Prediction models in first episode psychosis

Lee, Rebecca; Leighton, Samuel P.; Thomas, Lucretia; Gkoutos, Georgios; Wood, Stephen; Fenton, Sarah-Jane; Deligianni, Fani; Cavanagh, Jonathan; Mallikarjun, Pavan

DOI:
10.1192/bjp.2021.219

BJPsych | The British Journal of Psychiatry (2022)
Page 1 of 13. doi: 10.1192/bjp.2021.219

Review

Prediction models in first-episode psychosis: systematic review and critical appraisal

Rebecca Lee*, Samuel P. Leighton*, Lucretia Thomas, Georgios V. Gkoutos, Stephen J. Wood, Sarah-Jane H. Fenton, Fani Deligianni, Jonathan Cavanagh and Pavan K. Mallikarjun

Background

People presenting with first-episode psychosis (FEP) have heterogenous outcomes. More than 40% fail to achieve symptomatic remission. Accurate prediction of individual outcome in FEP could facilitate early intervention to change the clinical trajectory and improve prognosis.

Aims

We aim to systematically review evidence for prediction models developed for predicting poor outcome in FEP.

Method

A protocol for this study was published on the International Prospective Register of Systematic Reviews, registration number CRD42019156897. Following Preferred Reporting Items for Systematic Reviews and Meta-Analyses guidance, we systematically searched six databases from inception to 28 January 2021. We used the Checklist for Critical Appraisal and Data Extraction for Systematic Reviews of Prediction Modelling Studies and the Prediction Model Risk of Bias Assessment Tool to extract and appraise the outcome prediction models. We considered study characteristics, methodology and model performance.

Results

Thirteen studies reporting 31 prediction models across a range of clinical outcomes met criteria for inclusion. Eleven studies used logistic regression with clinical and sociodemographic predictor variables. Just two studies were found to be at low risk of bias. Methodological limitations identified included a lack of appropriate validation, small sample sizes, poor handling of missing data and inadequate reporting of calibration and discrimination measures. To date, no model has been applied to clinical practice.

Conclusions

Future prediction studies in psychosis should prioritise methodological rigour and external validation in larger samples. The potential for prediction modelling in FEP is yet to be realised.

Keywords

Schizophrenia; psychotic disorders; outcome studies; prediction; precision medicine.

Psychosis

Psychosis is a mental illness characterised by hallucinations, delusions and thought disorder. The median lifetime prevalence of psychosis is around 8 per 1000 of the global population.[1] Psychotic disorders, including schizophrenia, are in the top 20 leading causes of disability worldwide.[2] People with psychosis have heterogeneous outcomes. More than 40% fail to achieve symptomatic remission.[3] At present, clinicians struggle to predict long-term outcome in individuals with first-episode psychosis (FEP).

Prediction modelling

Prediction modelling has the potential to revolutionise medicine by predicting individual patient outcome.[4] Early identification of those with good and poor outcomes would allow for a more personalised approach to care, matching interventions and resources to those most at need. This is the basis of precision medicine. Risk prediction models have been successfully employed clinically in many areas of medicine; for example, the QRISK tool predicts cardiovascular risk in individual patients.[5] However, within psychiatry, precision medicine is not yet established within clinical practice. In FEP, precision medicine could enable rapid stratification and targeted intervention, thereby decreasing patient suffering and limiting treatment associated risks such as medication side-effects and intrusive monitoring.

Salazar de Pablo et al recently undertook a broad systematic review of individualised prediction models in psychiatry.[6] They found clear evidence that precision psychiatry has developed into an important area of research, with the greatest number of prediction models focusing on outcomes in psychosis. However, the field is hindered by methodological flaws such as lack of validation. Further, there is a translation gap, with only one study considering implementation into clinical practice. Systematic guidance for the development, validation and presentation of prediction models is available.[7] Further, the Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD) statement sets standards for reporting.[8] Models that do not adhere to these guidelines result in unreliable predictions, which may cause more harm than good in guiding clinical decisions.[9] Salazar de Pablo et al 's review was impressive in scope, but necessarily limited in detailed analysis of the specific models included.[6] Systematic reviews focusing on predicting the transition to psychosis[10,11] and relapse in psychosis have also been published.[12] In our present review, we focus on FEP with the aim to systematically review and critically appraise the prediction models for the prediction of poor outcomes.

* Joint first authors.

Method

We designed this systematic review in accordance with the Checklist for Critical Appraisal and Data Extraction for Systematic Reviews of Prediction Modelling Studies (CHARMS).[13] A protocol for this study was published with the International Prospective Register of Systematic Reviews (PROSPERO), under registration number CRD42019156897.

We developed the eligibility criteria under the Population, Index, Comparator, Outcome, Timing and Setting (PICOTS) guidance (see Supplementary Material available at https://doi.org/10.1192/bjp.2021.219). A study was eligible for inclusion if it utilised a prospective design, including patients diagnosed with FEP, and developed, updated or validated prognostic prediction models for any possible outcome, in any setting. We excluded non-English language studies, those where the full text was not available, those involving diagnostic prediction models and those where the outcome predicted was ≤3 months from baseline as we were interested in longer-term prediction.

We searched PubMed, PsycINFO, EMBASE, CINAHL Plus, Web of Science Core Collection and Google Scholar, from inception up to 28 January 2021. In addition, we manually checked references cited in the systematically searched articles. The search terms were based around three themes: 'Prediction', 'Outcome' and 'First Episode Psychosis' terms. The full search strategy is available in the Supplementary Material. Two reviewers (R.L. and L.T.) independently screened the titles and abstracts. Full-text screening was completed by three independent reviewers (R.L., P.K.M. and S.P.L.). Disagreements were resolved by consensus.

Data extraction was conducted independently by two reviewers (R.L. and S.P.L.), following recommendations in the CHARMS checklist.[13] From all eligible studies, we collected information on study characteristics, methodology and performance. Study characteristics collected included first author name, year, region, whether the study was multicentre, study type, setting, participant description, outcome, outcome timing, predictor categories and number of models presented. Methodology considered sample size, events per variable (EPV), number of events in validation data-set, number of candidate and retained predictors, methods of variable selection, presence and handling of missing data, modelling strategies, shrinkage, validation strategies (see below), whether models were recalibrated, if clinical utility was assessed and whether the full models were presented. Steyerberg and Harrell outline a hierarchy of validation strategies from apparent (which assesses model performance on the data used to develop it and will be severely optimistic) to internal (via cross-validation or bootstrapping), internal–external (e.g. validation across centres in the same study) and external validation (to assess if models generalise to related populations in different settings).[14] Apparent, internal and internal–external validation use the derivation data-set only, whereas external validation requires the addition of a validation data-set. Performance for the best-performing model per outcome in each article was considered by model validation strategy, including model discrimination (reported as the C-statistic, which is equal to the area under the receiver operating characteristic curve for binary outcomes), calibration, other global performance measures and classification metrics. If not reported, where possible, the balanced accuracy (sensitivity + specificity / 2) and the prognostic summary index (positive + negative predictive value – 1) were calculated.

Two reviewers (R.L. and S.P.L.) independently assessed the risk of bias in included studies by using the Prediction Model Risk Of Bias Assessment Tool (PROBAST), a risk-of-bias assessment tool designed for systematic reviews of diagnostic or prognostic prediction models.[15,16] We considered all models reported in each article and assigned an overall rating to the article. PROBAST uses a structured approach with signalling questions across four domains: 'participants', 'predictors', 'outcome' and 'statistical analysis'. Signalling questions are answered 'yes', 'probably yes', 'no', 'probably no' or 'no information'. Answering 'yes' indicates a low risk of bias, whereas answering 'no' indicates high risk of bias. A domain where all signalling questions are answered as 'yes' or 'probably yes' indicates low risk of bias. Answering 'no' or 'probably no' flags the potential for the presence of bias, and reviewers should use their personal judgement to determine whether issues identified have introduced bias. Applicability of included studies to the review question is also considered in PROBAST.

