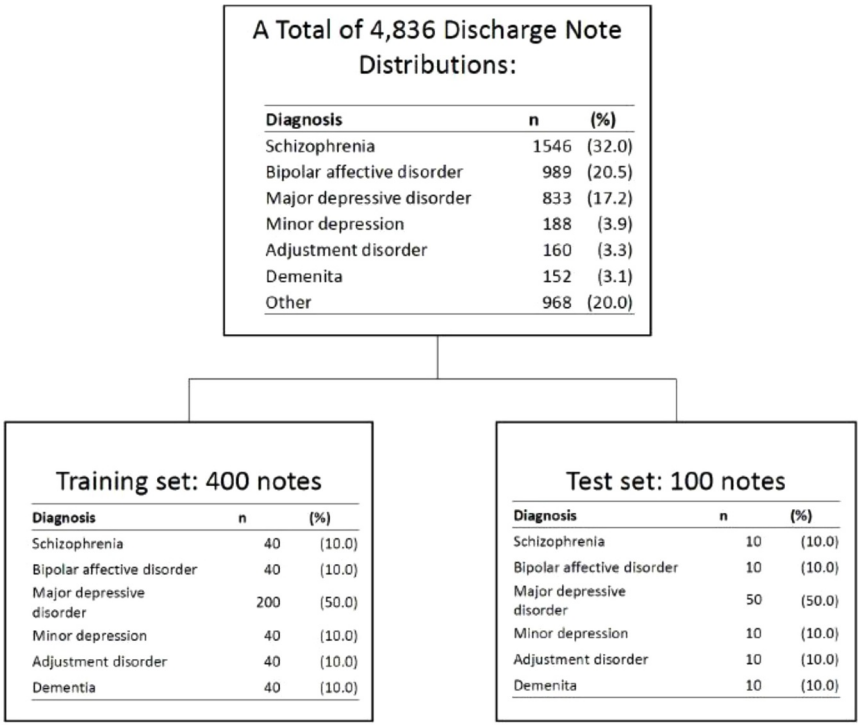

Fig. 1. Flowchart of selection process for study discharge notes.

2.5. Dictionary- and machine learning-based approach for information extraction

Based on the annotated discharge notes, 400 charts were selected as the training dataset and the other 100 charts were used as the test dataset. Two methods were applied to examine their effectiveness on the compiled dataset. The first was a dictionary-based approach. We compiled a symptom- and function-dictionary for nine symptoms and four dimensions of functional impairment by manually inspecting all texts annotated as an interest item within the training dataset. However, there were some negation descriptions in EHRs, including "The patient denied any symptom of depressive mood, anhedonia, or suicidal ideation." The dictionary-based method could not distinguish between presence or absence of given symptoms. Therefore, a machine learning-based approach was used to recognize items of interest and exclude those mentioned in the negation descriptions sentence. In this study, the linear-chain conditional random filed (CRF) model, trained to maximize a conditional probability of random variables (Lafferty et al., 2001), was used to recognize target items for a given tokenized textual sentence. We formulated the task as a sequential labeling task and used the BIO tag scheme to represent the boundaries of entities. The tag scheme used B or I tag, followed by an entity type to indicate the boundaries of an entity. If the word is not related to any symptom, it would be labeled as "O." Consider the following sequence of tokenized words as an example.

"However$_{/O}$,$_{/O}$ depressed$_{/B\text{-}DM}$ mood$_{/I\text{-}DM}$ persisted$_{/O}$ so$_{/O}$ venlafaxine$_{/O}$ was$_{/O}$ increased$_{/O}$ to$_{/O}$ 75$_{/O}$ mg$_{/O}$ BID$_{/O}$ and$_{/O}$ admission$_{/O}$ was$_{/O}$ suggested$_{/O}$.$_{/O}$"

Here, we annotate each word with BIO tags. The first word of a depressed mood (DM) entity would be annotated as B-DM. If the entity consists of more than one word, the other words would be annotated as I-DM. Here the O, B, and I tags indicate the outside, beginning, and end of the item of interest (depressed mood), respectively.

