nature mental health

Article

https://doi.org/10.1038/s44220-022-00001-z

Predicting treatment resistance from first-episode psychosis using routinely collected clinical information

Received: 20 May 2022

Accepted: 3 November 2022

Published online: 19 January 2023

Check for updates

A list of authors and their affiliations appears at the end of the paper

Around a quarter of people who experience a first episode of psychosis (FEP) will develop treatment-resistant schizophrenia, but there are currently no established clinically useful methods to predict this from baseline. We aimed to explore the predictive potential for clozapine use as a proxy for treatment-resistant schizophrenia of routinely collected, objective biomedical predictors at FEP onset, and to validate the model externally in a separate clinical sample of people with FEP. We developed and externally validated a forced-entry logistic regression risk prediction model for clozapine treatment, or MOZART, to predict up to 8-year risk of clozapine use from FEP using routinely recorded information including age, sex, ethnicity, triglycerides, alkaline phosphatase levels and lymphocyte counts. We also produced a least-absolute shrinkage and selection operator (LASSO) based model, additionally including neutrophil count, smoking status, body mass index and random glucose levels. The models were developed using data from two United Kingdom (UK) psychosis early intervention services and externally validated in another UK early intervention service. Model performance was assessed by discrimination and calibration. We developed the models in 785 patients and validated them externally in 1,110 patients. Both models predicted clozapine use well during internal validation (MOZART: *C* statistic, 0.70 (95% confidence interval, 0.63–0.76); LASSO: 0.69 (0.63–0.77)). At external validation, discrimination performance reduced (MOZART: 0.63 (0.58–0.69); LASSO: 0.64 (0.58–0.69)) but recovered after re-estimation of the lymphocyte predictor (0.67 (0.62–0.73)). Calibration plots showed good agreement between observed and predicted risk in the forced-entry model. We also present a decision-curve analysis and an online data visualization tool. The use of routinely collected clinical information including blood-based biomarkers taken at FEP onset can help to predict the individual risk of clozapine use, and should be considered equally alongside other potentially useful information such as symptom scores in large-scale efforts to predict psychiatric outcomes.

e-mail: efo22@cam.ac.uk

https://doi.org/10.1038/s44220-022-00001-z

Table 1 | Model comparisons including coefficients for development and external validation

Model type	Model predictors of TRS	Coefficients after shrinkage for optimism[a]	Pooled development sample performance statistics[b]	Shrinkage factor	Validation sample			Calibration plots for external validation
					Performance statistics[b]	New coefficients after model recalibration/ revision	Performance after model recalibration/ revision[b]	
Forced-entry model (MOZART)	Intercept	−2.827381	*C*: 0.70 (0.63–0.76) Brier score: 0.07	0.79	*C*: 0.63 (0.58–0.69) Brier score: 0.08	Lymphocyte coefficient: −0.695404405 Intercept: −1.336220 Slope: 1.0519963	*C*: 0.67 (0.62–0.73) Brier score: 0.08	Fig. 2a,b
	Sex	0.286466741						
	Age	−0.036205346						
	Black/African-Caribbean ethnicity	0.419614174						
	Asian ethnicity	−0.144147329						
	Triglycerides	0.149214138						
	ALP	0.006713513						
	Lymphocyte count	0.131215526						
LASSO-based model	Intercept	−2.736365	*C*: 0.69 (0.63–0.77) Brier score: 0.07	N/A	*C*: 0.64 (0.58–0.69) Brier score: 0.08	Lymphocyte coefficient: −0.03608036 Intercept: −2.706553 Slope: 1.3102021	*C*: 0.64 (0.58–0.69) Brier score: 0.08	Fig. 2c,d
	Sex	0.13205						
	Age	−0.248397						
	Black/African-Caribbean ethnicity	0.304147						
	Asian ethnicity	−0.002375						
	Triglycerides	0.139795						
	ALP	0.131153						
	Lymphocyte count	0.060623						
	Smoking status	0.057593						
	BMI	−0.026467						
	Random plasma glucose	−0.027369						
	Neutrophil count	−0.012826						

[a]The coefficients are relative to non-scaled values for forced-entry models, and to scaled and centred values for the LASSO model. [b]*C* is the *C* value (95% CI; Methods).

Schizophrenia spectrum disorders can have remarkably different life courses. Approximately half of people presenting with a first episode of psychosis (FEP) show good outcomes, such as remission[1] or no need for long-term secondary care[2]. However, ~23–24% of patients with a FEP go on to develop treatment-resistant schizophrenia (TRS)[3]. TRS is typically defined as resistance to two antipsychotic treatments, each given at an adequate dose for at least 6 weeks, with evidence of medication adherence[4]. TRS is associated with reduced quality of life, substantial societal burden and up to tenfold higher healthcare costs[5].

It is not currently possible to predict accurately whether someone with FEP will develop TRS. This is important because there is evidence that clozapine, the only treatment licensed for TRS[6], is more effective the sooner it is prescribed[7]. Yet, in clinical practice there are often long delays before clozapine is considered[8]. This highlights the need to identify treatment resistance as soon as possible.

Risk prediction in psychosis is a flourishing field (Extended Data Fig. 1). However, existing studies have commonly included predictors that are not easy to deploy in routine clinical practice (for example, neuroimaging[9] or genetic measures[10]); not routinely or reliably collected (for example, duration of untreated psychosis[11], substance misuse[12,13] or premorbid functioning[14]); or not available at FEP onset (for example, symptom patterns over time[12,15]). All these characteristics limit the potential clinical usefulness of existing efforts in TRS prediction.

In addition to limited clinical usefulness, most previous studies are limited by methodological difficulties or poor reporting practices, particularly a lack of assessment of model calibration, a lack of external validation to assess generalizability[16,17], limited consideration of sample size and the risk of overfitting, and the inclusion of variables that cannot be known at FEP onset, such as medication during follow-up[15].

Blood biomarkers, which are objective, are commonly used to predict clinical outcomes in routinely used, large-scale, risk-prediction algorithms based on the general population[18]. Indeed, biomarkers and clinical measures commonly taken at FEP onset can help predict metabolic outcomes in patients with psychosis[19]. Furthermore, inflammatory and metabolic alterations are already evident in antipsychotic-naïve patients with FEP, including impaired glucose tolerance, insulin resistance[20], hypertriglyceridaemia[21] and pro-inflammatory changes[22]. Biomarker alterations may additionally be associated with a more chronic psychiatric illness course[2,23].

