DOI:
10.1093/schbul/sbaa126

USING NATURAL LANGUAGE PROCESSING ON ELECTRONIC HEALTH RECORDS TO ENHANCE DETECTION AND PREDICTION OF PSYCHOSIS RISK

Jessica Irving[1], MSc; Rashmi Patel[1], MD, PhD; Dominic Oliver[1], MSc; Craig Colling[2], MSc; Megan Pritchard[2,3], MSc; Matthew Broadbent[2], MA; Helen Baldwin[1], MSc; Daniel Stahl, PhD[1]; Robert Stewart[2,3], MD, PHD; Paolo Fusar-Poli[1,2,5], MD, PhD

Affiliations

[1] Early Psychosis: Interventions and Clinical-detection (EPIC) lab, Department of Psychosis Studies, Institute of Psychiatry, Psychology & Neuroscience, King's College London, London, UK;
[2] South London and Maudsley NHS Foundation Trust, London, UK;
[3] Department of Psychological Medicine, Institute of Psychiatry, Psychology & Neuroscience, King's College London, London, UK;
[4] Department of Biostatistics & Health Informatics, Institute of Psychiatry, Psychology & Neuroscience, King's College London, London, UK;
[5] Department of Brain and Behavioral Sciences, University of Pavia, Pavia, Italy

Keywords: natural language processing; electronic health records; prevention; psychosis; machine learning; prediction

Correspondence to: Dr Paolo Fusar-Poli MD PhD, Department of Psychosis Studies, 5th Floor, Institute of Psychiatry, Psychology & Neuroscience, PO63, 16 De Crespigny Park, SE5 8AF London, UK. E-mail: paolo.fusar-poli@kcl.ac.

Abstract word count: 250

Main body word count: 3112

Competing interests: All authors have completed the ICMJE uniform disclosure form at www.icmje.org/coi_disclosure.pdf and declare: RS has received research funding from Roche, Janssen, GSK and Takeda. PFP has received grant funds from Lundbeck and honoraria from Lundbeck, Menarini and Angelini outside the current study. The other authors report no financial relationships with commercial interests.

Contributors: The study was conceived by PFP. Data extraction and statistical analysis was performed by JI with support from CC, RP, MP and DS. Reporting of findings were carried out by JI and PFP. All authors (JI, CC, DS, MP, RS, PFP, HB and RP) contributed to study design, manuscript preparation and approved the final version.

Ethics approval: The CRIS data resource received ethical approval as an anonymised dataset for secondary analyses from Oxfordshire REC C (Ref: 18/SC/0372).

Source of funding: HS, MP, RS and CC receive(d) funding from the National Institute for Health Research (NIHR) Biomedical Research Centre at South London and Maudsley NHS Foundation Trust and King's College London, which also supports the development and maintenance of the BRC Case Register. RP has received support from a Medical Research Council (MRC) Health Data Research UK Fellowship (MR/S003118/1) and a Starter Grant for Clinical Lecturers (SGL015/1020) supported by the Academy of Medical Sciences, The Wellcome Trust, MRC, British Heart Foundation, Arthritis Research UK, the Royal College of Physicians and Diabetes UK.

Role of funder: The views expressed are those of the authors and not necessarily those of the NHS, the NIHR or the Department of Health. The funders had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication.

Data sharing: The data accessed by CRIS remain within an NHS firewall and governance is provided by a patient-led oversight committee. Subject to these conditions, data access is encouraged and those interested should contact RS (robert.stewart@kcl.ac.uk), CRIS academic lead.

ABSTRACT

Background: Using novel data mining methods such as natural language processing (NLP) on Electronic Health Records (EHR) for screening and detecting individuals at risk for psychosis.

Method: The study included all patients receiving a first index diagnosis of nonorganic and nonpsychotic mental disorder within the South London and Maudsley (SLaM) NHS Foundation Trust between January 1, 2008, and July 28, 2018. LASSO-regularised Cox regression was used to refine and externally validate a refined version of a 5-item individualised, transdiagnostic, clinically based risk calculator previously developed (Harrell's C = 0.79) and piloted for implementation. The refined version included 14 additional NLP-predictors: tearfulness, poor appetite, weight loss, insomnia, cannabis, cocaine, guilt, irritability, delusions, hopelessness, disturbed sleep, poor insight, agitation and paranoia.

Results: A total of 92,151 patients with a first index diagnosis of nonorganic and nonpsychotic mental disorder within the SLaM Trust were included in the derivation (n = 28,297) or external validation (n = 54,716) data sets. Mean age was 33.6 years, 50.7% were women, and 67.0% were of white race/ethnicity. Mean follow-up was 1590 days. The overall 6-year risk of psychosis in secondary mental health care was 3.4 (95% CI, 3.3 – 3.6). External validation indicated strong performance on unseen data (Harrell's C 0.85, 95% CI 0.84–0.86), an increase of 0.06 from the original model.

Conclusions: Using NLP on EHRs can considerably enhance the prognostic accuracy of psychosis risk calculators. This can help identify patients at risk of psychosis who require assessment and specialized care, facilitating earlier detection and potentially improving patient outcomes.

