*DOI:*
10.1016/j.bpsc.2022.09.002

*Document Version*
Peer reviewed version

# Natural language processing: unlocking the potential of electronic health record data to support transdiagnostic psychiatric research

Rashmi Patel[1,2], Matthew Wickersham[3], Rudolf N. Cardinal[4], Paolo Fusar-Poli[1] and Christoph U. Correll[5,6,7]

1 Institute of Psychiatry, Psychology & Neuroscience, King's College London, London, UK
2 Holmusk Technologies Inc, New York, NY, USA
3 Weill-Cornell/Rockefeller/Sloan-Kettering Tri-Institutional MD-PhD Program, New York, NY, USA
4 Department of Psychiatry, University of Cambridge, Cambridge, UK; Cambridgeshire and Peterborough NHS Foundation Trust, Cambridge, UK; Cambridge University Hospitals NHS Foundation Trust, Cambridge, UK
5 Department of Child and Adolescent Psychiatry, Psychosomatic Medicine and Psychotherapy, Charité – Universitaetsmedizin Berlin, corporate member of Freie Universitaet Berlin, Humboldt Universitaet zu Berlin, and Berlin Institute of Health, Berlin, Germany
6 Department of Psychiatry, The Zucker Hillside Hospital, Northwell Health, Glen Oaks, NY, USA
7 Department of Psychiatry and Molecular Medicine, Zucker School of Medicine at Hofstra/Northwell, Hempstead, NY, USA

Correspondence to: Rashmi Patel (rashmi.patel@kcl.ac.uk)

King's College London, Institute of Psychiatry, Psychology & Neuroscience, London, UK

Running title: The application of natural language processing to electronic health record data in psychiatry

Keywords: transdiagnostic; natural language processing (NLP); electronic health records (EHR); electronic medical records (EMR); symptom profiles

Word count (1,500 max): 1,471

References (10 max): 10

Mental disorders bring a substantial burden of illness and disability, but their underlying etiology and pathophysiology remains poorly characterized. Here, we outline the potential for data derived from natural language processing (NLP) of electronic health records (EHRs) to support transdiagnostic approaches to characterize mental disorders.

Traditional classifications of mental disorders define them by the presence or absence of symptoms and observed signs persisting for a minimum period and causing distress and/or functional impairment. However, there is significant overlap in the underlying biological factors associated with different mental disorders.(1) This fact poses a significant challenge to therapeutic development and success. Pharmacological therapies target neurochemical processes irrespective of specific mental disorders.

Attention has moved towards the application of transdiagnostic approaches to develop novel treatments that target specific symptom clusters, rather than global treatments for a particular mental disorder.(2) In so doing, it may be possible to develop and deliver treatments that are better tailored towards an individual's own symptoms rather than a clinically heterogeneous diagnostic category.(3)

However, such a transdiagnostic approach requires detailed data from a population with a wide range of mental disorders. Such data are not typically available in randomized controlled trials, where findings are drawn from a relatively small group of individuals, not necessarily representative of patients seen in routine clinical practice. Such data are also not available in structured real-world data represented in national registries and insurance claim databases, which do not typically capture a broad range of symptom data or their dynamic course and interactions over time.

EHRs have been widely adopted in mental healthcare services and capture rich clinical information from large samples of patients, with the potential to support transdiagnostic psychiatric research. Clinicians use EHRs to document the clinical presentation and treatment plan of individuals receiving mental healthcare. De-identified EHR data have been assembled in research databases to support epidemiology and health economic/outcome research in psychiatry.(4) EHR data may be recorded in structured or unstructured data fields. Structured data fields include numerical values (e.g. date of birth or symptom scale scores) or a constrained, pre-defined range of categories (e.g. sex, race or ethnicity, or diagnostic codes), whereas unstructured fields contain free text typically entered by a clinician to describe the clinical presentation, treatment plan, or observed outcomes.

Due to time constraints, structured symptom rating scales are seldom recorded in real-world clinical practice, and where they are recorded, their utility to support research is often limited due to missing data. Most rich clinical data encompassing symptoms, signs, and outcomes are repeatedly recorded as free text in unstructured data fields. However, manual extraction of relevant information from unstructured data is cumbersome and time consuming, and often impractical, especially at lager scale.

