Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Externally validated prediction models for incident mental health disorders in young adults: A systematic review of methods, performance and clinical utility

Anna Tsaligopoulou

ucjuat5@ucl.ac.uk

University College London https://orcid.org/0000-0002-0570-4299

Layla Sadek

UCL

Jennifer Dykxhoorn

UCL

Naomi Launders

UCL

Joseph Hayes

UCL https://orcid.org/0000-0003-2286-3862

Analysis

Keywords:

Posted Date: November 12th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10819677/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: Yes there is potential Competing Interest. JFH has received consultancy fees from Juli Health, Wellcome and Swiss Re. He is a co-founder and shareholder of Juli Health. Juli Health has a patent pending.

Externally validated prediction models for incident mental health disorders in young adults: A systematic review of methods, performance and clinical utility

Author Information:

Anna Tsaligopoulou, Division of Psychiatry, University College London, London, UK.

Layla Sadek, Division of Psychiatry, University College London, London, UK.

Jennifer Dykxhoorn, Division of Psychiatry, University College London, London, UK.

Naomi Launders, Division of Psychiatry, University College London, London, UK.

Joseph Hayes, Division of Psychiatry, University College London, London, UK.

Abstract

Mental disorders constitute a substantial public health concern globally. Prediction models may improve early identification of high-risk individuals. We conducted a systematic review to identify externally validated prediction models of depression, anxiety, bipolar and psychotic disorders. We searched MEDLINE, APA PsycINFO and Embase up to April 2026 along with backward and forward citation tracking. Out of 692 records yielded, 14 unique prediction models were identified for anxiety (n = 1), depression (n = 3), bipolar disorder (n = 1) or psychosis (n = 9). Most models used Cox proportional hazards (n= 6) and 13 models used solely psychosocial predictors. Model discrimination ranged from poor (AUC = 0.51) to considerable performance (AUC = 0.82). Only 1 model tested clinical utility. To date, there is a lack of accurate and clinically applicable models for predicting incident mental disorders.

Introduction

Common mental disorders (CMD) (such as depression and anxiety disorders) and severe mental illness (SMI) (e.g. bipolar disorder and psychotic disorders) constitute a substantial public health concern, with 1 in 7 individuals living with a mental health condition worldwide (World Health Organization: WHO, 2025). Many experience delayed diagnosis, are untreated or undertreated, or lack access to effective care, worsening long-term prognosis (Memon et al., 2016; Vigo et al., 2025). Adolescence and emerging adulthood (WHO, 2025; Arnett et al., 2014) is a sensitive period for early identification and targeted early interventions, given that 75% of mental health problems emerge between the ages of 12-25 (McGorry and Mei, 2018).

Early intervention results in better mental health outcomes (Catania et al., 2011), reducing symptom severity and hospital admissions (Bird et al., 2010). Current prevention and early intervention strategies usually focus on reducing risk factors and enhancing protective factors for mental health (Kirkbride et al., 2024). Therefore, a critical aspect of modifying mental health outcomes is accurately identifying the early-life experiences and exposures that shape individuals' susceptibility to mental health problems in adulthood (Venn et al., 2025). Early life risk factors for both CMD and SMI have been identified across domains, including biological (such as genetic markers (Grotzinger et al., 2026), brain structure and activity (Xu et al., 2022), and proteomics (Ding et al., 2025)) and psychosocial

factors (such as adverse childhood experiences (Gu et al., 2022), education, neighbourhood, and socioeconomic status (Salem and Robenson, 2025)).

Identifying mental health risk as early as possible could be achieved by implementing risk prediction models (Saelmans et al., 2025). By developing and implementing accurate risk prediction models, we should be able to successfully identify individuals with a propensity for specific mental health conditions.

Although numerous mental health prediction models have been developed, relatively few are implementable. This is often due to methodological limitations, such as poor quality of datasets, small sample sizes or lack of internal validation (Wynants et al., 2020). Ramspek et al. (2021) found that few prognostic prediction models were externally validated, which is a key limitation given that external validation is essential for assessing a model's reproducibility and transportability across different settings. Without robust external validation, prediction models have limited potential for implementation into routine clinical practice. While some models are externally validated, there is limited understanding of how close any existing CMD or SMI risk prediction models are to deployment.

While prediction models using traditional statistical methods have provided valuable insights, their reliance on regression-based techniques potentially overlooks the complex, non-linear relationships among predictors and the volume of existing healthcare data (Atias et al., 2025). This limitation has increasingly been tackled through the use of new machine learning (ML) methods in the development of prediction models. However, it is still unclear whether these methods perform better than traditional modelling approaches.

To our knowledge, there is currently no comprehensive review of the methodology, performance or clinical utility of externally validated prediction models for CMD or SMI. To address this issue, this review aimed to systematically identify and compare externally validated models for the prediction of incident depression, anxiety, psychosis and bipolar disorder in young people.

Results

Our search yielded 692 records; after removing duplicates, 566 unique records were included for abstract and title screening. Of these 146 were considered suitable for full text screening. After data extraction and matching models with their corresponding external

validation reports, 31 eligible manuscripts met our inclusion criteria (45 reports in total). These 45 reports represent the internal and external validation of 14 unique risk prediction models, with 31 external validation reports (1-4 per model). We identified 4 models for predicting CMD (1 model for anxiety disorders (2 manuscripts), 3 models for depression (8 manuscripts)) and 10 models for SMI (9 models for psychosis (21 manuscripts), 1 model for bipolar disorder (1 manuscript)). One manuscript (Osborne et al., 2019) reported on the external validation of 2 different models. The most common reasons for exclusion were manuscripts which did not include childhood/adolescent exposures or adolescent/young-adult outcomes (n = 24 at full-text screening) and lack of external validation (n = 54 at data extraction) (see PRISMA diagram, Supplementary 1). The interrater agreement for title and abstract screening between the two reviewers was high (Cohen's $\kappa = 0.85$; 95%CI: 0.77 – 0.93; 96.1% agreement) and for the full-text screening was substantial (Cohen's $\kappa = 0.76$; 95%CI: 0.60 – 0.93; 88.9% agreement). Minor deviations from the PROSPERO protocol are reported in Supplementary 2.

Model development

Included risk prediction models were developed and validated between 2008-2025, with a median publication year of 2017. Of the 14 models, 12 were developed using cohort data and 2 using case-control data. In terms of population, 9 models were developed on European populations, 2 were developed on North America populations, 1 in China, 1 in Brazil and 1 in Turkey. The sample size varied from 34 to 299,993 study participants. The mean participant age (where reported, n = 13) ranged from 11.9 to 49.2 years old, with an outcome follow-up duration (where reported, n = 13) from 6 months to 10 years (Table 1). In 2 reports that did not report mean age, participant ages ranged from 15 to 75 years. All models predicting anxiety and depression (n=4) were developed in the general population, while only one model predicting SM (model 13, predicting psychosis) was. The remaining 9 models predicting SMI were developed in clinical high-risk populations (including subthreshold bipolar disorder, recent-onset depression, and clinical samples with prodromal psychosis symptoms/psychotic-like experiences). Variable missingness was reported in 9 models. Approaches to handling missing data were either conducting data imputation (multiple imputation n = 5 or K-nearest neighbour n = 1) or complete case analysis (n = 3).

Between 5 and 28 predictors were included in the models. 12 of the development models used a combination of psychosocial predictors (e.g. prodromal symptoms/index diagnosis, ACEs, family history of mental health diagnoses), while one used structural MRI

features (Zarogianni et al. 2017) and another used a combination of sMRI and psychosocial features (Arslan et al., 2025). 11 of the models included demographic variables as a predictive feature (e.g. age, sex, ethnicity etc.) while 6 models included clinical variables (e.g. index diagnosis and prodromal symptom severity). The most frequent methods of feature selection were clinical knowledge/existing literature (n = 9 models), machine learning methods (n = 4 models) and univariable analysis (n = 4 models). Half (n=7) of these methods were used in combination e.g. clinical knowledge/existing literature and machine learning or another method. Models used Cox proportional hazards regression (n = 6), logistic regression (n = 6), support vector machine (n= 2) methods (Table 1).

External validation

Similarly to model development, most external validation reports (n = 27/31) used longitudinal cohort design to externally validate the models compared to 4 reports that used case-control design. Of the 31 external validation reports, 13 were validated in Europe (e.g. UK, Switzerland, Estonia), 8 in North America (USA and Canada), 4 in Latin America (Chile and Brazil), and 6 in other populations (Nepal, Nigeria, New Zealand, China, and Turkey). The sample size varied from 35 – 2,430,333 study participants. The mean participant age (where reported, n = 22) ranged from 11.8 to 47 years old, with a mean follow-up (where reported, n =27) of outcome collection ranged from 6 months to 8.19 years (Table 1). There were more reports validating models in clinical high risk (n = 17) compared to general population samples (n = 13), and one study validated a model on child soldier population. Missingness in the datasets was usually reported (n = 22), where they either conducted data imputation (multiple imputation n = 7, nearest neighbour n = 3, mean/mode n= 1) or complete-case analysis (n = 11).

Table 1. Summary of the 45 reports and the methods, performance and clinical utility of the 14 respective models.

