Natural language processing for structuring clinical text data on depression using UK-CRIS

Nemanja Vaci [1], Qiang Liu,[1] Andrey Kormilitzin,[1] Franco De Crescenzo,[1,2] Ayse Kurtulmus,[1,3] Jade Harvey,[2] Bessie O'Dell,[1] Simeon Innocent,[1] Anneka Tomlinson,[1] Andrea Cipriani [1,2] Alejo Nevado-Holgado[1,4,5]

[1]Department of Psychiatry, University of Oxford, Oxford, UK
[2]Research and Development, Oxford Health NHS Foundation Trust, Oxford, UK
[3]Department of Psychiatry, Istanbul Medeniyet University Goztepe Research and Training Hospital, Istanbul, Turkey
[4]Big Data Institute, University of Oxford, Oxford, United Kingdom
[5]Artificial intelligence, Akrivia Health, Oxford, United Kingdom

Correspondence to
Dr Nemanja Vaci, Department of Psychiatry, University of Oxford, Oxford, Oxfordshire OX3 7JX, UK; nemanja.vaci@psych.ox.ac.uk

Received 29 November 2019
Accepted 6 January 2020

To cite: Vaci N, Liu Q, Kormilitzin A, *et al*. *Evid Based Ment Health* 2020;**23**:21–26.

ABSTRACT

Background Utilisation of routinely collected electronic health records from secondary care offers unprecedented possibilities for medical science research but can also present difficulties. One key issue is that medical information is presented as free-form text and, therefore, requires time commitment from clinicians to manually extract salient information. Natural language processing (NLP) methods can be used to automatically extract clinically relevant information.

Objective Our aim is to use natural language processing (NLP) to capture real-world data on individuals with depression from the Clinical Record Interactive Search (CRIS) clinical text to foster the use of electronic healthcare data in mental health research.

Methods We used a combination of methods to extract salient information from electronic health records. First, clinical experts define the information of interest and subsequently build the training and testing corpora for statistical models. Second, we built and fine-tuned the statistical models using active learning procedures.

Findings Results show a high degree of accuracy in the extraction of drug-related information. Contrastingly, a much lower degree of accuracy is demonstrated in relation to auxiliary variables. In combination with state-of-the-art active learning paradigms, the performance of the model increases considerably.

Conclusions This study illustrates the feasibility of using the natural language processing models and proposes a research pipeline to be used for accurately extracting information from electronic health records.

Clinical implications Real-world, individual patient data are an invaluable source of information, which can be used to better personalise treatment.

INTRODUCTION

Depression is one of the major causes of global disease burden, with approximately 350 million people affected worldwide.[1] Randomised control trials (RCTs) generally support the therapeutic effect of antidepressants,[2] which are routinely prescribed to treat major depressive disorder.[3] However, RCTs often focus on a select group of patients without medical or psychiatric comorbidities, and who are closely followed for a short period of time. Routinely collected observational data, such as electronic health records (EHRs), can complement the clinical picture by providing us with the relevant information about the efficacy of treatments in real world.[4]

Large data have become increasingly used in many scientific areas of research, as it enables the investigation of complex behaviours[5] measured over time. This approach is also prevalent in medical sciences, with the development of algorithms capable of personalising treatments for mental health disorders. Such approaches to treatment have aimed at increasing both prognostic and diagnostic accuracies.[6] This approach has materially improved recently by making large collections of EHRs more ubiquitous and accessible for clinical research.[7] EHRs cover a wide variety of longitudinal information, from medication prescriptions to environmental variables.[8] Therefore, enabling the investigation of complex interactions between treatment effects and auxiliary variables, such as information on symptoms or number of previous episodes.

