Applying deep neural networks to unstructured text notes in electronic medical records for phenotyping youth depression

Joseph Geraci,[1,2,3] Pamela Wilansky,[1] Vincenzo de Luca,[1] Anvesh Roy,[1] James L Kennedy,[1] John Strauss[1,3,4]

[1]Centre for Addiction and Mental Health, Toronto, Ontario, Canada; [2]Department of Pathology and Molecular Medicine, Queen's University, Kingston, New York, Canada; [3]Shannon Centennial Informatics Lab, Centre for Addiction and Mental Health, Toronto, Ontario, Canada; [4]Cundill Centre for Child and Youth Depression, Centre for Addiction and Mental Health, Toronto, Ontario, Canada
Correspondence to John Strauss, Cundill Centre for Child and Youth Depression, Centre for Addiction and Mental Health, Toronto, ON M6J1H4, Canada; john.strauss@camh.ca

ABSTRACT

Background We report a study of machine learning applied to the phenotyping of psychiatric diagnosis for research recruitment in youth depression, conducted with 861 labelled electronic medical records (EMRs) documents. A model was built that could accurately identify individuals who were suitable candidates for a study on youth depression.

Objective Our objective was a model to identify individuals who meet inclusion criteria as well as unsuitable patients who would require exclusion.

Methods Our methods included applying a system that coded the EMR documents by removing personally identifying information, using two psychiatrists who labelled a set of EMR documents (from which the 861 came), using a brute force search and training a deep neural network for this task.

Findings According to a cross-validation evaluation, we describe a model that had a specificity of 97% and a sensitivity of 45% and a second model with a specificity of 53% and a sensitivity of 89%. We combined these two models into a third one (sensitivity 93.5%; specificity 68%; positive predictive value (precision) 77%) to generate a list of most suitable candidates in support of research recruitment.

Conclusion Our efforts are meant to demonstrate the potential for this type of approach for patient recruitment purposes but it should be noted that a larger sample size is required to build a truly reliable recommendation system.

Clinical implications Future efforts will employ alternate neural network algorithms available and other machine learning methods.

BACKGROUND

Recruitment of clinical research participants is routinely disappointing with traditional methods failing to identify up to 60% of possible participants.[1 2] Substantial institutional and departmental expense is incurred and little scientific benefit is gained by low-enrolling studies, which made up 31% of the studies at one institution over a single year.[3] Evidence indicates that dramatic increases, up to fourfold, in recruitment are possible with automated recruitment.[4 5] Such approaches are scalable in research settings—some research institutions have linked, with proper privacy safeguards in place, electronic medical records (EMRs) data together with genotype data for discovery in large-scale databases and virtual cohorts.[6] EMR analysis has been suggested as a useful means of measuring outcomes and defining disorder subpopulations.[7]

Research inclusion criteria in psychiatry often use diagnosis. Structured diagnosis codes are sometimes available in EMR clinical notes, but are frequently missing. Procedures such as natural language processing (NLP) and machine learning (ML) methods have been used to extract clinical information from EMRs' unstructured text. The eMERGE group has used NLP extensively, with improved accuracy of their phenotyping algorithms[8]—examples include determining colorectal cancer screening status[9] and diagnosing rheumatoid arthritis.[10] NLP methods have also been applied to EMRs to boost the efficiency of manual chart abstraction for breast cancer recurrence with 92% sensitivity and 96% sensitivity.[11] More recently, NLP has been used to identify adverse drug events including extrapyramidal side effects in psychiatric patients[12] and to phenotype children at risk for Kawasaki disease in emergency department notes.[13] In another investigation, ML classification algorithms were used to identify rheumatoid arthritis patients with coronary artery disease—NLP was used to detect features in clinical notes and outperformed features selected by experts.[14]