Given an input sequence of tokens W, a linear-chain CRF model computed the probability associated with its corresponding hidden labelled sequence Y as

$$p_\lambda(Y|W) = \frac{1}{Z(W)} \exp\left(\sum_{c\in C}\sum_i \lambda_i f_i(y_{c-1}, y_c, W, c)\right)$$

where $Z(W)$ is the normalization factor that makes the probability of all state sequences sum up to one, C is the set of all cliques in this textual sequence, and c is a single clique that reflects the position of the current token. The function $f_i(y_{c\text{-}1}, y_c, W, c)$ is a binary-valued feature function whose weight is λ_i. Large positive values of λ_i indicate a preference for such corresponding feature (Dai et al., 2015; Kerr et al., 2012).

For each word, a set of feature functions was defined, and their feature values were extracted and trained with the CRF model to build the item recognizer. In addition to the symptom- and function-dictionary, features like the part of speech of each word, such as noun, verb, adjective, adverb, etc. were extracted to train the CRF model. Finally, we used three types of feature: (1) words, (2) part-of-speech tags, (3) symptom- and function-dictionary in the CRF model.

2.6. Identification of major depressive disorder by text mining

We further examined the accuracy of the diagnoses of major depressive disorders made by text mining. Based on the DSM-IV diagnostic criteria, patients with major depressive disorder should have at least 5 symptoms. However, individuals with dysthymic disorder, mild depressive disorder, or other mental disorders should have 4 or less symptoms. Therefore, we used 5 depressive symptoms as the cutoff point and explored the accuracy of major depressive disorders made by extracted depressive symptom number. Functional impairment is also a criterion for the diagnosis of major depressive disorder. The performance using 5 symptoms plus any domain of functional impairment was evaluated.

2.7. Evaluation measures

The results from medical chart review were considered as the gold standard to access the accuracy of ICD-9 diagnostic codes assigned for each discharge note and to evaluate the performance of the developed

Table 1
Inter-annotator agreement for then clinical diagnoses, symptomatology, and functional impairment.

	Observed agreement (%)	Cohen's kappa (95%CI)
Clinical diagnosis		
Major depressive disorder	91	0.82 (0.70, 0.93)
Minor depression	82	0.74 (0.57, 0.91)
Bipolar affective disorder	96	0.73 (0.47, 0.98)
Schizophrenia	98	0.90 (0.76, 1.00)
Dementia or organic mental disorders	96	0.91 (0.78, 1.00)
Symptomatology		
Depressive mood	94	0.78 (0.61, 0.95)
Loss of interest	98	0.96 (0.94, 1.00)
Sleep disturbance	98	0.94 (0.87, 1.00)
Appetite change	96	0.91 (0.82, 1.00)
Fatigue	98	0.96 (0.90, 1.00)
Psychomotor retardation/ agitation	96	0.89 (0.79, 1.00)
Poor concentration/ indecisiveness	96	0.91 (0.83, 1.00)
Negative thinking	92	0.84 (0.73, 0.95)
Suicidality	96	0.90 (0.81, 1.00)
Functional impairment		
Occupational/academic	85	0.64 (0.48, 0.80)
Social life/ interpersonal	89	0.74 (0.60, 0.88)
Family	88	0.61 (0.42, 0.81)
General /unspecific	86	0.66 (0.50, 0.82)

Table 2
The distribution of clinical diagnosis, symptomatology, and functional impairment among the training dataset.

	Overall (n = 500)	Training (n = 400)	Test (n = 100)
Clinical diagnosis*,%			
Major depressive disorder	51.8	51.5	53.0
Minor depression	18.8	19.3	17.0
Bipolar affective disorder	11.2	11.5	10.0
Schizophrenia	12.2	12.3	12.0
Dementia or organic mental disorders	10.2	9.8	12.0
Symptomatology.%			
Depressive mood	83.4	84.8	78.0
Loss of interest	54.0	55.0	50.0
Sleep disturbance	79.4	80.8	74.0
Appetite change	60.8	61.0	60.0
Fatigue	42.4	43.5	38.0
Psychomotor retardation/ agitation	22.4	22.3	23.0
Poor concentration/ indecisiveness	37.2	38.8	31.0
Negative thinking	59.0	59.8	56.0
Suicidality	74.6	74.0	77.0
Functions,%			
Occupational/academic	18.8	19.8	15.0
Social life/ interpersonal	33.6	35.5	26.0
Family	28.6	28.8	28.0
General /unspecific	35.2	35.5	34.0

* not mutual exclusive, some patients have dual diagnosis.

text mining system in terms of precision, recall, and F-measure.[1] Precision, also called positive predictive value, indicates "how correct of the text mining approaches are". Recall, also known as sensitivity, indicates "how complete the approaches are". F-measure, which is calculated as the harmonic mean between precision and recall, is a summary score, which gives equal importance to precision and recall. The values of all of the above evaluation metrics fall in the range of [0…1]. A higher value indicates better performance. In general, state-of-the-art text mining systems can achieve an average F-score of at least 0.8 or more for the task of named entity recognition (Dai et al., 2019a; Devlin et al., 2019).