The major challenge in using EHRs is that 80% of medical information is recorded in the form of natural text, as opposed to coded data.[9] This makes the extraction of the information problematic, as the manual review of EHRs requires an extensive time commitment from numerous highly skilled professionals who are trained to identify and interpret the information of interest. An algorithmic approach results in comparable accuracy to manual approaches,[10] while improving efficacy and cost-effectiveness. Previous studies have shown the utility of using natural language processing (NLP) models when extracting information from EHRs on psychotic[11] and suicidal behaviours.[12] However, to reach an acceptable level of accuracy and flexibility, NLP models also require input from trained medical staff in the form of annotated medical documents. Statistical learning models infer relations from data but require a certain level of data preparation and structuring to be able to do so.[13] This structuring is performed by adding supplemental information to medical texts, where clinicians highlight spans of text that detail and describe medical concepts of interest. The amount of input required is considerably smaller in comparison with the manual reviews, whereas the utilisation of the state-of-the-art NLP procedures further reduces the need for in-depth annotation process.[14]

Direct involvement is required from medical practitioners when utilising EHRs. Their first-hand experience as data generators (ie, both coded fields and text are generated by them) is essential in deciding how relevant medical information is going to be extracted from the EHRs. The development and usage of NLP methods aim to alleviate clinicians' involvement when screening EHRs. In this study, we develop a procedure to extract and structure raw medical information from EHRs. We use a case study (see the next section) to illustrate the complete research pipeline. In the first step of the study, we define the types of variables that we are aiming to extract from EHRs and illustrate the annotating process. In the second step, we show how we develop NLP models. We illustrate the process of model training with the goal of automatically extracting information. Importantly, we also demonstrate how we reduce clinician workload by combining initial high-quality input from clinicians with the state-of-the-art methods in the NLP research. Our study highlights that the combination of methods results in the accurate identification and extraction of the medical terms and concepts from EHRs.

OBJECTIVE

Our aim is to use NLP to capture real-world data on individuals with depression from the Clinical Record Interactive Search (CRIS) clinical text to foster the use of electronic healthcare data in mental health research.

METHODS

Case study and data sources

The PETRUSHKA project (Personalise Efficacy and Tolerability of antidepRessants in Unipolar depreSsion combining individual cHoices, risKs and big dAta)[15] aims to develop and test a precision medicine approach to the pharmacological treatment of unipolar depression. The project aims to achieve this by combining data from RCTs with real-world observational datasets, and by incorporating the treatment preferences of patients, carers and clinicians. Furthermore, PETRUSHKA focuses on the influence of effect modifiers and prognostic factors, such as demographic and clinical characteristics (ie, age, gender, severity of illness and number of previous episodes) on the effects of medication to personalise antidepressant treatment in depression.

To estimate how the effect modifiers change and stratify the effectiveness of the antidepressant medication, researchers plan to use a large collection of EHRs. The required data are accessible through the UK-CRIS system that provides a means of searching and analysing de-identified clinical case records from 12 National Health Service Mental Health Trusts (https://crisnetwork.co/). This system allows access to the wealth of data recorded in routine clinical practice, from structured information, such as diagnosis and demographics information (ie, International Classification of Diseases (ICD-10) codes), to unstructured text information such as clinical summary notes and written assessments. Clinical notes contain rich textual information on patient's history of mental health disorders, as well as cognitive and health score measurements, current and past medication taken by patients and any other relevant information for clinical practice. In this study, we focus on the clinical notes documented for patients with a clinical diagnosis of depression (ICD-10 codes 32 and 33). The final dataset used in our study contains information collected from 13 000 patients that contribute over 1 800 000 clinical documents.

Overview of the NLP pipeline

Our NLP pipeline consisted of multiple steps (figure 1). First, we defined the variables and calculated their frequencies in the medical notes. Second, we developed the annotation schema and use frequency counts to choose the sample of documents for the annotation task. The annotation schema was then developed as an iterative process coupled with annotation of the first 10 clinical notes (exercise notes). If, for example, certain symptoms (eg, anxiety) were frequently reported in the clinical documents (while not being included in the annotation schema), then the schema was updated to include anxiety following feedback from clinicians. Third, the annotated documents were divided into *developmental corpus* (gold data; used to build the NLP model) and *validation corpus* (used to test the model). The development of the model was also iteratively performed, as we used active learning procedures to fine-tune it.[16] We used the NLP systems that allowed for the identification of ambiguous sentences, which when annotated, improved the performance of the model considerably. Finally, the fine-tuned model was tested against the validation dataset and deployed on the complete UK-CRIS data.