In recent years, mental health researchers in South London and Maudsley NHS Trust have begun using EMRs for research recruitment.[15 16] For phenotyping, a small number of studies have focused on extracting depression diagnoses from unstructured EMR text. Early work on diabetes outpatient records compared diagnosis by coding versus by NLP—NLP improved detection of depression diagnosis by almost a third.[17] Researchers developed and tested NLP in patients with a billing code of major depressive disorder to characterise symptom remission and treatment resistance, and found that adding NLP resulted in higher area under receiver operating characteristic curve than billing data only (0.85–0.88 vs 0.54–0.55) for classification of mood state.[18] NLP has been used for categorisation of publicly available Twitter data into several mental health diagnoses, including depression and bipolar disorder.[19] A later publication identified patients with depression from free-text discharge summaries: a combination of NLP and ML algorithms was used, with the best performance coming from Medical Text Extraction, Reasoning and Mapping System's[20] knowledge-based decision tree method, yielding an F-measure of 89.6%.[21] To summarise the rationale for extracting diagnosis inclusion criteria from unstructured EMR using NLP and ML, it is known that research recruitment supported by automation is more successful; further, that NLP and ML can be useful for information extraction from unstructured text notes, and that such methods have been applied with some degree of success to depression-related phenotypes.

Since structured diagnosis codes had limited availability in our EMR, we used NLP and ML on EMR notes data to extract our diagnostic inclusion criteria, in this case Diagnostic and Statistical Manual of Mental Disorders (DSM)-IV depression diagnoses, to support recruitment for a cognitive-genomic study of youth depression. This article summarises the NLP and ML processes and results. The core purpose of this report is to present a model that identifies youth with a depression diagnosis and without specific exclusion comorbidities—a model evaluated via cross-validation and an independent test data set, based on deep neural networks.

METHODS

Deidentification

Clinical documents commonly contain sensitive information about individuals; accordingly, in this Research Ethics Board-approved study, we deidentified the corpus to remove personal identifying information (PII). For this task, we created a suite of programs that made use of the freely available Perl-based

software package De-id V.1.1.[22] With these programs, we performed the following tasks:

- Inserted the necessary text tags at the beginning and end of each document so that it would be recognised by De-id
- Converted all the documents to. txt files so that the format conformed to the De-id specifications
- Looped the De-id algorithm over the whole document corpus to remove PII and thereby coded them
- Translated the coded documents into. csv files to get them ready for the training and testing protocols of our supervised learning methods.

Clinical documents for youth psychiatric patients often contain important free-text information regarding a patient's lifestyle, activities and clinical impressions, including diagnosis; however, often a discrete/structured diagnosis is missing. Our aim is to use NLP and ML to identify our phenotype of interest: youth patients ages 12–18 with DSM-IV defined Major Depressive Disorder or Dysthymic Disorder. Exclusion criteria included schizophrenia, bipolar disorder, autism, epilepsy, personality disorder, developmental delay and traumatic brain injury. From our EMR, we obtained a corpus consisting of 861 physician documents on 366 patients ages 12–18 years for a 6-month period, and deidentified them as noted above; the documents were predominantly progress notes, with character counts ranging from 533 to 24 803 (without spaces) and a median character count of ~4300. Almost all the child and adolescent patient population at the Centre for Addiction and Mental Health is outpatient in nature. Of the corpus, 60% of documents were on females.

This specific phenotyping effort requires a model that is capable of rejecting documents of individuals manifesting the exclusion criteria, but simultaneously requires a model capable of including suitable participants' documents.

We used two distinct approaches for this task: (i) a brute force search method based on specific terms stored in dictionaries and (ii) an ML protocol known as neural networks.[23] Both methods relied on NLP packages/methods available through the R programming language: (wordnet, RKEA, tm, SDMTools). These methods take the EMR clinical document corpus and translate it into a structure that allows a machine to efficiently compute the frequency structure of the words used in each document; the term frequencies are recorded in the Document Term Matrix (DTM) (see table 1).[24] The NLP methods assure that only meaningful words are used by performing functions such as stripping grammatical articles from the text. For the neural network algorithms we present, the DTM (table 1) is the data that is being used by the brute force and neural network algorithms to find potential study participants.