In this work, we aimed to predict clozapine use (as a proxy for TRS) up to eight years after FEP onset, using routinely collected, objective and measurable biomedical predictors at baseline, with the aim of producing the most parsimonious prediction model with the potential for clinical use. We used patient data from three UK early intervention in psychosis services (EIPs) to investigate the predictive potential of

https://doi.org/10.1038/s44220-022-00001-z

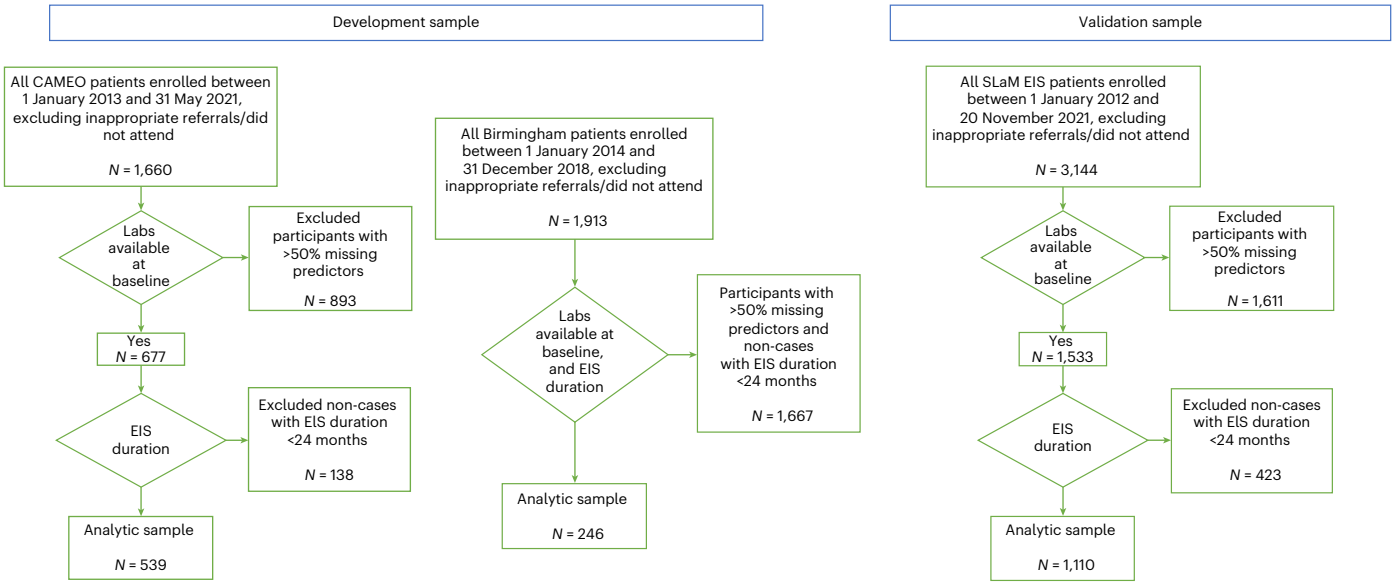


Fig. 1 | Patient selection flow charts, by cohort. The flow charts describe the application of inclusion and exclusion criteria for each cohort, starting from the sampling frames, and up to the analytic samples.

sociodemographic, lifestyle and biological data routinely recorded at FEP baseline. We aimed to follow best practice by including an external validation step to examine generalizability. We followed the Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD) guidelines (Supplementary Table 1).

Results

Development of a TRS prediction model

The coefficients for MOZART and for the LASSO model are presented in Table 1. Histograms of predicted outcome probabilities are provided as Extended Data Figs. 2 and 3. Univariable logistic regression coefficients (clozapine-predictor) are presented in Supplementary Table 2.

Internal validation

Measures of pooled internal validation performance of the models over 100 imputed datasets are shown in Table 1. The *C* statistic for the forced-entry model (MOZART) was 0.70 (95% confidence interval (CI): 0.63–0.76), while that for the LASSO model was 0.69 (0.63–0.77). Calibration plots showed good agreement between observed and expected risk at most predicted probabilities for both models, although the LASSO model showed slight overprediction of risk at lower predicted probabilities (Extended Data Figs. 4 and 5).

External validation

The external validation sample comprised 1,110 patients from the South London and Maudsley NHS Foundation Trust (SLaM) EIP (Table 1 and Fig. 1).

Applying the models developed in the joint development sample to the SLaM sample, the *C* statistic for MOZART was 0.63 (95% CI: 0.58–0.69), while that for the LASSO model was 0.64 (0.58–0.69; Table 1).

The calibration plot for MOZART showed good agreement between observed and expected risk (Fig. 2a), while that for the LASSO model showed evidence of mild overprediction of risk at higher predicted probabilities and of slight overprediction for very low risk (Fig. 2c). In all models, the 95% CIs widened as predicted probabilities became higher, owing to lower numbers of participants.

External validation after model recalibration and revision

We applied logistic recalibration to both main models in the external validation sample. Additionally, the coefficient for lymphocyte count was selected for revision as the sign of the coefficient was reversed between the development and validation samples.

Table 1 shows that, after MOZART's recalibration/revision, the *C* statistic was restored to values close to internal validation performance (0.67 (95% CI: 0.62–0.73)). The same procedure performed on the LASSO model, however, did not produce any improvement on the original model performance statistics.

The calibration plots for both recalibrated models are shown in Fig. 2b,d. Both showed good agreement between observed and expected risk.

Decision curve analysis and data visualization tool

Decision curve analysis for MOZART (Fig. 3) suggests that at propensity-to-intervene thresholds greater than 0.05 (revised model) or 0.06 (original model), the models provided greater net benefit than the competing extremes of treating all patients or none. The recalibrated model provided higher net benefit at most, if not all, thresholds over 0.05 than the original model.

Numerical decision curve analysis results (net benefit, standardized net benefit, sensitivity and specificity) are shown in Supplementary Table 3 across a range of propensity-to-intervene thresholds. For example, if a low-risk intervention such as close monitoring for TRS was considered suitable above a propensity-to-intervene threshold of 0.10 (>10% risk of clozapine use), the recalibrated model would provide a net benefit of 2% (95% CI: 1–4%), meaning that an additional 24% of patients could be closely monitored for the presence of TRS (standardized net benefit). However, for a potentially more invasive intervention such as starting clozapine treatment, at a propensity-to-intervene threshold of 0.50, the same model would provide no net benefit, owing to insufficient sensitivity.

We also developed an online data visualization tool for both the original and recalibrated MOZART models, which allows interactive exploration of the effect of each predictor and their combinations on the risk of clozapine use based on the predictors included in this study (https://eosimo.shinyapps.io/trs_app/).

Sensitivity analysis with iterative improvements

To examine the added benefit of selected demographic and biological predictors, we examined iterative improvements in the forced entry model. Model 1 (M1) included sex as the only predictor; M2 included