We reported our results according to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) 2020 Statement (see Supplementary Material).[17]

Results

Systematic review of the literature yielded 2353 records from database searches and 67 from additional sources. After removal of duplicates, 1543 records were screened. Of these, 82 full texts were reviewed, which resulted in 13 studies meeting criteria for inclusion in our qualitative synthesis (Fig. 1).[18–30]

Study characteristics are summarised in Table 1. The 13 included studies, comprising a total of 19 different patient cohorts, reported 31 different prediction models. Dates of publication ranged from 2006 to 2021. Twelve studies (92%) recruited participants from Europe, with two studies (15%) also recruiting participants from Israel and one study (8%) from Singapore. Over two-thirds ($n = 9$) of studies were multicentre. Ten studies (77%) included participants from cohort studies, three studies (23%) included participants from randomised controlled trials and two studies (15%) included participants from case registries. Two studies (15%) included only out-patients, four (31%) included in-patients and out-patients, and the rest did not specify their setting. Cohort sample size ranged from 47 to 1663 patients. The average age of patients ranged from 21 to 28 years, and 49–77% of the cohorts were male. Where specified, the average duration of untreated psychosis ranged from 34 to 106 weeks. Ethnicity was reported in eight studies (62%), with the percentage of Black and minority ethnic patients in the cohorts ranging from 4 to >75%. The definition of FEP was primarily non-affective psychosis in the majority of patient cohorts, with the minority also including affective psychosis, and two cohorts also including drug-induced psychosis. All but one study (92%) considered solely sociodemographic and clinical predictors. A wide range of outcomes were assessed across the 13 included studies, including symptom remission in five studies (38%), global functioning in five studies (38%), vocational functioning in three studies (23%), treatment resistance in two studies (15%), hospital readmission in two studies (15%) and quality of life in one study (8%). All of the outcomes were binary. The follow-up period of included studies ranged from 1 to 10 years.

Study prediction-modelling methodologies are outlined in Table 2. Nine (69%) studies pertained solely to model development, with the highest level of validation reported being apparent validity in four of the studies, internal validity in three of the studies and internal–external validity (via leave-one-site-out cross-validation) in two of the studies. The remaining four (31%) studies also included a validation cohort and reported external validity. High dimensionality was common across the study cohorts, with the majority having a very low EPV ratio and up to 258 candidate predictors considered. Some form of variable selection was used in the majority (62%) of studies. The number of events in the external validation cohort ranged from 23 to 173. All of the studies had missing data. Six studies (46%) used complete-case analysis, five (38%) studies used single imputation and the remaining two (15%) studies applied multiple imputation.

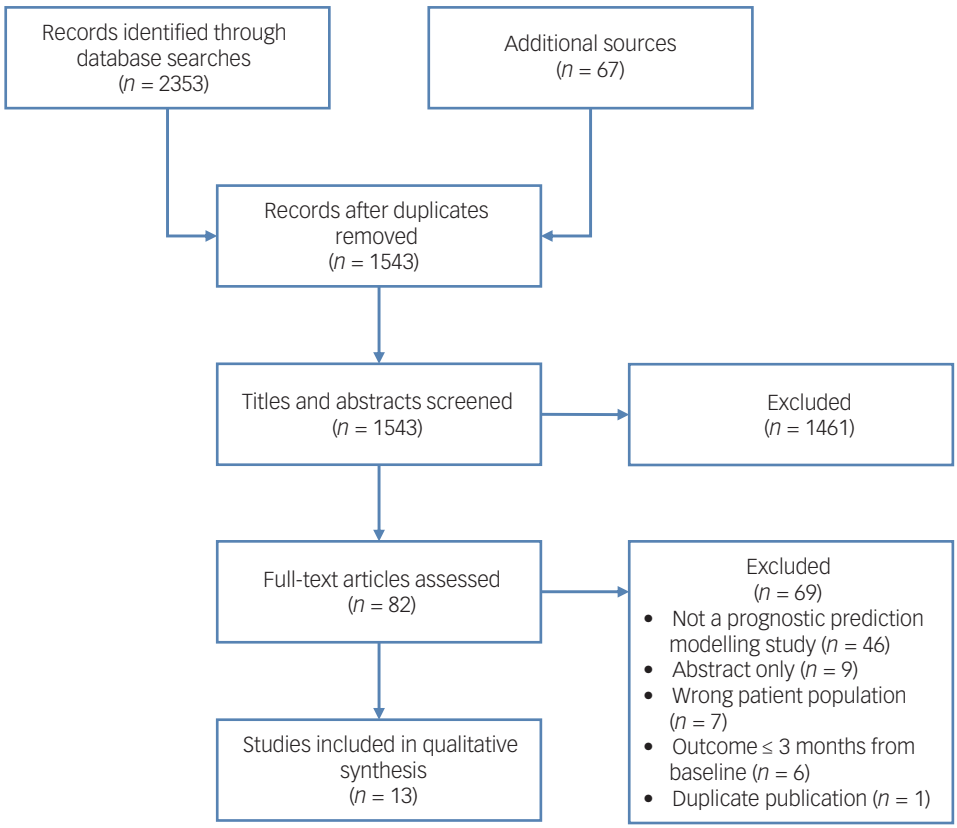

Fig. 1 Preferred Reporting Items for Systematic Reviews and Meta-Analyses flow diagram.

The most common modelling methodology was logistic regression fitted by maximum likelihood estimation, followed by logistic regression with regularisation. Only two studies used machine learning methods, both via support vector machines. Just over half of the studies (54%) did not use any variable shrinkage, and only three (23%) studies recalibrated their models based on validation to improve performance. The full model was presented in seven (54%) studies. Only two (15%) studies assessed clinical utility.

The performance of the best model per study outcome grouped by method of validation to allow for appropriate comparisons is reported in Table 3. For the five studies (38%) reporting only apparent validity, two reported a measure of discrimination and only one considered calibration. For the seven (54%) studies reporting internal validation performance, four reported discrimination with a C-statistic ranging from 0.66 to 0.77, and four reported calibration. For the three (23%) studies reporting internal–external validation, only one study considered discrimination with a C-statistic, which ranged from 0.703 to 0.736 across each of its four models. None of the studies reporting internal–external validation considered any measure of calibration. All four (31%) studies reporting external validation considered model discrimination, with C-statistics ranging from 0.556 to 0.876. However, only two of these studies considered calibration. Table 3 also records any global performance metrics, including the Brier score and McFadden's pseudo-R^2, both of which incorporate aspects of discrimination and calibration. Various classification metrics were reported across the study models, but it is difficult to make any meaningful comparisons between these alone, without considering the models' corresponding discrimination and calibration metrics, which were not universally reported.

We applied the PROBAST tool to the 31 different prediction models across the 13 studies in our systematic review, and determined an overall risk-of-bias rating for each study, as summarised in Supplementary Table 1. The majority (85%) of studies had an overall 'high' risk of bias. In each of these studies, the risk of bias was rated 'high' in the analysis domain, with one study also having a 'high' risk of bias in the predictors domain. The main reasons for the 'high' risk of bias in the analysis domain were insufficient participant numbers and consequently low EPV, inappropriate methods of variable selection including via univariable analysis, a lack of appropriate validation with only apparent validation, an absence of reported measures of discrimination and calibration, and inappropriate handling of missing data by either complete-case analysis or single imputation. Two studies, Leighton et al[29] and Puntis et al,[30] were rated overall 'low' risk of bias. These studies considered symptom remission and psychiatric hospital readmission outcomes, respectively. Both studies externally validated their prediction model and considered its clinical utility. However, neither study considered the implementation of the prediction model into actual clinical practice. When we assessed the 13 included studies according to PROBAST applicability concerns, all of the studies were considered overall 'low' concern. This is indicative of the broad scope of our systematic review.