The DTM records information about how often a word is encountered in a document but, in our case, it also includes information about how often it is found in the full document set. In table 1, each row represents a document from a patient, and the columns are words. We use the tf-idf computation (term frequency-inverse document frequency) which captures information about how often words show up in a document but it also adjusts for the effect of high-frequency words. A word such as 'the' or 'diagnosis' may add little information and their influence appropriately minimised. This approach allows our algorithms to focus on terms that 'stand out'.

We used a supervised learning paradigm—we applied labels to the documents, for the algorithm to learn from, that is, suitable research participant candidates or not. To label a data set of EMR documents, two fully qualified psychiatrists (AR and JS) independently annotated 900 patient documents, which resulted in 861, after omitting 39 unclassifiable documents. Agreement between annotators was 98% based on 100 documents annotated by both psychiatrists. Of this set, there were 126 documents that were classified as belonging to patients that would meet the above inclusion criteria, and not meet any criteria for exclusion.

Table 1 Example of the Document Term Matrix data used to train our models

Patient	Frequency of 'responded'	Frequency of 'responding'	Frequency of 'response'	Frequency of 'restless'
1	0	0	0.014249584	0.02089797
3	0	0	0	0.000758773
4	0	0.01683432	0	0
5	0.00742017	0	0	0

Each column provides a frequency measure for the given word. The most predictive words make their way into the neural network model.

Box 1 Positive dictionary: a dictionary of terms to help identify depression

- Major depressive disorder
- Major depression
- Double depression
- Dysthymic disorder
- Persistent depressive disorder
- Depressive disorder
- Depression
- MDD

Brute force

The brute force method attempted to identify suitable participants by looking for certain keywords that would cause the machine to either reject or accept a particular document as belonging to a patient that would make a suitable participant. The method used a *positive dictionary* (PD) for inclusion criteria diagnoses and a *negative dictionary* (ND) for exclusion criteria diagnoses, along with a subalgorithm that looked at words that come before or after the specific PD or ND words. The words in the PD would increase a score of acceptance for an EMR and words in the ND would have the opposite effect. The subalgorithm that looked at the surrounding words would decide if the words in either the PD or ND should be negated, for example, 'it is unlikely Samantha has major depressive disorder' (box 1, 2).

Neural networks

Neural networks have received more attention in recent years mainly due to advancements in methodology and access to affordable powerful computation platforms. The popularity of what are known as deep neural networks stems from their ability to robustly identify images.[23] Advances in the last decade have been very impressive for image classification[25] in addition to NLP.[26] We decided to use the deep learning paradigm (DL) because of the expected non-linear relationships that exist between the language used within the EMRs and DL's ability to learn several representations simultaneously for distinguishing between suitable participants and not. Deep neural networks encode information to make a prediction in a way that uses several layers of information by making non-linear inferences between the variables—in this

Box 2 Negative dictionary: terms that would indicate that someone is not suitable

- Bipolar disorder
- Schizophrenia
- Bipolar II
- Bipolar I
- Traumatic brain injury
- Developmental delay
- Personality disorder
- Borderline personality disorder
- Hypomanic
- Autism
- Epilepsy

Figure 1 The more sensitive DL1 method was initially applied. Following DL1, the more specific DL0 model was then used on the documents selected with DL1. DL, deep learning paradigm.

case the frequencies of used words and co-occurrences of used words. If the two groups of patients were linearly separable, then such a sophisticated method would not be necessary, and indeed, for a subset of patients this is true as some documents contain clear diagnoses. However, we are using our ability to move beyond a simple search, as was implemented in the brute force approach, via deep neural networks.