INTRODUCTION

The burden of psychotic disorders is substantial: for example, schizophrenia accounted for 13 million years lived with disability in 2017(1), with the most recent estimate reporting a European economic burden of €93.9 billion.(2) Existing treatments have little impact on illness course in established psychosis.(3,4) Primary indicated prevention in individuals at clinical high risk for psychosis (CHR-P), however, has the potential to reduce the duration of untreated psychosis and alter its course.(5,6) Effective preventive intervention is reliant on the successful identification of individuals at risk for psychosis for referral to specialised CHR-P clinical services;(7–9) these individuals tend to present with attenuated psychotic symptoms and overall functional impairment.(10,11) Detection of at-risk individuals currently relies on help-seeking behaviours(12) and idiosyncratic referral pathways initiated on suspicion of psychosis risk.(13) Emerging evidence suggests that current detection strategies are highly inefficient,(13) with only 5%(14)-12%(15) of first episode cases intercepted by CHR-P clinical services. To tackle these challenges, we previously developed an individualised, clinically-based transdiagnostic risk calculator, using clinical and demographic predictors widely available in Electronic Health Records (EHRs): age, sex, age by sex, ethnicity, and index ICD-10 diagnosis or CHR-P designation.(16) The transdiagnostic risk calculator has been externally validated in two separate large EHR datasets, demonstrating adequate prognostic performance (n = 54,716, Harrell's C = 0.79(14); n = 13,702, Harrell's C = 0.73).(17) This transdiagnostic risk calculator has already undergone pilot testing for clinical implementation.(18) The calculator's potential for implementation, combined with its unique position to enhance large-scale detection of at-risk individuals, underscores the importance of improving its prognostic accuracy. Given the replication crisis in psychiatry and science(19,20), improving existing, previously validated risk prediction models represents a more efficient approach than redeveloping new models.(21)

While EHRs offer some information (notably on sociodemographic characteristics) in structured fields, the majority of information is recorded in free text such as event notes and uploaded attachments, representing an enormous reservoir of untapped information.(22) For example, information on symptomatology and substance use is not routinely recorded in a structured way.(22) Natural language processing (NLP) techniques have recently been developed to mine structured data from free text. These offer an unprecedented opportunity to incorporate more granular predictors closer to the pathophysiology of psychosis onset into a model. By applying NLP to EHRs, this study aims to improve on the prognostic accuracy achieved by the previously published transdiagnostic risk calculator(16), further supporting the efficient detection of individuals at risk for psychosis.

METHODS AND MATERIALS

Setting

The South London and Maudsley (SLaM) NHS Trust is one of Europe's largest secondary mental healthcare providers.(23) Its main catchment area covers four socioeconomically diverse South London boroughs: Croydon, Lambeth, Lewisham and Southwark, alongside tertiary referrals from the rest of London and the United Kingdom. The Clinical Record Interactive Search (CRIS) system facilitates interrogation of de-identified EHRs held by the Trust, which adopted a system of electronic recording in 2007.(22,23) SLaM now has EHRs for over 400,000 individuals, providing data on their sociodemographic and clinical characteristics.

Study population

We extracted data for all individuals accessing the SLaM NHS Trust between January 1, 2008 and July 28, 2018. Inclusion and exclusion criteria followed that of the original analysis and development of the psychosis risk calculator: namely, all individuals who received a first index primary diagnosis of any nonorganic and nonpsychotic mental disorder.(16)

Model Specifications

Original model

As detailed in Fusar-Poli et al(16), the original transdiagnostic, clinically-based and individualised risk calculator was developed using a retrospective cohort study leveraging EHRs from the SLaM boroughs of Lambeth and Southwark. Cox regression was used to predict the hazard ratio (HR) of developing any psychotic disorder over time (defined in Supplementary eMethods 1). Predictors included age (at the time of index diagnosis), sex, age by sex, self-assigned ethnicity, and cluster index diagnosis or CHR-P designation (defined in eTables 2 and 3). The model was externally validated first in the SLaM boroughs of Croydon and Lewisham and later in Camden and Islington NHS Trust.(16,17) In this retrospective version of the risk calculator, individuals who developed psychosis within three months following their index diagnosis were excluded. However, implementing this diagnostic lag prospectively in the subsequent implementation study would have resulted in delays for referral to assessment. Therefore, a refined version of the risk calculator without the lag period, which demonstrated similar external prognostic accuracy, was optimised for prospective use and is considered in the current study (for details see eTable 4).