https://doi.org/10.1038/s44220-022-00001-z

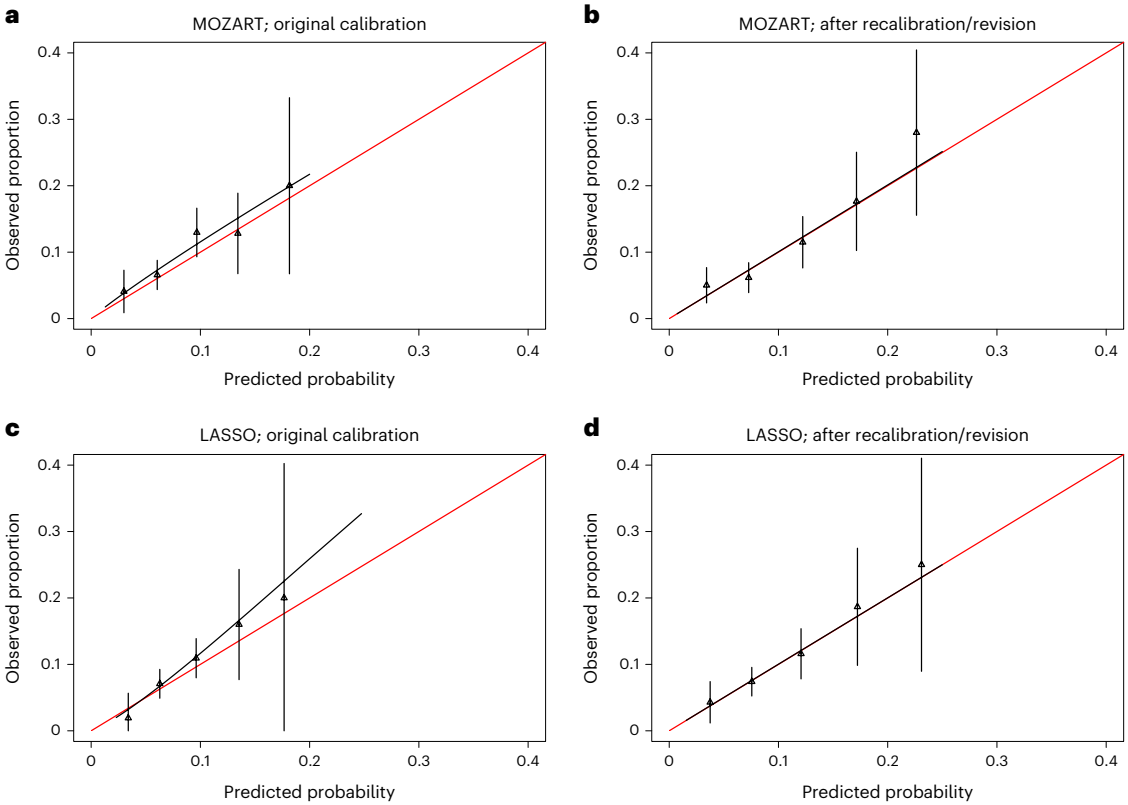


Fig. 2 | Calibration plots for the main models based on the external validation sample. a–d, Model calibration is the extent to which outcomes predicted by the model are similar to those observed in the validation dataset. Calibration plots illustrate agreement between the observed proportion of participants developing TRS (*y* axis) and predicted risk of TRS (*x* axis). Perfect agreement would trace the red line. Model calibration is shown by the continuous black line. Triangles denote grouped observations for participants at deciles of predicted risk, with 95% CIs indicated by the vertical black lines. Axes range between 0 and 0.4 because very few individuals received predicted probabilities greater than 0.3. **a**,**b**, Calibration plots based on the external validation sample for the forced-entry model (MOZART) before (**a**) and after (**b**) recalibration/revision. **c**,**d**, External validation calibration plots for the LASSO model before (**c**) and after (**d**) recalibration/revision. $N = 1{,}110$ participants in external validation sample.

all demographics; M3 included demographics plus triglyceride levels; M4 additionally included ALP. The internal coefficients and shrinkage factors for each model are presented in Supplementary Table 4. The *C* statistic increased from 0.56 (95% CI: 0.50–0.62) for M1 to 0.69 (0.62–0.76) for M4. Calibration plots showed good agreement between observed and expected risk at most predicted probabilities for M3 and M4 (shown, alongside histograms of predicted outcome probabilities, in Extended Data Figs. 6–9).

Discussion

We examined the predictive potential of routinely collected sociodemographic, lifestyle and clinical information, obtained at the start of a FEP, for the risk of clozapine use, as a proxy for developing TRS. We developed two models: MOZART, based on forced-entry logistic regression, and another based on LASSO for coefficient generation and shrinkage. The two models performed adequately in both internal and external validation. MOZART performed better than LASSO in external validation, and its performance improved following recalibration/revision.

Decision curve analysis revealed that MOZART shows clinical utility at lower propensity-to-intervene thresholds, such as 10–20%. This model cannot yet be recommended for clinical use and requires prospective validation in larger samples, health technology assessment and regulatory approval. However, in future our model could allow implementation of low-risk strategies, for example, stratifying patients at higher risk of developing antipsychotic resistance for closer monitoring of TRS. These strategies have very low risk of causing harm and might show potential for earlier recognition and treatment of TRS. Clozapine is more effective when given soon after treatment resistance is established, although in clinical practice there are long delays to starting it[7,8]; therefore, starting treatment early might show potential in reducing symptoms and improving quality of life in people with unrecognized TRS. However, given the higher risk and licensing conditions of clozapine, and the lower sensitivity of the model at higher risk thresholds, this model alone will not be useful for initiating higher risk interventions, such as starting clozapine.

In the future, inclusion of genetic risk scores and other predictors might make clozapine prediction models more accurate, and therefore more clinically useful. Two existing studies found that polygenic risk scores for schizophrenia did not produce significant increases in predictive power of a model for TRS[24,25]. However, the publication since then of larger genome-wide association studies (GWAS) for schizophrenia[26] and of a specific TRS GWAS[27] will likely make the approach more powerful.

MOZART extends existing research by using only seven common predictors available at FEP baseline; by including an external validation analysis, a crucial step to demonstrate generalizability; and by following best practice guidelines[28,29].

We show that simple blood-based biomarkers measured at the onset of psychosis can explain part of the variance of the risk of clozapine use, as demonstrated by the increased *C* statistic for the incremental model including biomarkers. This suggests that the variance of a psychiatric phenotype (development of TRS) may be explained, at least in part, by inflammatory, fat and liver biomarkers.

https://doi.org/10.1038/s44220-022-00001-z

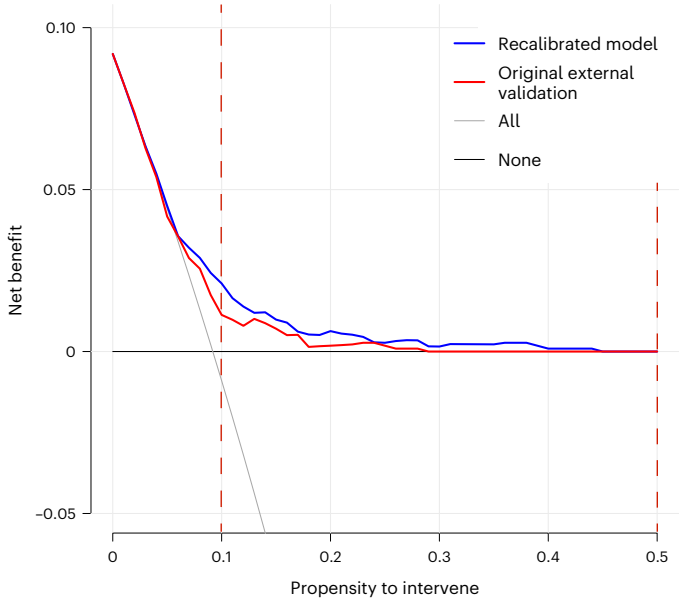


Fig. 3 | Decision curve analysis plot for forced-entry original and recalibrated models. The plot reports the net benefit (*y* axis) of the forced-entry (MOZART) original and recalibrated models across a range of propensity-to-intervene thresholds (*x* axis), compared with intervening in all patients and intervening in no patients. The dashed red vertical lines represent the two thresholds we selected a priori to study the potential clinical value of low-risk (for example, monitoring) and high-risk (for example, starting clozapine) interventions.