We used an R language implementation of the H2O.ai package, which includes a multilayer, feedforward deep neural network for the purpose of prediction under a supervised protocol. For more details, please refer to H2O open-source software.[25] We used the 861 documents as two main data sets of 758 and 103 documents, respectively: (i) a training data set consisted of 758 documents, with 101 suitable participants and 657 unsuitable participants; and (ii) a test data set consisted of 103 documents with 25 of them belonging to suitable participants and 78 unsuitable participants. Our training phase resulted in two models that we shall refer to as DL1 and DL0: DL1 is capable of accurately identifying suitable participants but is poor at identifying unsuitable participants and DL0 has the opposite capabilities, as it is very accurate at correctly rejecting participants. Test statistics will be provided in the 'Results' section.

These two models were combined into a single protocol that takes patient documents as input and provides a list of patients for inclusion in our study. We shall refer to this model as DL1+0, which works by first passing a new group of patients to evaluate through DL1. The DL1+0 method will then provide a label for each patient by evaluating the corresponding document. At this stage DL1+0 will capture a good proportion of the true candidates but it will likely label many unsuitable candidates as suitable, so it then passes this new smaller list of documents through DL0, which then removes documents of patients that it deems to be unsuitable, thus ending up with a list of proposed true potential participants. See figure 1 for a synopsis of this process.

FINDINGS

For information regarding De-id performance, please refer to Neamatullah and colleagues.[22] We customised De-id for our purposes to include a larger set of proper nouns including names and regional institutional names for more optimal deidentification. The performance statistics presented here relate to individual documents, not patients.

The brute force method was capable of performing well on some data sets but it did not generalise well. On some independent test sets (training on 761 and testing on another 100 documents), we achieved the following: sensitivity=80%, specificity=88%, with a total proportion correct of 86%. However, this model performed poorly in general, that is, when evaluated via a cross-validation. More specifically, performance on some of the leave-out sets was poor with a sensitivity and specificity around 50% and thus not predictive at all.

We trained two neural networks (DL0 and DL1) and combined them to construct an aggregate predictor (DL1+0). We first report the topologies of the two component deep neural network models and then their independent performances, and finally, the performance of DL1+0.

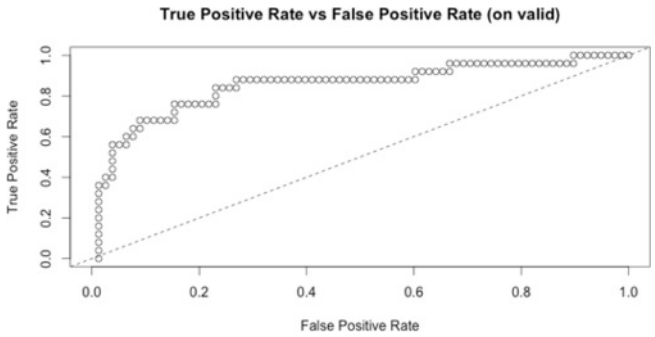

Figure 2 A typical receiver operating characteristic (ROC) curve for DL0 models derived from a fivefold cross validation. The reason the area under the ROC (AUC) curve is relatively high compared with the AUC for DL1 is because there are a large number of true 0s captured by this model. DL, deep learning paradigm.

DL0 was trained with 758 labelled documents: 657 documents that belonged to patients annotated as unsuitable and 101 that belonged to suitable patients. The input layer had 758 nodes (not related to the 758 documents; 758 is the number of input variables for DL0). The three hidden rectifier layers each have 200 nodes (we experimented with tanh layers and with several other topologies including a decreasing number of nodes, and more layers with no significant improvements), and the output layer used softmax so that there were two outputs, being a 0 or 1, that is, reject or accept. DL1's input layer had 102 input nodes, but was trained with 100 0s and 101 1s (figures 2, 3).

In order to evaluate our models, we used a fivefold cross-validation (performance was stable over other cross-validations ranging from 5-fold to 20-fold), and we performed an independent data test set evaluation. Cross-validation is a standard practice, which theoretically determines how generalisable our models are—a protocol is used which leaves out a data set for testing, trains on the complement of the data and repeats this a number of times to generate statistics regarding sensitivity and specificity. The performance of each of these models is given in table 2 and table 3.