3. Results

A total of 500 discharge notes, including 15,271 sentences, were manually annotated to create the gold standard. Among them, 100 discharge notes were double annotated, yielding an average Cohen's kappa range from 0.61 to 0.96 (see Table 1). Generally, the consistency of the diagnosis of major depressive disorder, schizophrenia, and dementia was almost perfect and that of bipolar disorder and minor depression was also acceptable (ranged from 0.73 and 0.74). In terms of 9 depressive symptoms, the consistency was also almost perfect. However, the consistency for 4 dimensional functional impairments was less satisfactory, ranging from 0.61 to 0.74.

Based on the manual annotations from medical chart review, the distributions of clinical diagnoses, symptom profiles, and functional impairment in training and test dataset were shown in Table 2. Given the inconsistency between ICD-9 diagnostic codes and diagnoses from the medical chart review, the distribution of clinical diagnosis was slightly different from the initial sampling. Regarding symptom profiles, we found the depressive mood, sleep disturbance, and suicidality were the most common symptoms. Psychomotor retardation or agitation were the least frequent symptoms recorded (22.4%). Only 66.9% of discharge notes contained records related to functional impairment. The general or unspecific functional impairment was the most common dimension of functional impairment.

Using the diagnoses made from medical chart review as gold standard, the accuracy of the ICD-9 code clinical diagnoses in the claims database was calculated (see Table 3). The recall and precision for major depressive disorder were both 0.864. The accuracy of schizophrenia was substantial. The recall for dementia was 0.897 but precision was only 0.686. However, the precision and recall for bipolar disorder and minor depression were less satisfactory. The F-score for these two diagnoses were only 0.687 and 0.623, respectively.

The performance of dictionary- and machine learning-based approaches on extracting symptoms and function was reported on the sentence level on the test set[2] (see in Table 4). Overall, the performance on symptom identification was excellent. The precision of the 4 dimensions of functional impairment was also higher than 0.80; however, all of the recalls were less than or equal to 0.75. Compared with the dictionary-based approach, the machine-learning approach generally had better precisions but poorer recall. Both approaches to identify functional impairment were still not satisfactory.

We further examined whether we could use the number of depressive symptoms to diagnose major depressive disorder. Using the definition, major depressive disorders having 5 or more depressive symptoms by dictionary-based approach, we found the recall was 0.853, precision was 0.691, and F-score was 0.764. Using the number of depressive symptoms by machine learning approach, the recall, precision, and F-score was 0.846, 0.689, and 0.759, respectively. We further modified the criteria for 5 or more depressive symptoms with any domain of dysfunction; however, these modified criteria did not improve but worsened the recall and F-score (the recall, precision, and F-score using dictionary-based approach were 0.595, 0.703, and 0.645, respectively).

4. Discussion

Generally, this study had well-defined annotation guidelines. The inter-annotator consistency was satisfactory, and the generation of gold standard annotation was reliable. The accuracy of the ICD-9 diagnostic

[1] Please refer to supplementary Fig. 1 for the mathematical definition.

[2] The results of the 10 fold cross validation on the training set are available at the supplementary file Table 1 and 2.

Table 3
Accuracy of diagnostic code in claims records in NHIRD[a].

	ICD-9-CM code in claims data	Recall	Precision	F-score
Major depressive disorders	296.2 or 296.3	0.864	0.864	0.864
Bipolar affective disorder	296.0, 296.1, 296.4–296.8	0.739	0.642	0.687
Schizophrenia-spectrum disorder	295.x	0.735	0.818	0.774
Minor depressive disorders	300.4, 311, 309.0, 309.1	0.635	0.610	0.623
Dementia	290.x, 294.0, 294.1, 331.0, 331.1	0.897	0.686	0.778

The results of medical chart review as gold standard.

Table 4
The accuracy of dictionary-based and CRF-based approaches for symptoms and functional impairments.