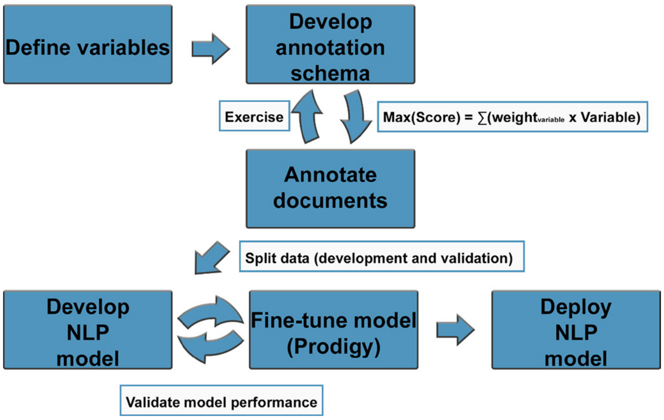
Define variables
Develop annotation schema
Exercise
Max(Score) = ∑(weight$_{variable}$ x Variable)
Annotate documents
Split data (development and validation)
Develop NLP model
Fine-tune model (Prodigy)
Deploy NLP model
Validate model performance

Figure 1 Illustration of the natural language processing pipeline. The full-colour boxes indicate main parts of the data or model development: (1) definition of the variables, (2) development of annotation schema, (3) annotation of the documents, (4) development of initial model, (5) fine-tuning of the model using active learning procedures and (6) deployment of the model on all clinical notes. The outlined boxed indicate processing of information and calculations between main steps: (1) calculation of the quality of notes and using ones with maximum quality, (2) changing schema through exercise with clinicians, (3) splitting the gold corpus on development and validation part and (4) validation of performance for developed and fine-tuned model.

Variables and annotation schema

For this study, we focused on the variables previously observed to have a moderating influence on medication effect.[15] We divided information of interest into seven categories: diagnosis, history, symptoms, clinical assessment tools (rating scores), medication, response to medication and adverse side-effects of medication (figure 2). Each parent category contained multiple subcategories and types of information that we were aiming to extract from the EHRs.

For the diagnosis, we focused on the diagnosis of depression, bipolar disorder, schizophrenia and any mention of physical illness that can influence the clinician's choice of medication prescription. The history of the patient was divided into childhood and adulthood history (life events). Childhood history (younger than 18 years) contained mentions on physical abuse from family members, neglect, emotional and sexual abuse, loss of parents, caregivers with psychiatric history and family violence. In the case of adulthood history (equal to or older than 18 years), we focused on references to divorce, death of an immediate family member, loss of job and domestic violence. Symptoms were divided into five groups: mood (low mood,