Total (N = 45)	**Development (N = 14 models)**	**Validation (N = 31 reports)**
Predicted outcome		
Common Mental Disorders	4 models	14 reports
Depression	3 models	10 reports
Anxiety	1 model	4 reports
Severe Mental Illness	10 models	17 reports
Psychosis	9 models	16 reports

Bipolar	1 model	1 report
Study design		
Cohort study	12 models	27 reports
Case control study	2 models	4 reports
Geographical region		
Europe (e.g. UK, Switzerland, Spain etc.)	9 models	13 reports
North America (e.g. US and Canada)	2 models	8 reports
Latin America (e.g. Brazil and Chile)	1 model	4 reports
Other (e.g. China, Nigeria, Nepal, New Zealand, Turkey)	2 models	6 reports
Sample size range	34 – 299,993	35 – 2,430,333
Missing data		
Imputation	6 models	11 reports
Complete case analysis	3 models	11 reports
Not reported	5 models	9 reports
Range of mean age of participants	11.9 – 49.2	11.8 – 47
Range of study follow-up	6 months – 10 years	6 months – 8.19 years
Population		
General	5 models	10 reports
Clinical high risk	9 models	20 reports
Other (e.g. child soldiers)		1 report
Model total features	5 – 28	5 – 28
Features used		
Demographics*	11 models	27 reports
Psychosocial	13 models	24 reports
Biological (e.g. structural MRI)	2 models	3 reports
Clinical symptoms	6 models	12 reports
Feature selection method		
Clinical knowledge/past literature	9 models	
Machine learning (e.g. LASSO, Elastic Net, RFE)	4 models	
Univariable analysis	4 models	
Backward/forward elimination	3 models	
Combination of methods	7 models	
Model type		

Cox proportional hazards regression	6 models	
Logistic regression	5 models	
Support Vector Machine	2 models	
Model discrimination (AUC/c-index)	0.66 – 0.83	0.51 – 0.83
Mean for internal validation	0.77 (95%CI: 0.72-0.81)	
Mean for external validation		0.69 (95%CI: 0.61-0.77)
Not reported	1 model	1 report
Model calibration		
Calibration slope	0.759 – 1	0.93 – 0.996
Calibration in the large (CITL)	-2.405 – 0.02	-
Not reported	8 models	16 reports

*Demographics included age, sex/gender, ethnicity/skin color, education, marital status and socioeconomic status, age x gender interaction.

Model development performance

To measure discrimination performance, 13 of the 14 models reported either Area Under the Curve (AUC) (n = 4), Harrel's c-index (n = 9). One model reported balanced accuracy (BAC; Zarogianni et al., 2017). 4 models did not report the confidence intervals for the model discrimination. Only 6 models reported calibration statistics (calibration slope, calibration in the large and Hosmer-Lemeshow test). 7 models reported sensitivity and specificity values (see Supplementary 3 and 4).

External validation performance

In external validation, 14 of the 31 reports reported AUC, 16 reported Harrel's c-index and one reported balanced accuracy (Zarogianni et al., 2019). Of the 24 manuscripts detailing external validation studies, 6 did not report the confidence intervals for the model discrimination. Only 11 manuscripts reported calibration statistics (calibration slope and CITL) and 8 manuscripts reported sensitivity and specificity values (see Supplementary 3 and 4).

Common Mental Health Disorders

Anxiety Disorders. One prediction model was developed for anxiety disorders (Predict-A model) (Model 1: King et al., 2011). The model was developed in a general

(primary care) European population (UK, Spain, Slovenia, Portugal) with 9 psychosocial and demographic features using logistic regression. The model had fair discrimination abilities (0.70-0.79) (c-index = 0.78; 95%CI: 0.74 - 0.81) (Figure 1). No calibration was reported. The Predict-A model had four external validation analyses (King et al., 2011; Nigatu et al., 2019) in general population in Netherlands, Estonia, Chile and US. The external validation ranged from considerable discrimination in the Netherlands (0.80-0.89) (c-index = 0.81; 95%CI: 0.74 - 0.89) (King et al., 2011) to poor discrimination in the US (c-index = 0.62; 95%CI: 0.61 - 0.64) (Nigatu et al., 2019) (see Figure 1). Calibration statistics were only reported for the latter report (poor calibration ($p < 0.001$); Supplementary 3).

Depression. Three models were developed for depression (Model 2: PredictD, Model 3: PSRA, and Model 4: IDEA-RS). The PSRA model had considerable discrimination abilities (0.80-0.89) (c-index = 0.82; 95%CI: 0.79 - 0.84; Bellon et al., 2011), compared to PredictD (c-index = 0.79; 95%CI: 0.77 – 0.81; King et al., 2008) and IDEA-RS (c-index = 0.78; 95%CI: 0.73 - 0.82; Rocha et al., 2021) that showed fair discrimination (0.70-0.79). Hence, the best performing model for depression is PSRA (Figure 2). Calibration scores were only reported for IDEA-RS where the model was near perfect (calibration slope: 1; CITL: 0.02 (Supplementary 3)). The PSRA, IDEA-RS and PredictD models were externally validated 2-6 times. Similarly to its internal validation, the PSRA model had the highest mean discrimination abilities (c-index = 0.73; range: 0.70 - 0.76; 95%CI: 0.70 - 0.77), compared to PredictD (c-index = 0.71; range: 0.708 - 0.710; 95%CI: 0.68 - 0.73) with one study reporting poor calibration (Nigatu et al., 2016), and IDEA-RS (c-index = 0.66; range: 0.59 - 0.83; 95%CI: 0.63 - 0.73) that had poor discrimination (0.60-0.69). Calibration was reported for every external validation report of IDEA-RS indicating the model was overfitting in external datasets (see Supplementary 3). The PSRA model was the best performing overall, with the highest discrimination abilities in both internal and external validation. This was a logistic regression model using 13 psychosocial and demographic features (PSRA model; Bellon et al. 2011). It was developed in a general Spanish population (primary care) and validated in Europe (UK, Estonia, Netherlands, Portugal, Slovenia) and Chile.

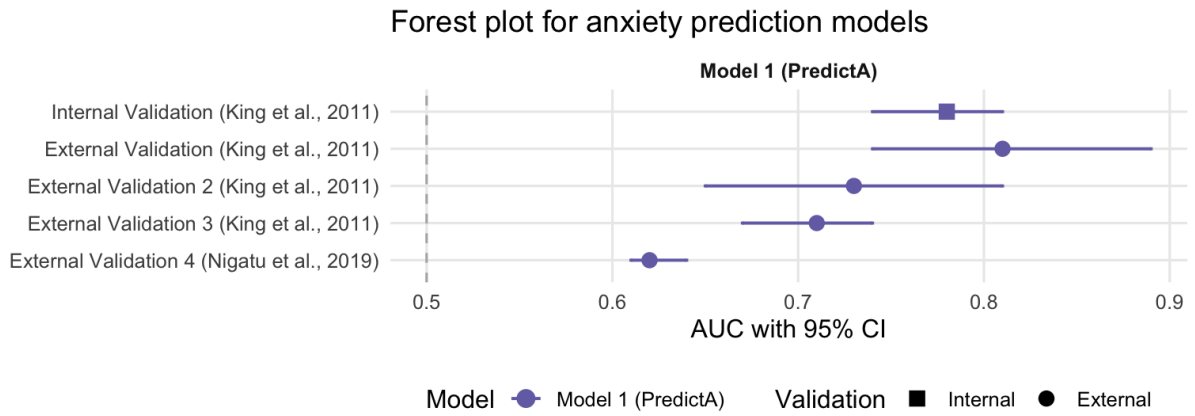


Figure 1. Forest plot for anxiety prediction models.

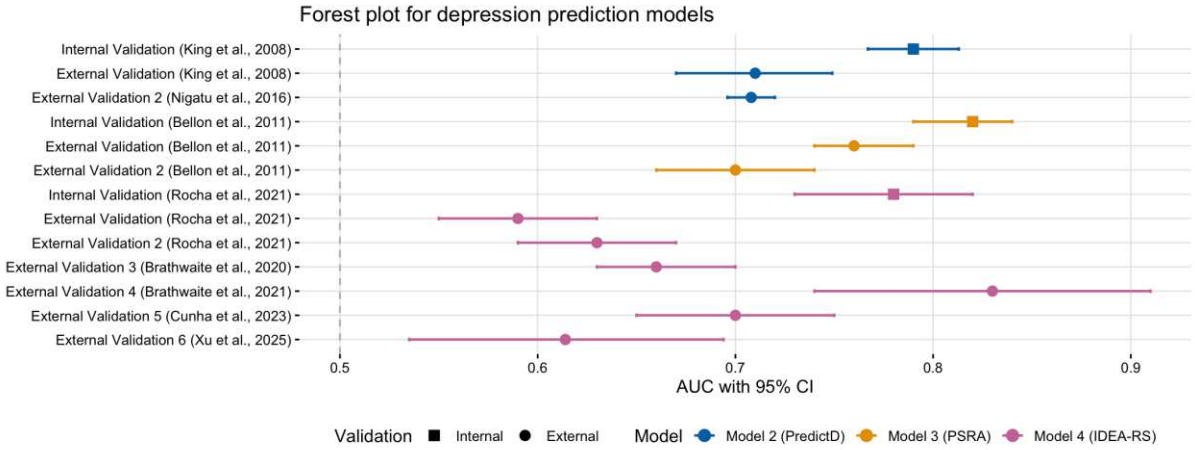


Figure 2. Forest plot for depression prediction models.