Model refinement with NLP data

In the present study we refined the prospective version of the transdiagnostic model using all original predictors plus additional NLP-derived predictors. NLP tools were used to extract symptom and substance use data from free text recorded by clinicians within the six months prior to the index diagnosis. This time period was chosen to ensure that predictor data did not overlap with our outcome variables, and because symptom assessment tends to precede formal diagnosis. We employed CRIS-specific NLP algorithms that convert unstructured information from free text into structured and quantifiable data. Details on symptom algorithm development can be found in Jackson *et al.*(22); in general, these were developed using cross-validated support vector machines on a gold-standard, human-annotated training corpus for each symptom. A regularly updated algorithm library, with comprehensive detail on keywords used and validation efforts, can be found on the CRIS website.(24) NLP algorithm performance is mainly measured in terms of precision (proportion of true positive instances of total NLP-labelled positive instances) and recall (proportion of true positive instances of all positive instances available in the text). As EHRs provide multiple opportunities for term detection, we favoured precision over recall, using only NLP algorithms with at least 80% precision (see eMethods 2 and eTable 5). We also excluded predictors with near-zero variance, which can cause model instability across validation folds.(25) This resulted in 14 NLP symptom and substance use predictors with a mean (SD) precision of 0.91 (0.06): tearfulness, poor appetite, weight loss, insomnia, cannabis use, cocaine use, guilt, irritability, delusions, hopelessness, disturbed sleep, poor insight, agitation and paranoia. We dichotomised NLP symptom and substance use predictors as trigger terms tend to be repeated within and across records; treating predictors as continuous would otherwise have resulted in the model erroneously interpreting them as a linear reflection of severity.(26–28) The value ‘0’ indicated that a given symptom or substance was *not mentioned* in a patient’s EHR. We retained individuals without NLP-derivable symptom or substance data prior to index diagnosis, treating NLP data as bonus information where available.

Statistical analysis

This EHR clinical register-based study is reported according to the RECORD and STROBE statements (see eTable 1).(29) Model development and validation followed the methodological guidelines of Royston and Altman,(30) Steyerberg *et al*(31) and the Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD).(32)

We performed descriptive analyses of baseline clinical and sociodemographic characteristics of the sample, obtaining means and frequencies for continuous and categorical variables respectively. The Kaplan-Meier failure function (1-survival) and Greenwood 95% confidence intervals were used to describe the cumulative risk of psychosis onset in SLaM patients. The primary outcome measure was model prognostic discrimination performance measured through Harrell’s C, a recommended measure for external validation of Cox models.(33) A Harrell’s C value of 0.9–1.0 is considered outstanding, 0.8–0.9 excellent and 0.7–0.8 acceptable.(34) Model development and validation analyses were conducted in R version 3.6.1.

Model development

We first divided our cohort into derivation and validation samples using non-random, geographic split-sampling, which is one of the recommended approaches to model building (32). Mirroring the original analysis, the derivation sample comprised all cases from Lambeth and Southwark until 31st December 2015.(16) The validation sample comprised all cases from the same two boroughs from January 2016 plus cases from all other boroughs, constituting temporal and geographic forms of external validation. These samples differ on several sociodemographic characteristics.(23)

We then trained a Cox proportional-hazards model on our derivation sample using the refined transdiagnostic model. Since adding large numbers of predictors can result in overfitting,(21) we regularised our model using the Least Absolute Shrinkage and Selection Operator (LASSO) penalty, implemented via the *glmnet* package in R. The LASSO algorithm performs feature selection by shrinking the coefficients of redundant predictors towards zero. This penalty requires selection of a tuning parameter lambda, which controls the number of coefficients estimated to be non-zero. We used 10-fold cross-validation to select the optimal lambda, choosing the minimum lambda value that maximised partial likelihood from a range of possible values. With the resultant model coefficients, we developed a prognostic index in the derivation dataset by generating prognostic risk scores for each individual.

External model validation

We applied the regression coefficients from the derivation data set to each case in the external validation set to generate the prognostic index for the validation dataset. Model prognostic performance (Harrell's C, which captures discrimination)(31) was the primary outcome measure. We further assessed overall model performance using the Brier score (the average mean squared difference between predicted probabilities and actual outcomes), which captures calibration and discrimination aspects.(31) A lower score indicates higher precision and less bias. Calibration (the agreement between observed outcomes and predictions) was assessed with the regression slope of the prognostic index.(35) Finally, we performed a sensitivity analysis to assess whether our model would perform better in temporal or geographic external validation by splitting our external validation set into these two groups.

RESULTS

Sociodemographic and clinical characteristics

Of 108,211 individuals receiving a first SLaM diagnosis of nonorganic and nonpsychotic mental disorder within SLaM in the period between 1st Jan 2008 and 28th July 2017, 92,151 had complete data across all original predictors (see Figure 1). Mean (SD) age was 33.6 (19.0); individuals were almost evenly split by sex (female 50.7%, male 49.3%), most were of White ethnicity (67.0%), and anxiety disorders were the most frequent index diagnoses (27.5%, Table 1). With respect to the new NLP predictors, 44,368 (41%) individuals had no symptom or substance data in the six months prior to their index diagnosis. Derivation (n = 28,297) and validation sets (n = 63,854, Figure 1) showed notable differences in terms of ethnic make-up and the spread of index diagnoses (e.g., substance use disorder was more prevalent in the derivation set, see eTable 6). Mean (SD) follow-up was 1,590 (721) days with a significant difference between derivation and validation sets (derivation: mean (SD), 1896 (463); validation: mean (SD), 1455 (772)). Overall 6-year risk of developing a psychotic disorder was 0.034 (95% CI, 0.033 – 0.036). Cumulative incidence (Kaplan-Meier failure function) for risk of development of psychotic disorders is presented in eFigure 1.