One can compute the specificity and sensitivity from the tables above. For DL0, the specificity is 97% and the sensitivity is 44.5%. In contrast for DL1, the specificity is 53% but the sensitivity is 89%. This means that we have one model that consistently performs well when classifying 0s and another model that performs well when classifying 1s. By experimenting with the topology of the neural network, it was possible to trade in a loss of specificity for DL0 to gain some sensitivity. It is worth mentioning that a model similar to DL0, which we shall call DL0_2, was trained that performed quite well in general.

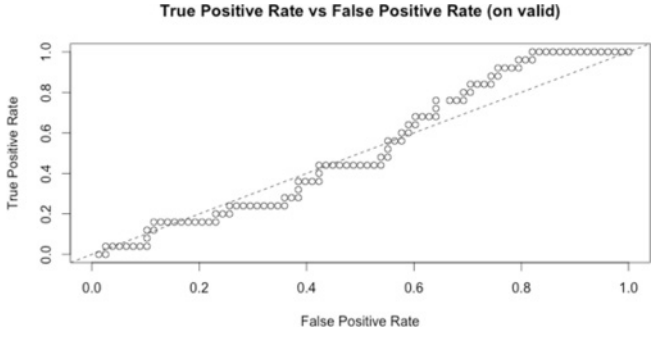

Figure 3 A typical receiver operating characteristic (ROC) curve for DL1 models derived from a fivefold cross-validation. The number of true 0s and true 1s in the data set used to train DL1 is balanced and thus the area under the ROC curve is quite poor despite the fact that this model is excellent at predicting true 1s. DL, deep learning paradigm.

Table 2 Performance of DL0 considering a fivefold cross-validation

	Predicted 0s	Predicted 1s
True 0s	639	18
True 1s	56	45

Sensitivity 44.5%; specificity 97%.

Note that it performs very well with rejecting unsuitable patients accurately, but it does not perform well with predicting suitable participants (the true 1s).

DL, deep learning paradigm.

It had a specificity of 87% and a sensitivity of 75%. It was trained and tested on the same data as DL0 via a similar cross-validation process (refer Table 4).

As described, we used cross-validation to produce and tune a set of models that we combined into a set of DL1+0 models. For replication, we tested these models on a second, completely separate, independent test set of 103 documents (which included 25 true candidates, ie, documents labelled as 1) that were not included with the original 758 documents mentioned above. The DL1+0 algorithm yields as output a set of documents that correspond to patients that it considers suitable participants; next we report on how it performed on the second independent test set of 103 documents.

- The sets of patients that DL1+0 identified as suitable participants ranged depending on the training of the neural networks. Many of the models generated returned predictions of 15 documents of which 13 were correct, giving a positive predictive value (precision) of 87%. Another set of models returned predictions for 22 out of the 25 possible suitable participants, 77% (ie, 17/22) of which were correctly identified as suitable, in terms of precision (positive predictive value). In practice, one could choose a model that would reveal many suitable candidates accurately, but that would miss many possible patients. Alternatively, one could use a model that returned more suitable candidates but it would include some patients that would not be suitable Table 5.
- An actual output example is given here for two of the these models:
  - Input: 103 documents, 25 of which are annotated as suitable participants.
  - Output of DL1+0 (called suitable by DL1+0) = (41,43,44,45,46,47,48,55,56,57,58,60,62,66,67,70,72,73,74,75,77,99). Of these 22 documents, 62,66,67,75 and 77 were not annotated as suitable, which means that the output returned 17/22 (77%) correct calls.
  - Output of DL1+0 (called suitable by DL1+0_short) = (41,44,45,47,48,57,58,60,62,70,72,74,77,99,102). Of these 15 documents, 62 and 102 were not annotated as suitable, which means that 13/15 were correct calls.
- DL1+0 is excellent at rejecting patients correctly with a worst-case score of 90% specificity, which occurs when sensitivity is 68%.
- Though statistically this model appears very similar to the single-shot neural network model DL0_2, the user can be more certain of the reliability of the output list of recommended patients due to an increase in precision (positive predictive value). After several tests, DL1+0 consistently returns lists that are more conservative but more precise than DL1, DL0 or DL0_2 alone.