Discussion

Our systematic review identified 13 studies reporting 31 prognostic prediction models for the prediction of a wide range of clinical outcomes. The majority of models were developed via logistic regression. There were several methodological limitations identified, including a lack of appropriate validation, issues with handling missing data and a lack of reporting of calibration and discrimination measures. We identified two studies with models at low risk

Table 1 Study characteristics

						Participants included in modelling					Outcome			
Study	Country	Multicentre	Recruitment dates	Type of study	Setting	Gender (% male)	Age (mean years)	Ethnicity	DUP (mean weeks)	FEP definition	Definition	Timing	Predictor categories	Number of models
Ajnakina et al, 2020[18]	UK	No	Dec 2005 to Oct 2010	Cohort	In-patients and out-patients	67.5%	27.2 (at baseline)	39.9% White, 60.1% Black	34.3	Non-affective	Early treatment resistance from illness onset, later treatment resistance	Follow-up for 5 years	Sociodemographic, clinical	4
Bhattacharyya et al, 2021[19]	UK	No	Sample 1: 1 Apr 2006 to 31 Mar 2012; sample 2: 12 Apr 2002 to 26 Jul 2013	Sample 1: case registry; sample 2: cohort	Sample 1: out-patients; sample 2: out-patients	Sample 1: 63.9%; sample 2: 60%	Sample 1: 24.4 (at onset); sample 2: 28.1 (at onset)	Sample 1: 31.1% White, 50.6% Black; sample 2: 34.2% White, 54.2% Black	Not reported	Sample 1: non-affective and affective; sample 2: non-affective and affective	Psychiatric hospital readmission	Follow-up for 2 years	Sociodemographic, clinical	3
Chua et al, 2019[20]	Singapore	No	2001–2012	Cohort	Not reported	49.2%	27.5 (at baseline)	76.7% Chinese	65.4	Non-affective	EET status	At 2 years	Sociodemographic, clinical	2
Demjaha et al, 2017[21]	UK	Yes	Sep 1997 to Aug 1999	Cohort	Not reported	58.4%	28.9 (at onset)	48.2% White, 39.8% Black	Not reported	Non-affective and affective	Early treatment resistance from illness onset	Follow-up for 10 years	Sociodemographic, clinical	1
De Nijs, 2019[22]	The Netherlands and Belgium	Yes	8 Jan 2004 to 6 Feb 2008	Cohort	In-patients and out-patients	76.9%	27.6 (at baseline)	85.9% White	Not reported	Non-affective	Andreasen symptom remission (6-month duration) GAF ≥65	At 3 years and 6 years	Sociodemographic, clinical, genetic, environmental	8
Derks et al, 2010[23]	Austria, Belgium, Bulgaria, Czech Republic, Germany, France, Israel, Italy, The Netherlands, Poland, Rumania, Spain, Sweden and Switzerland	Yes	23 Dec 2002 to 14 Jan 2006	Randomised controlled trial	Not reported	56.5%	26.0 (at baseline)	Not reported	Not reported	Non-affective	Andreasen symptom remission (6-month duration)	Follow-up for 1 year	Sociodemographic, clinical	1
Flyckt et al, 2006[24]	Sweden	Yes	1 Jan 1996 to 31 Dec 1997	Cohort	Not reported	52.9%	28.8 (at baseline)	Not reported	62.4	Non-affective and affective (with mood-incongruent delusions)	Global functioning (independent living, EET status and GAF score ≥60)	At mean of 5.4 years	Sociodemographic, clinical	1
González-Blanch et al, 2010[25]	Spain	No	Feb 2001 to Feb 2005	Cohort	Not reported	62%	26.6 (at baseline)	Not reported	66.6	Non-affective	Global functioning (EET status and DAS score ≤1)	At 1 year	Sociodemographic, clinical	1
Koutsouleris et al, 2016[26]	Austria, Belgium, Bulgaria, Czech Republic, Germany, France, Israel, Italy, The Netherlands, Poland, Rumania, Spain, Sweden and Switzerland	Yes	23 Dec 2002 to 14 Jan 2006	Randomised controlled trial	Not reported	56%	26.1 (at baseline)	Not reported	Not reported	Non-affective	GAF score ≥65	At 1 year	Sociodemographic, clinical	1

(*Continued*)

Table 1 (*Continued*)

						Participants included in modelling					Outcome			
Study	Country	Multicentre	Recruitment dates	Type of study	Setting	Gender (% male)	Age (mean years)	Ethnicity	DUP (mean weeks)	FEP definition	Definition	Timing	Predictor categories	Number of models
Leighton et al, 2019[27]	UK	Yes	Development sample: 2011 to 2014; validation sample: 1 Sep 2006 to 31 Aug 2009	Development sample: cohort; validation sample: cohort	Development sample: in-patients and out-patients; validation sample: in-patients and out-patients	Development sample: 66%; validation sample: 68%	Development sample: 25.2 (at baseline); validation sample: 24.6 (at baseline)	Development sample: 81% White; validation sample: 96% White	Not reported	Development sample: non-affective and affective; validation sample: non-affective and affective	EET status, Andreasen symptom remission (no duration criteria), Andreasen symptom remission (6 months duration)	At 1 year	Sociodemographic, clinical	3
Leighton et al, 2019[28]	UK and Denmark	Yes	Development sample: Aug 2005 to Apr 2009; validation sample UK: 1 Sep 2006 to 31 Aug 2009 and 2011–2014; validation sample Denmark: Jan 1998 to Dec 2000	Development sample: cohort; validation sample UK: 2 cohort studies; validation sample Denmark: randomised controlled trial	Development sample: not reported; validation sample UK: in-patients and out-patients; validation sample Denmark: in-patients and out-patients	Development sample: 69%; validation sample UK: 67%; validation sample Denmark: 59%	Development sample: 21.3 (at baseline); validation sample UK: 24.9 (at baseline); validation sample Denmark: 26.6 (at baseline)	Development sample: 73% White; validation sample UK: 88% White; validation sample Denmark: 94% White	Development sample: 44; validation sample UK: 44.4; validation sample Denmark: 106	Development sample: non-affective, affective and drug-induced; validation sample UK: non-affective and affective; validation sample Denmark: non-affective	EET status, GAF score ≥65, Andreasen symptom remission (6-month duration, quality of life	At 1 year	Sociodemographic, clinical	4
Leighton et al, 2021[29]	UK	Yes	Development sample: Aug 2005 to Apr 2009; validation sample: Apr 2006 to Feb 2009	Development sample: cohort; validation sample: cohort	Not reported	Development sample: 68.8%; validation sample: 61.8%	Development sample: 22.6 (at baseline); validation sample: 25.0 (at baseline)	Not reported	Development sample: 41.3; validation sample: 48.9	Development sample: non-affective, affective and drug-induced; validation sample: non-affective, affective and drug-induced	Andreasen symptom remission (6-month duration)	At 1 year	Sociodemographic, clinical	1
Puntis et al, 2021[30]	UK	Yes	Development sample: 1 Jan 2011 to 8th Oct 2019; validation sample: 31 Jan 2006 to 18 Jun 2019	Development sample: case registry; validation sample: case registry	Development sample: out-patients; validation sample: out-patients	Development sample: 63%; validation sample: 63%	Development sample: 25.6 (at baseline); validation sample: 26.7 (at baseline)	Development sample: 74.8% White; validation sample: 35.4% White	Not reported	Not reported	Psychiatric hospital admission after discharge from early intervention	Follow-up for 1 year	Sociodemographic, clinical	1

DUP, duration of untreated psychosis; FEP, first-episode psychosis; EET, employment, education or training; GAF, Global Assessment of Functioning; DAS, Disability Assessment Schedule.