Model development

There were 1,060 transitions to psychosis in the derivation dataset (raw counts stratified per index diagnosis are available in eTable 7), of which 55 were observed in the CHR-P group (5%). The refined risk prediction model significantly predicted psychosis onset (likelihood ratio χ_2 test, 2769; $p < .001$, Table 2). No variables were selected out of the model via LASSO regularisation. The LASSO penalty improves model performance at the expense of bias in parameter estimates (which reduces coefficient interpretability), therefore significance testing for individual predictors would not be appropriate.(36) Paranoia, delusions and agitation were the strongest positive NLP-derived predictors of psychosis while hopelessness was the strongest negative one. The transdiagnostic model refined with NLP predictors showed good apparent prognostic performance (Harrell's C index, 0.86, Table 3), an increase of 0.05 from the Harrell's C of the original model.

External model validation

There were 1,662 transitions to psychosis in the external validation dataset, which far exceeds the minimum value of 100 events required for robust external validation.(37) The transdiagnostic model refined with NLP predictors still retained good prognostic performance when applied to unseen data (Harrell's C 0.85), an increase of 0.06 from the Harrell's C of the original model. The calibration slope coefficient of 1.06 (95% CI 1.03 – 1.09) indicated no major miscalibration issues. A sensitivity analysis found stronger model discriminatory performance in temporal external validation (Harrell's C = 0.91) than geographic (Harrell's C = 0.86; eTable 8).

DISCUSSION

To our knowledge this is the first study demonstrating the potential of applying automated methods such as NLP to EHRs to detect individuals at risk for psychosis. By incorporating NLP-derived data on symptoms and substance use, we refined a previously validated transdiagnostic, individualised and clinically based risk prediction model. Model refinement considerably improved the external prognostic accuracy of the model to a good level (Harrell's C 0.85) compared with an adequate level when using structured field data alone (Harrell's C 0.79).

Efficient detection of individuals at CHR-P has been a neglected area of research despite its necessity for successful early intervention. This study progresses the field by demonstrating, for the first time, that advanced data-mining NLP methods(38) can improve prognostication of psychosis risk at large scale. Our NLP-refined model confers a substantial increase in external prognostic accuracy, with a Harrell's C increment of 0.06. Harrell's C indexes the probability that for any given case-control pair, the model will generate a higher predicted risk score for the case. Our refinement has effected an improvement from adequate to good, a level of prognostic accuracy exceeding that of the original CHR-P instruments (C-statistic 0.79).(39) Compared with CHR-P instruments, the NLP-based risk calculator produces reasonably well-calibrated and individualised estimates of risk (as opposed to group-level estimates), is automatable in EHRs and can be applied in large datasets. Furthermore, our risk calculator detects psychosis risk transdiagnostically outside the CHR-P designation.(40) This is crucial given recent evidence that a third of first episode psychosis cases do not evolve through a previous CHR-P stage.(13) Indeed, the original risk calculator has already been externally validated, and shown to perform well, in other NHS Trusts that do not have CHR-P services.(17) A recent review of psychosis risk prediction models found that clinical variables such as paranoia and unusual thought content consistently appear as significant predictors of psychosis.(41) NLP techniques can extract symptom data at a fraction of the cost of individual patient recruitment. Prognostic performance of our NLP-based refinement of the transdiagnostic risk calculator also exceeds that achieved using harder-to-obtain neuroanatomical predictors (e.g. grey matter volume), with accuracies ranging from 0.50 to 0.63.(42) In a previous study we found that the use of machine learning *per se* is not associated with improved prognostic accuracy.(43)

The first step towards translating NLP tools into clinical benefits for patients is to apply these methods in larger risk cohorts to test their reproducibility.(44) This study represents the largest application of NLP in the area of predicting conversion to psychosis to date (the second largest study includes only 59 patients).(44) Our promising findings align with the existing NLP literature. For example, NLP-based automated speech analysis has been used to measure subtle, clinically relevant mental state changes in emerging psychosis.(45) Other studies have found that NLP-derived tools can identify symptom distributions in clinician notes beyond those captured by ICD codes, and that these domains usefully map onto Research Domain Criteria.(46) Our group has also confirmed that CRIS-NLP algorithms can reliably extract data on typically complex symptomatic domains such as negative psychotic symptoms or insight,(26,47) which has been replicated by an independent team.(28) These algorithms can also extract substance use data that predict longitudinal outcomes.(48) The accuracy of NLP-based estimates compared with gold-standard domains is further corroborated by their robust prognostic (i.e. predicting the course of a condition) and predictive (i.e. predicting the response to treatment) value.(28,46,49) NLP-derived data can also be incorporated into dynamic risk prediction models such as those recently used to predict psychosis onset risk.(50) For example, one could use NLP to extract dynamic treatment data relevant to CHR-P populations, such as exposure to cognitive behavioural therapy, which is not routinely recorded in structured fields.(51) This information could dynamically flow into risk predictions that are updated every time a patient’s record is updated with new information.