Table 3 Performance of DL1 considering a fivefold cross-validation

	Predicted 0s	Predicted 1s
True 0s	47	53
True 1s	11	90

Sensitivity 89%; specificity 53%.

In contrast to model DL0, this model is excellent at accurately predicting participants (true 1s) but is poor at rejecting inappropriate patients.

DL, deep learning paradigm.

Table 4 Performance of DL0_2 considering a fivefold cross-validation

	Predicted 0s	Predicted 1s
True 0s	570	87
True 1s	25	76

Sensitivity 75%; specificity 87%.

DL, deep learning paradigm.

DISCUSSION

To summarise, we deidentified a corpus of EMR documents from a set of patients, annotated it using a set of inclusion and exclusion criteria, and used brute force and deep neural network approaches to phenotype potential research participants. Performance of the brute force method was inconsistent. We constructed a recommendation system by first training two deep neural networks, one that accurately recognises patients who are not suitable and another that accurately recognises patients who are suitable. We combined the two deep neural network models into a single model to augment a researcher's ability to recruit suitable participants. By missing many potential participants, we have found that this algorithm can return document lists that are up to 87% accurate. This was validated on an independent test set after tuning each component with a fivefold cross-validation protocol.

The current investigation has several limitations. The most important potential limitation is the phenotype itself: in the DSM-5 field trials, the kappa for Major Depressive Disorder was 0.28 for both adult and child versions.[27] Further, we were recently reminded that Major Depressive Disorder is an index of something and that we should not take an index of something as the thing itself.[28] Psychiatric symptoms have successfully been extracted from EMR data on patients with serious mental illness,[29] and this may be an alternative approach; to improve on a symptom-based phenotyping method, a network/complex dynamic system model may also be informative.[30]

We were not able to make use of structured diagnosis codes as are commonly available in most EMRs. It may be argued that discrete DSM or Systematized Nomenclature of Medicine (SNOMED) codes being available would render our deep neural network approach unnecessary; however, published evidence suggests that NLP/ML methods improve information extraction tasks in non-psychiatric phenotypes[9 10 12 13 31] and depression.[17 18 21]

The poor performance of the brute force method led us to abandon this approach. The suspected reason for this fluctuation in performance is based on the understanding that the content of the EMR documents could vary substantially. Some documents have clear diagnoses, while others have clear narrative, and then others would be too ambiguous for the brute force approach to capture reliable information. Improving the PD and ND may help, but if these terms are too comprehensive, it will limit the types of patients that are recommended. An immediate step to improve this approach would be to apply some Bayesian methods—probabilistic methods that can capture a distribution of responses and make the procedure more flexible to variation.

Several supervised ML techniques are available and we used feedforward deep neural networks trained directly on document term matrices. It is important to note that we attempted to use singular value decomposition

Table 5 Performance of DL1+0 considering a fivefold cross-validation

	Predicted 0s	Predicted 1s
True 0s	73	5
True 1s	8	17

Sensitivity 93.5%; specificity 68%; positive predictive value (precision) 77%.

At first it appears that there is not a significant improvement obtained via this model but the user can be more certain that the output recommended candidates are more reliable than DL1 or DL0 alone.

DL, deep learning paradigm.

techniques to reduce noise but in our case it reduced the performance of our models. We should mention however that our algorithms already reduce noise by only considering words that occur above some threshold. Our modest sample size is the greatest contributor to the low sensitivity of our results. We used neural networks because we wished to experiment with a method that is truly generalisable for our task—our EMR documents are a heterogeneous and complex data set, representing several distinct psychiatric patient populations—a larger corpus may have yielded more accurate results. In total, there were 861 documents on 366 patients for the 6-month period. This limitation reminds us to treat this effort as a proof of concept. However, our models did replicate on a fully independent data set, suggesting our methods have some merit. In the future, we will use a larger training set in addition to a more powerful variant of neural networks known as recursive deep networks which have shown promise for natural language efforts.[32] Future experiments will involve other ML techniques such as gradient boosting.[33]

Acknowledgements The authors would like to acknowledge the support of Mr Chris Wakefield.