Table 2 Study methodology

Study	Sample Size	EPV	Number of events in validation data-set	Number of candidate predictors	Number of retained predictors	Variable selection	Missing data per predictor	Handling of missing data	Modelling method	Shrinkage	Validation method reported	Re-calibration performed	Full model presented	Clinical usefulness assessed
Ajnakina et al, 2020[18]	Recruited: 283; included in modelling: 190 to 222	2 to 4	No external validation	13	12 to 13	Full model approach or LASSO	up to 59.9%	Single imputation	Logistic regression via ridge and LASSO	Penalised estimation and then uniform	Internal	Yes	Yes	No
Bhattacharyya et al, 2021[19]	Sample 1: 1738 recruited, 1663 included in modelling; sample 2: 240 recruited, 240 included in modelling	4 to 62	No external validation	10 to 21	10 to 21	Full model approach	Sample 1: up to 4.3%; sample 2: none	Complete-case analysis	Logistic regression via MLE	None	Apparent and internal	No	Yes	No
Chua et al, 2019[20]	Recruited: 1724; included in modelling: 1177	16	No external validation	22	22	Full model approach	Yes but not reported	Complete-case analysis	Logistic regression via MLE	None	Apparent	No	No	No
Demjaha et al, 2017[21]	Recruited: 557; included in modelling: 286	8	No external validation	8	6	LASSO	Yes but not reported	Complete-case analysis	Logistic regression via LASSO	Penalised estimation	Internal	No	Yes	No
De Nijs, 2019[22]	Recruited: 1100; included in modelling: 442 to 523	2	No external validation	258	119 to 152	Recursive feature elimination	up to 20%	Single imputation	Linear support vector machine	None	Internal and internal–external	No	No	No
Derks et al, 2010[23]	Recruited: 498; included in modelling: 297	9 to 18	No external validation	10 to 20	10 to 20	Full model approach	Yes but not reported	Complete-case analysis	Logistic regression via MLE	None	Apparent	No	No	No
Flyckt et al, 2006[24]	Recruited 175; included in modelling: 111	2	No external validation	32	5	Forward selection	Yes but not reported	Complete-case analysis	Logistic regression via MLE	None	Apparent	No	Yes	No
González-Blanch et al, 2010[25]	Recruited: 174; included in modelling: 92	4	No external validation	23	2	Univariate significance testing ($P < 0.1$) then forward selection	Yes but not reported	Complete-case analysis	Logistic regression via MLE	None	Apparent	No	Yes	No
Koutsouleris et al, 2016[26]	Recruited: 498; included in modelling: 334	<1	No external validation	189	Not reported	Forward selection	up to 20%	Single imputation	Nonlinear support vector machine	None	Internal and internal–external	No	No	No
Leighton et al, 2019[27]	Development sample: 83 recruited, 67 to 75 included in modelling; validation sample: 79 recruited, 64 to 67 included in modelling	<1	27 to 46	56	5 to 13	Elastic net	Development sample: up to 13%; validation sample: up to 37%	Single imputation	Logistic regression via elastic net	Penalised estimation	External	No	No	No
Leighton et al, 2019[28]	Development sample: 1027 recruited, 673 to 829 included in modelling; validation sample UK: 162 recruited, 47 to 142 included; validation sample Denmark: 578 recruited, 226 to 553 included	1 to 2	23 to 173	163	17 to 26	Elastic net	Development sample: up to 20%; validation sample: yes but not reported	Single imputation	Internal validation: logistic regression via elastic net; external validation: logistic regression via MLE	Internal–external validation: penalised estimation; external validation: none	Internal–external and external	No	No	No
Leighton et al, 2021[29]	Development sample: 1027 recruited, 673 included in modelling; validation sample: 399 recruited, 191 included	25	103	14	14	Full model approach	Development sample: up to 14.9%; validation sample: up to 56.5%	Multiple imputation	Logistic regression via MLE	Uniform	Internal and external	Yes	Yes	Yes
Puntis et al, 2021[30]	Development sample: recruited not reported; 831 included in modelling; validation sample: recruited not reported; 1393 included	10	162	8	8	Full model approach	Development sample: up to 15.4%; validation sample: up to 5.5%	Multiple imputation	Logistic regression via MLE	Uniform	Internal and external	Yes	Yes	Yes

EPV, events per variable; LASSO, least absolute shrinkage and selection operator; MLE, maximum likelihood estimation.

Table 3 Performance metrics for best model per outcome in each study

Study	Outcome	Discrimination C-statistic	Calibration	Other global performance metrics	Classification metrics
Studies reporting apparent validity					
Bhattacharyya et al, 2021[19]	Psychiatric hospital readmission	0.749	Calibration plot only; No α or β	Brier score 0.192	Not reported
Chua et al, 2019[20]	EET status at 2 years	0.759 (95% CI 0.728–0.790)	Not reported	Not reported	Classification accuracy 0.759; PPV 0.64; NPV 0.78; PSI 0.42
Derks et al, 2010[23]	Andreasen symptom remission (6-month duration) with 1 year follow-up	Not reported	Not reported	Not reported	Classification accuracy 0.63; balanced accuracy 0.665; sensitivity 0.73; specificity 0.60; PPV 0.73; NPV 0.61; PSI 0.34
Flyckt et al, 2006[24]	Global functioning (independent living, EET status, GAF score ≥60) at mean 5.4 years	Not reported	Not reported	Not reported	Classification accuracy 0.81; balanced accuracy 0.805; sensitivity 0.84; specificity 0.77
González-Blanch et al, 2010[25]	Global functioning (EET status, DAS score ≤1) at 1 year	Not reported	Hosmer–Lemeshow test $P \geq 0.05$	Not reported	Classification accuracy 0.750; balanced accuracy 0.587; sensitivity 0.261; specificity 0.913; PPV 0.500; NPV 0.788; PSI 0.288
Studies reporting internal validity					
Ajnakina et al, 2020[18]	Early treatment resistance from illness onset with 5-year follow-up	0.77	$\alpha = 0.028$; $\beta = 1.264$; no calibration plot	Not reported	Balanced accuracy 0.5; sensitivity 0; specificity 1.00; PPV 0.48, NPV 0.84; PSI 0.32
	Later treatment resistance with 5-year follow-up	0.77	$\alpha = 0.504$; $\beta = 1.838$; no calibration plot	Not reported	Balanced accuracy 0.81; sensitivity 0.62; specificity 1.00; PPV 0.42; NPV 1.00; PSI 0.42
Bhattacharyya et al, 2021[19]	Psychiatric hospital readmission	0.66	Calibration plot only; no α or β	Brier score 0.232	Not reported
Demjaha et al, 2017[21]	Early treatment resistance from illness onset with 10-year follow-up	Not reported	Not reported	Brier score 0.146; McFadden pseudo R^2 0.1	Not reported
De Nijs, 2019[22]	Andreasen symptom remission (6-month duration) at 3 years	Not reported	Not reported	Not reported	Balanced accuracy 0.644; sensitivity 0.76; specificity 0.50; PPV 0.722; NPV 0.548; PSI 0.27
	GAF score ≥65 at 3 years	Not reported	Not reported	Not reported	Balanced accuracy 0.676; sensitivity 0.749; specificity 0.584; PPV 0.701; NPV 0.642; PSI 0.343
	Andreasen symptom remission (6-month duration) at 6 years	Not reported	Not reported	Not reported	Balanced accuracy 0.647; sensitivity 0.787; specificity 0.465; PPV 0.690; NPV 0.590; PSI 0.28
	GAF score ≥65 at 6 years	Not reported	Not reported	Not reported	Balanced accuracy 0.676; sensitivity 0.818; specificity 0.477; PPV 0.718; NPV 0.616; PSI 0.334
Koutsouleris et al, 2016[26]	GAF score ≥65 at 1 year	Not reported	Not reported	Not reported	Balanced accuracy 0.738; sensitivity 0.667; specificity 0.809; PPV 0.515; NPV 0.888; PSI 0.403
Leighton et al, 2021[29]	Andreasen symptom remission (6-month duration) at 1 year	0.74 (95% CI 0.73–0.75)	$\beta = 0.84$ (95% CI 0.81–0.86); no calibration plot	Not reported	Not reported
Puntis et al, 2021[30]	Psychiatric hospital admission after discharge from early intervention	0.76 (95% CI 0.75–0.77)	$\alpha = 0.01$ (95% CI: −0.25 to 0.24); $\beta = 0.89$ (95% CI 0.88–0.89); Calibration plot	Brier score 0.078	Not reported