The NLP-refined transdiagnostic risk calculator is well suited for implementation in routine clinical care. First, our calculator represents the only available, pragmatic option for improving detection of individuals at risk of psychosis in secondary mental health care. The only existing alternative is to conduct extensive outreach campaigns that promote referrals based on clinicians' suspicion of psychosis risk. This approach is inefficient because it dilutes risk enrichment (i.e. refers more patients with only low risk for psychosis).(52) Second, NLP-derived data were available for most at-risk individuals. Third, the NLP-refined model performed well in external validation efforts both temporally and geographically. The NLP algorithms used can be transferred to other sites with electronic health registers interrogable via CRIS for further external validation (for example, the Oxford Health NHS Foundation Trust). Fourth, this study refines an already well-performing model. Enhancing the benefits of clinical implementation through continual refinement of a given prognostic model is preferable to repeatedly developing new models from scratch that may never enter clinical routine.(21) As the original risk calculator has already been piloted for implementation,(18) the new NLP-refined model can be easily absorbed into clinical practice. Finally, this refined risk prediction model is designed to work around the way clinicians and mental health professionals enter text into EHRs. Future developments could implement an automated algorithm to trigger a prompt when all five baseline predictors are entered, taking into account any NLP symptom or substance use data entered prior to this point. For those (41%) without symptom or substance use information recorded prior to formal diagnosis, the original risk calculator can still be used. We have recently integrated the original version of this risk calculator in EHRs and prospectively piloted a real-time and real-world psychosis risk detection and alerting system.(18,53) This method leverages the CogStack platform, which is an open-source text extraction system.(53) The CogStack platform offers full-text search of clinical data, real-time calculation of psychosis risk, early risk alerts to clinicians and visual monitoring of patients over time.(53) This method is highly transportable and can be easily deployed in NHS Trusts with a CRIS or CogStack platform. So far, the CRIS platform — including consenting procedures — is under expansion across 12 NHS Trusts in the UK, harnessing over 2 million deidentified patient records (https://crisnetwork.co/). This study also provides further empirical evidence supporting the expansion of EHRs in clinical psychiatry to facilitate precision and stratified medicine approaches on a global scale. Study limitations are appended in the eLimitations section.

CONCLUSIONS

Applying NLP techniques to EHR data recorded during routine clinical practice can facilitate robust research in large, representative samples of patients. NLP can add value to precision psychiatry by enhancing the prognostic accuracy of risk prediction models. Automatic text extraction from EHRs through NLP enhanced the transdiagnostic prognostic power achieved by a previously developed psychosis risk calculator. This can help to facilitate earlier detection of patients at risk of developing psychosis who require an assessment and specialised care, potentially improving outcomes of psychosis.

REFERENCES

1. GBD 2017 Disease and Injury Incidence and Prevalence Collaborators SL, Abate D, Abate KH, Abay SM, Abbafati C, Abbasi N, *et al.* (2018): 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. *Lancet (London, England)* 392: 1789–1858.
2. Gustavsson A, Svensson M, Jacobi F, Allgulander C, Alonso J, Beghi E, *et al.* (2011): Cost of disorders of the brain in Europe 2010. *Eur Neuropsychopharmacol*. https://doi.org/10.1016/j.euroneuro.2011.08.008
3. Jaaskelainen E, Juola P, Hirvonen N, McGrath JJ, Saha S, Isohanni M, *et al.* (2013): A Systematic Review and Meta-Analysis of Recovery in Schizophrenia. *Schizophr Bull* 39: 1296–1306.
4. Millan MJ, Andrieux A, Bartzokis G, Cadenhead K, Dazzan P, Fusar-Poli P, *et al.* (2016): Altering the course of schizophrenia: progress and perspectives. *Nat Rev Drug Discov* 15: 485–515.
5. Fusar-Poli P, Bauer M, Borgwardt S, Bechdolf A, Correll CU, Do KQ, *et al.* (2019): European college of neuropsychopharmacology network on the prevention of mental disorders and mental health promotion (ECNP PMD-MHP). *Eur Neuropsychopharmacol* 29: 1301–1311.
6. Oliver D, Davies C, Crossland G, Lim S, Gifford G, McGuire P, Fusar-Poli P (2018): Can We Reduce the Duration of Untreated Psychosis? A Systematic Review and Meta-Analysis of Controlled Interventional Studies. *Schizophr Bull* 44: 1362–1372.
7. Fusar-Poli P, Tantardini M, De Simone S, Ramella-Cravaro V, Oliver D, Kingdon J, *et al.* (2017): Deconstructing vulnerability for psychosis: Meta-analysis of environmental risk factors for psychosis in subjects at ultra high-risk. *Eur Psychiatry* 40: 65–75.
8. Oliver D, Reilly TJ, Baccaredda Boy O, Petros N, Davies C, Borgwardt S, *et al.* (2019): What Causes the Onset of Psychosis in Individuals at Clinical High Risk? A Meta-analysis of Risk and Protective Factors. *Schizophr Bull*. https://doi.org/10.1093/schbul/sbz039
9. Radua J, Ramella-Cravaro V, Ioannidis JPA, Reichenberg A, Phiphopthatsanee N, Amir T, *et al.* (2018): What causes psychosis? An umbrella review of risk and protective factors. *World Psychiatry* 17: 49–66.
10. Fusar-Poli P, Rocchetti M, Sardella A, Avila A, Brandizzi M, Caverzasi E, *et al.* (2015): Disorder, not just state of risk: Meta-analysis of functioning and quality of life in people at high risk of psychosis. *Br J Psychiatry* 207: 198–206.
11. Fusar-Poli P, Byrne M, Badger S, Valmaggia LR, McGuire PK (2013): Outreach and support in South London (OASIS), 2001–2011: Ten years of early diagnosis and treatment for young individuals at high clinical risk for psychosis. *Eur Psychiatry* 28: 315–326.
12. Falkenberg I, Valmaggia L, Byrnes M, Frascarelli M, Jones C, Rocchetti M, *et al.* (2015): Why are help-seeking subjects at ultra-high risk for psychosis help-seeking? *Psychiatry Res*. https://doi.org/10.1016/j.psychres.2015.05.018