Previous studies using regression-based methods have shown that elevated triglycerides are associated with a worse psychiatric clinical outcome in psychosis at the group level[2,23]. We extend these findings by showing that elevated triglycerides at the individual level could aid in TRS prediction. We included ALP owing to the increasing importance that liver dysfunction is thought to play in the psychosis spectrum[30]. In particular, elevated ALP might relate to the primary dysglycaemic and dysmetabolic phenotype of FEP[20,31,32], or it might be its consequence (hyperlipidaemia leading to non-alcoholic fatty liver disease[33], a phenotype that has been found in FEP[30]). Elevated ALP may also capture some of the variance of substance use in a more objective manner than self-reporting[34,35].

As a proxy of inflammation we selected lymphocyte count as predictor, because the data is widely available across samples. In a previous analysis of mostly White participants with FEP, lymphocytes were associated with a worse psychiatric outcome[2]. However, cross-sectional studies have not found lymphocyte elevations in FEP[36,37], and a recent Mendelian randomization study did not find evidence for a causal association with schizophrenia[38], potentially discounting the likelihood of a causal association of elevated lymphocytes with schizophrenia. Further, we found that the drop in discrimination performance for the forced-entry model from internal to external validation was mostly due to differences in the lymphocyte predictor, with the sign of the coefficient switching direction between samples. In model updating, the *C* statistic could be partially preserved by updating the coefficient for lymphocytes. This might be explained by the different ethnic mix between the development sample (mainly White) and the external validation sample (mainly Black). It is known that inflammatory markers, including lymphocytes, show different distributions in different ethnic groups[39,40]. This might encourage repetition of the analysis using different inflammatory markers, such as C-reactive protein (CRP), in future research. We could not include CRP, because it was most often sampled in the included cohorts when there was suspicion of infection; therefore, data were only available for a small subset and likely showed strong selection bias.

The use of longitudinal EIP cohort data is the main strength of this study. Enrolment into an EIP fosters confidence in the psychiatric phenotype of included participants and into the naturalistic nature of the sample including many consecutive referrals with little possibility of selection bias from the sampling frame. Most EIPs in the UK NHS (National Health Service), including all three in this analysis, are the only treatment providers for FEP in each geographical area, thus covering a large proportion of all incident cases of FEP. Specifically, the Cambridgeshire and Peterborough Assessing, Managing and Enhancing Outcomes (CAMEO) EIP, used to develop our model, accepts people presenting with confirmed psychotic symptoms from any cause, including drug-induced psychoses and affective psychoses (including International Classification of Diseases (ICD)-10 codes F06.0-2, F20-F31, F32.3, F33.3 and F53.1). Therefore, MOZART is shown to work in real-life samples of FEP, which predisposes the results to be more clinically applicable (that is, to any patient presenting with a FEP). Because this study is based on real-life patient data from electronic health records (EHRs) from different regions, we were unable to address potential secular and regional trends in monitoring, laboratory testing and prescribing practice that could have biased results. However, using completely separate development and validation samples is required to adhere to best prediction modelling practice, which requires external validation on separate participants to avoid 'high risk of bias'[29].

Among the limitations of this study, we used clozapine treatment—a proxy measure for TRS—as the outcome, as in several previous studies[14]. Prevalence of clozapine use in our samples was lower than the expected prevalence of 13% (see calculation in Methods). In the UK, clozapine should be offered to all patients with TRS[41]. However, a recent national audit showed that only 52% of patients with FEP who have not responded adequately to at least two antipsychotics are offered clozapine[42]. Furthermore, EIPs accept patients with psychotic symptoms from any cause, thus including bipolar and unipolar mood disorders; this diagnostically inclusive nature of our FEP cohort might partially explain the relatively low rate of TRS. However, while our outcome definition may have a reduced sensitivity for capturing treatment resistance, the specificity is likely to be high. Indeed, the UK National Institute for Health and Care Excellence (NICE) guidance is that prescription of clozapine is reserved for those with schizophrenia in whom two trials of antipsychotics have failed[43], and the only UK indication for clozapine other than TRS is Parkinson's disease, which would be extremely rare in FEP cohorts only including adults up to 65 (mean age of 28 years; Table 2). Further, the literature suggests that clozapine in the UK is used 'off label' for treating refractory mania, psychotic depression, aggression in psychotic patients, the reduction of tardive dyskinesia symptoms and borderline personality disorder[44], therefore the presence of such diagnoses among the cases cannot be excluded, and is a limitation of this study. However, a UK-based systematic investigation of off-label antipsychotic use in secondary care established that clozapine is the least likely to be used outside its approved indications, with only 1 of 46 patients (~2%) in the study using it off label[45], which might be a consequence of the very strict regulations in place for clozapine use. Another UK-based study of TRS, including 14,299 patients, both inpatient and community-based, undergoing mandatory clozapine blood-monitoring, found 56 off-label clozapine prescriptions (0.4%)[46]. While these studies included any patient on antipsychotics, our cohorts are based on UK EIP teams, which only accept young patients with a FEP, and therefore it is likely that off-label clozapine use in this group is even rarer. Further, not all cohorts could provide information about time of clozapine initiation, and therefore time-to-event analysis could not be performed. Moreover, follow-up data were available for up to eight years following a FEP. This means that we might not have been able to capture 'late onset' TRS, which might develop after a number of relapses and over a number of years[47]; this might also help to explain the relatively low clozapine rate in our samples.

Table 2 | Predictor comparisons between samples used in model development and internal/external validation

		Development		External validation
Predictor	**CAMEO EIP**	**Birmingham EIP**	**Pooled development sample**	**SLaM EIP validation sample**
Final included sample size, *N*	539	246	785	1,110
Male sex, *N* (%)	328 (60.9%)	146 (59.3%)	474 (60.4%)	692 (62.3%)
Age (years), mean (s.d.)	30.23 (12.00)	23.86 (4.87)	28.24 (10.73)	28.82 (9.94)
Age (years), min, max	14, 65	15, 37	14, 65	17.5, 64
White or unrecorded ethnicity, *N* (%)	449 (83.3%)	70 (28.4%)	519 (66.1%)	378 (34.0%)
Black or African-Caribbean ethnicity, *N* (%)	21 (3.9%)	57 (23.2%)	78 (9.9%)	507 (45.7%)
Asian ethnicity, *N* (%)	69 (12.8%)	119 (48.4%)	188 (23.9%)	225 (20.3%)
Triglycerides (mmol l^{-1}), mean (s.d.)	1.42 (1.07)	1.55 (1.30)	1.46 (1.15)	1.25 (0.96)
ALP (U l^{-1}), mean (s.d.)	78.58 (25.55)	82.67 (25.78)	79.86 (25.68)	75.03 (22.01)
Lymphocyte count (billion l^{-1}), mean (s.d.)	1.91 (0.69)	2.22 (0.79)	2.01 (0.74)	1.98 (0.64)
Smoking, *N* (%)	201 (37.3%)	124 (50.4%)	325 (41.4%)	468 (42.2%)
BMI (kg m^{-2}), mean (s.d.)	24.68 (6.65)	25.74 (5.78)	25.01 (6.41)	24.07 (5.58)
Random plasma glucose (mmol l^{-1}), mean (s.d.)	5.25 (1.73)	4.87 (1.28)	5.13 (1.61)	5.11 (1.80)
Neutrophil count (billion l^{-1}), mean (s.d.)	4.60 (2.00)	4.14 (2.06)	4.46 (2.03)	4.01 (1.98)
Follow-up time (years), mean (s.d.)	4.45 (1.57)	3.55 (0.58)	4.17 (1.40)	4.41 (1.76)
Follow-up time (years), min, max	0.58[a], 8.50	2.67, 4.58		0.75[a], 8.75
TRS at follow-up, *N* (%)	35 (6.5%)	23 (9.3%)	58 (7.4%)	102 (9.2%)

[a]Participants were required to have at least 2 years of follow-up, with the exception of TRS cases.