Funding University of Toronto McLaughlin Centre, grant number: MC 2014-18. This work was supported by a McLaughlin Accelerator Grant in Genomic Medicine (PW, JS).

Competing interests None declared

Provenance and peer review Not commissioned; externally peer reviewed.

OPEN ACCESS

doi:10.1136/eb-2017-102688

Received 10 April 2017; Revised 12 June 2017; Accepted 21 June 2017

REFERENCES

1. **Carlson RW,** Tu SW, Lane NM, *et al*. Computer-based screening of patients with HIV/AIDS for clinical-trial eligibility. *Online J Curr Clin Trials* 1995;**Doc No 179**:179.
2. **Fink E,** Kokku PK, Nikiforou S, *et al*. Selection of patients for clinical trials: an interactive web-based system. *Artif Intell Med* 2004;**31**:241–54.
3. **Kitterman DR,** Cheng SK, Dilts DM, *et al*. The prevalence and economic impact of low-enrolling clinical studies at an academic medical center. *Acad Med* 2011;**86**:1360–6.
4. **Schmickl CN,** Li M, Li G, *et al*. The accuracy and efficiency of electronic screening for recruitment into a clinical trial on COPD. *Respir Med* 2011;**105**:1501–6.
5. **Weng C,** Batres C, Borda T, *et al*. A real-time screening alert improves patient recruitment efficiency. *AMIA Annu Symp Proc* 2011;**2011**:1489–98.
6. **Salem RM,** Pandey B, Richard E, *et al*. The VA hypertension Primary Care Longitudinal Cohort: electronic Medical records in the post-genomic era. *Health Informatics J* 2010;**16**:274–86.
7. **Simon GE,** Perlis RH. Personalized medicine for depression: can we match patients with treatments? *Am J Psychiatry* 2010;**167**:1445–55.
8. **Denny JC.** Surveying recent themes in translational bioinformatics: big Data in EHRs, omics for drugs, and Personal Genomics. *Yearb Med Inform* 2014;**9**:199–205.
9. **Denny JC,** Peterson JF, Choma NN, *et al*. Extracting timing and status descriptors for colonoscopy testing from electronic medical records. *J Am Med Inform Assoc* 2010;**17**:383–8.
10. **Carroll RJ,** Thompson WK, Eyler AE, *et al*. Portability of an algorithm to identify rheumatoid arthritis in electronic health records. *J Am Med Inform Assoc* 2012;**19**:e162–e169.
11. **Carrell DS,** Halgrim S, Tran DT, *et al*. Using natural language processing to improve efficiency of manual chart abstraction in research: the case of breast Cancer recurrence. *Am J Epidemiol* 2014;**179**:749–58.
12. **Iqbal E,** Mallah R, Jackson RG, *et al*. Identification of adverse drug events from Free text electronic patient Records and Information in a large Mental Health Case Register. *PLoS One* 2015;**10**:e0134208.
13. **Doan S,** Maehara CK, Chaparro JD, *et al*. Pediatric Emergency Medicine Kawasaki Disease Research Group. building a natural language processing Tool to identify patients with high clinical suspicion for Kawasaki Disease from Emergency Department Notes. *Acad Emerg Med* 2016;**23**:628–36.
14. **Yu S,** Liao KP, Shaw SY, *et al*. Toward high-throughput phenotyping: unbiased automated feature extraction and selection from knowledge sources. *J Am Med Inform Assoc* 2015;**22**:993–1000.
15. **Callard F,** Broadbent M, Denis M, *et al*. Developing a new model for patient recruitment in mental health services: a cohort study using Electronic Health Records. *BMJ Open* 2014;**4**:e005654.