	Training dataset						Test dataset					
	Dictionary-based approach			CRF-based approach			Dictionary-based approach			CRF-based approach		
	Recall	Precision	F-score	Recall	Precision	F-score	Recall	Precision	F-score	Recall	Precision	F-score
Symptomatology												
Depressive mood	0.982	0.961	0.972	0.980	0.972	0.976	0.963	0.963	0.963	0.959	0.974	0.966
Loss of interest	0.981	0.948	0.964	0.959	0.968	0.964	0.863	0.940	0.900	0.863	0.984	0.920
Sleep disturbance	0.989	0.972	0.981	0.984	0.974	0.979	0.940	0.981	0.960	0.934	0.981	0.957
Appetite change	0.974	0.958	0.966	0.964	0.971	0.968	0.943	0.954	0.949	0.943	0.976	0.960
Fatigue	0.992	0.974	0.983	0.989	0.974	0.981	0.984	1.000	0.992	0.967	1.000	0.983
Psychomotor retardation/ agitation	0.930	0.960	0.944	0.922	0.975	0.948	0.931	0.964	0.947	0.931	0.964	0.947
Poor concentration/ Indecisiveness	0.981	0.929	0.954	0.972	0.950	0.961	0.881	0.974	0.925	0.881	0.949	0.914
Negative thinking	0.995	0.942	0.968	0.985	0.965	0.975	0.961	0.970	0.966	0.951	0.970	0.960
Suicidality	0.993	0.911	0.950	0.988	0.961	0.974	0.988	0.941	0.964	0.980	0.969	0.975
Overall (micro-average)	0.985	0.948	0.966	0.980	0.969	0.975	0.954	0.961	0.958	0.948	0.974	0.961
Functions												
Occupational/academic	0.848	0.694	0.764	0.707	0.814	0.757	0.750	0.833	0.789	0.450	0.900	0.600
Social life/ interpersonal	0.800	0.933	0.861	0.811	0.933	0.868	0.629	0.815	0.710	0.686	0.800	0.738
Family	0.849	0.916	0.881	0.816	0.930	0.869	0.633	0.912	0.747	0.633	0.912	0.747
General /unspecific	0.964	0.926	0.945	0.932	0.937	0.935	0.750	0.971	0.846	0.727	0.970	0.831
Overall (micro-average)	0.872	0.884	0.878	0.836	0.917	0.875	0.682	0.894	0.774	0.649	0.897	0.753

code in discharge notes was generally acceptable, especially for the diagnosis of major depressive disorder and schizophrenia. Using dictionary-based and machine learning approaches to extracting depressive symptoms, the accuracy was almost perfect. Compared with the dictionary-based approach, the machine-learning approach generally had better precision but poor recall. Both approaches to identify functional impairment were still not satisfactory. Using the number of depressive symptoms to identify major depressive disorder, we found the recall and precision were not satisfactory.

4.1. Accuracy of ICD-9 diagnostic codes

To the best of our knowledge, there were no documents to validate the accuracy of major psychiatric illness in Taiwan's NHIRD. We found the accuracy of ICD-9 diagnostic codes for schizophrenia, and major depressive disorder were reliable. In addition, the diagnostic code for dementia had high recall (sensitivity); however, the precision was poor. It might be caused by some patients possibly did not have severe symptoms to fit diagnostic criteria of dementia. The accuracy of bipolar disorder were not satisfactory either. This might be due to the diagnostic codes, including pure manic, mixed episodes, bipolar II disorder, and unspecific bipolar disorder. The differential diagnosis between bipolar-spectrum disorder, atypical depressive disorder, and the manifestation of personality disorder is difficult based on chart review only. We also found the accuracy of minor depressive disorder was low. In addition to previously mentioned reasons, low accuracy was also attributed to some patients with schizophrenia or dementia having minor depression. It might be difficult to distinguish if the depressive symptoms were derived from underlying psychiatric illness or another independent depressive disorder. Finally, some clinical diagnoses were controversial. For example, antidepressant-induced manic episodes might be coded as major depressive disorder or unspecified bipolar disorder in the ICD-9 diagnostic code system. This inconsistency reflected the gray area in clinical judgement. Inconsistency in bipolar disorder and minor depression between inter-annotators was also obvious. Although we have established a clear annotation guideline, the Cohen's kappa was only 0.74 for minor depression and 0.73 for bipolar affective disorder.