13. Fusar-Poli P, Sullivan SA, Shah JL, Uhlhaas PJ (2019): Improving the Detection of Individuals at Clinical Risk for Psychosis in the Community, Primary and Secondary Care: An Integrated Evidence-Based Approach. *Front Psychiatry* 10: 774.
14. Fusar-Poli P (2017): Extending the Benefits of Indicated Prevention to Improve Outcomes of First-Episode Psychosis. *JAMA Psychiatry* 74: 667.
15. McGorry PD, Hartmann JA, Spooner R, Nelson B (2018): Beyond the “at risk mental state” concept: transitioning to transdiagnostic psychiatry. *World Psychiatry*. https://doi.org/10.1002/wps.20514
16. Fusar-Poli P, Rutigliano G, Stahl D, Davies C, Bonoldi I, Reilly T, McGuire P (2017): Development and validation of a clinically based risk calculator for the transdiagnostic prediction of psychosis. *JAMA Psychiatry*. https://doi.org/10.1001/jamapsychiatry.2017.0284
17. Fusar-Poli P, Werbeloff N, Rutigliano G, Oliver D, Davies C, Stahl D, *et al.* (2019): Transdiagnostic risk calculator for the automatic detection of individuals at risk and the prediction of psychosis: Second replication in an independent national health service trust. *Schizophr Bull*. https://doi.org/10.1093/schbul/sby070
18. Fusar-Poli P, Oliver D, Spada G, Patel R, Stewart R, Dobson R, McGuire P (2019): Real world implementation of a transdiagnostic risk calculator for the automatic detection of individuals at risk of psychosis in clinical routine: Study protocol. *Front Psychiatry*. https://doi.org/10.3389/fpsyt.2019.00109
19. Loken E, Gelman A (2017): Measurement error and the replication crisis. *Science*. https://doi.org/10.1126/science.aal3618
20. Begley CG, Ioannidis JPA (2015): Reproducibility in science: Improving the standard for basic and preclinical research. *Circulation Research*. https://doi.org/10.1161/CIRCRESAHA.114.303819
21. Fusar-Poli P, Hijazi Z, Stahl D, Steyerberg EW (2018): The Science of Prognosis in Psychiatry. *JAMA Psychiatry* 75: 1289.
22. 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: e012012.
23. Perera G, Broadbent M, Callard F, Chang C-K, Downs J, Dutta R, *et al.* (2016): Cohort profile of the South London and Maudsley NHS Foundation Trust Biomedical Research Centre (SLaM BRC) Case Register: current status and recent enhancement of an Electronic Mental Health Record-derived data resource. *BMJ Open* 6: e008721.
24. CRIS Natural Language Processing Applications Library (n.d.): Retrieved July 23, 2020, from https://www.maudsleybrc.nihr.ac.uk/facilities/clinical-record-interactive-search-cris/cris-natural-language-processing/
25. Kuhn M, Johnson K (2013): Applied predictive modeling. *Applied Predictive Modeling*.

https://doi.org/10.1007/978-1-4614-6849-3

26. Ramu N, Kolliakou A, Sanyal J, Patel R, Stewart R (2019): Recorded poor insight as a predictor of service use outcomes: Cohort study of patients with first-episode psychosis in a large mental healthcare database. *BMJ Open*. https://doi.org/10.1136/bmjopen-2019-028929
27. Patel R, Jayatilleke N, Broadbent M, Chang C-K, 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: e007619.
28. Downs J, Dean H, Lechler S, Sears N, Patel R, Shetty H, *et al.* (2018): Negative Symptoms in Early-Onset Psychosis and Their Association With Antipsychotic Treatment Failure. *Schizophr Bull*. https://doi.org/10.1093/schbul/sbx197
29. Benchimol EI, Smeeth L, Guttmann A, Harron K, Moher D, Petersen I, *et al.* (2015): The REporting of studies Conducted using Observational Routinely-collected health Data (RECORD) Statement. *PLOS Med* 12: e1001885.
30. Royston P, Altman DG (2013): External validation of a Cox prognostic model: Principles and methods. *BMC Med Res Methodol*. https://doi.org/10.1186/1471-2288-13-33
31. Steyerberg EW, Vergouwe Y (2014): Towards better clinical prediction models: Seven steps for development and an ABCD for validation. *European Heart Journal*. https://doi.org/10.1093/eurheartj/ehu207
32. Collins GS, Reitsma JB, Altman DG, Moons KGM (2015): Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): The TRIPOD Statement. *Eur Urol*. https://doi.org/10.1016/j.eururo.2014.11.025
33. Uno H, Cai T, Pencina MJ, D'Agostino RB, Wei LJ (2011): On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. *Stat Med* 30: 1105–17.
34. Hosmer DW, Lemeshow S, May S (2011): Applied Survival Analysis: Regression Modeling of Time to Event Data: Second Edition. *Applied Survival Analysis: Regression Modeling of Time to Event Data: Second Edition*. https://doi.org/10.1002/9780470258019
35. Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N, *et al.* (2010): Assessing the performance of prediction models: a framework for traditional and novel measures. *Epidemiology* 21: 128–38.
36. Hastie T, Tibshirani R, Wainwright M (2015): Statistical learning with sparsity: The lasso and generalizations. *Statistical Learning with Sparsity: The Lasso and Generalizations*. https://doi.org/10.1201/b18401
37. Collins GS, Ogundimu EO, Altman DG (2016): Sample size considerations for the external validation of a multivariable prognostic model: A resampling study. *Stat Med*. https://doi.org/10.1002/sim.6787
38. Maynard D, Bontcheva K (2014): Natural language processing. *Perspectives on Ontology Learning*. https://doi.org/10.4018/ijssoe.2014010105