(*Continued*)

Table 3 (*Continued*)

Study	Outcome	Discrimination C-statistic	Calibration	Other global performance metrics	Classification metrics
Studies reporting internal–external validity					
De Nijs, 2019[22]	Andreasen symptom remission (6-month duration) at 3 years	Not reported	Not reported	Not reported	Balanced accuracy 0.638; sensitivity 0.629; specificity 0.647; PPV 0.758; NPV 0.485; PSI 0.243
	GAF score ≥65 at 3 years	Not reported	Not reported	Not reported	Balanced accuracy 0.648; sensitivity 0.658; specificity 0.638; PPV 0.727; NPV 0.565; PSI 0.292
	Andreasen symptom remission (6-month duration) at 6 years	Not reported	Not reported	Not reported	Balanced accuracy 0.625; sensitivity 0.685; specificity 0.565; PPV 0.743; NPV 0.493; PSI 0.236
	GAF score ≥65 at 6 years	Not reported	Not reported	Not reported	Balanced accuracy 0.640; sensitivity 0.718; specificity 0.561; PPV 0.732; NPV 0.553; PSI 0.285
Koutsouleris et al, 2016[26]	GAF score ≥65 at 1 year	Not reported	Not reported	Not reported	Balanced accuracy 0.711; sensitivity 0.641; specificity 0.781; PPV 0.472; NPV 0.877; PSI 0.349
Leighton et al, 2019[28]	EET status at 1 year	0.736 (95% CI 0.702–0.771)	Not reported	Not reported	Classification accuracy 0.693 (95% CI 0.660–0.725); balanced accuracy 0.694 (95% CI 0.562–0.812); sensitivity 0.722 (95% CI 0.573–0.821); specificity 0.666 (95% CI 0.550–0.803); PPV 0.719 (95% CI 0.673–0.785); NPV 0.668 (95% CI 0.606–0.736); PSI 0.387 (95% CI 0.279–0.521)
	GAF score ≥65 at 1 year	0.731 (95% CI 0.697–0.765)	Not reported	Not reported	Classification accuracy 0.687 (95% CI 0.657–0.718); balanced accuracy 0.691 (95% CI 0.541–0.825); sensitivity 0.722 (95% CI 0.487–0.778); specificity 0.660 (95% CI 0.594–0.871); PPV 0.650 (95% CI 0.616–0.769); NPV 0.726 (95% CI 0.655–0.766); PSI 0.376 (95% CI 0.271–0.535)
	Andreasen symptom remission (6-month duration) at 1 year	0.703 (95% CI 0.664–0.742)	Not reported	Not reported	Classification accuracy 0.670 (95% CI 0.636–0.703); balanced accuracy 0.668 (95% CI 0.518–0.827); sensitivity 0.584 (95% CI 0.491–0.827); specificity 0.751 (95% CI 0.544–0.827); PPV 0.679 (95% CI 0.601–0.739); NPV 0.667 (95% CI 0.631–0.734); PSI 0.346 (95% CI 0.232–0.473)
	Quality of life at 1 year	0.704 (95% CI 0.667–0.742)	Not reported	Not reported	Classification accuracy 0.668 (95% CI 0.632–0.704); balanced accuracy 0.667 (95% CI 0.532–0.789); sensitivity 0.623 (95% CI 0.512–0.774); specificity 0.711 (95% CI 0.551–0.803); PPV 0.633 (95% CI 0.575–0.701); NPV 0.700 (95% CI 0.659–0.759); PSI 0.333 (95% CI 0.234–0.460)
Studies reporting external validity					

(*Continued*)

Table 3 (*Continued*)

Study	Outcome	Discrimination C-statistic	Calibration	Other global performance metrics	Classification metrics
Leighton et al, 2019[27]	EET status at 1 year	0.876 (95% CI 0.864–0.887)	Not reported	Not reported	Classification accuracy 0.851; balanced accuracy 0.845; sensitivity 0.815; specificity 0.875; PPV 0.815; NPV 0.875; PSI 0.690
	Andreasen symptom remission (no duration criteria) at 1 year	0.652 (95% CI 0.635–0.670)	Not reported	Not reported	Classification accuracy 0.612; balanced accuracy 0.623; sensitivity 0.578; specificity 0.667; PPV 0.794; NPV 0.424; PSI 0.218
	Andreasen symptom remission (6-month duration) at 1 year	0.630 (95% CI 0.612-0.647)	Not reported	Not reported	Classification accuracy 0.625; balanced accuracy 0.626; sensitivity 0.606; specificity 0.645; PPV 0.645; NPV 0.606; PSI 0.251
Leighton et al, 2019,[28] validated in UK	EET status at 1 year	0.867 (95% CI 0.805–0.930)	Not reported	Not reported	Classification accuracy 0.838 (95% CI 0.775–0.894); balanced accuracy 0.853 (95% CI 0.740–0.935); sensitivity 0.898 (95% CI 0.780–0.966); specificity 0.807 (95% CI 0.699–0.904); PPV 0.766 (95% CI 0.679–0.867); NPV 0.911 (95% CI 0.840–0.971); PSI 0.677 (95% CI 0.519–0.838)
	Andreasen symptom remission (6-month duration) at 1 year	0.680 (95% CI 0.587–0.773)	Not reported	Not reported	Classification accuracy 0.695 (95% CI 0.618–0.771); balanced accuracy 0.695 (95% CI 0.535–0.841); sensitivity 0.621 (95% CI 0.455–0.773); specificity 0.769 (95% CI 0.615–0.908); PPV 0.729 (95% CI 0.636–0.854); NPV 0.667 (95% CI 0.593–0.759); PSI 0.396 (95% CI 0.229–0.613)
	Quality of life at 1 year	0.679 (95% CI 0.522–0.836)	Not reported	Not reported	Classification accuracy 0.702 (95% CI 0.596–0.809); balanced accuracy 0.729 (95% CI 0.407–0.917); sensitivity 0.957 (95% CI 0.564–1.000); specificity 0.500 (95% CI 0.250–0.833); PPV 0.640 (95% CI 0.561–0.800); NPV 0.900 (95% CI 0.643–1.000); PSI 0.540 (95% CI 0.204–0.800)

(*Continued*)

Table 3 (*Continued*)