Severe Mental Illnesses

Bipolar Disorder. One prediction model was developed for bipolar disorder, the COBY risk calculator (Model 5: Birmaher et al., 2018). The model was developed in the clinical high risk COBY sample (United States) with 11 psychosocial and demographic features using Cox proportional hazards. The model had fair discrimination abilities (0.70-0.79) (AUC = 0.71; 95%CI: 0.67 - 0.74). The Hosmer-Lemeshow goodness-of-fit test indicated no significant deviation between observed and predicted outcomes (χ^2_{28} = 6.79, p = .56), confirming adequate model calibration (see Figure 3). The COBY risk calculator (Birmaher et al., 2018) had one external validation in the clinical high risk BIOS sample (United States). External validation also showed fair discrimination in the clinical high-risk

population (AUC = 0.75; 95%CI: 0.59 - 0.91) (see Figure 3). However, calibration was not reported for the validation analyses.

Psychosis. Nine models were developed for psychosis outcomes (Models 6-14). Three models showed considerable discrimination (0.80 - 0.89): AUC = 0.83; 95% CI: 0.73 - 0.93 (Model 12; Koutsouleris et al., 2020), c-index = 0.81; 95%CI: 0.80-0.83 (Model 11; Fusar-Poli et al., 2019) and c-index = 0.80; 95%CI: 0.79 - 0.82 (Model 9; Fusar-Poli et al., 2017). Model 12 reported no calibration, but Model 11 and Model 9 showed perfect calibration (calibration slope: 1; CITL: 0). Four models demonstrated fair discrimination (0.70-0.79): c-index = 0.77; 95%CI: 0.75 - 0.80 (Model 13: P-risk; Sullivan et al., 2022), AUC: 0.75; 95% CI: 0.67 - 0.83 (Model 14; Arslan et al., 2025), c-index = 0.74; 95%CI: 0.69 - 0.80 (Model 10: SIPS-RC; Zhang et al., 2019), c-index = 0.71; 95%CI: 0.64 - 0.78 (Model 6: NAPLS-2; Cannon et al., 2016). These models did not report any calibration statistics. However, Model 7 (Fusar-Poli et al., 2016) had poor discrimination (c-statistic = 0.66; 95%CI: 0.59–0.73) with a calibration of slope of 0.759 and calibration in the large of -2.405 (Figure 4). Model 8 using structural MRI features did not report c-index/AUC but instead reported balanced accuracy (BAC = 88.2%) (Zarogianni et al., 2017).

Each model had between 1-4 external validation reports (Figure 4). 3 models demonstrated fair discrimination (0.70-0.79): c-index = 0.79 (95%CI: 0.78 - 0.81; Model 11), c-index = 0.787 (95% CI: 0.781-0.794; Model 13: P-risk), average c-index = 0.72 (95%CI: 0.57 - 0.86; Model 6: NAPLS-2). 4 models showed poor discrimination (0.60-0.69): average c-index = 0.68 (Model 9; 95%CI: 0.67 - 0.70), average c-index = 0.675 (Model 12; 95%CI: 0.52 - 0.83), c-index = 0.65 (95% CI: 0.45 - 0.85; Model 10: SIPS-RC), average c-index = 0.615 (Model 14; 95%CI: 0.45 - 0.78). However, Model 7 had poor discrimination AUC = 0.51 (95% CI: 0.43–0.59) with a calibration slope of −0.164. Model 8 which used structural MRI features did not report c-index/AUC but instead reported balanced accuracy (BAC = 74.2%) (Zarogianni et al., 2019). Model 11, Model 13:P-risk and Model 6:NAPLS-2 were able to discriminate amongst incident psychosis populations. However, Model 13:P-risk does not report calibration, whereas Model 11 showed an excellent calibration slope of 0.996 and NAPLS-2 showed excellent calibration across predicted probabilities of 0.0 – 0.4 (Supplementary 4).

Although Model 12 demonstrated the strongest performance in internal validation, it was developed in a clinical high-risk population (PRONIA) and then reduced, via feature selection, to 7 clinical and neurocognitive features before undergoing external validation

Consequently, the internal validation performance does not fully translate to the external validation, where performance is attenuated when only the condensed predictor set is used. Model 12 also had a subsequent externally validated model using only sMRI features, however, this model was not included in this analysis as it was trained on the inner pooled cross-validation cycle of the clinical neurocognitive risk calculator. The overall best performing model for predicting psychosis was Model 11. Using 5 features (age, gender, ethnicity, age x gender interaction, index diagnosis) with Cox proportional hazards regression, the model was developed and validated on two different clinical high-risk population in the UK.

Overall, the discrimination of all models at the internal validation stage ranged from 0.66 - 0.83 with a mean AUC = 0.77 with narrow confidence intervals (95%CI: 0.72-0.81)). The discrimination of all models at the external validation stage ranged from 0.51 - 0.83, with a mean AUC of 0.69 with wide confidence intervals (95%CI: 0.61-0.77).

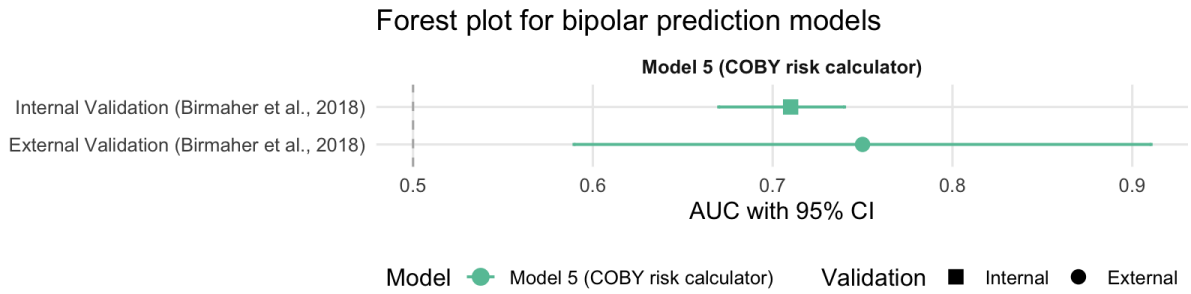


Figure 3. Forest plot for bipolar disorder prediction models.

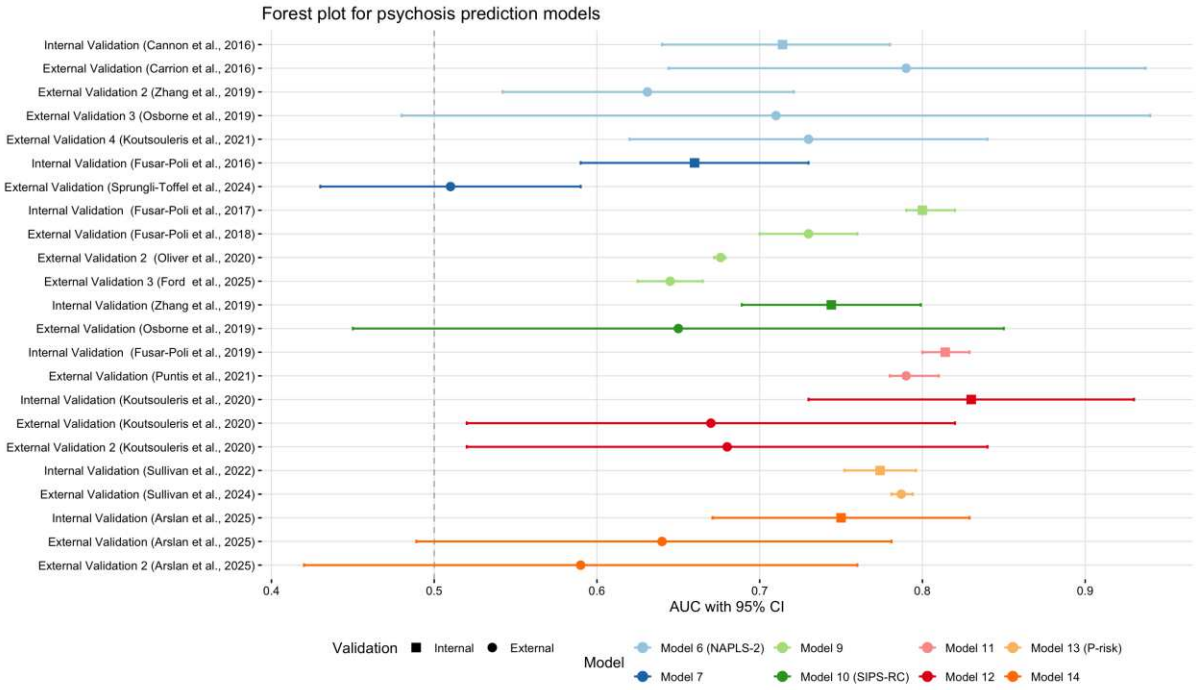


Figure 4. Forest plot for psychosis prediction models.*

*Model 8 not shown as AUC not reported.

PROBAST+AI quality assessment

Risk of bias assessment was conducted for each model's internal and external validation reports separately. The interrater agreement between the two assessors was high (Cohen's $\kappa = 0.913$,;95% CI: 0.80 – 1.00; 95.7% agreement). When development and best quality validation reports were considered together, 7 models (4 for psychosis: Model 6, Model 9, Model 11, Model 13: P-risk, and 3 for depression: Model 2: PredictD, Model 3: PRSA, Model 4: IDEA-RS) were marked as having an overall low risk of bias and high applicability, whereas 7 of them were marked as overall unclear(Supplementary 5). Every model appropriately reported on predictor assessments, preprocessing and timing, hence receiving a low risk of bias score. Participant and data sources section was reported as low risk and high applicability, except the COBY model (Model 5: Birmaher et al.; 2018). In addition, 1 model (Model 14) was marked as unclear in the Outcome section, as they had limited information on whether psychosis was determined in a similar way for all participant and whether they conducted the predictor assessment without knowledge of the outcome. Moreover, 7 models were marked as having unclear risk of bias in the Analysis section, due to lack of evidence on including a reasonable sample size, lack of model evaluation beyond apparent performance (e.g. internal validation), and/or missing information on how missing data were handled.