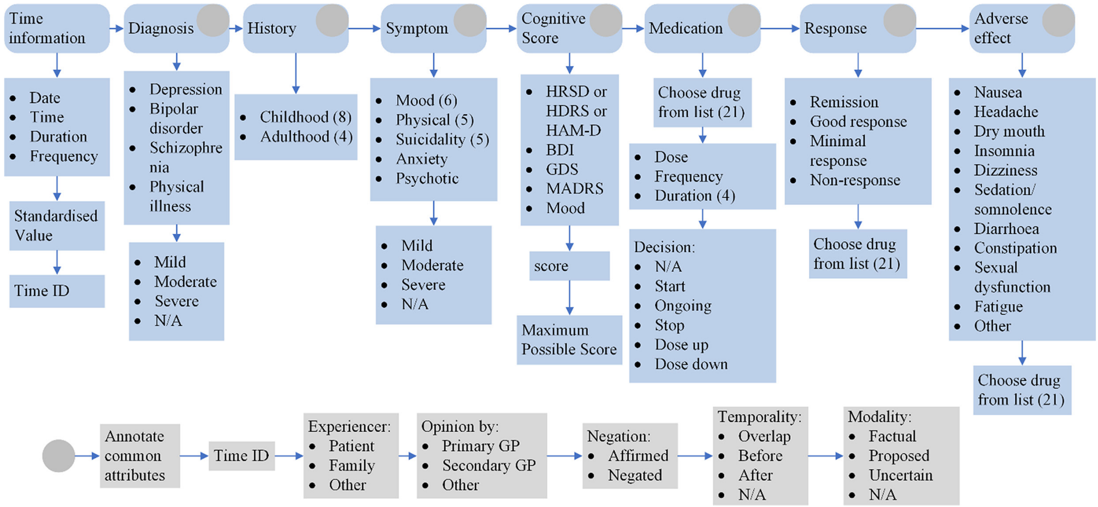

Figure 2 Illustration of the schema developed for the annotation of the RAW clinical text. BDI, Beck Depression Inventory; GDS, Geriatric Depression Scale; GP, general practitioner; HRSD, Hamilton Rating Scale of Depression; MADRS, Montgomery-Asberg Depression Rating Scale.

anhedonia, guilt, concentration problems, reduced self-esteem and pessimistic thoughts), physical symptoms (change in appetite, weight loss, change in sleep, change in activity and fatigue), suicidality (suicide attempt, thoughts of suicide, suicide plan, self-harm and completed suicide), anxiety and psychotic behaviour. The prescribed medication was one of the major outcomes and, we therefore, focused on extracting data pertaining to 21 antidepressants, including dosage, frequency and route of administration (see appendix A for the full list of medications).

The main outcome of interest was a response to treatment, which was categorised as either remission, good response, minimal response and no response. Further side effects indicate potential reasons for treatment discontinuation, and as such we searched specifically for reference to nausea, headache, dry mouth, insomnia, dizziness, sedation/somnolence, diarrhoea, constipation, sexual dysfunction and fatigue. Finally, we focused on extracting information pertaining to performance on clinical assessments, such as measures of Hamilton Rating Scale of Depression (HRSD), Beck Depression Inventory (BDI), Geriatric Depression Scale (GDS), Montgomery-Asberg Depression Rating Scale (MADRS) and subjective mood estimation (on scale from 1 to 10), as reported in EHRs.

Three clinical experts (authors AyK, JH and SI) developed a gold dataset, that is, developmental corpus, that was used to build the NLP model. The clinical expert 'annotators' manually read the medical documents and highlighted word spans which corresponded to the defined variables. We used GATE (https://gate.ac.uk/) software for the annotation task that allowed for the use of a programmable graphical interfacefigure 3. More importantly, this software offers a possibility for subcategorical annotations, allowing for the annotation of nested categories and attributes of the medical concepts. Using GATE, medical professionals annotate the time information, that is, the date when the medical event occurred (eg, date of sertraline prescription). They indicate who is experiencing this event (eg, patient, family and other), by whom it is prescribed (secondary care service, primary GP or other), how valid the information is in the text (factual, proposed and uncertain), and whether it is affirmed or negated. The final developmental gold data consisted of 526 unique clinical notes.

To maximise the amount of information annotated in the clinical notes, three authors (FDC, AT and AC) predefine several potential keywords for each variable category based on their clinical knowledge. We then applied a string-matching algorithm based on the Levenshtein distance to calculate and measure the differences between selected keywords and all the words in each medical note.[17] The Levenshtein distance between two words is the minimum number of single-character edits required to change one word into the other. If the distance was identified as being within a predefined threshold (two characters edit), then we considered it the case that a keyword had appeared in a medical note. This medical note would then be allocated one point for the corresponding category. In addition, we calculated the frequency of mentions for the seven categories on a random sample of 10 000 clinical notes and used them as a weighting factor. Thus, less frequent mentions receive higher weights and vice versa. These weights are multiplied with the scores calculated using Levenshtein distances, resulting in an informativeness score for each note. The notes with the highest scores were selected for the annotators to work on.