Study	Outcome	Discrimination C-statistic	Calibration	Other global performance metrics	Classification metrics
Leighton et al, 2019,[28] validated in Denmark	EET status at 1 year	0.660 (95% CI 0.610–0.710)	Not reported	Not reported	Classification accuracy 0.680 (95% CI 0.609–0.725); balanced accuracy 0.655 (95% CI 0.516–0.774); sensitivity 0.584 (95% CI 0.457–0.723); specificity 0.726 (95% CI 0.574–0.824); PPV 0.490 (95% CI 0.421–0.563); NPV 0.793 (95% CI 0.760–0.831); PSI 0.283 (95% CI 0.181–0.394)
	GAF score ≥65 at 1 year	0.573 (95% CI 0.504–0.643)	Not reported	Not reported	Classification accuracy 0.456 (95% CI 0.328–0.817); balanced accuracy 0.589 (95% CI 0.234–0.926); sensitivity 0.781 (95% CI 0.233–0.945); specificity 0.396 (95% CI 0.234–0.906); PPV 0.179 (95% CI 0.158–0.333); NPV 0.914 (95% CI 0.876–0.967); PSI 0.093 (95% CI 0.034–0.300)
	Andreasen symptom remission (6-month duration) at 1 year	0.616 (95% CI 0.553–0.679)	Not reported	Not reported	Classification accuracy 0.618 (95% CI 0.524–0.704); balanced accuracy 0.621 (95% CI 0.342–0.864); sensitivity 0.612 (95% CI 0.306–0.843); specificity 0.629 (95% CI 0.378–0.885); PPV 0.476 (95% CI 0.412–0.636); NPV 0.742 (95% CI 0.687–0.829); PSI 0.217 (95% CI 0.099–0.465)
	Quality of life at 1 year	0.556 (95% CI 0.481–0.631)	Not reported	Not reported	Classification accuracy 0.589 (95% CI 0.540–0.637); balanced accuracy 0.589 (95% CI 0.312–0.845); sensitivity 0.876 (95% CI 0.419–0.947); specificity 0.301 (95% CI 0.204–0.743); PPV 0.559 (95% CI 0.527–0.642); NPV 0.706 (95% CI 0.555–0.841); PSI 0.265 (95% CI 0.081–0.483)
Leighton et al, 2021[29]	Andreasen symptom remission (6-month duration)	0.73 (95% CI 0.71–0.75)	$\alpha = 0.12$ (95% CI 0.02–0.22); $\beta = 0.98$ (95% CI 0.85–1.11); calibration plot	Not reported	Not reported
Puntis et al, 2021[30]	Psychiatric hospital admission after discharge from early intervention	0.70 (95% CI 0.66–0.75)	$\alpha = -0.01$ (95% CI −0.17 to 0.167); $\beta = 1.00$ (95% CI 0.78–1.22); calibration plot	Brier score 0.094	Not reported

PPV, positive predictive value; NPV, negative predictive value; PSI, prognostic summary index; EET, employment, education or training; GAF, Global Assessment of Functioning; DAS, Disability Assessment Schedule.

of bias as assessed with PROBAST, both of which externally validated their models.

Principal findings in context

Our systematic review found no consistent definition of FEP across the different cohorts used for developing and validating prediction models. A lack of an operational definition for FEP within clinical and research settings has previously been identified as major a barrier to progress.[31] The majority of cohorts in our systematic review included only individuals with non-affective psychosis, with a minority also including affective psychosis. In contrast, early intervention services typically do not make a distinction between affective and non-affective psychosis in those that they accept onto their service.[32] As such, there may be issues with generalisability of prediction models developed in cohorts with solely non-affective psychosis to real-world clinical practice.

A wide range of different outcomes were predicted by the FEP models, including symptom remission, global functioning, vocational functioning, treatment resistance, hospital readmission and quality-of-life outcomes. This is reflective of the fact that recovery from FEP is not readily distilled down to a single factor such as symptom remission. Meaningful recovery is represented by a constellation of multidimensional outcomes unique to each individual.[33] We should engage people with lived experience, to ensure that prediction models are welcomed and are predicting outcomes most relevant to the people they are for.

All of the prediction models were developed in populations from high-income countries, and only three studies included participants from countries outside of Europe, an issue not unique to FEP research. Consequently, it is currently unknown how prediction models for FEP would generalise to low-income countries. Prediction models may have considerable benefit in low-income countries, where almost 80% of patients with FEP live, but where mental health support is often scarce.[34] Prediction models could help prioritise the appropriate utilisation of limited healthcare resources.

Only one study considered predictor variables other than clinical or sociodemographic factors. In this study, the additional predictors did not add significant value.[22] In recent years, substantial progress has been made in elucidating the pathophysiological mechanisms underpinning the development of psychosis. We now recognise important roles for genetic factors, neurodevelopmental factors, dopamine and glutamate.[35] Prediction model performance may be improved by the incorporation of these biologically relevant disease markers as predictor variables. However, the cost–benefit aspect of adding more expensive and less accessible disease markers must be carefully considered, especially if models are to be utilised in settings where resources are more limited.

Machine learning can be operationally defined as 'models that directly and automatically learn from data'. This is in contrast to regression models, which 'are based on theory and assumptions, and benefit from human intervention and subject knowledge for model specification'.[36] Just two studies used machine learning techniques for their modelling.[22,26] The rest of the studies used logistic regression. We were unable to make any comparison between the discrimination and calibration ability of the two studies that used machine learning and the other studies, because these metrics were not provided. However, a recent systematic review found no evidence of superior performance of clinical prediction models that use machine learning methods over logistic regression.[36] In any case, the distinction between regression models and machine learning has been viewed to be artificial. Instead, algorithms may exist 'along a continuum between fully human-guided to fully machine-guided data analysis'.[37] An alternative comparison may be between linear and non-linear classifiers. Only one study used a non-linear classifier,[26] but again we were unable to gain meaningful insights into its relative performance because appropriate metrics were not provided.

A principal finding from our systematic review is the presence of methodological limitations across the majority of studies. Steyerberg et al outline four key measures of predictive performance that should be assessed in any prediction-modelling study: two measures of calibration (the model intercept (A) and the calibration slope (B)), discrimination via a concordance statistic (C) and clinical usefulness with decision-curve analysis (D).[7] Model calibration is the level of agreement between the observed outcomes and the predictions. For example, if a model predicts a 5% risk of cancer, then, according to such a prediction, the observed proportion should be five cancers per 100 people. Discrimination is the ability of a model to distinguish between a patient with the outcome and one without.[7] Our review found that only seven studies (54%) reported discrimination and just five (38%) reported any measure of calibration. The remaining studies reported only classification metrics, such as accuracy or balanced accuracy. The problem with solely reporting classification metrics is that they vary both across models and across different probability thresholds for the same model. This renders the comparison between models less meaningful. It is further argued that setting a classification threshold for a probability-generating model is premature. Rather, a clinician may choose to set different probability thresholds for the same prediction model, depending on the situation at hand, to optimise the balance between false positives and false negatives. For example, in the case of a model predicting cancer, a clinician may choose a lower probability threshold to offer a non-invasive screening test and a higher probability threshold to suggest an invasive and potentially harmful biopsy. Further, without any measure of model calibration, we are unable to assess if the model can make unbiased estimates of outcome.[38] The final key step in assessing the performance of a prediction model is to determine its clinical usefulness – that is, can better decisions be made with the model than without? Decision-curve analysis considers the net benefit (the treatment threshold weighted sum of true- minus false-positive classifications) for a prediction model compared with the default strategy of treating all or no patients, across an entire range of treatment thresholds.[39] Only two studies (15%) included in our review considered whether the model was clinically useful. Without proper validation of the prediction models, the reported performances are likely to be overly optimistic. Four studies (31%) reported only apparent validity. Just four studies (31%) reported external validation, which is considered essential before applying a prediction model to clinical practice.[14]

Altogether, just two studies (15%) had an overall 'low' risk of bias according to PROBAST, reflecting these methodological limitations. Neither study considered real-world implementation. To progress with implementation, impact studies are required. These would involve a cluster randomised trial comparing patient outcomes between a group with treatment informed by a clinical prediction model and a control group.[40] We are not aware of any such study having been carried out within the field of psychiatry. However, Salazar de Pablo et al suggest that PROBAST thresholds for considering a study to be a 'low' risk of bias may be too strict.[6] Indeed, in the field of machine learning, multiple imputation is frequently computationally infeasible, and single imputation may be viewed as sufficient. This is especially true in larger data-sets or in the presence of relatively few missing values.[41]

Strengths and limitations

Our review had a number of strengths. We provide the first systematic overview of prediction-modelling studies for use in patients

with FEP. We offer a detailed critique of the study characteristics, their methodologies and model performance metrics. Further, our review adheres to gold-standard guidance for extracting data from prediction models and for assessing bias, namely the CHARMS checklist and PROBAST.