Model Development. Overall, 9 models were high quality and had high applicability for each assessment section (Participants, Predictors, Outcomes, Analysis). The rest were marked as unclear, either due to their analysis (n = 4), outcomes (n = 1) or participants and data sources (n = 1).

Model Evaluation. We assessed the risk of bias for 31 external validation reports, however, only the best external validation report for each model was included for final risk of bias assessment (n = 14). 7 models showed overall low risk of bias and high applicability. Similarly to model development, most models were marked as unclear in the Analysis section (n = 7), 2 had unclear outcomes and 1 was unclear in their data sources.

Clinical utility

Decision curve analysis was only tested in one manuscript (Model 9; Fusar-Poli et al., 2017). The manuscript reported on the exchange rate based on the number of individuals needed to treat and conducted decision curve analysis to evaluate net benefit across a range of clinically plausible thresholds (exchange rate: 1/13 = 7.69%). Although decision-analytic methods were used, threshold selection remained partly assumption-based. The rest of the manuscripts either discussed clinical utility through sensitivity/specificity thresholds (n = 18), or made no mention of clinical usefulness (n = 12) (Supplementary 3 and 4). Therefore, most models were evaluated solely on discrimination performance without assessing the potential impact on clinical decision-making.

Discussion

This review highlights the growing interest in prediction modelling for CMD and SMI in the last 10 years (as half of these models were published after 2017), while demonstrating that the literature remains limited by the small number of robustly developed and externally validated models. Only 14 unique externally validated prediction models met the inclusion criteria. Across the included reports, most models were developed using cohort data. Predictors were predominantly psychosocial and demographics features, with limited use of biological features. Although ML has received increased attention in recent years, few externally validated prediction models employed these methods. Model performance during development varied considerably across reports, however, discrimination was generally fair (mean AUC = 0.77). Comparatively, performance declined following external validation (mean AUC = 0.69), with many models demonstrating poor to fair discrimination.

Interpretation of these results was further limited by the inconsistent reporting of calibration statistics. Finally, clinical utility was tested in only one manuscript. These findings highlight a substantial gap between model development and evidence for implementation in clinical settings.

The highest performing model was the Model 3: PSRA which was developed for predicting incident depression by Bellon et al. (2011) with low risk of bias at internal validation. This is a promising model for depression as it also retained the highest mean external validation performance (AUC = 0.73) compared to Model 2: PredictD and Model 4: IDEA-RS. However, the lack of calibration statistics hinders the generalizability of the model as it's unclear how closely it predicts a true event of depression. In addition, the relatively older mean age of participants (49.2 years) may have contributed to the model's apparent performance, as incident depression in this age group could reflect a clinically distinct subgroup. Similarly, some "incident" cases may represent recurrent rather than first-onset depression. Even though an online version of the risk calculator was developed available for use in 2011 (not available anymore), clinical utility was not actually tested.

The lowest performing model for both internal and external validation was Model 7 (Fusar-Poli et al., 2016; Sprungli-Toffel et al., 2025). This model was developed to predict psychosis in a clinical population with relatively small sizes (710 and 518). The 8 predictors were selected by LASSO and only included demographics along with referral year and referral source. Compared to the best performing psychosis model (Model 11), this model used two more demographics (marital status and socioeconomic status), in addition to using referral year and source instead of index diagnosis. Although the model was based on simple predictors that are easily collected in routine clinical practice according to the authors, there was no further clinical utility testing conducted. These findings are further supported by the risk of bias assessment, which identified methodological uncertainty in the analysis section of these reports.

Furthermore, the two best performing models for predicting psychosis (Model 9 and Model 11) represent a variation of the same underlying model. Both were developed using the same clinical high-risk dataset and included the same 5 predictors: age, gender, ethnicity, age x gender interaction and index diagnosis; and implemented cox proportional hazards method. The key difference was the pre-processing of the age predictor: Model 9 modelled age as a linear variable, whereas Model 11 modelled age nonlinearly. This change resulted in a slight improvement in predictive performance for Model 11. However, both models

demonstrated strong performance in internal and external validation and perfect internal calibration, suggesting that the core model provides robust discrimination for psychosis risk prediction. This is also supported by the overall low risk of bias assessment for Model 11. Although, Model 9 was tested for clinical utility, indicating promising results for clinical practice, further testing is required to ensure the same for Model 11. The same model also implemented ML for the same predictors (Fusar-Poli et al., 2019), however, as the model was not externally validated, we had to exclude it from the final data extraction. The best performing psychosis models suggest that simple models with fewer predictors might benefit clinical populations. By contrast, predictions in the general population may require more complex models with more features to capture the diversity of cases. This is consistent with the drop of discrimination performance when externally validating Model 9 in a general population (c-index = 0.676; Oliver et al., 2020).

A notable model was Model 8 (Zarogianni et al., 2017), the only externally validated model that was developed using exclusively sMRI features and machine learning methods. While AUC and c-index were not reported for this model, preventing direct comparison with other models in this review, both internal and external validation reports reported good predictive accuracy (BAC > 70%). However, these findings should be interpreted with caution as both reports used small datasets and did not report calibration which limits model generalizability. This is consistent with their unclear risk of bias.

An additional methodological consideration is the inconsistency in predictor inclusion between internal and external validation reports for two models (Model 6: NAPLS-2 and Model 4: IDEA-RS). The NAPLS-2 model was originally developed using 8 predictors, but every external validation study was conducted using a reduced set of 6 predictors, excluding stressful life events and childhood traumas predictors. Similarly, IDEA-RS was initially developed with 11 predictors, however, at external validation predictors varied between 7-11 predictors instead. Specifically, the model was excluding skin color, variables relating to parent-child relationship or replacing skin color with ethnicity or caste system to better reflect the cultural aspects of the validation dataset. These adaptations complicate the interpretation because the validated models no longer correspond exactly to the original model. This is due to uncertainty about whether the adapted predictors capture the same underlying risk construct or introduce additional sources of bias that could partly explain the differences in model performance across settings. This uncertainty highlights the constraints of developing

models based on non-harmonised datasets as it wouldn't be directly applicable or easily externally validated.

The distinction between models developed in clinical high-risk (CHR) and general population samples deserves emphasis, as these represent fundamentally different prediction tasks. CHR samples are risk-enriched through help-seeking behaviour and sequential referral filters, such that much of the predictive work has been performed before any model is applied (Fusar-Poli et al., 2016). Within these pre-selected samples, transition rates are sufficiently high that even modest discrimination can translate into clinically meaningful positive predictive values, and simple models based on a handful of routinely recorded variables may suffice. In the general population, by contrast, the low incidence of psychotic disorders means that identical discrimination yields many false positives for each true case identified; models intended for population-level use must therefore achieve substantially better discrimination and, critically, well-calibrated absolute risk estimates before screening could be contemplated. The decline in performance when Model 9 was transported to a general population sample (Oliver et al., 2020) illustrates that CHR-derived models cannot be assumed to generalise beyond the risk-enriched settings in which they were developed, because both baseline risk and predictor-outcome associations differ. The field may therefore require parallel strategies: parsimonious models embedded within specialist early intervention services, and separate population-level models, most plausibly developed in large routinely collected healthcare datasets, designed for earlier detection at much lower prevalence.

Our findings also illustrate a tension inherent in external validation itself. Validation performance cannot be interpreted without reference to the relatedness of the development and validation samples: validation in a population closely resembling the development sample (reproducibility) is more likely to be favourable but tests little beyond the stability of the original estimates, whereas validation in a genuinely different setting (transportability) is a stronger and more informative test but conflates deficiencies of the model with real differences in case-mix, predictor measurement and outcome ascertainment (Justice et al., 1999; Debray et al., 2015). The IDEA-RS validations exemplify this trade-off: substituting culturally appropriate predictors improved local applicability at the cost of no longer strictly validating the original model. A single favourable external validation should therefore not be treated as conferring permanently "validated" status; performance is specific to setting, population and time, and can drift as incidence, measurement practices and care pathways

change (Van Calster et al., 2023). Reframing external validation as an ongoing programme of evaluation, recalibration and updating, rather than a one-off hurdle to be cleared, would better serve eventual implementation.

A related theme concerns the balance between model complexity and clinical utility. Discrimination alone is a poor guide to usefulness: whether a model improves decision-making depends on calibration and on the consequences of true and false positive classifications at clinically sensible thresholds, which are best assessed using decision-analytic measures such as net benefit (Vickers and Elkin, 2006). Calibration, which was inconsistently reported across the included models, is arguably the more important property for individual-level decision-making, as miscalibrated risk estimates systematically mislead clinicians and patients even when discrimination is adequate (Van Calster et al., 2019). It is notable that the best performing psychosis models in this review used only five simple predictors, whereas the sole neuroimaging-based model, although promising, relied on data that could not feasibly be collected at scale. Each additional predictor or data modality carries real costs, including measurement burden, missing data and degraded transportability across settings with different data infrastructures, which must be offset by demonstrable gains in net benefit rather than marginal improvements in AUC. Parsimonious models built on routinely recorded variables may therefore represent the most credible route to implementation, with complex multimodal models reserved for contexts in which their incremental value has been formally demonstrated.