Predictor availability was limited to those markers that were available in all three study cohorts. No cohort included a symptom or severity measure, such as the Positive and Negative Syndrome Scale (PANSS); we could therefore not include symptoms at baseline as a predictor. The number of predictors that we could include was also limited by our sample size, although we took particular care in predictor selection and this may have helped to prevent model overfitting[28,48]. It must be pointed out that this work did not aim to make any assumptions about whether the included predictors might be causal to TRS: variables were selected if they were known to be associated – that is, likely capturing part of the outcome's variance.

A further limitation of this work is the potential for the inclusion of patients already taking antipsychotic medication at baseline. Antipsychotics could influence the levels of the biomarkers. However, most patients admitted to an EIP are medication naïve or minimally treated. Bloods tests were only used for prediction if performed within 100 days of referral to the EIP; it is likely that some patients were started on antipsychotic medication during this time, though the duration of treatment is likely to have been relatively short. However, participants were excluded if the outcome (starting clozapine) predated baseline blood collection.

Conclusions

In conclusion, we report that, based on three large samples of patients, routinely recorded demographics and biomarkers measured at presentation with a FEP could be useful in the individualized prediction of the risk of clozapine use (as a proxy for developing TRS) up to eight years later. Subject to further external validation and regulatory approval, MOZART appears useful at predicting the risk of TRS at lower propensity-to-intervene thresholds, thus potentially allowing implementation of low-risk strategies such as closer psychiatric monitoring for TRS in at-risk populations. This could potentially speed up the time from FEP onset to clozapine start, thus reducing delays in TRS recognition and treatment, and consequently reducing suffering and improving quality of life.

We suggest that future efforts in TRS risk prediction should seek to consider such routinely collected data. Doing so may improve both model predictive performance and likely clinical usefulness, both of which are crucial for the future routine deployment of a risk prediction model into clinical practice.

Methods

Model development

We used a forced-entry logistic regression model as the most parsimonious way to predict a binary outcome (such as clozapine use) from a small number of predictors. However, to explore whether additional predictors may improve performance in a manner that reduces risk of overfitting, we also used a LASSO-based selection model, which has the benefit of independently shrinking the predictors' coefficients up to excluding them, and can be more robust to a slightly larger number of predictors.

Data from 785 patients were included in the pooled development sample: 539 from CAMEO and 246 from the Birmingham EIP (Table 2), following EHR searches and application of inclusion and exclusion criteria (Fig. 1, and a description of the included and excluded samples in Supplementary Table 5). Included patients had a mean age of 28.2 years, an average BMI of 25, and were 66% White and 41% smokers. In the pooled development sample, 58 (7.4%) patients were treated with clozapine.

Ethical approval

All research complied with relevant ethical regulations and underwent the local approval process in each of the three cohorts. CAMEO data were identified by anonymously searching for all EIP patients using the Cambridgeshire and Peterborough NHS Foundation Trust (CPFT) Research Database[49]—approved under UK NHS Research Ethics Service references 12/EE/0407, 17/EE/0442. Anonymized data for all patients enrolled in the Birmingham EIP were collected as part of the National Clinical Audit of Psychosis Quality Improvement

Programme, and were enhanced locally with biomarker data; the work conformed to the Health Research Authority definition of service evaluation (confirmed by Birmingham Women's and Children's Hospital NHS Foundation Trust). The Clinical Records Interactive Search (CRIS) resource was used to capture anonymized data from SLaM EIP—approved under UK NHS Research Ethics Service references 18/SC/0372 and 08/H0606/71+5; National Institute for Health Research (NIHR) Biomedical Research Centre (BRC) CRIS Oversight Committee reference 20-005.

Samples

EIP currently represents the gold standard of care for people presenting with a FEP[50]. EIPs are built around a specialized, multidisciplinary team, which can deliver both pharmacological and psychological interventions, family and social support, support with employment and physical healthcare checks for up to five years after patients are diagnosed with a FEP. This model of care has been shown to be highly effective[1], and, given the longitudinal follow-up by a highly specialized team, to guarantee a high degree of confidence that enrolled patients suffer from a FEP, including the first manifestations of both primary psychotic conditions such as schizophrenia and schizoaffective disorders, and psychotic mood disorders such as psychotic depression or mania.

Model development sample. We developed a risk prediction model using pooled longitudinal data from patients enrolled in the CAMEO psychosis EIP, searching for patients enrolled between 1 January 2013 and 31 May 2021 (sampling frame $n = 1{,}660$), or the Birmingham EIP, searching for patients enrolled between 1 January 2014 and 31 December 2018 (sampling frame $n = 391$). This was selected as the development sample for the present study as CAMEO data were recently used to examine group-level associations between mean biomarker levels and psychiatric outcomes[2].

Predictors were assessed within 100 days of patient EIP enrolment. We excluded any participant who had missing data on more than 50% of predictor variables, and non-cases (patients who did not use clozapine) who had less than two years of follow-up to reduce the probability of including future TRS cases as non-cases. Further details on missing data management can be found in the Supplementary Notes. All patients who developed TRS were included regardless of duration of follow-up. As predictors must predate outcomes, we also excluded all cases where the outcome start date (clozapine treatment start date) predated the earliest available baseline bloods in the CAMEO cohort (and in the SLaM cohort, see the Model external validation sample section), or participants who started taking clozapine within 100 days of baseline in the Birmingham cohort. Patients were excluded if they died or moved out of the Trust's catchment area during follow-up.

Model external validation sample. We used the CRIS resource to capture anonymized data from SLaM. Our sampling frame included 3,012 EIP patients, all those enrolled between 1 January 2012 and 20 November 2021. Patients were excluded and predictors and outcomes were assessed as for the development sample.

Outcome

Owing to data availability, we adopted a pragmatic definition of TRS: patients were defined as having TRS if they had been treated with clozapine at any point during the follow-up period. Clozapine is the only clinically approved treatment for TRS in the UK, and provides an objective, easily quantifiable measure of TRS[41]. We calculated an expected prevalence of clozapine use of 13%. This was calculated as follows: starting from a population prevalence of 23%[3,14,51], we expected to capture mostly 'early onset' cases, which represent ~84% of cases[11]. From previous literature, clozapine is given in ~68% of TRS cases[11], so the expected prevalence was $0.23 \times 0.84 \times 0.68 = 0.13$.

Predictor variables

Routinely used clinical predictors were included based on a balance of clinical knowledge, existing research and likely clinical usefulness. Demographic variables were considered if they had shown evidence of potential predictive ability for TRS in existing prognosis research[16,24]. Biomarkers and clinical measures were considered if they showed evidence from past longitudinal association studies of biological measures at FEP using long-term clinical outcomes[2,23]. Predictors were only included if they were part of the suite of measurements that should be collected at baseline as part of local or national guidelines, to avoid ascertainment bias. We did not include variables that may only be recorded in specific circumstances, such as CRP, which may only be recorded when an infection is suspected. All predictors needed to be available in all three EIP samples. Therefore, we considered the following parameters, measured within 100 days of EIP start: sex (female or male); age (years); ethnicity (categorical: White European or not recorded (reference), Black or African-Caribbean, Asian, or other); triglyceride concentration (mmol l^{-1}); lymphocyte and neutrophil blood cell counts (billion l^{-1}); ALP levels (units l^{-1}), smoking status (binary, at least one cigarette on average daily); BMI (kg m^{-2}); and random glucose levels (mmol l^{-1}).