4.2. Accuracy of symptoms and functional profiles

We found the accuracy of symptom identification was excellent. Most of the depressive symptoms were included in the dictionary, and the machine learning-based approach could exclude negation sentences effectively. However, the accuracy of identification for functional impairment was still less satisfactory. The judgement for functional impairment is somewhat subjective in nature. Thus, there was no clear cutoff point to distinguish whether there is functional impairment or not. The low inter-rater agreement also reflected the difficulty in identifying functional impairment. In addition, loss of job might be due to psychiatric illness or other causes, such as company bankruptcy or family issues. The presented approaches cannot clarify the underlying precipitating factors which might be described in Chinese or even be omitted. Furthermore, the description for functional impairment was more diverse than symptoms. Not all terms related to functional impairment could be included in the dictionary. We thought machine learning-based approach might be more flexible and have a better performance; however, the current results didn't support our assumption. More training data were needed to establish a more accurate model.

4.3. Accuracy of the diagnoses of major depressive disorders using text mining method

The results showed how with accurate diagnosis major depressive disorder through a text mining method, the recall was acceptable but

only fair in precision. Adding dysfunction criteria did not improve the recall because the functional profile could not be exactly identified using the current model. Theoretically, patients with major depressive disorder should have substantial functional impairment. However, the medical chart did not always describe the patient's functional status. Some patients with major depressive disorder were admitted to the hospital due to high suicidal risk even if they only had mild functional impairment. Therefore, functional impairment was not mentioned comprehensively. In current diagnostic criteria for major depressive disorder, ICD-9 or DSM-5, other requirements include duration of symptoms for more than two weeks and exclusion of secondary causes such as other psychiatric disorder, medical conditions, substance use, and bereavement. Further investigation is needed to develop a model that effectively used the text mining method to identify these related conditions.

5. Limitations

There were several limitations in this study. First, we used the discharge notes from a single university hospital. In Taiwan, clinicians must narratively note down their observations in the section of mental status examination of EHRs without the assistance of click tabs or check boxes. Most medical charts in medical centers or university hospitals are written in English. Nevertheless, it somehow may contain a few Chinese descriptions mixed with English descriptions. In particular, for psychiatric notes, clinicians may use Chinese descriptions to provide supplementary information because the use of the native language enable them to provide more appropriate and comprehensive information to reflect the status of patients. Although we excluded sentences containing Chinese descriptions in our corpora and analysis, for local hospitals, the code-mixed sentences involving Chinese and English may occur frequently or even recorded in purely Chinese. In addition, the format of narrative medical notes might vary across different hospitals. Therefore, our results might not be reproducible at other hospitals. In the future, we would like to include more medical charts from different hospitals as well as those sentences containing code-mixed descriptions for developing our text mining systems. Secondly, when using the number of depressive symptoms and /or functional impairment to identify major depressive disorder, the performance was still unsatisfactory. This study did not look at other medical conditions, substance use, bereavement, and other psychotic and manic symptoms. Thus, we could not use these symptoms or medical conditions to increase diagnostic accuracy. We are planning to extend the same approach to identify medical illness, and psychotic, manic, and cognitive symptom in the future. Thirdly, because the annotated corpora contained annotations for affirmative symptoms only, the machine-learning approaches could exploit surrounding context to exclude symptoms or functional impairments in negation descriptions. The results shown in Table 4 demonstrate that the overall precision of the machine learning-based methods outperformed the dictionary-based approaches by 0.003 ~ 0.033 indicating the effectiveness. There were several methods to deal with the issues of negation descriptions (Roque et al., 2011; Thomas et al., 2014). In the future, we will implement the aforementioned methods to identify negation scopes and encode the negative information as features in our model for comparison. Finally, we used the CRF approach, which is one of the best methods for named entity recognition. Whether or not using other machine learning algorithms could improve the performance needs further investigations.

5.1. Clinical and research implications

We have demonstrated that the proposed text mining approaches could identify depressive symptom profiles accurately. In term of clinical implications, text mining approaches have be found to improve the prediction of suicide (Velupillai et al., 2019) or treatment response (Carrillo et al., 2018). We believed that by integrating real world or even real time data like posts in social media with the data in routine clinical practice captured by EHRs can provide cues for future clinical decision support systems to improve the quality of clinical care. Taiwan's NHIRD now already includes EHRs if the privacy issue can be addressed in the near future, the proposed method can be used to identify more symptom profile like manic, psychotic, or cognitive symptoms and overcome current research study limitations.