Finally, this review suggests that the anticipated impact of machine learning (ML) on the field remains largely unrealised. Few externally validated models employed ML methods, while 19 of the development studies excluded at data extraction were ML-based models without published external validation, indicating that the recent wave of ML model development has not yet been matched by rigorous evaluation. There are reasons for both caution and optimism. In the low-dimensional clinical data used by most included models, systematic evidence indicates no performance benefit of ML over logistic regression (Christodoulou et al., 2019), and flexible algorithms are data hungry, prone to optimism and frequently poorly calibrated when trained on modest samples (van der Ploeg et al., 2014). The plausible advantage of ML lies elsewhere: in exploiting high-dimensional and multimodal data, including neuroimaging, genomics, digital phenotyping and free text from electronic health records, where non-linear and interactive structure cannot be pre-specified (Rajkomar et al., 2019; Chekroud et al., 2021). Realising this potential will require ML

models to be held to the same standards as regression-based approaches, including protocol registration, transparent reporting according to TRIPOD+AI (Collins et al., 2024), and assessment of calibration, clinical utility and external validity.

This review has several strengths. We searched 3 of the most relevant databases and included every relevant report up to April 2026. A second reviewer (LS) was involved at every stage of screening and data extraction to ensure result accuracy with high interrater agreement. Included prediction models were all critically appraised, extracted and transparently reported according to CHARMS checklist and TRIPOD-SRMA statement. All models were externally validated and were assessed for risk of bias through PROBAST-AI, which increases our confidence of the generalisability and methodological quality of those models. Our findings are consistent with previous work on mental health prediction models in several ways (Hunt et al., 2024; Senior et al., 2021). Similar imbalance was found in terms of psychosis prediction models having more papers and calibration statistics missing (Meehan et al., 2022). However, it also extends as this is the first report looking at externally validated models and assessing clinical utility.

Our review focused on externally validated models. While this meant that we included the most clinically relevant models to ensure their generalisability beyond its development sample and its use in clinical practice, focusing on externally validated models can have two downsides. First, recently developed models using ML methods may be excluded because their corresponding external validation has not been published yet (19 out of 54 excluded manuscripts were using ML). Second, older models might also be excluded as external validation was not standard practice at the time they were developed. Similarly, models using biological features were also frequently excluded due to a cross-sectional design.

A lack of consistent reporting meant that Model 8 was unable to be compared to the rest of the models due to missing measures of AUC. 10 confidence intervals had to be calculated approximately as they were not reported in the original study. Finally, while we aimed to assess clinical utility, measures of clinical utility were only reported for 1 model.

A key limitation of this literature is that models were developed using data spanning a wide age range (8-75) which makes it difficult to detangle whether models were discriminating mental health-disorders for age-specific groups. In addition, evidence was uneven for different outcomes with only one model for anxiety and one for bipolar disorder.

This limited evidence makes it impossible to compare the model methods and performance with alternative approaches.

These findings provide a comprehensive summary of externally validated prediction models for CMD and SMI which could be useful to both researchers and clinicians. For researchers, these findings highlight the importance of moving beyond model development and acknowledging the importance of external validation and clinical utility assessments; limited studies currently achieve this. Importantly, relatively few models adopted a truly biopsychosocial approach, underscoring the need to uncover the balance between using complex multimodal data and creating simply explainable models to be used in everyday practice. A focus on feasible models is necessary by consulting clinicians during the development phase. For clinicians, our review highlights existing models appear promising for implementation, but none have been rigorously tested for use in routine clinical practice. Implementation will require clear decision thresholds, integration into existing care pathways, and further external validation or updating. Future assessments of clinical utility should move beyond discrimination to decision-analytic measures such as net benefit, evaluated at thresholds agreed with clinicians and service users (Vickers and Elkin, 2006).

To conclude, there is ever-growing interest to develop mental health prediction models in recent years. This review summarises the diverse methodologies used to develop and validate prediction models for CMD and SMI. Performance is variable with discrimination measures as the best comparator, while calibration measures remain scarcely reported. Most of those models have not been adequately evaluated for use in clinical settings. Future work should focus on 1. externally validating already developed models, especially for anxiety and bipolar disorder 2. validate state-of-art machine learning models 3. test clinical utility of existing models 4. report calibration and decision-analytic measures of clinical utility as standard.

Online Methods

PROSPERO registration

This systematic review is reported according to the Transparent Reporting of multivariable prediction models for Individual Prognosis Or Diagnosis checklist for Systematic Reviews and Meta-Analyses (TRIPOD-SRMA; Snell et al., 2023; Supplementary 6). The review protocol was registered with PROSPERO “A systematic review of prediction models for the onset of mental health disorders” (CRD420251021730).

Searching and selection of articles

To identify primary articles that reported on risk prediction models for CMD (depression, anxiety) and SMI (bipolar disorder and psychosis), this review searched three databases via the Ovid platform including MEDLINE (1946 to current), APA PsycINFO (1806 to current), Embase (1980 to current) and up to April 2026. We used psychosis as an umbrella term to encompass any models for schizophrenia spectrum disorders and first episode psychosis. To identify models across the psychosis spectrum, the search strategy included both "psychos?s" and "schizo*" as search terms. Forward and backward citation searching was used to identify any additional risk prediction models as well as to match internal and external validation reports for each identified modele. Key search terms concepts included "common and severe mental health disorders", "incidence/onset", "prediction models", "external validation", and "model performance" (Full search strategy is available in Supplementary 7).

Eligibility criteria

Primary articles were included if they reported on the development and/or external validation of risk prediction models for the onset of depression, anxiety, bipolar disorder, or psychosis. The models were included only if the disorders were diagnosed based on the ICD or DSM criteria. Included models measured predictive features from the prenatal, childhood, or adolescent period (<18 years). Only models developed to predict the onset of depression, anxiety, psychosis and bipolar disorder in young adults (<30 years) were included.

We excluded any reports that developed models where the minimum age of the study population was above 30 years old. Reports that included young adults (<30 years), even if the upper age range extended beyond 30), were retained as the disorders of interest commonly have their first onset in young people (Arnett et al., 2014). Cross-sectional models were excluded as there was no temporal relationship between the predictors and the outcome. However, retrospective cross-sectional studies were included if they recorded historic variables such as childhood trauma, and retained the temporal relationship of the variables. Prediction models developed for treatment response or relapse outcomes were also excluded, along with any systematic reviews, meta-analyses or qualitative studies. The results were restricted to articles available in the English language and human studies.

Screening and data extraction

All search results were imported into Covidence software. Two independent reviewers (AT (100%) and LS (40%)) screened the study title and abstracts to identify reports

meeting the inclusion criteria. Full-text papers were retrieved for reports deemed eligible and assessed based on the inclusion and exclusion criteria by AT (100%) and LS (40%). Where the reviewers disagreed on eligibility, the article was jointly reassessed to reach a unanimous decision. In cases where consensus could not be achieved, a discussion with a third independent reviewer (JH, JD or NL) was held. Excluded reports were coded to document the reasons for exclusion. We used the Critical appraisal and data extraction for systematic reviews of prediction modelling studies checklist (The CHARMS checklist; Moon et al., 2014; Supplementary 8) to extract data from manuscripts meeting the inclusion criteria. The checklist includes key categories relating to prediction models such as data sources, model development, model performance and model evaluation. Reports were examined for measures of discrimination and calibration, and data were extracted for comparison between models where possible. Where manuscripts reported discrimination measures without reporting confidence intervals (95% CI), confidence intervals were calculated using the AUC value, standard error and sample sizes for event vs non-event. Cohen's kappa was used to calculate interrater reliability for title/abstract screening and full-text screening.

Following the screening, model development reports were matched to their external validation reports. Any model development reports that didn't conduct external validation were excluded from this review. External validation was defined as testing the model in new samples that don't overlap in participants (geographical variations, temporal variations etc.) (Riley et al., 2024). To ensure methodological quality, internal-external validation and cross validation methods were not considered as "external validation" due to sample overlap. We implemented the external validation exclusion for all prediction models to focus on reproducible and generalizable prediction models (Rampesk et al., 2021).

Narrative synthesis

Prediction models were categorised by predicted clinical outcome (CMD and SMI). A narrative synthesis of findings is provided for each prediction model based on predicted clinical outcome (e.g. depression, psychosis etc.), predictors included, data sources, methods of model development and external validation, methods performance, and clinical utility. This is complemented by a summary of methodological features for each model according to the CHARMS checklist.

To compare model performance, we assessed discrimination (e.g. AUC, Harrel's c-index etc.) and calibration (e.g. accuracy, precision etc.). We used the AUC guidelines below (Çorbacıoğlu & Aksel, 2023): fail (0.5 - 0.6), poor discrimination (0.6 - 0.7), fair

discrimination (0.7 - 0.8), considerable discrimination (0.8 - 0.9), and excellent discrimination (> 0.9). Any AUC value above 0.8 is generally considered clinically useful (Çorbacıoğlu & Aksel, 2023). AUC was reported as the universal measure of the models as it's the most commonly used metric in prediction tool literature (Cabot and Ross, 2023). Models that did not report AUC or c-index were not compared in the model performance as they are not comparable to AUC/c-index. We compared model performance for each disorder as well as the differences within CMD and SMI models. A meta-analysis was considered but deemed infeasible due to the substantial heterogeity of included prediction models; consequently, a pooled AUC would not have been clinically interpretable.