See Supplementary Notes for full rationale and details of data extraction.

Statistical analysis

All data analyses were conducted in R (v. 4.2.1)[52]. We performed sample size calculations using the R package pmsampsize (v. 1.1.2)[53]; for details on sample size calculations, please see the Supplementary Notes. For data imputation we used the MICE package (v. 3.14)[54]. For logistic modelling we used base R and the pROC package (v. 1.18)[55]. For calibration plots we used the CalibrationCurves package (v. 0.1.5)[56]. For LASSO model development, we used the MAMI package (v. 0.9.13)[57]. Finally, for coefficient shrinking we used the psfmi package (v. 1.0)[58].

Primary analysis. We performed sample size calculations using the R package pmsampsize[48,53]. The sample size required was estimated from the estimated outcome prevalence, the a priori estimated R^2 of the model, and the estimated required model shrinkage. For 11 predictors, the minimum sample required was 412. We did not consider non-linear terms or interactions to reduce the risk of overfitting. See Supplementary Notes for detailed sample size calculations.

We used multiple imputation using chained equations for missing data and pooled estimates using Rubin's rules (see Supplementary Notes for details about predictor missingness). Internal validation involved bootstrap resampling (500 bootstraps) to obtain an estimate of the corrected calibration slope. The resulting pooled, corrected C slope was then used as a shrinkage factor for our coefficients. After this step, predictive performance was assessed.

We developed the risk calculator using two alternative model selection methods:

1. A forced-entry logistic regression model, including all sociodemographic and three biological predictors (one lipid, one inflammatory and one liver marker), based on a balance of clinical knowledge, past research and likely clinical usefulness.
2. A LASSO-based selection model, after predictor scaling and centring, including all 11 pre-selected sociodemographic, lifestyle and biological predictors. The inclusion of additional variables was enabled by LASSO including a predictor selection step, and by its more efficient coefficient shrinkage, leading to less risk of model overfit[59]. For the LASSO model we used 100-fold cross-validation to tune the penalty parameter in the development sample as implemented in glmnet[60].

Both methods involved variable pre-selection, after ruling out predictor multi-collinearity to minimize risk of overfitting, as is recommended for smaller datasets[61].

The models were applied to the external validation sample. The distribution of predicted outcome probabilities was inspected using histograms.

Model performance was assessed primarily with measures of discrimination (the ability of the model to distinguish participants with the outcome from those without), such as the *C* statistic, and calibration (the extent to which the outcome probabilities predicted by the model in specified risk-defined subgroups are similar to those observed in the validation dataset), assessed by inspection of calibration plots (presented as figures, e.g. Fig. 2).

The discrimination of the models was assessed using the *C* statistic. For binary outcomes, this is equivalent to the area under the receiver operating characteristic curve[61], which plots sensitivity against 1 minus specificity. The *C* statistic normally ranges from 0.5 to 1, with a value of 1 representing perfect discrimination and a value of 0.5 representing discrimination no better than chance. *C* statistics were determined in relation to the observed binary outcomes (subsequent clozapine use or not).

We also recorded calibration intercepts (ideally close to 0) and Brier scores (an overall measure of model performance, ideally close to 0, with scores >0.25 generally indicating a poor model). For further details of our prediction methods, see ref. 19.

Model recalibration/revision. Additionally, where performance at external validation differed from internal validation performance, we considered two recalibration approaches. First, we considered logistic recalibration. This method is used where the coefficients of the original model may have been over-fitted, affecting calibration performance. Logistic recalibration assumes similar relative effects of the predictors, but allows for a larger or smaller absolute effect of the predictors[62]. Further details are in Supplementary Notes. Second, where there was evidence of a clear difference in the association of a predictor with clozapine use between the development and validation samples, we considered logistic recalibration plus revising a single predictor in the model. We limited this model revision approach to a maximum of one model predictor, to preserve as much of the character of an external validation analysis as possible, though we note that all recalibrated/revised models will require a further external validation in an additional unseen sample.

Decision curve analysis. Decision curve analysis was performed to assess potential clinical benefit[63]. Clinical net benefit of the prediction model is calculated against offering an intervention to all or no patients. This can be calculated at a range of propensity-to-intervene thresholds. Net benefit is defined as the minimum probability of clozapine use at which the intervention would be warranted, as net benefit = sensitivity × prevalence − (1 − specificity) × (1 − prevalence) × w, where w is the odds at the propensity-to-intervene threshold[64]. In decision curve analysis, it is usual to only consider the range of propensity-to-intervene thresholds that may be clinically relevant; these depend on how risky the intervention being offered might be.

For starting clozapine, we selected a priori a propensity-to-intervene threshold of 0.50, representing a greater than 50% risk of developing TRS. We believe that such a threshold would represent a good balance between the potential positives of early clozapine initiation and relatively rare risks of clozapine. We also selected a lower propensity-to-intervene threshold of 0.10 (>10% risk of developing TRS) for defining a 'TRS-at risk population' who may be eligible for close monitoring.

The decision curve plot is presented as Fig. 3, to visualize the net benefit of both model versions (forced-entry original and recalibrated) over varying propensity-to-intervene thresholds compared with treating all patients or no-one. Classical decision theory proposes that at a chosen propensity-to-intervene threshold, the choice with the greatest net benefit should be preferred[63].

Sensitivity analysis. To examine the added benefit of selected demographic and biological predictors, we examined iterative improvements of the model. The first model included only a single demographic predictor, sex; the second added all demographics; the third included all demographics plus a single biological predictor (triglycerides); the last model included all the above plus a second biological predictor (ALP). We did not externally validate the incremental models.

Visual representation of the model

We developed an online data visualization tool using shiny for R[65], allowing interactive exploration of the effect of sociodemographic, lifestyle, and clinical variables and their combinations on TRS risk in people with FEP. The tool is not yet suitable for clinical use.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The source data for this work is anonymized patient records from three UK NHS Trusts: the CPFT Research Database, SLaM CRIS and the Birmingham EIP. Data from these Trusts are only available to clinicians and clinical researchers with clinical contracts with the Trusts. The data are securely held on clinical systems and available following ethical approval to preserve patient confidentiality. Therefore, the raw data cannot be shared. However, we developed an online data visualization tool (https://eosimo.shinyapps.io/trs_app/) for both the original and recalibrated MOZART models, which allows interactive exploration of the effect of each predictor and their combinations on the risk of clozapine use based on the predictors included in this study.

Code availability

R code for data extraction and analysis is available upon request to the corresponding author.