Development and fine-tuning of the NLP model

To identify and extract events and entities from the clinical documents we developed a named entity recognition (NER) neural network. The NLP model was developed and trained on a set of 526 documents. The model uses GloVe embeddings for the representation of words[18] and is followed by bi-directional LSTM neural network.[19]

In the second stage of the model development, we used an active learning tool to rapidly annotate a vast amount of information and fine-tune the NER model.[20] The active learning is a special case of supervised machine learning where we maximise useful information derived from the annotated data. This is done by choosing data points or sentences where predictions of the model result in high ambiguity, that is, lower confidence of the prediction. For the active learning aspect of the pipeline, we combined the NER model developed on the rich annotations with the Prodigy active learning tool.[21] In comparison with the GATE software, Prodigy identifies spans of texts that are most informative to the model and presents this information to the annotators. They decide whether the decision from the model is accurate and in an iterative process inform the model on misclassifications. The fourth author (AyK) used the Prodigy tool to fine-tune the model, where 4779 spans of text were additionally annotated that were split into the fine-tuning corpus (3836 spans) and validation spans (938 spans). In the final step of model development and fine-tuning, we repeated the estimation of the model performance on the testing set of the documents.

Measures of performance

To estimate how well the NLP model extracts information of interest, as well as, the level of quality of annotations (interannotator agreement), we used a combination of methodologies from 'Message Understanding Conference' (MUC)[22] and the 'International Workshop on Semantic Evaluation' (SemEval).[23] Based on MUC categories, the extractions from NLP model can be correct (COR), incorrect (INC), partial (PAR—not identical agreement between gold standard and extraction), missing extraction (MIS—model did not extract the concept) or spurious extraction (SPU—extracted concept that does not exist in the gold standard). Based on this categorisation, we calculated the number of possible annotations (POS) in the corpus that contribute to the validation score by summarising correct, incorrect, partial and missing outcomes (true positive +false negative). Equally, we calculated total or actual (ACT) number of annotations that our NLP model produced by summarising correct, incorrect, partial and spurious outcomes (true positives+false positives). Using these two measures, we estimate precision and recall of the system. The precision tells us how many extractions were correct out of the total number of extracted concepts and it is calculated as the ratio between correct (COR) extractions and the actual number of annotations (ACT). The recall indicates the percentage of entities correctly identified in the corpus and is calculated as the ratio between correct (COR) and all possible outcomes (POS). Finally, we also calculated the overall performance of the model by using harmonic mean between these two values, where F1=(2*Recall*Precision)/(Recall+Precision).

RESULTS

Interannotator agreement

Besides developing the gold corpus of 526 documents, clinical experts annotated 12 documents that were used to calculate the interannotator agreement. This score is calculated only on the main seven categories, excluding subcategories such as time information of medication, due to the small number of mentions in the text. Results show a considerable amount of overlap between the three annotators (table 1). The low recall measures indicate that they often annotate different parts of the clinical document with information of interest. In particular, they annotate identical information on average 54% of the cases. For these overlapping concepts, annotators show the almost perfect interpretation of the information (precision). In other words, all identified medical concepts shared among the annotators is identically labelled.

Initial validation of NLP performance

Initial validation of the NLP model was performed on the model developed on rich annotations using GATE software. Results show that the model performs well on the categories of medication extraction, where the route, dosage and medication are extracted with the highest accuracy (table 2). The NLP model identifies almost all mentions of these concepts in the raw text, while correctly extracts approximately 80% of them. This is not the case with other concepts, such as response to medication, clinical assessment scores and symptoms. In these cases, the model almost perfectly extracts the information (precision), but only when it recognises the mentions in the text (recall). However, this recognition does not occur often.

Final validation of the fine-tuned NLP model

After using active learning pipelines to fine-tune the NLP model, we tested the overall performance of the model on the annotations from the validation corpus. Results show considerable improvement in the accuracy of the model extractions. The extraction of medication information stays equally accurate, while most improvement we see in previously underperforming categories. The recall of symptoms, assessment scales, and response improved considerably without major decreases in the precision of the extractions. This similarly applies to the extraction of time when clinical events occur, which increase from an F1 score of 0.06 to 0.75. The only decrease in the performance we observe in the case of medication route and history of life-events (childhood and adulthood history).