Risk of bias assessment

To assess the quality of the articles included, we used the Prediction model Risk Of Bias ASsessment Tool (PROBAST+AI) tool (Moons et al., 2025). This tool is organized into 4 domains: Participants and data sources, Predictors, Outcome and Analysis. The assessment differs between model development and model evaluation phase as model development includes signalling questions tailored to development and assesses the Quality and Applicability of the model. However, model evaluation includes signalling questions tailored to validation and assesses Risk of bias and Applicability. Signalling questions are answered as yes, probably yes, no, probably no, no information, or, when appropriate, not applicable. Quality, risk of bias and applicability are assessed as "Low", "Unclear" or "High". In this study, we also computed overall risk of the model based on quality, applicability and risk of bias concerns. In case of more than one external validation reports, we have only included the best quality report. This approach provides a synthesis of the evidence for the model's accuracy, its generalisability and the overall risk of bias from model development and external validation. This assessment was also conducted in conjunction with a second reviewer (LS; 50%). Cohen's kappa was used to calculate interrater reliability for risk of bias assessment.

References

Arnett, J. J., Žukauskienė, R., & Sugimura, K. (2014). The new life stage of emerging adulthood at ages 18–29 years: Implications for mental health. *The Lancet Psychiatry*, *1*(7), 569–576. https://doi.org/10.1016/S2215-0366(14)00080-7

Arslan, S., Kaşıkçı, M., Dağ, O., Şahin-Çevik, D., Çakmak, I. B., Vassos, E., Van Den Heuvel, M., & Toulopoulou, T. (2025). Multimodal prediction of psychotic-like experiences using elastic net modeling: External validation in a clinical sample. *Psychological Medicine*, *55*, e346. https://doi.org/10.1017/S0033291725102201

Atias, D., Ashri, S., Goldbourt, U., Benyamini, Y., Gilad-Bachrach, R., Hasin, T., Gerber, Y., & Obolski, U. (2025). Machine learning in epidemiology: An introduction, comparison with traditional methods, and a case study of predicting extreme longevity. *Annals of Epidemiology*, *110*, 23–33. https://doi.org/10.1016/j.annepidem.2025.07.024

Bellón, J. Á., Conejo-Cerón, S., Moreno-Peral, P., King, M., Nazareth, I., Martín-Pérez, C., Fernández-Alonso, C., Ballesta-Rodríguez, M. I., Fernández, A., Aiarzaguena, J. M., Montón-Franco, C., Ibanez-Casas, I., Rodríguez-Sánchez, E., Rodríguez-Bayón, A., Serrano-Blanco, A., Gómez, M. C., LaFuente, P., Del Mar Muñoz-García, M., Mínguez-Gonzalo, P., … De Dios Luna, J. (2013). Preventing the onset of major depression based on the level and profile of risk of primary care attendees: Protocol of a cluster randomised trial (the predictD-CCRT study). *BMC Psychiatry*, *13*(1), 171. https://doi.org/10.1186/1471-244X-13-171

Bird, V., Premkumar, P., Kendall, T., Whittington, C., Mitchell, J., & Kuipers, E. (2010). Early intervention services, cognitive–behavioural therapy and family intervention in early psychosis: Systematic review. *British Journal of Psychiatry*, *197*(5), 350–356. https://doi.org/10.1192/bjp.bp.109.074526

Birmaher, B., Merranko, J. A., Goldstein, T. R., Gill, M. K., Goldstein, B. I., Hower, H., Yen, S., Hafeman, D., Strober, M., Diler, R. S., Axelson, D., Ryan, N. D., & Keller, M. B. (2018). A Risk Calculator to Predict the Individual Risk of Conversion From Subthreshold Bipolar Symptoms to Bipolar Disorder I or II in Youth. *Journal of the*

American Academy of Child & Adolescent Psychiatry, *57*(10), 755-763.e4. https://doi.org/10.1016/j.jaac.2018.05.023

Brathwaite, R., Rocha, T. B.-M., Kieling, C., Gautam, K., Koirala, S., Mondelli, V., Kohrt, B., & Fisher, H. L. (2021). Predicting the risk of depression among adolescents in Nepal using a model developed in Brazil: The IDEA Project. *European Child & Adolescent Psychiatry*, *30*(2), 213–223. https://doi.org/10.1007/s00787-020-01505-8

Brathwaite, R., Rocha, T. B.-M., Kieling, C., Kohrt, B. A., Mondelli, V., Adewuya, A. O., & Fisher, H. L. (2020). Predicting the risk of future depression among school-attending adolescents in Nigeria using a model developed in Brazil. *Psychiatry Research*, *294*, 113511. https://doi.org/10.1016/j.psychres.2020.113511

Cabot, J. H., & Ross, E. G. (2023). Evaluating prediction model performance. *Surgery*, *174*(3), 723–726. https://doi.org/10.1016/j.surg.2023.05.023

Cannon, T. D., Yu, C., Addington, J., Bearden, C. E., Cadenhead, K. S., Cornblatt, B. A., Heinssen, R., Jeffries, C. D., Mathalon, D. H., McGlashan, T. H., Perkins, D. O., Seidman, L. J., Tsuang, M. T., Walker, E. F., Woods, S. W., & Kattan, M. W. (2016). An Individualized Risk Calculator for Research in Prodromal Psychosis. *American Journal of Psychiatry*, *173*(10), 980–988. https://doi.org/10.1176/appi.ajp.2016.15070890

Carrión, R. E., Cornblatt, B. A., Burton, C. Z., Tso, I. F., Auther, A. M., Adelsheim, S., Calkins, R., Carter, C. S., Niendam, T., Sale, T. G., Taylor, S. F., & McFarlane, W. R. (2016). Personalized Prediction of Psychosis: External Validation of the NAPLS-2 Psychosis Risk Calculator With the EDIPPP Project. *American Journal of Psychiatry*, *173*(10), 989–996. https://doi.org/10.1176/appi.ajp.2016.15121565

Catania, L. S., Hetrick, S. E., Newman, L. K., & Purcell, R. (2011). Prevention and early intervention for mental health problems in 0–25 year olds: Are there evidence-based models of care? *Advances in Mental Health*, *10*(1), 6–19. https://doi.org/10.5172/jamh.2011.10.1.6

Chekroud, A. M., Bondar, J., Delgadillo, J., Doherty, G., Wasil, A., Fokkema, M., Cohen, Z., Belgrave, D., DeRubeis, R., Iniesta, R., Dwyer, D., & Choi, K. (2021). The promise of machine learning in predicting treatment outcomes in psychiatry. *World Psychiatry*, *20*(2), 154–170. https://doi.org/10.1002/wps.20882

Christodoulou, E., Ma, J., Collins, G. S., Steyerberg, E. W., Verbakel, J. Y., & Van Calster, B. (2019). A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *Journal of Clinical Epidemiology*, *110*, 12–22. https://doi.org/10.1016/j.jclinepi.2019.02.004

Coie, J. D., Lochman, J. E., Terry, R., & Hyman, C. (1992). Predicting early adolescent disorder from childhood aggression and peer rejection. *Journal of Consulting and Clinical Psychology*, *60*(5), 783–792. https://doi.org/10.1037/0022-006X.60.5.783

Collins, G. S. (2024). TRIPOD+AI statement: Updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ*, q902. https://doi.org/10.1136/bmj.q902

Cunha, G. R., Caye, A., Pan, P., Fisher, H. L., Pereira, R., Ziebold, C., Bressan, R., Miguel, E. C., Salum, G. A., Rohde, L. A., Kohrt, B. A., Mondelli, V., Kieling, C., & Gadelha, A. (2023). Identifying Depression Early in Adolescence: Assessing the performance of a risk score for future onset of depression in an independent Brazilian sample. *Brazilian Journal of Psychiatry*. https://doi.org/10.47626/1516-4446-2022-2775

Debray, T. P. A., Vergouwe, Y., Koffijberg, H., Nieboer, D., Steyerberg, E. W., & Moons, K. G. M. (2015). A new framework to enhance the interpretation of external validation studies of clinical prediction models. *Journal of Clinical Epidemiology*, *68*(3), 279–289. https://doi.org/10.1016/j.jclinepi.2014.06.018

Ding, C., Qi, W., Tu, H., Wang, Y., Zhang, T., & Sun, H. (2025). Proteomic profiling of human plasma for anxiety and depression: Discovery of potential biomarkers and mechanistic insights. *Journal of Affective Disorders*, *391*, 120067. https://doi.org/10.1016/j.jad.2025.120067

Ford, E., Stone, J., Oliver, D., Fell, B., Roque, G., Robertson, S., Fusar-Poli, P., & Greenwood, K. (2025). Local adaptation and validation of a transdiagnostic risk calculator for first episode psychosis using mental health patient records. *Frontiers in Psychiatry*, *16*, 1584719. https://doi.org/10.3389/fpsyt.2025.1584719

Fusar-Poli, P., Davies, C., Rutigliano, G., Stahl, D., Bonoldi, I., & McGuire, P. (2019). Transdiagnostic Individualized Clinically Based Risk Calculator for the Detection of Individuals at Risk and the Prediction of Psychosis: Model Refinement Including Nonlinear Effects of Age. *Frontiers in Psychiatry*, *10*, 313. https://doi.org/10.3389/fpsyt.2019.00313

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*, *74*(5), 493. https://doi.org/10.1001/jamapsychiatry.2017.0284