6. Conclusion

In this study, we found the accuracy of ICD-9 diagnostic code in discharge notes was acceptable. These findings support the utilization of psychiatric diagnoses in Taiwan's NHIRD. We found that both dictionary-based and CRF approaches could identify depressive symptoms accurately. However, the functional level assessment using this approach remains unsatisfactory. More study sample size or using other machine learning approaches might be needed to improve accuracy.

CRediT authorship contribution statement

Chi-Shin Wu: Conceptualization, Data curation, Formal analysis, Writing - original draft. **Chian-Jue Kuo:** Conceptualization, Visualization, Writing - review & editing. **Chu-Hsien Su:** Conceptualization, Formal analysis, Visualization, Writing - review & editing. **Shi-Heng Wang:** Conceptualization, Formal analysis, Visualization, Writing - review & editing. **Hong-Jie Dai:** Conceptualization, Formal analysis, Writing - original draft, Visualization, Writing - review & editing.

Declaration of Competing Interest

All the authors have no conflict of interest.

Role of the Funding source

This study was supported by grants from the Ministry of Science and Technology, Taiwan (MOST-106-2314-B-002-105, MOST-107-2314-B-002-216 and MOST-106-2221-E-143-007-MY3). The funder had no role in study design, data collection/analysis, decision to publish, or preparation of the manuscript.

Acknowledgement

We thank the staff of the Department of Medical Research, National Taiwan University Hospital for the Integrated Medical Database (NTUH-IMD).

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.jad.2019.09.044.

Reference

Abbe, A., Grouin, C., Zweigenbaum, P., Falissard, B., 2015. Text mining applications in psychiatry: a systematic literature review. Int. J. Methods Psychiatr. Res. 25, 86–100.

Carrillo, F., Sigman, M., Slezak, D.F., Ashton, P., Fitzgerald, L., Stroud, J., Nutt, D.J., Carhart-Harris, R.L., 2018. Natural speech algorithm applied to baseline interview data can predict which patients will respond to psilocybin for treatment-resistant depression. J. Affect Disord. 230, 84–86.

Chen, Y.-.C., Yeh, H.-.Y., Wu, J.-.C., Haschler, I., Chen, T.-.J., Wetter, T., 2010. Taiwan's national health insurance research database: administrative health care database as study object in bibliometrics. Scientometrics 86, 365–380.

Chien, I.-.C., Chou, Y.-.J., Lin, C.-.H., Bih, S.-.H., Chou, P., 2004. Prevalence of psychiatric disorders among National Health Insurance enrollees in Taiwan. Psychiatr. Serv. 55, 691–697.

Dai, H.-.J., Jonnagaddala, J., 2018. Assessing the severity of positive valence symptoms in initial psychiatric evaluation records: should we use convolutional neural networks? PLoS ONE 13, e0204493.

Dai, H.-.J., Su, C.-.H., Wu, C.-.S., 2019a. Adverse drug event and medication extraction in electronic health records via a cascading architecture with different sequence labeling models and word embeddings. J. Am. Med. Inf. Assoc..

Dai, H.-.J., Syed-Abdul, S., Chen, C.-.W., Wu, C.-.C., 2015. Recognition and evaluation of clinical section headings in clinical documents using token-based formulation with conditional random fields. Biomed. Res. Int. 2015.

Dai, H.-.J., Wang, C.-.K., Chang, N.-.W., Huang, M.-.S., Jonnagaddala, J., Wang, F.-.D., Hsu, W.-.L., 2019b. Statistical principle-based approach for recognizing and normalizing microrna s described in scientific literature. Database.

Dai, H.-.J., Wu, C.-.Y., Tsai, R.T.-H., Hsu, W.-.L., 2014. Chapter 12: tText mining in biomedicine and healthcare, biological data mining and its applications in healthcare, pp. 325–372.

Devlin, J., Chang, M.-.W., Lee, K., Toutanova, K., 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Association for Computational Linguistics, Minneapolis, Minnesota, pp. 4171–4186.

Goodwin, R.D., Jacobi, F., Bittner, A., Wittchen, H.-.U., 2006. Epidemiology of mood disorders. The American Psychiatric Publishing Textbook of Mood Disorders, Arlington, VA.