NLP can be used to extract information that enables the automated generation of structured data from unstructured free text. Its application to EHR data allows the rapid identification of rich clinical data that has the potential to support transdiagnostic research, across large numbers of patients, that would otherwise be unfeasible through prospective data collection.(5)

Developing NLP models to extract information from EHR data requires a dataset (corpus) of unstructured clinical records documented during routine mental healthcare practice. Once a corpus has been generated, several pre-processing steps are typically required for NLP model development.(6) In a typical sequence, the text is first segmented into sentences, which are then tokenized into their individual components (e.g. words, spaces, punctuation). Based on tokens' position and content, the parts of speech of each token can be ascertained. Once parts of speech are determined, the speech is lemmatized such that pluralities and verb tenses are deconstructed to the simplest form. Common words that do not contribute to the text can be removed, while relevant words can undergo dependency parsing to examine how they relate to one another. With these associations, named entity recognition can be applied to detect and label the text with likely meanings based on word content and grouping. Named entity recognition systems can, for example, tag names, locations, medications, diseases, and symptoms. Unstructured text data may also be assembled as word embeddings, in which words are represented as vectors that may be analyzed to identify semantically related words or phrases within the document. A vector representation may better represent the complexity and variability of unstructured EHR data than a keyword-based approach.

Following initial pre-processing steps, NLP models may be developed through rules-based or machine learning approaches, or a combination of both.(6) For rules-based NLP models, predefined filters are applied to determine if a sentence contains words, or combinations of words, that indicate the presence of the clinical construct of interest. A structured data output is thus generated, which could simply be a binary output indicating the presence of the construct of interest, or a numerical measure that has been documented in the text. A disadvantage of a rules-based approach is the time required to develop and test the model's rules, given the complexity of unstructured psychiatric data recorded in EHRs, in which clinicians may represent the same construct through a

large variety of keywords and phrases. Nonetheless, rules-based approaches may be particularly helpful when clinicians record well-structured concepts (e.g. test results or standardized diagnostic codes) in free text.

Supervised machine learning methods, such as deep learning, enable the development of NLP models and their application to complex unstructured documents without requiring the generation of predefined rules for entity recognition and classification.(6) Therefore, machine learning-based approaches are often well suited to NLP development in EHR data. The underlying principle of supervised machine learning-based NLP is to develop a model that can accurately classify examples of unstructured text by training an algorithm using text that has been pre-classified by human annotators. The resulting model is tested for accuracy against a separate reference dataset, also pre-classified by humans. The reference dataset should ideally be annotated by multiple individuals, to assess for inter-rater reliability in ascertaining the clinical construct of interest. This process is particularly important in the development of supervised machine learning-based NLP models for EHR data in psychiatry, as the documentation of symptoms and clinical features relevant to mental disorders is complex and variable; if a reasonable degree of inter-annotator agreement is not reached, then the resulting NLP model may have limited validity when applied to unseen data. Validation should ideally occur across data from multiple clinical services to reduce the possibility that a model is classifying based on words or phrases that are specific to a particular data source. Once a suitably accurate NLP model is developed, it can be applied to the entire corpus to generated structured data rapidly and at scale for subsequent analysis.

The opportunity to assemble large volumes of rich clinical information captured from unstructured EHR data using NLP has enabled transdiagnostic research that would otherwise be unfeasible through prospective data collection, and that would not be possible using data ascertained from

structured fields in EHRs. NLP-derived EHR data have been used to demonstrate the associations of negative(7) and cognitive(8) symptoms of schizophrenia with poor clinical outcomes. Rules-based and machine learning-based NLP classifiers have been developed to identify documented suicidal ideation and suicide attempts.(9) NLP-derived EHR symptom data have also been shown to enhance the accuracy of predictive models to detect the risk of psychosis beyond structured EHR data alone.(10)

There are important limitations to consider in the application of NLP to unstructured EHR data. EHR data are recorded as part of routine clinical practice, and certain aspects are likely not recorded as frequently, or comprehensively, as structured symptom rating scale data would be in a prospective clinical trial or observational study. The absence of recorded EHR data does not necessarily indicate the absence of a particular clinical feature of interest. Additionally, EHR data is normally recorded by a clinician, and so may often be unsuitable for linguistic analysis of patients' speech, except where direct quotations can be clearly distinguished from clinicians' text. Nonetheless, it is likely that important clinical data, such as the presence of suicidality or key symptoms of a particular mental disorder, are recorded. A further limitation is NLP model accuracy, which can be affected by poor inter-rater reliability in the annotation process, and be due to clinician/service variation in the recording of data within and across EHRs.

Despite some limitations discussed above, the application of NLP to unstructured EHR data has expanded the breadth of clinical data that is available to generate evidence to support a better understanding of transdiagnostic features in psychiatry. Structured data sources, such as clinical trial, registry, and insurance claims datasets, may be insufficient to generate evidence using rich clinical information in large sample sizes that are representative of the wider clinical population. Given the increasing focus on targeting symptom profiles within and beyond individual mental

disorder diagnoses, the application of NLP to EHR data could substantially enhance the generation of evidence to support novel therapeutic developments in psychiatry.