Fusar-Poli, P., Rutigliano, G., Stahl, D., Schmidt, A., Ramella-Cravaro, V., Hitesh, S., & McGuire, P. (2016). Deconstructing Pretest Risk Enrichment to Optimize Prediction of

Psychosis in Individuals at Clinical High Risk. *JAMA Psychiatry*, *73*(12), 1260. https://doi.org/10.1001/jamapsychiatry.2016.2707

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. *Translational Psychiatry*, *9*(1), 259. https://doi.org/10.1038/s41398-019-0600-9

Fusar-Poli, P., Werbeloff, N., Rutigliano, G., Oliver, D., Davies, C., Stahl, D., McGuire, P., & Osborn, D. (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. *Schizophrenia Bulletin*, *45*(3), 562–570. https://doi.org/10.1093/schbul/sby070

Grotzinger, A. D., Werme, J., Peyrot, W. J., Frei, O., De Leeuw, C., Bicks, L. K., Guo, Q., Margolis, M. P., Coombes, B. J., Batzler, A., Pazdernik, V., Biernacka, J. M., Andreassen, O. A., Anttila, V., Børglum, A. D., Breen, G., Cai, N., Demontis, D., Edenberg, H. J., … Smoller, J. W. (2026). Mapping the genetic landscape across 14 psychiatric disorders. *Nature*, *649*(8096), 406–415. https://doi.org/10.1038/s41586-025-09820-3

Gu, W., Zhao, Q., Yuan, C., Yi, Z., Zhao, M., & Wang, Z. (2022). Impact of adverse childhood experiences on the symptom severity of different mental disorders: A cross-diagnostic study. *General Psychiatry*, *35*(2), e100741. https://doi.org/10.1136/gpsych-2021-100741

Hill, E. D., Kashyap, P., Raffanello, E., Wang, Y., Moffitt, T. E., Caspi, A., Engelhard, M., & Posner, J. (2025). Prediction of mental health risk in adolescents. *Nature Medicine*, *31*(6), 1840–1846. https://doi.org/10.1038/s41591-025-03560-7

Hunt, A., Law, H., Carney, R., Mulholland, R., Flores, A., Tudur Smith, C., Varese, F., Parker, S., Yung, A. R., & Bonnett, L. J. (2024). Systematic review of clinical prediction models for psychosis in individuals meeting At Risk Mental State criteria. *Frontiers in Psychiatry*, *15*, 1408738. https://doi.org/10.3389/fpsyt.2024.1408738

Justice, A. C., Covinsky, K. E., & Berlin, J. A. (1999). Assessing the Generalizability of Prognostic Information. *Annals of Internal Medicine*, *130*(6), 515–524. https://doi.org/10.7326/0003-4819-130-6-199903160-00016

Kaplow, J. B., & Widom, C. S. (2007). Age of onset of child maltreatment predicts long-term mental health outcomes. *Journal of Abnormal Psychology*, *116*(1), 176–187. https://doi.org/10.1037/0021-843X.116.1.176

King, M., Bottomley, C., Bellón-Saameño, J. A., Torres-Gonzalez, F., Švab, I., Rifel, J., Maaroos, H.-I., Aluoja, A., Geerlings, M. I., Xavier, M., Carraça, I., Vicente, B., Saldivia, S., & Nazareth, I. (2011). An international risk prediction algorithm for the onset of generalized anxiety and panic syndromes in general practice attendees: predictA. *Psychological Medicine*, *41*(8), 1625–1639. https://doi.org/10.1017/S0033291710002400

King, M., Walker, C., Levy, G., Bottomley, C., Royston, P., Weich, S., Bellón-Saameño, J. Á., Moreno, B., Švab, I., Rotar, D., Rifel, J., Maaroos, H.-I., Aluoja, A., Kalda, R., Neeleman, J., Geerlings, M. I., Xavier, M., Carraça, I., Gonçalves-Pereira, M., … Nazareth, I. (2008). Development and Validation of an International Risk Prediction Algorithm for Episodes of Major Depression in General Practice Attendees: The PredictD Study. *Archives of General Psychiatry*, *65*(12), 1368. https://doi.org/10.1001/archpsyc.65.12.1368

Kirkbride, J. B., Anglin, D. M., Colman, I., Dykxhoorn, J., Jones, P. B., Patalay, P., Pitman, A., Soneson, E., Steare, T., Wright, T., & Griffiths, S. L. (2024). The social determinants of mental health and disorder: Evidence, prevention and recommendations. *World Psychiatry*, *23*(1), 58–90. https://doi.org/10.1002/wps.21160

Koutsouleris, N., Dwyer, D. B., Degenhardt, F., Maj, C., Urquijo-Castro, M. F., Sanfelici, R., Popovic, D., Oeztuerk, O., Haas, S. S., Weiske, J., Ruef, A., Kambeitz-Ilankovic, L., Antonucci, L. A., Neufang, S., Schmidt-Kraepelin, C., Ruhrmann, S., Penzel, N., Kambeitz, J., Haidl, T. K., … PRONIA Consortium. (2021). Multimodal Machine Learning Workflows for Prediction of Psychosis in Patients With Clinical High-Risk Syndromes and Recent-Onset Depression. *JAMA Psychiatry*, *78*(2), 195. https://doi.org/10.1001/jamapsychiatry.2020.3604

Koutsouleris, N., Worthington, M., Dwyer, D. B., Kambeitz-Ilankovic, L., Sanfelici, R., Fusar-Poli, P., Rosen, M., Ruhrmann, S., Anticevic, A., Addington, J., Perkins, D. O., Bearden, C. E., Cornblatt, B. A., Cadenhead, K. S., Mathalon, D. H., McGlashan, T., Seidman, L., Tsuang, M., Walker, E. F., … Cannon, T. D. (2021). Toward Generalizable and Transdiagnostic Tools for Psychosis Prediction: An Independent Validation and Improvement of the NAPLS-2 Risk Calculator in the Multisite PRONIA Cohort. *Biological Psychiatry*, *90*(9), 632–642. https://doi.org/10.1016/j.biopsych.2021.06.023

McGorry, P. D., & Mei, C. (2018). Early intervention in youth mental health: Progress and future directions. *Evidence Based Mental Health*, *21*(4), 182–184. https://doi.org/10.1136/ebmental-2018-300060

Meehan, A. J., Lewis, S. J., Fazel, S., Fusar-Poli, P., Steyerberg, E. W., Stahl, D., & Danese, A. (2022). Clinical prediction models in psychiatry: A systematic review of two decades

of progress and challenges. *Molecular Psychiatry*, *27*(6), 2700–2708. https://doi.org/10.1038/s41380-022-01528-4

Memon, A., Taylor, K., Mohebati, L. M., Sundin, J., Cooper, M., Scanlon, T., & De Visser, R. (2016). Perceived barriers to accessing mental health services among black and minority ethnic (BME) communities: A qualitative study in Southeast England. *BMJ Open*, *6*(11), e012337. https://doi.org/10.1136/bmjopen-2016-012337

Moons, K. G. M., Damen, J. A. A., Kaul, T., Hooft, L., Andaur Navarro, C., Dhiman, P., Beam, A. L., Van Calster, B., Celi, L. A., Denaxas, S., Denniston, A. K., Ghassemi, M., Heinze, G., Kengne, A. P., Maier-Hein, L., Liu, X., Logullo, P., McCradden, M. D., Liu, N., … Van Smeden, M. (2025). PROBAST+AI: An updated quality, risk of bias, and applicability assessment tool for prediction models using regression or artificial intelligence methods. *BMJ*, *388*, e082505. https://doi.org/10.1136/bmj-2024-082505

Moons, K. G. M., De Groot, J. A. H., Bouwmeester, W., Vergouwe, Y., Mallett, S., Altman, D. G., Reitsma, J. B., & Collins, G. S. (2014). Critical Appraisal and Data Extraction for Systematic Reviews of Prediction Modelling Studies: The CHARMS Checklist. *PLoS Medicine*, *11*(10), e1001744. https://doi.org/10.1371/journal.pmed.1001744

Nigatu, Y. T., Liu, Y., & Wang, J. (2016). External validation of the international risk prediction algorithm for major depressive episode in the US general population: The PredictD-US study. *BMC Psychiatry*, *16*(1), 256. https://doi.org/10.1186/s12888-016-0971-x

Nigatu, Y. T., & Wang, J. (2019). External validation of the International Risk Prediction Algorithm for the onset of generalized anxiety and/or panic syndromes (The Predict A) in the US general population. *Journal of Anxiety Disorders*, *64*, 40–44. https://doi.org/10.1016/j.janxdis.2019.03.004

Oliver, D., Wong, C. M. J., Bøg, M., Jönsson, L., Kinon, B. J., Wehnert, A., Jørgensen, K. T., Irving, J., Stahl, D., McGuire, P., Raket, L. L., & Fusar-Poli, P. (2020). Transdiagnostic individualized clinically-based risk calculator for the automatic detection of individuals at-risk and the prediction of psychosis: External replication in 2,430,333 US patients. *Translational Psychiatry*, *10*(1), 364. https://doi.org/10.1038/s41398-020-01032-9

Osborne, K. J., & Mittal, V. A. (2019). External validation and extension of the NAPLS-2 and SIPS-RC personalized risk calculators in an independent clinical high-risk sample. *Psychiatry Research*, *279*, 9–14. https://doi.org/10.1016/j.psychres.2019.06.034