39. Oliver D, Kotlicka-Antczak M, Minichino A, Spada G, McGuire P, Fusar-Poli P (2018): Meta-analytical prognostic accuracy of the Comprehensive Assessment of at Risk Mental States (CAARMS): The need for refined prediction. *Eur Psychiatry* 49: 62–68.
40. Fusar-Poli P, Solmi M, Brondino N, Davies C, Chae C, Politi P, *et al.* (2019): Transdiagnostic psychiatry: a systematic review. *World Psychiatry*. https://doi.org/10.1002/wps.20631
41. Worthington MA, Cao H, Cannon TD (2019): Discovery and Validation of Prediction Algorithms for Psychosis in Youths at Clinical High Risk. *Biol Psychiatry Cogn Neurosci Neuroimaging*. https://doi.org/10.1016/J.BPSC.2019.10.006
42. Vieira S, Gong Q, Pinaya WHL, Scarpazza C, Tognin S, Crespo-Facorro B, *et al.* (2020): Using Machine Learning and Structural Neuroimaging to Detect First Episode Psychosis: Reconsidering the Evidence. *Schizophr Bull* 46: 17–26.
43. Fusar-Poli P, Stringer D, M. S. Durieux A, Rutigliano G, Bonoldi I, De Micheli A, Stahl D (2019): Clinical-learning versus machine-learning for transdiagnostic prediction of psychosis onset in individuals at-risk. *Transl Psychiatry* 9: 259.
44. Corcoran CM, Carrillo F, Fernández-Slezak D, Bedi G, Klim C, Javitt DC, *et al.* (2018): Prediction of psychosis across protocols and risk cohorts using automated language analysis. *World Psychiatry* 17: 67–75.
45. Bedi G, Carrillo F, Cecchi GA, Slezak DF, Sigman M, Mota NB, *et al.* (2015): Automated analysis of free speech predicts psychosis onset in high-risk youths. *npj Schizophr* 1: 15030.
46. McCoy TH, Castro VM, Rosenfield HR, Cagan A, Kohane IS, Perlis RH (2015): A Clinical Perspective on the Relevance of Research Domain Criteria in Electronic Health Records. *Am J Psychiatry* 172: 316–320.
47. Patel R, Jayatilleke N, Broadbent M, Chang C-K, 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: e007619.
48. Patel R, Wilson R, Jackson R, Ball M, Shetty H, Broadbent M, *et al.* (2015): Cannabis use and treatment resistance in first episode psychosis: a natural language processing study. *Lancet (London, England)* 385 Suppl 1: S79.
49. Patel R, Wilson R, Jackson R, Ball M, Shetty H, Broadbent M, *et al.* (2016): Association of cannabis use with hospital admission and antipsychotic treatment failure in first episode psychosis: An observational study. *BMJ Open*. https://doi.org/10.1136/bmjopen-2015-009888
50. Studerus E, Beck K, Fusar-Poli P, Riecher-Rössler A (2019): Development and Validation of a Dynamic Risk Prediction Model to Forecast Psychosis Onset in Patients at Clinical High Risk. *Schizophr Bull*. https://doi.org/10.1093/schbul/sbz059
51. Colling C, Evans L, Broadbent M, Chandran D, Craig TJ, Kolliakou A, *et al.* (2017): Identification of the

delivery of cognitive behavioural therapy for psychosis (CBTp) using a cross-sectional sample from electronic health records and open-text information in a large UK-based mental health case register. *BMJ Open* 7: e015297.

52. Fusar-Poli P, Schultze-Lutter F, Cappucciati M, Rutigliano G, Bonoldi I, Stahl D, *et al.* (2016): The Dark Side of the Moon: Meta-analytical Impact of Recruitment Strategies on Risk Enrichment in the Clinical High Risk State for Psychosis. *Schizophr Bull* 42: 732–743.

53. Wang T, Oliver DAP, Msosa YJ, Colling C, Spada G, Roguski Ł, *et al.* (2019): A real-time psychosis risk detection and alerting system based on electronic health records using CogStack. *J Vis Exp*.