## Acknowledgments

Sources of funding: RP has received funding from an NIHR Advanced Fellowship (NIHR301690) and a Medical Research Council (MRC) Health Data Research UK Fellowship (MR/S003118/1). RNC's research is supported by the MRC (MR/W014386/1); all research at the Department of Psychiatry in the University of Cambridge is supported by the UK National Health Service (NHS) National Institute for Health Research (NIHR) Cambridge Biomedical Research Centre (BRC-1215-20014) and the NIHR East of England Applied Research Centre; the views expressed are those of the authors and not necessarily those of the NHS, the NIHR, or the Department of Health and Social Care.

## Disclosures

RP has received grant funding from Janssen and personal fees from Holmusk. MW reports no conflicts of interest. RNC consults for Campden Instruments Ltd and receives royalties from Cambridge University Press, Cambridge Enterprise, and Routledge. PFP has received research funds or personal fees from Lundbeck, Angelini, Menarini, Sunovion, Boehringer Ingelheim, Mindstrong, Proxymm Science, outside the current study. CUC has been a consultant and/or advisor to or has received honoraria from: AbbVie, Acadia, Alkermes, Allergan, Angelini, Aristo, Boehringer-Ingelheim, Cardio Diagnostics, Cerevel, CNX Therapeutics, Compass Pathways, Darnitsa, Gedeon Richter, Hikma, Holmusk, IntraCellular Therapies, Janssen/J&J, Karuna, LB Pharma, Lundbeck, MedAvante-ProPhase, MedInCell, Merck, Mindpax, Mitsubishi Tanabe Pharma, Mylan, Neurocrine, Newron, Noven, Otsuka, Pharmabrain, PPD Biotech, Recordati, Relmada, Reviva, Rovi, Seqirus, SK Life Science, Sunovion, Sun Pharma, Supernus, Takeda, Teva, and Viatris. He provided expert testimony for Janssen and Otsuka. He served on a Data Safety Monitoring Board for Lundbeck, Relmada, Reviva, Rovi, Supernus, and Teva. He has received grant support from Janssen and Takeda. He received royalties from UpToDate and is also a stock option holder of Cardio Diagnostics, Mindpax, LB Pharma and Quantic.

## References

1. Lee PH, Anttila V, Won H, Feng YCA, Rosenthal J, Zhu Z, *et al.* (2019): Genomic Relationships, Novel Loci, and Pleiotropic Mechanisms across Eight Psychiatric Disorders. *Cell* 179: 1469-1482.e11.

2. Pahwa M, Sleem A, Elsayed OH, Good ME, El-Mallakh RS (2021): New Antipsychotic Medications in the Last Decade. *Curr Psychiatry Rep* 23: 87.

3. McGorry P, Nelson B (2016): Why we need a transdiagnostic staging approach to emerging psychopathology, early diagnosis, and treatment. *JAMA Psychiatry* 73: 191–192.

4. Patel R, Wee SN, Ramaswamy R, Thadani S, Tandi J, Garg R, *et al.* (2022): NeuroBlu, an electronic health record (EHR) trusted research environment (TRE) to support mental healthcare analytics with real-world data. *BMJ Open* 12: e057227.

5. Jackson RG, Patel R, Jayatilleke N, Kolliakou A, Ball M, Gorrell G, *et al.* (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. https://doi.org/10.1136/bmjopen-2016-012012

6. Henry S, Yetisgen M, Uzuner O (2021): Chapter 13: Natural Language Processing in Mental Health Research and Practice. In: Tenenbaum JD, Ranallo PA, editors. *Mental Health Informatics: Enabling a Learning Mental Healthcare System*. Cham: Springer International Publishing, pp 317–353.

7. Patel R, Jayatilleke N, Broadbent M, Chang CK, Foskett N, Gorrell G, *et al.* (2015): Negative symptoms in schizophrenia: A study in a large clinical sample of patients using a novel automated method. *BMJ Open* 5. https://doi.org/10.1136/bmjopen-2015-007619

8. Mascio A, Stewart R, Botelle R, Williams M, Mirza L, Patel R, *et al.* (2021): Cognitive Impairments in Schizophrenia: A Study in a Large Clinical Sample Using Natural Language Processing. *Frontiers in Digital Health*, vol. 3. https://doi.org/10.3389/fdgth.2021.711941

9. Fernandes AC, Dutta R, Velupillai S, Sanyal J, Stewart R, Chandran D (2018): Identifying Suicide Ideation and Suicidal Attempts in a Psychiatric Clinical Research Database using Natural Language Processing. *Sci Rep* 8: 7426.

10. Irving J, Patel R, Oliver D, Colling C, Pritchard M, Broadbent M, *et al.* (2021): Using Natural Language Processing on Electronic Health Records to Enhance Detection and Prediction of Psychosis Risk. *Schizophr Bull* 47: 405–414.