Puntis, S., Oliver, D., & Fusar-Poli, P. (2021). Third external replication of an individualised transdiagnostic prediction model for the automatic detection of individuals at risk of psychosis using electronic health records. *Schizophrenia Research*, *228*, 403–409. https://doi.org/10.1016/j.schres.2021.01.005

Rajkomar, A., Dean, J., & Kohane, I. (2019). Machine Learning in Medicine. *New England Journal of Medicine*, *380*(14), 1347–1358. https://doi.org/10.1056/NEJMra1814259

Ramspek, C. L., Jager, K. J., Dekker, F. W., Zoccali, C., & Van Diepen, M. (2021). External validation of prognostic models: What, why, how, when and where? *Clinical Kidney Journal*, *14*(1), 49–58. https://doi.org/10.1093/ckj/sfaa188

Riley, R. D., Archer, L., Snell, K. I. E., Ensor, J., Dhiman, P., Martin, G. P., Bonnett, L. J., & Collins, G. S. (2024). Evaluation of clinical prediction models (part 2): How to undertake an external validation study. *BMJ*, *384*, e074820. https://doi.org/10.1136/bmj-2023-074820

Rocha, T. B.-M., Fisher, H. L., Caye, A., Anselmi, L., Arseneault, L., Barros, F. C., Caspi, A., Danese, A., Gonçalves, H., Harrington, H. L., Houts, R., Menezes, A. M. B., Moffitt, T. E., Mondelli, V., Poulton, R., Rohde, L. A., Wehrmeister, F., & Kieling, C. (2021).

Identifying Adolescents at Risk for Depression: A Prediction Score Performance in Cohorts Based in 3 Different Continents. *Journal of the American Academy of Child & Adolescent Psychiatry*, *60*(2), 262–273. https://doi.org/10.1016/j.jaac.2019.12.004

Saelmans, A., Seinen, T., Pera, V., Markus, A. F., Fridgeirsson, E., John, L. H., Schiphof-Godart, L., Rijnbeek, P., Reps, J., & Williams, R. (2025). Implementation and Updating of Clinical Prediction Models: A Systematic Review. *Mayo Clinic Proceedings: Digital Health*, *3*(3), 100228. https://doi.org/10.1016/j.mcpdig.2025.100228

Salem, M., & Robenson, J. (2025). The Impact of Socioeconomic Factors on Mental Health: A Conceptual Framework. *Cureus*. https://doi.org/10.7759/cureus.88244

Senior, M., Fanshawe, T., Fazel, M., & Fazel, S. (2021). Prediction models for child and adolescent mental health: A systematic review of methodology and reporting in recent research. JCPP Advances, 1(3), e12034. https://doi.org/10.1002/jcv2.12034

Snell, K. I. E., Levis, B., Damen, J. A. A., Dhiman, P., Debray, T. P. A., Hooft, L., Reitsma, J. B., Moons, K. G. M., Collins, G. S., & Riley, R. D. (2023). Transparent reporting of multivariable prediction models for individual prognosis or diagnosis: Checklist for systematic reviews and meta-analyses (TRIPOD-SRMA). *BMJ*, *381*, e073538. https://doi.org/10.1136/bmj-2022-073538

Sprüngli-Toffel, E., Studerus, E., Curtis, L., Conchon, C., Alameda, L., Bailey, B., Caron, C., Haase, C., Gros, J., Herbrecht, E., Huber, C. G., Riecher-Rössler, A., Conus, P., Solida, A., Armando, M., Kapsaridi, A., Ducommun, M. M., Klauser, P., Plessen, K. J., … Andreou, C. (2025). Individualized pretest risk estimates to guide treatment decisions in patients with clinical high risk for psychotic disorders. *Spanish Journal of Psychiatry and Mental Health*, *18*(2), 133–140. https://doi.org/10.1016/j.sjpmh.2024.09.001

Sullivan, S. A., Kounali, D., Morris, R., Kessler, D., Hamilton, W., Lewis, G., Lilford, P., & Nazareth, I. (2022). Developing and internally validating a prognostic model (P Risk) to improve the prediction of psychosis in a primary care population using electronic health records: The MAPPED study. *Schizophrenia Research*, *246*, 241–249. https://doi.org/10.1016/j.schres.2022.06.031

Sullivan, S. A., Morris, R., Kounali, D., Kessler, D., Hamilton, W., Lewis, G., Lilford, P., & Nazareth, I. (2024). External validation of a prognostic model to improve prediction of psychosis: A retrospective cohort study in primary care. *British Journal of General Practice*, *74*(749), e854–e860. https://doi.org/10.3399/BJGP.2024.0017

Van Calster, B., McLernon, D. J., Van Smeden, M., Wynants, L., Steyerberg, E. W., & On behalf of Topic Group 'Evaluating diagnostic tests and prediction models' of the STRATOS initiative. (2019). Calibration: The Achilles heel of predictive analytics. *BMC Medicine*, *17*(1), 230. https://doi.org/10.1186/s12916-019-1466-7

Van Calster, B., Steyerberg, E. W., Wynants, L., & Van Smeden, M. (2023). There is no such thing as a validated prediction model. *BMC Medicine*, *21*(1), 70. https://doi.org/10.1186/s12916-023-02779-w

Van Der Ploeg, T., Austin, P. C., & Steyerberg, E. W. (2014). Modern modelling techniques are data hungry: A simulation study for predicting dichotomous endpoints. *BMC Medical Research Methodology*, *14*(1), 137. https://doi.org/10.1186/1471-2288-14-137

Venn, R., Sigrist, C., Rudzki, S., D'Cunha, N. M., Buchhorn, R., Koenig, J., Northey, J. M., Naumovski, N., & McKune, A. J. (2025). Mental health consequences of early life stress in children – The autonomic nervous system as a potential target for early detection and intervention? *Psychoneuroendocrinology*, *179*, 107505. https://doi.org/10.1016/j.psyneuen.2025.107505

Vickers, A. J., & Elkin, E. B. (2006). Decision Curve Analysis: A Novel Method for Evaluating Prediction Models. *Medical Decision Making*, *26*(6), 565–574. https://doi.org/10.1177/0272989X06295361

Vigo, D. V., Stein, D. J., Harris, M. G., Kazdin, A. E., Viana, M. C., Munthali, R., Munro, L., Hwang, I., Kessler, T. L., Manoukian, S. M., Sampson, N. A., Kessler, R. C., World Mental Health Survey Collaborators, Aguilar-Gaxiola, S., Alonso, J., Andrade, L. H., Benjet, C., Borges, G., Bruffaerts, R., … Xavier, M. (2025). Effective Treatment for Mental and Substance Use Disorders in 21 Countries. *JAMA Psychiatry*. https://doi.org/10.1001/jamapsychiatry.2024.4378

World Health Organization: WHO. (2025, September 30). *Mental disorders*. https://www.who.int/news-room/fact-sheets/detail/mental-disorders

Xu, E. Y., MacSweeney, N., Thng, G., Barbu, M. C., Shen, X., Kwong, A. S. F., Romaniuk, L., McIntosh, A., Lawrie, S. M., & Whalley, H. C. (2025). Assessing a prediction model for depression risk using an early adolescent sample with self-reported depression. *JCPP Advances*, *5*(2), e12276. https://doi.org/10.1002/jcv2.12276

Xu, M., Zhang, W., Hochwalt, P., Yang, C., Liu, N., Qu, J., Sun, H., DelBello, M. P., Lui, S., & Nery, F. G. (2022). Structural connectivity associated with familial risk for mental illness: A meta-analysis of diffusion tensor imaging studies in relatives of patients with severe mental disorders. *Human Brain Mapping*, *43*(9), 2936–2950. https://doi.org/10.1002/hbm.25827

Zarogianni, E., Storkey, A. J., Borgwardt, S., Smieskova, R., Studerus, E., Riecher-Rössler, A., & Lawrie, S. M. (2019). Individualized prediction of psychosis in subjects with an at-risk mental state. *Schizophrenia Research*, *214*, 18–23. https://doi.org/10.1016/j.schres.2017.08.061

Zarogianni, E., Storkey, A. J., Johnstone, E. C., Owens, D. G. C., & Lawrie, S. M. (2017). Improved individualized prediction of schizophrenia in subjects at familial high risk, based on neuroanatomical data, schizotypal and neurocognitive features. *Schizophrenia Research*, *181*, 6–12. https://doi.org/10.1016/j.schres.2016.08.027

Zhang, T., Li, H., Tang, Y., Niznikiewicz, M. A., Shenton, M. E., Keshavan, M. S., Stone, W. S., McCarley, R. W., Seidman, L. J., & Wang, J. (2018). Validating the Predictive Accuracy of the NAPLS-2 Psychosis Risk Calculator in a Clinical High-Risk Sample From the SHARP (Shanghai At Risk for Psychosis) Program. *American Journal of Psychiatry*, *175*(9), 906–908. https://doi.org/10.1176/appi.ajp.2018.18010036

Zhang, T., Xu, L., Tang, Y., Li, H., Tang, X., Cui, H., Wei, Y., Wang, Y., Hu, Q., Liu, X., Li, C., Lu, Z., McCarley, R. W., Seidman, L. J., Wang, J., & on behalf of the SHARP (ShangHai At Risk for Psychosis) Study Group. (2019). Prediction of psychosis in prodrome: Development and validation of a simple, personalized risk calculator. *Psychological Medicine*, *49*(12), 1990–1998. https://doi.org/10.1017/S0033291718002738

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- NMHSupplementaryInfoTsaligopolouAugust2026Copy.docx