DISCUSSION

In this study, we show how a state-of-the-art NLP procedure, in combination with the initial input from medical professionals, can be leveraged to develop algorithms that accurately extract salient clinical information.[24] EHRs contain a multitude of information related to treatments and diagnosis, but also contain auxiliary variables such as education, leisure activities and history of life events. This information, coded in textual format, cannot be used straightforwardly and requires commitment from medical staff to screen and extract variables of interest.

We illustrate the complete methodological pipeline devised to structure information derived from EHRs, from the definition of key variables to the fine-tuning of the NLP model. The results obtained using these methodologies show accurate identification of the medical concepts in the raw clinical texts. The identification of the patient's prescribed medication is shown to be particularly accurate, as the linguistic variety is relatively low for reporting

Table 1 Inter-annotator agreement for seven main categories

Variable name	Annotated	Correct	Incorrect	Missed	Precision	Recall	F1
Diagnosis	66	44	0	22	1.00	0.62	0.76
Medication	126	77	0	49	1.00	0.61	0.75
Assessment scales	15	8	0	7	1.00	0.53	0.69
History	66	34	1	31	0.97	0.51	0.67
Symptoms	630	322	2	306	0.99	0.51	0.67
Adverse effects	69	38	4	27	0.90	0.55	0.70
Response	21	9	0	12	1.00	0.47	0.63
TOTAL					**0.98**	**0.54**	**0.69**

Table 2 Accuracy of the NLP model all categories

		NLP model (Gate)			Fine-tuned NLP model (Prodigy)		
Variable name	Spans	Precision	Recall	F1	Precision	Recall	F1
Dosage	75	0.84	1.0	0.91	0.94	0.93	0.93 (0.02)
Route	6	0.79	1.0	0.88	0.40	0.33	0.36 (0.52)
Medication	152	0.69	0.92	0.79	0.90	0.91	0.90 (0.11)
Frequency	88	0.73	0.64	0.68	0.80	0.70	0.75 (0.07)
Duration	60	0.51	0.71	0.59	0.59	0.66	0.62 (0.03)
History	9	1.0	0.40	0.57	0.00	0.00	0.00 (0.57)
Form	24	0.60	0.38	0.46	0.84	0.91	0.86 (0.40)
Diagnosis	33	1.0	0.17	0.29	0.53	0.48	0.50 (0.21)
Assessment scales	6	1.0	0.13	0.22	0.66	0.33	0.44 (0.22)
Symptoms	272	0.93	0.08	0.14	0.65	0.71	0.68 (0.54)
Time info	173	1.0	0.03	0.06	0.73	0.78	0.75 (0.72)
Adverse effects	10	0.00	0.00	0.00	0.25	0.10	0.14 (0.14)
Response	30	0.00	0.00	0.00	0.47	0.30	0.36 (0.36)
TOTAL		0.73	0.43	0.54	0.74	0.73	**0.74 (0.20)**

Total precision, recall and F1 are a micro-averaged measures
NLP, natural language processing.

of medications in the text.[25] Space of possible brand and generic names used to report the medication prescription are well documented, whereas medication mentions are usually followed with the information on dosage, route and frequency of prescription. Because of the strict rules and low linguistic variations, the NLP models tend to outperform human annotators when identifying mentions of medication prescriptions. This is indeed outcome in the case of recall measures, where NLP models identify all medication prescriptions in the document, whereas clinicians often miss some of the mentions. The same results are not replicated in categories such as symptoms and adverse side effects. Depending on the context and general history of the patients, information on other categories can significantly vary when reported in documents.[21] Our study shows that a model developed on the gold annotated corpus underperforms on the more complex medical categories, where only textual cases similar or identical to the ones observed during the model training are accurately identified and extracted.