References

1. Menezes, N., Arenovich, T. & Zipursky, R. A systematic review of longitudinal outcome studies of first-episode psychosis. *Psychol. Med.* **36**, 1349–1362 (2006).
2. Osimo, E. F. et al. Inflammatory and cardiometabolic markers at presentation with first episode psychosis and long-term clinical outcomes: A longitudinal study using electronic health records. *Brain Behav. Immun.* **91**, 117–127 (2021).
3. Siskind, D. et al. Rates of treatment-resistant schizophrenia from first-episode cohorts: systematic review and meta-analysis. *Br. J. Psychiatry* **220**, 115–120 (2022).
4. Howes, O. D., Thase, M. E., & Pillinger, T. Treatment resistance in psychiatry: state of the art and new directions. *Mol. Psychiatry* **27**, 58–72 (2021).
5. Kennedy, J. L., Altar, C. A., Taylor, D. L., Degtiar, I. & Hornberger, J. C. The social and economic burden of treatment-resistant schizophrenia: a systematic literature review. *Int. Clin. Psychopharmacol.* **29**, 63–76 (2014).
6. Mizuno, Y., McCutcheon, R. A., Brugger, S. P. & Howes, O. D. Heterogeneity and efficacy of antipsychotic treatment for schizophrenia with or without treatment resistance: a meta-analysis. *Neuropsychopharmacology* **45**, 622–631 (2020).
7. Howes, O. D. et al. Adherence to treatment guidelines in clinical practice: study of antipsychotic treatment prior to clozapine initiation. *Br. J. Psychiatry* **201**, 481–485 (2012).
8. Barnes, T. R. et al. Evidence-based guidelines for the pharmacological treatment of schizophrenia: updated recommendations from the British Association for Psychopharmacology. *J. Psychopharmacol.* **34**, 3–78 (2020).

9. McGuire, P. & Dazzan, P. Does neuroimaging have a role in predicting outcomes in psychosis? *World Psychiatry* **16**, 209–210 (2017).
10. Wimberley, T. et al. Predictors of treatment resistance in patients with schizophrenia: a population-based cohort study. *Lancet Psychiatry* **3**, 358–366 (2016).
11. Demjaha, A. et al. Antipsychotic treatment resistance in first-episode psychosis: prevalence, subtypes and predictors. *Psychol. Med.* **47**, 1981–1989 (2017).
12. Chan, S. et al. Predictors of treatment resistant schizophrenia-spectrum disorder: 10-year retrospective study of first-episode psychosis (A56). *Early Interv. Psychiatry* **8**, 78 (2014).
13. Bozzatello, P., Bellino, S. & Rocca, P. Predictive factors of treatment resistance in first episode of psychosis: a systematic review. *Front. Psychiatry* **10**, 67 (2019).
14. Lally, J. et al. Two distinct patterns of treatment resistance: clinical predictors of treatment resistance in first-episode schizophrenia spectrum psychoses. *Psychol. Med.* **46**, 3231–3240 (2016).
15. Üçok, A. et al. Correlates of clozapine use after a first episode of schizophrenia: results from a long-term prospective study. *CNS Drugs* **30**, 997–1006 (2016).
16. Smart, S., Kępińska, A., Murray, R. & MacCabe, J. Predictors of treatment resistant schizophrenia: a systematic review of prospective observational studies. *Psychol. Med.* **51**, 44–53 (2021).
17. Dwyer, D. B., Falkai, P. & Koutsouleris, N. Machine learning approaches for clinical psychology and psychiatry. *Annu Rev. Clin. Psychol.* **14**, 91–118 (2018).
18. 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* **357**, j2099 (2017).
19. Perry, B. I. et al. Development and external validation of the Psychosis Metabolic Risk Calculator (PsyMetRiC): a cardiometabolic risk prediction algorithm for young people with psychosis. *Lancet Psychiatry* **8**, 589–598 (2021).
20. Perry, B. I., McIntosh, G., Weich, S., Singh, S. & Rees, K. The association between first-episode psychosis and abnormal glycaemic control: systematic review and meta-analysis. *Lancet Psychiatry* **3**, 1049–1058 (2016).
21. Pillinger, T., Beck, K., Stubbs, B. & Howes, O. D. Cholesterol and triglyceride levels in first-episode psychosis: systematic review and meta-analysis. *Br. J. Psychiatry* **211**, 339–349 (2017).
22. Pillinger, T. et al. A meta-analysis of immune parameters, variability, and assessment of modal distribution in psychosis and test of the immune subgroup hypothesis. *Schizophr. Bull.* **45**, 1120–1133 (2019).
23. Nettis, M. A. et al. Metabolic-inflammatory status as predictor of clinical outcome at 1-year follow-up in patients with first episode psychosis. *Psychoneuroendocrinology* **99**, 145–153 (2019).
24. Legge, S. et al. Clinical indicators of treatment-resistant psychosis. *Br. J. Psychiatry* **216**, 259–266 (2020).
25. Wimberley, T. et al. Polygenic risk score for schizophrenia and treatment-resistant schizophrenia. *Schizophr. Bull.* **43**, 1064–1069 (2017).
26. Trubetskoy, V. et al. Mapping genomic loci implicates genes and synaptic biology in schizophrenia. *Nature* **604**, 502–508 (2022).
27. Pardiñas, A. F. et al. Interaction testing and polygenic risk scoring to estimate the association of common genetic variants with treatment resistance in schizophrenia. *JAMA Psychiatry* **79**, 260–269 (2022).
28. Steyerberg, E. W. et al. Prognosis Research Strategy (PROGRESS) 3: prognostic model research. *PLoS Med.* **10**, e1001381 (2013).
29. Wolff, R. F. et al. PROBAST: a tool to assess the risk of bias and applicability of prediction model studies. *Ann. Intern. Med.* **170**, 51–58 (2019).
30. Morlán-Coarasa, M. J. et al. Incidence of non-alcoholic fatty liver disease and metabolic dysfunction in first episode schizophrenia and related psychotic disorders: a 3-year prospective randomized interventional study. *Psychopharmacology* **233**, 3947–3952 (2016).
31. Perry, B. I. et al. Dysglycaemia, inflammation and psychosis: findings from the UK ALSPAC birth cohort. *Schizophr. Bull.* **45**, 330–338 (2019).
32. Pillinger, T. et al. Impaired glucose homeostasis in first-episode schizophrenia: a systematic review and meta-analysis. *JAMA Psychiatry* **74**, 261–269 (2017).
33. Machado, M. V. & Diehl, A. M. Pathogenesis of nonalcoholic steatohepatitis. *Gastroenterology* **150**, 1769–1777 (2016).
34. Dix, H. M., Robinson, E. M. & Dillon, J. F. in *Textbook of Addiction Treatment* (eds. el-Guebaly, N., et al.) 1099–1111 (Springer, 2021).
35. Van de Mortel, T. F. Faking it: social desirability response bias in self-report research. *Aust. J. Adv. Nurs.* **25**, 40–48 (2008).
36. Moody, G. & Miller, B. J. Total and differential white blood cell counts and hemodynamic parameters in first-episode psychosis. *Psychiatry Res.* **260**, 307–312 (2018).
37. Garcia-Rizo, C. et al. Blood cell count in antipsychotic-naive patients with non-affective psychosis. *Early Interv. Psychiatry* **13**, 95–100 (2019).
38. Perry, B. I. et al. Associations of immunological proteins/traits with schizophrenia, major depression and bipolar disorder: a bi-directional two-sample Mendelian randomization study. *Brain Behav. Immun.* **97**, 176–185 (2021).
39. Bunders, M., Cortina-Borja, M. & Newell, M.-L. Age-related standards for total lymphocyte, CD4+ and CD8+ T cell counts in children born in Europe. *Pediatr. Infect. Dis. J.* **24**, 595–600 (2005).
40. Lang, X. et al. Differences in patterns of metabolic abnormality and metabolic syndrome between early-onset and adult-onset first-episode drug-naive schizophrenia patients. *Psychoneuroendocrinology* **132**, 105344 (2021**)**.
41. *Psychosis and Schizophrenia in Adults: Prevention and Management* CG178 (National Institute for Health and Care Excellence, 2014). https://www.nice.org.uk/guidance/cg178
42. *National Clinical Audit of Psychosis – National Report for the Early Intervention in Psychosis Audit 2019/2020. London* (Royal College of Psychiatrists, 2020). www.rcpsych.ac.uk/NCAP
43. *Psychosis and Schizophrenia in Adults* QS80 (National Institute for Health and Care Excellence, 2015). https://www.nice.org.uk/guidance/qs80
44. Haw, C. & Stubbs, J. Off-label use of antipsychotics: are we mad? *Expert Opin. Drug Saf.* **6**, 533–545 (2007).
45. Hodgson, R. & Belgamwar, R. Off-label prescribing by psychiatrists. *Psychiatric Bull.* **30**, 55–57 (2006).
46. Pardiñas, A. F. et al. Common schizophrenia alleles are enriched in mutation-intolerant genes and in regions under strong background selection. *Nat. Genet.* **50**, 381–389 (2018).
47. Chan, S. K. W. et al. Predictors of treatment-resistant and clozapine-resistant schizophrenia: a 12-year follow-up study of first-episode schizophrenia-spectrum disorders. *Schizophrenia Bull.* **47**, 485–494 (2021).
48. Riley, R. D. et al. Minimum sample size for developing a multivariable prediction model: PART II—binary and time-to-event outcomes. *Stat. Med.* **38**, 1276–1296 (2019).
49. Cardinal, R. N. Clinical records anonymisation and text extraction (CRATE): an open-source software system. *BMC Med. Inf. Decis. Making* **17**, 50 (2017).
50. McGorry, P. D. Early intervention in psychosis: obvious, effective, overdue. *J. Nerv. Ment. Dis.* **203**, 310–318 (2015).
51. Meltzer, H. Y. Treatment-resistant schizophrenia-the role of clozapine. *Curr. Med. Res. Opin.* **14**, 1–20 (1997).