Harpe, S.E., 2009. Using secondary data sources for pharmacoepidemiology and outcomes research. Pharmacotherapy 29, 138–153.

Jackson, R.G., Patel, R., Jayatilleke, N., Kolliakou, A., Ball, M., Gorrell, G., Roberts, A., Dobson, R.J., Stewart, R., 2017. Natural language processing to extract symptoms of severe mental illness from clinical text: the clinical record interactive search comprehensive data extraction (CRIS-CODE) project. BMJ Open 7, e012012.

Jensen, P.B., Jensen, L.J., Brunak, S., 2012. Mining electronic health records: towards better research applications and clinical care. Nat. Rev. Genet. 13, 395–405.

Kerr, W.T., Lau, E.P., Owens, G.E., Trefler, A., 2012. The future of medical diagnostics: large digitized databases. Yale J. Biol. M. 85, 363–377.

Lafferty, J., McCallum, A., Pereira, F., 2001. Conditional random fields: probabilistic models for segmenting and labeling sequence data. In: Proceedings of the eighteenth international conference on machine learning, ICML, pp. 282–289.

McCoy, T.H., Castro, V.M., Roberson, A.M., Snapper, L.A., Perlis, R.H., 2016. Improving prediction of suicide and accidental death after discharge from general hospitals with natural language processing. JAMA Psychiatry 73, 1064–1071.

Meystre, S.M., Savova, G.K., Kipper-Schuler, K.C., Hurdle, J.F., 2008. Extracting information from textual documents in the electronic health record: a review of recent research. Yearb. Med. Inform. 35, 128–144.

Perlis, R., Iosifescu, D., Castro, V., Murphy, S., Gainer, V., Minnier, J., Cai, T., Goryachev, S., Zeng, Q., Gallagher, P., 2012. Using electronic medical records to enable large-scale studies in psychiatry: treatment resistant depression as a model. Psychol. Med. 42, 41–50.

Roque, F.S., Jensen, P.B., Schmock, H., Dalgaard, M., Andreatta, M., Hansen, T., Søeby, K., Bredkjær, S., Juul, A., Werge, T., 2011. Using electronic patient records to discover disease correlations and stratify patient cohorts. PLoS Comput. Biol. 7, e1002141.

Schneeweiss, S., Avorn, J., 2005. A review of uses of health care utilization databases for epidemiologic research on therapeutics. J. Clin. Epidemio. 58, 323–337.

Sheehan, D., Harnett-Sheehan, K., Raj, B., 1996. The measurement of disability. Int. Clin. Psychopharmacol. 11, 89–95.

Smoller, J.W., 2018. The use of electronic health records for psychiatric phenotyping and genomics. Am. J. Med. Genet. B Neuropsychiatr. Genet. 177, 601–612.

Thomas, C.E., Jensen, P.B., Werge, T., Brunak, S., 2014. Negation scope and spelling variation for text-mining of danish electronic patient records. In: The 5th International Workshop on Health Text Mining and Information Analysis (Louhi)@ EACL 2014. Association for Computational Linguistics.

Velupillai, S., Hadlaczky, G., Baca-Garcia, E., Gorrell, G.M., Werbeloff, N., Nguyen, D., Patel, R., Leightley, D., Downs, J., Hotopf, M., 2019. Risk assessment tools and data-driven approaches for predicting and preventing suicidal behavior. Front. Psychiatry 10.

Wu, C.-.S., Shau, W.-.Y., Chan, H.-.Y., Lai, M.-.S., 2013. Persistence of antidepressant treatment for depressive disorder in Taiwan. Gen. Hosp. Psychiatry 35, 279–285.

Wu, C.-.S., Yu, S.-.H., Lee, C.-.Y., Tseng, H.-.Y., Chiu, Y.-.F., Hsiung, C.A., 2017. Prevalence of and risk factors for minor and major depression among community-dwelling older adults in Taiwan. Int. Psychogeriatr. 29, 1113–1121.

Wu, C.S., Shau, W.Y., Chan, H.Y., Lee, Y.C., Lai, Y.J., Lai, M.S., 2012. Utilization of antidepressants in Taiwan: a nationwide population-based survey from 2000 to 2009. Pharmacoepidemiol. Drug Saf. 21, 980–988.