The success of NLP methods relies on quality-labelled training data. However, we can see that even with 500 in-depth annotated documents, models fail to achieve satisfactory performance on more complex medical entities. One popular solution to this problem is an active learning approach,[14] which maximises learning accuracy while minimising annotation efforts. The active learning procedure uses uncertainty sampling to find spans of text with the lowest probability of prediction. Once annotated, the new spans carry a large amount of information that improves the overall performance of the model. In this study, we used Prodigy for active learning to fine-tune the developed NLP model. Results show improvement in overall accuracy (F1 score) for almost all categories (ie, variables of interest). Improvements are especially observable in underperforming categories such as time identification and symptoms. As the active learning procedure focuses on the sentences that are estimated to bring the largest increase in the model performance when annotated, we introduce a large amount of variability in the training dataset and adapt the model to the new patterns of textual information.[16] This consequently helps the model to better generalise and identify symptoms and other variables with a greater degree of accuracy. However, we can also see that active learning procedures decrease the accuracy of identification for some medical events (eg, route of the medication and life events indicative for depression diagnosis). One possible reason behind this is that some of the categories overlap and by adding more information, the model misclassifies originally accurate classifications. In addition, the two categories are sparsely represented in the documents, where a model trained on rich annotations learns to perfectly identify a few instances of life-event history or medication routes. However, as we added more information, the statistical weights for these categories are weakened and the model underperforms on recalling the subsequent information. Adding more data often resolves both issues, as the model learns to disentangle concepts from each other or strengthens the weights for these categories.

In summary, we show how we develop NLP models for the extraction of highly complex medical information reported in EHRs. The reported NLP models show high accuracy with regard to drug-related information but demonstrate much lower accuracy levels on the auxiliary variables. In combination with state-of-the-art active learning paradigms, the performance of the model increases considerably and illustrates the feasibility of the research pipeline to be used for the extraction of the EHRs.

Twitter Andrea Cipriani @And_Cipriani

Acknowledgements The authors would like to acknowledge the work and support of the Oxford CRIS Team, Tanya Smith, CRIS Manager, Adam Pill, CRIS Academic Support and Information Analyst, Suzanne Fisher, CRIS Academic Support and Information Analyst, and Lulu Kane CRIS Administrator.

Contributors NV, QL, AK and AN-H: designed the study and analysed the data. AyK, JH, SI: prepared input data. BO, FDC, AT, AC: advised on the scope of clinical variables in the study. NV, BO, FDC, AT, AC and AN-H: wrote the manuscript. All authors edited the manuscript.

Funding This project was funded by the MRC Pathfinder Grant (MC-PC-17215), by the National Institute for Health Research (NIHR) Oxford Health Biomedical Research Centre (BRC-1215-20005) and EPSRC-NIHR HTC Partnership Award 'Plus': NewMind - Partnership with the MindTech HTC (EP/N026977/1). This work was supported by the UK Clinical Record Interactive Search (UK-CRIS) system data and systems of the NIHR Oxford Health Biomedical Research Centre (BRC-1215-20005). AC is supported by the NIHR Oxford Cognitive Health Clinical Research Facility, by an NIHR Research Professorship (grant RP-2017-08-ST2-006) and by the NIHR Oxford Health Biomedical Research Centre (grant BRC-1215-20005).

Competing interests None declared.

Patient consent for publication Not required.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data may be obtained from a third party and are not publicly available. The electronic health records record patient identifiable information and therefore cannot be shared publicly. The data can be used and re-used by applying through UK-CRIS Oxford NHS Trust (https://crisnetwork.co/uk-cris-programme).

ORCID iDs

Nemanja Vaci http://orcid.org/0000-0002-8094-0902
Andrea Cipriani http://orcid.org/0000-0001-5179-8321

REFERENCES

1 James SL, Abate D, Abate KH, *et al*. Global, regional, and national incidence, prevalence, and years lived with disability for 354 diseases and injuries for 195 countries and territories, 1990–2017: a systematic analysis for the global burden of disease study 2017. *The Lancet* 2018;392:1789–858.

2 Cipriani A, Furukawa TA, Salanti G, *et al*. Comparative efficacy and acceptability of 21 antidepressant drugs for the acute treatment of adults with major depressive disorder: a systematic review and network meta-analysis. *The Lancet* 2018;391:1357–66.