Table 1. Sample characteristics

Variable	**No (%)**		
	Study population **(n = 92,150)**	**Derivation dataset** **(n = 28,297)**	**Validation dataset** **(n = 63,853)**
Age, mean (SD) [a]	33.6 (19.0)	34.8 (18.8)	33.0 (19.1)
Sex[a]			
Female	46741 (50.7%)	13861 (49.0%)	32880 (51.5%)
Male	45410 (49.3%)	14436 (51.0%)	30974 (48.5%)
Ethnicity[a]			
White	61711 (67.0%)	16700 (59.0%)	45011 (70.5%)
Asian	4549 (4.94%)	1030 (3.64%)	3519 (5.51%)
Black	15187 (16.5%)	6029 (21.3%)	9158 (14.3%)
Mixed	3805 (4.13%)	1183 (4.18%)	2622 (4.11%)
Other	6899 (7.49%)	3355 (11.9%)	3544 (5.55%)
Index diagnosis			
CHR-P	445 (0.48%)	238 (0.84%)	207 (0.32%)
Anxiety disorders	25323 (27.5%)	6765 (23.9%)	18558 (29.1%)
Acute and transient psychotic disorders	1568 (1.70%)	552 (1.95%)	1016 (1.59%)
Bipolar disorders	3149 (3.42%)	1018 (3.60%)	2131 (3.34%)
Childhood onset disorders	12332 (13.4%)	3351 (11.8%)	8981 (14.1%)
Developmental disorders	4645 (5.04%)	923 (3.26%)	3722 (5.83%)
Mental retardation	1640 (1.78%)	609 (2.15%)	1031 (1.61%)
Nonbipolar affective disorders	15965 (17.3%)	5240 (18.5%)	10725 (16.8%)
Personality disorders	3524 (3.82%)	1071 (3.78%)	2453 (3.84%)
Physiological disorders	6806 (7.39%)	1958 (6.92%)	4848 (7.59%)
Substance use disorders	16754 (18.2%)	6572 (23.2%)	10182 (15.9%)
Tearfulness	20214 (21.9%)	13835 (21.7%)	6379 (22.5%)
Appetite loss	13653 (14.8%)	9322 (14.6%)	4331 (15.3%)
Weight loss	8623 (9.36%)	6002 (9.40%)	2621 (9.26%)
Insomnia	5115 (5.55%)	3401 (5.33%)	1714 (6.06%)
Poor insight	17089 (18.5%)	12000 (18.8%)	5089 (18.0%)
Guilt	9953 (10.8%)	6665 (10.4%)	3288 (11.6%)
Irritability	9049 (9.82%)	6259 (9.80%)	2790 (9.86%)
Delusions	5352 (5.81%)	3649 (5.71%)	1703 (6.02%)
Hopelessness	8883 (9.64%)	6117 (9.58%)	2766 (9.77%)
Disturbed sleep	25786 (28.0%)	17576 (27.5%)	8210 (29.0%)
Agitation	12916 (14.0%)	9054 (14.2%)	3862 (13.6%)
Paranoia	13212 (14.3%)	9201 (14.4%)	4011 (14.2%)
Substance use			
Cannabis	13604 (14.8%)	9271 (14.5%)	4333 (15.3%)
Cocaine	10229 (11.1%)	6554 (10.3%)	3675 (13.0%)

[a] Missingness values for ethnicity, sex and age were 11.9%, 0.06% and 0.02% respectively.

Table 2. Characteristics of the refined, individualised and transdiagnostic clinically-based risk prediction model employing NLP predictors to detect individuals at risk for psychosis in EHRs. Coefficients obtained via LASSO-regularised, multivariable Cox proportional hazards regression using the derivation dataset (*n* = 28,297).

	Predictor	Hazard ratio
Original model	Male sex	1.29
	Age	1.01
	Sex * Age	0.99
	Ethnicity	
	White	Ref
	Asian	1.57
	Black	2.16
	Mixed	1.20
	Other	1.18
	Primary diagnosis	
	CHR-P	Ref
	Anxiety disorders	0.16
	Acute and transient psychotic disorders	1.26
	Bipolar disorders	0.38
	Childhood-onset disorders	0.06
	Developmental disorders	0.07
	Mental retardation	0.07
	Nonbipolar affective disorders	0.22
	Personality disorders	0.17
	Physiological disorders	0.11
	Substance use disorders	0.15
New NLP symptoms and substance use		
	Agitation	1.64
	Appetite loss	1.06
	Cannabis	1.13
	Cocaine	0.87
	Delusions	2.10
	Disturbed sleep	1.12
	Guilt	0.93
	Hopelessness	0.70
	Insomnia	1.05
	Irritability	1.05
	Loss of insight	1.02
	Paranoia	2.62
	Tearfulness	0.93
	Weight loss	1.14

Table 3. Performance measures for the original transdiagnostic individualised risk prediction model vs the NLP model refinement

Performance Measure	**Original transdiagnostic model**		**Refined model including NLP predictors**	
	Derivation	**Validation**	**Derivation**	**Validation**
Overall				
Brier	0.028	0.021	0.085	0.061
R_2	0.746	0.719	0.885	0.885
Discrimination				
Harrell’s C (95% CI)	0.809 (0.795 – 0.822)	0.790 (0.775 – 0.806)	0.861 (0.849 – 0.873)	0.848 (0.838 – 0.858)
Calibration				
Calibration slope	1	0.968	1	1.059