There were several limitations. Our initial aim was to perform a meta-analysis of any prediction model that was validated across different settings and populations. However, no meta-analysis was possible because no single prediction model was validated more than once. In addition, as a consequence of poor reporting of discrimination and calibration performance across the studies, it was often difficult to make meaningful comparison between the prediction models. Also, the lack of consensus as to the most important outcome measure in FEP, with six different outcomes considered across only 13 included studies, further hindered efforts at drawing meaningful comparisons between the included studies and their respective prediction models. Likewise, if more studies had considered the same outcome measures, this may have afforded the opportunity to validate existing prediction models rather than necessitating the creation of additional new models. All published prediction-modelling studies in FEP reported significant positive findings. It is possible that studies that had negative findings were held back from publication, reflecting the possibility of publication bias. We originally intended to evaluate the overall certainty in the body of evidence by using the Grading of Recommendations Assessment, Development and Evaluation (GRADE) framework.[42] GRADE was originally designed for reviews of intervention studies, but has not yet been adapted for use in systematic reviews of prediction models. Consequently, in its current form, we did not find GRADE to be a suitable tool for our review and decided not to use it. Future research should consider how to adapt GRADE for use in systematic reviews of prediction models.

Implications for future research

It is clear that there is a growing trend for the development of prediction models in FEP.[6] FEP is an illness that responds best to an early intervention paradigm.[43] Prediction models have the potential to optimise the allocation of time-critical interventions, like clozapine for treatment resistance.[44] However, several steps are necessary before meaningful implementation into real-world clinical practice. The field must prioritise external validation and replication of existing prediction models in larger sample sizes, to increase the EPV. This is best accomplished by an emphasis on data-sharing and open collaboration. Prediction studies should include FEP cohorts from low-income countries, where there is considerable potential for benefit by helping to prioritise limited resources to those most in need. Harmonisation of data collection across the field, both in terms of predictors and outcomes measured, would facilitate validation efforts. There should be a greater consideration of biologically relevant and cognitive predictors based on our growing understanding of disease mechanisms, which could optimise prediction model performance. Finally, our review highlights considerable methodological pitfalls in much of the current literature. Future prediction-modelling studies should focus on methodological rigour with adherence to accepted best-practice guidance.[9,14,38] Our goal in psychiatry should be to develop an innovative approach to care by using prediction models. Application of these approaches into clinical practice would enable rapid and targeted intervention, thereby limiting treatment-associated risks and reducing patient suffering.

Rebecca Lee, Institute for Mental Health, University of Birmingham, UK; **Samuel P. Leighton**, Institute of Health and Wellbeing, University of Glasgow, UK; **Lucretia Thomas**, Birmingham Medical School, University of Birmingham, UK; **Georgios V. Gkoutos**, Institute of Cancer and Genomic Sciences, University of Birmingham, UK; **Stephen J. Wood**, Orygen Youth Health Research Centre, National Centre of Excellence in Youth Mental Health, Australia; School of Psychological Sciences, University of Melbourne, Australia; and School of Psychology, University of Birmingham, UK; **Sarah-Jane H. Fenton**, Institute for Mental Health, University of Birmingham, UK; **Fani Deligianni**, School of Computing Science, University of Glasgow, UK; **Jonathan Cavanagh**, Institute of Infection, Immunity and Inflammation, University of Glasgow, UK; **Pavan K. Mallikarjun**, Institute for Mental Health, University of Birmingham, UK

Correspondence: Samuel P. Leighton. Email: samuel.leighton@glasgow.ac.uk

First received 31 Aug 2021, final revision 23 Nov 2021, accepted 9 Dec 2021

Supplementary material

Supplementary material is available online at https://doi.org/10.1192/bjp.2021.219.

Data availability

Data is available from the corresponding author, S.P.L., upon reasonable request.

Author contributions

P.K.M. and R.L. formulated the research question and designed the study. R.L., S.P.L., L.T. and P.K.M. collected the data. R.L., S.P.L. and P.K.M. analysed the data and drafted the manuscript. L.T., G.V.G., S.J.W., S.-J.H.F., F.D. and J.C. critically evaluated and revised the manuscript.

Funding

R.L. is funded by the Institute for Mental Health Priestley Scholarship, University of Birmingham. S.P.L. is funded by a clinical academic fellowship from the Chief Scientist Office, Scotland (CAF/19/04). S.J.W. is funded by the Medical Research Council, UK (grant MR/K013599).

Declaration of interest

G.V.G. has received support from Horizon 2020 E-Infrastructures (H2020-EINFRA), the National Institute for Health Research (NIHR) Birmingham Experimental Cancer Medicine Centre (ECMC), NIHR Birmingham Surgical Reconstruction Microbiology Research Centre (SRMRC), the NIHR Birmingham Biomedical Research Centre, and the Medical Research Council Health Data Research United Kingdom (MRC HDR UK), an initiative funded by UK Research and Innovation, Department of Health and Social Care (England), the devolved administrations and leading medical research charities. J.C. has received grants from the Wellcome Trust and Sackler Trust, and honorariums from Johnson & Johnson. P.K.M. has received honorariums from Sunovion and Sage, and is a Director of Noux Technologies Limited. All other authors declare no competing interests.

References

1 Moreno-Küstner B, Martín C, Pastor L. Prevalence of psychotic disorders and its association with methodological issues. A systematic review and meta-analyses. *PLoS One* 2018; **13**: e0195687.

2 Institute for Health Metrics and Evaluation (IHME). *GBD Compare Data Visualization*. IHME, University of Washington, 2021 (http://vizhub.healthdata.org/gbd-compare).

3 Lally J, Ajnakina O, Stubbs B, Cullinane M, Murphy KC, Gaughran F, et al. Remission and recovery from first-episode psychosis in adults: systematic review and meta-analysis of long-term outcome studies. *Br J Psychiatry* 2017; **211**: 350–8.

4 Darcy AM, Louie AK, Roberts LW. Machine learning and the profession of medicine. *JAMA* 2016; **315**(6): 551.

5 Hippisley-Cox J, Coupland C, Brindle P. Development and validation of QRISK3 risk prediction algorithms to estimate future risk of cardiovascular disease: prospective cohort study. *BMJ* 2017; **357**: j2099.

6 Salazar de Pablo G, Studerus E, Vaquerizo-Serrano J, Irving J, Catalan A, Oliver D, et al. Implementing precision psychiatry: a systematic review of individualized prediction models for clinical practice. *Schizophr Bull* 2021; **47**(2): 284–97.

7 Steyerberg EW, Vergouwe Y. Towards better clinical prediction models: seven steps for development and an ABCD for validation. *Eur Heart J* 2014; **35**(29): 1925–31.

8 Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *BMJ* 2015; **350**: g7594.

9 Wynants L, Van Calster B, Collins GS, Riley RD, Heinze G, Schuit E, et al. Prediction models for diagnosis and prognosis of Covid-19: systematic review and critical appraisal. *BMJ* 2020; **369**: 26.

10 Studerus E, Ramyead A, Riecher-Rössler A. Prediction of transition to psychosis in patients with a clinical high risk for psychosis: a systematic review of methodology and reporting. *Psychol Med* 2017; **47**: 1163–78.