3 Chisholm D, Sweeny K, Sheehan P, *et al*. Scaling-up treatment of depression and anxiety: a global return on investment analysis. *Lancet Psychiatry* 2016;3:415–24.

4 Bombardier C, Maetzel A. Pharmacoeconomic evaluation of new treatments: efficacy versus effectiveness studies? *Ann Rheum Dis* 1999;58:i82–5.

5 Vaci N, Cocić D, Gula B, *et al*. Large data and Bayesian modeling-aging curves of NBA players. *Behav Res Methods* 2019;51:1544–64.

6 Zullig LL, Blalock D, Dougherty S, *et al*. The new landscape of medication adherence improvement: where population health science meets precision medicine. *Patient Preference and Adherence* 2018;12:1225–30.

7 Pagliari C, Detmer D, Singleton P. Potential of electronic personal health records. *BMJ* 2007;335:330–3.

8 Menachemi N, Collum TH. Benefits and drawbacks of electronic health record systems. *Risk Manag Healthc Policy* 2011;4:47–55.

9 Murdoch TB, Detsky AS. The inevitable application of big data to health care. *JAMA* 2013;309:1351–2.

10 McConnell RA, Kane SV. The potential and pitfalls of using the electronic health record to measure quality. *Am J Gastroenterol* 2018;113:1111–3.

11 Kadra G, Stewart R, Shetty H, *et al*. Extracting antipsychotic polypharmacy data from electronic health records: developing and evaluating a novel process. *BMC Psychiatry* 2015;15:166.

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

13 Pustejovsky J, Stubbs A. *Natural language annotation for machine learning: a guide to corpus-building for applications*. O'Reilly Media, Inc, 2012.

14 Settles B, Craven M, Friedland L. Active learning with real annotation costs. Proceedings of the NIPS workshop on cost-sensitive learning, 2008:1–10.

15 Tomlinson A, Furukawa TA, Efthimiou O, *et al*. Personalise antidepressant treatment for unipolar depression combining individual choices, risks and big data (PETRUSHKA): rationale and protocol. *Evid Based Ment Health* 2019. doi:10.1136/ebmental-2019-300118. [Epub ahead of print: 23 Oct 2019].

16 Dredze M, Crammer K. Active learning with confidence. Proceedings of ACL-08: HLT, Short Papers, 2008:233–6.

17 Linckels S, Meinel C. *Natural language processing. E-Librarian service*. Berlin, Heidelberg: Springer, 2011: 61–79.

18 Pennington J, Socher R, Manning C. Glove: global vectors for word representation. Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), 2014:1532–43.

19 Schuster M, Paliwal KK. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing* 1997;45:2673–81.

20 Prodigy: A new annotation tool for radically efficient machine teaching", Ines Montani and Matthew Honnibal (to appear in Artificial Intelligence), 2018. Available: https://explosion.ai/blog/prodigy-annotation-tool-active-learning [Accessed 15 Nov 2019].

21 Gligic L, Kormilitzin A, Goldberg P, *et al*. Named entity recognition in electronic health records using transfer learning bootstrapped neural networks. *Neural Networks* 2020;121:132–9.

22 Sundheim BM. Tipster/MUC-5: information extraction system evaluation. Proceedings of the 5th conference on Message understanding, 1993:27–44.

23 Uzzaman N, Llorens H, Derczynski L, *et al*. SemEval-2013 task 1: TEMPEVAL-3: evaluating time expressions, events, and temporal relations, 2013. Available: https://bitbucket.org/leondz/te3-platinum [Accessed 15 Nov 2019].

24 Hofer M, Kormilitzin A, Goldberg P, *et al*. Few-shot learning for named entity recognition in medical text. *arXiv preprint arXiv:1811.05468* 2018.

25 Patrick J, Li M. High accuracy information extraction of medication information from clinical notes: 2009 i2b2 medication extraction challenge. *J Am Med Inform Assoc* 2010;17:524–7.