16. **Patel R,** Oduola S, Callard F, *et al*. What proportion of patients with psychosis is willing to take part in research? A mental health electronic case register analysis. *BMJ Open* 2017;**7**:e013113.
17. **Fischer LR,** Rush WA, Kluznik JC, *et al*. Abstract C-C1-06: Identifying depression among diabetes patients using natural language processing of office notes. *Clin Med Res* 2008;**6**:125–6.
18. **Perlis RH,** Iosifescu DV, Castro VM, *et al*. Using electronic medical records to enable large-scale studies in psychiatry: treatment resistant depression as a model. *Psychol Med* 2012;**42**:41–50.
19. **Coppersmith G,** Dredze M, Harman C. Quantifying Mental Health signals in Twitter. *ACL Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality* 2014:51–60 http://www.aclweb.org/anthology/W14-3207 (accessed June 2017) .
20. **Zhou L,** Plasek JM, Mahoney LM, *et al*. Using medical text extraction, reasoning and mapping system (MTERMS) to process medication information in outpatient clinical notes. *AMIA Annu Symp Proc* 2011;**2011**:1639–48 http://www.ncbi.nlm.nih.gov/pubmed/22195230.
21. **Zhou L,** Baughman AW, Lei VJ, *et al*. Identifying patients with depression using Free-text clinical documents. *Stud Health Technol Inform* 2015;**216**:629–33.
22. **Neamatullah I,** Douglass MM, Lehman LW, *et al*. Automated de-identification of free-text medical records. *BMC Med Inform Decis Mak* 2008;**8**:32.
23. **Hinton GE,** Osindero S, Teh YW. A fast learning algorithm for deep belief nets. *Neural Comput* 2006;**18**:1527–54.
24. **Awati K.** A gentle introduction to text mining using R. *Web Blog 'Eight to Late'* 2015 https://eight2late.wordpress.com/2015/05/27/a-gentle-introduction-to-text-mining-using-r/ (accessed 08 apr 2017).
25. **Gulshan V,** Peng L, Coram M, *et al*. Development and Validation of a Deep Learning Algorithm for detection of Diabetic Retinopathy in retinal fundus photographs. *JAMA* 2016;**316**:2402–10.
26. **Sarikaya R,** Hinton GE, Deoras A. Application of Deep Belief Networks for Natural Language Understanding. *IEEE/ACM Trans Audio Speech Lang Process* 2014;**22**:778–84.
27. **Freedman R,** Lewis DA, Michels R, *et al*. The initial field trials of DSM-5: new blooms and old thorns. *Am J Psychiatry* 2013;**170**:1–5.
28. **Kendler KS.** The Phenomenology of Major depression and the Representativeness and Nature of DSM Criteria. *Am J Psychiatry* 2016;**173**:771–80.
29. **Jackson RG,** Patel R, Jayatilleke N, *et al*. 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* 2017;**7**:e012012.
30. **Cramer AO,** van Borkulo CD, Giltay EJ, *et al*. Major depression as a complex Dynamic System. *PLoS One* 2016;**11**:e0167490.
31. **Ford E,** Carroll JA, Smith HE, *et al*. Extracting information from the text of electronic medical records to improve case detection: a systematic review. *J Am Med Inform Assoc* 2016;**23**:1007–15.
32. **Zazo R,** Lozano-Diez A, Gonzalez-Dominguez J, *et al*. Language identification in short utterances using long Short-Term memory (LSTM) Recurrent neural Networks. *PLoS One* 2016;**11**:e0146917.
33. **Crick C,** Ambati S. H2O – Open source software for big data analysis. http://www.h2o.ai. - http://docs.h2o.ai/h2o/latest-stable/h2odocs/booklets/DeepLearningBooklet.pdf?_ga=1.228385053.122063005.1471206028 (accessed 08 Apr 2017).