11 Rosen M, Betz LT, Schultze-Lutter F, Chisholm K, Haidl TK, Kambeitz-Ilankovic L, et al. Towards clinical application of prediction models for transition to psychosis: a systematic review and external validation study in the PRONIA sample. *Neurosci Biobehav Rev* 2021; **125**: 478–92.

12 Sullivan S, Northstone K, Gadd C, Walker J, Margelyte R, Richards A, et al. Models to predict relapse in psychosis: a systematic review. *PLoS One* 2017; **12**(9): e0183998.

13 Moons KGM, de Groot JAH, Bouwmeester W, Vergouwe Y, Mallett S, Altman DG, et al. Critical appraisal and data extraction for systematic reviews of prediction modelling studies: the CHARMS checklist. *PLoS Med* 2014; **11**(10): e1001744.

14 Steyerberg EW, Harrell FE Jr. Prediction models need appropriate internal, internal-external, and external validation. *J Clin Epidemiol* 2016; **69**: 245–7.

15 Wolff RF, Moons KGM, Riley RD, Whiting PF, Westwood M, Collins GS, et al. PROBAST: a tool to assess the risk of bias and applicability of prediction model studies. *Ann Intern Med* 2019; **170**(1): 51–8.

16 Moons KGM, Wolff RF, Riley RD, Whiting PF, Westwood M, Collins GS, et al. PROBAST: a tool to assess risk of bias and applicability of prediction model studies: explanation and elaboration. *Ann Intern Med* 2019; **170**(1): W1–33.

17 Page MJ, Moher D, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. PRISMA 2020 explanation and elaboration: updated guidance and exemplars for reporting systematic reviews. *BMJ* 2021; **372**: n160.

18 Ajnakina O, Agbedjro D, Lally J, Forti M, Trotta A, Mondelli V, et al. Predicting onset of early- and late-treatment resistance in first-episode schizophrenia patients using advanced shrinkage statistical methods in a small sample. *Psychiatry Res* 2020; **294**: 113527.

19 Bhattacharyya S, Schoeler T, Patel R, di Forti M, Murray RM, McGuire P. Individualized prediction of 2-year risk of relapse as indexed by psychiatric hospitalization following psychosis onset: model development in two first episode samples. *Schizophr Res* 2021; **228**: 483–92.

20 Chua YC, Abdin E, Tang C, Subramaniam M, Verma S. First-episode psychosis and vocational outcomes: a predictive model. *Schizophr Res* 2019; **211**: 63–8.

21 Demjaha A, Lappin JM, Stahl D, Patel MX, MacCabe JH, Howes OD, et al. Antipsychotic treatment resistance in first-episode psychosis: prevalence, subtypes and predictors. *Psychol Med* 2017; **47**(11): 1981–9.

22 de Nijs J. *The Outcome of Psychosis*. Utrecht University, 2019 (https://dspace.library.uu.nl/bitstream/1874/376436/1/22_01_3_jessica_de_nijs_compleet_final.pdf).

23 Derks EM, Fleischhacker WW, Boter H, Peuskens J, Kahn RS. Antipsychotic drug treatment in first-episode psychosis should patients be switched to a different antipsychotic drug after 2, 4, or 6 weeks of nonresponse? *J Clin Psychopharmacol* 2010; **30**(2): 176–80.

24 Flyckt L, Mattsson M, Edman G, Carlsson R, Cullberg J. Predicting 5-year outcome in first-episode psychosis: construction of a prognostic rating scale. *J Clin Psychiatry* 2006; **67**(6): 916–24.

25 González-Blanch C, Perez-Iglesias R, Pardo-García G, Rodríguez-Snchez JM, Martínez-García O, Vázquez-Barquero JL, et al. Prognostic value of cognitive functioning for global functional recovery in first-episode schizophrenia. *Psychol Med* 2010; **40**(6): 935–44.

26 Koutsouleris N, Kahn RS, Chekroud AM, Leucht S, Falkai P, Wobrock T, et al. Multisite prediction of 4-week and 52-week treatment outcomes in patients with first-episode psychosis: a machine learning approach. *Lancet Psychiatry* 2016; **3**(10): 935–46.

27 Leighton SP, Krishnadas R, Chung K, Blair A, Brown S, Clark S, et al. Predicting one-year outcome in first episode psychosis using machine learning. *PLoS One* 2019; **14**(3): e0212846.

28 Leighton SP, Upthegrove R, Krishnadas R, Benros ME, Broome MR, Gkoutos GV, et al. Development and validation of multivariable prediction models of remission, recovery, and quality of life outcomes in people with first episode psychosis: a machine learning approach. *Lancet Digit Heal* 2019; **1**(6): e261–70.

29 Leighton SP, Krishnadas R, Upthegrove R, Marwaha S, Steyerberg EW, Broome MR, et al. Development and validation of a non-remission risk prediction model in first episode psychosis: an analysis of two longitudinal studies. *Schizophr Bull Open* 2021; **2**(1): sgab041.

30 Puntis S, Whiting D, Pappa S, Lennox B. Development and external validation of an admission risk prediction model after treatment from early intervention in psychosis services. *Transl Psychiatry* 2021; **11**: 35.

31 Breitborde NJK, Srihari VH, Woods SW. Review of the operational definition for first-episode psychosis. *Early Interv Psychiatry* 2009; **3**: 259–65.

32 National Institute for Health and Care Excellence (NICE). *Implementing the Early Intervention in Psychosis Access and Waiting Time Standard: Guidance*. NICE, 2016 (https://www.nice.org.uk/guidance/qs80/resources/implementing-the-early-intervention-in-psychosis-access-and-waiting-time-standard-guidance-2487749725).

33 Jääskeläinen E, Juola P, Hirvonen N, McGrath JJ, Saha S, Isohanni M, et al. A systematic review and meta-analysis of recovery in schizophrenia. *Schizophr Bull* 2013; **39**(6): 1296–306.

34 Singh SP, Javed A. Early intervention in psychosis in low- and middle-income countries: a WPA initiative. *World Psychiatry* 2020; **19**: 122.

35 Lieberman JA, First MB. Psychotic disorders. *N Engl J Med* 2018; **379**(3): 270–80.

36 Christodoulou E, Ma J, Collins GS, Steyerberg EW, Verbakel JY, Van Calster B. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *J Clin Epidemiol* 2019; **110**: 12–22.

37 Beam AL, Kohane IS. Big data and machine learning in health care. *JAMA* 2018; **319**: 1317–8.

38 Harrell FE Jr. *Regression Modeling Strategies*. Springer International Publishing, 2015.

39 Vickers AJ, van Calster B, Steyerberg EW. A simple, step-by-step guide to interpreting decision curve analysis. *Diagnostic Progn Res* 2019; **3**: 18.

40 Moons KGM, Kengne AP, Grobbee DE, Royston P, Vergouwe Y, Altman DG, et al. Risk prediction models: II. External validation, model updating, and impact assessment. *Heart* 2012; **98**: 691–8.

41 Steyerberg EW. *Clinical Prediction Models* 2nd ed. Springer International Publishing, 2019.

42 Schünemann HJ, Oxman AD, Brozek J, Glasziou P, Jaeschke R, Vist GE, et al. Grading quality of evidence and strength of recommendations for diagnostic tests and strategies. *BMJ* 2008; **336**(7653): 1106–10.

43 Birchwood M, Todd P, Jackson C. Early intervention in psychosis: the critical period hypothesis. *Br J Psychiatry* 1998; **172**(S33): 53–9.

44 Farooq S, Choudry A, Cohen D, Naeem F, Ayub M. Barriers to using clozapine in treatment-resistant schizophrenia: systematic review. *BJPsych Bull* 2019; **43**(1): 8–16.