52. R Core Team *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, 2022).
53. Ensor, J., Martin, E. C. & Riley, R. D. pmsampsize (2021). https://cran.r-project.org/web/packages/pmsampsize/index.html
54. Van Buuren, S. & Groothuis-Oudshoorn, K. mice: multivariate imputation by chained equations in R. *J. Stat. Softw.* **45**, 1–67. (2011).
55. Robin, X. et al. pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinf.* **12**, 77 (2011).
56. Van Calster, B. et al. A calibration hierarchy for risk models was defined: from utopia to empirical data. *J. Clin. Epidemiol.* **74**, 167–176 (2016).
57. Schomaker, M. & Heumann, C. Model selection and model averaging after multiple imputation. *Comput. Stat. Data Anal.* **71**, 758–770 (2014).
58. Eekhout, I., Van De Wiel, M. A. & Heymans, M. W. Methods for significance testing of categorical covariates in logistic regression models after multiple imputation: power and applicability analysis. *BMC Med. Res. Method.* **17**, 129 (2017).
59. Radchenko, P. & James, G. M. Variable inclusion and shrinkage algorithms. *J. Am. Stat. Assoc.* **103**, 1304–1315 (2008).
60. Friedman, J., Hastie, T. & Tibshirani, R. Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.* **33**, 1–22 (2010).
61. Steyerberg, E. W. *Clinical Prediction Models* (Springer, 2019).
62. Steyerberg, E. W., Borsboom, G. J., van Houwelingen, H. C., Eijkemans, M. J. & Habbema, J. D. F. Validation and updating of predictive logistic regression models: a study on sample size and shrinkage. *Stat. Med.* **23**, 2567–2586 (2004).
63. Vickers, A. J. & Elkin, E. B. Decision curve analysis: a novel method for evaluating prediction models. *Med. Decis. Making* **26**, 565–574 (2006).
64. Vickers, A. J., van Calster, B. & Steyerberg, E. W. A simple, step-by-step guide to interpreting decision curve analysis. *Diagn. Progn. Res.* **3**, 18 (2019).
65. Chang, W. et al. shiny: Web Application Framework for R v.1.7.2 (2022). https://cran.r-project.org/web/packages/shiny/index.html

Acknowledgements

This work was funded by a Clinical PhD Fellowship to E.F.O. jointly funded by the NIHR Imperial BRC and the UK Research and Innovation Medical Research Council London Institute of Medical Sciences. B.I.P. acknowledges funding support from the NIHR (doctoral research fellowship, DRF-2018-11-ST2-018). R.U. received funding support from the NIHR (HTA grant 127700) and Medical Research Council (Therapeutic Target Validation in Mental Health grant MR/S037675/1). G.M.K. received funding support from the Wellcome Trust (grant 201486/Z/16/Z), the MQ: Transforming Mental Health (grant MQDS17/40), the Medical Research Council UK (grants MC_PC_17213; MR/S037675/1; and MR/W014416/1), and the British Medical Association Foundation (J Moulton grant 2019). R.N.C. acknowledges support from the Medical Research Council (grants MC_PC_17213, MR/W014386/1). This research was supported in part by the NIHR Imperial BRC and NIHR Cambridge BRC (BRC-1215-20014); J.P. and P.B.J. acknowledge funding from the NIHR ARC EoE; the views expressed are those of the authors and not necessarily those of the NHS, the NIHR or the Department of Health and Social Care. The funding bodies had no role in design or conduct of the study; collection, management, analysis or interpretation of the data; preparation, review or approval of the manuscript; or the decision to submit the manuscript for publication.

Author contributions

E.F.O. and B.I.P. designed the study and selected the predictors and outcome variables under the supervision of R.U. and G.M.K., and with input from P.M., G.K.M., J.P., P.B.J., R.N.C. and O.D.H.; E.F.O. had access to all datasets, collected the data and performed the statistical analyses, in close discussion with B.I.P. and the wider supervisory team. M.P. and O.D.H. supported E.F.O. in data collection and analysis for the SLaM cohort. J.L. and R.N.C. supported E.F.O. in data collection and analysis for the Cambridge cohort. A.K. and R.U. supported E.F.O. in data collection and analysis for the Birmingham cohort. E.F.O. wrote the first draft of the manuscript, with constant support from B.I.P. All other authors contributed to the drafting, re-drafting and perfecting of the manuscript, including responses to reviewers' comments.

Competing interests

O.D.H. is a part-time employee of H. Lundbeck A/S. He has received investigator-initiated research funding from and/or participated in advisory/speaker meetings organized by Angelini, Autifony, Biogen, Boehringer Ingelheim, Eli Lilly, Heptares, Global Medical Education, Invicro, Janssen, H. Lundbeck A/S, Neurocrine, Otsuka, Sunovion, Recordati, Roche and Viatris/Mylan. O.D.H. has a patent for the use of dopaminergic imaging. R.N.C. consults for Campden Instruments and receives royalties from Cambridge Enterprise, Routledge and Cambridge University Press. The other authors declare no competing interests.

Additional information

Extended data is available for this paper at https://doi.org/10.1038/s44220-022-00001-z.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s44220-022-00001-z.

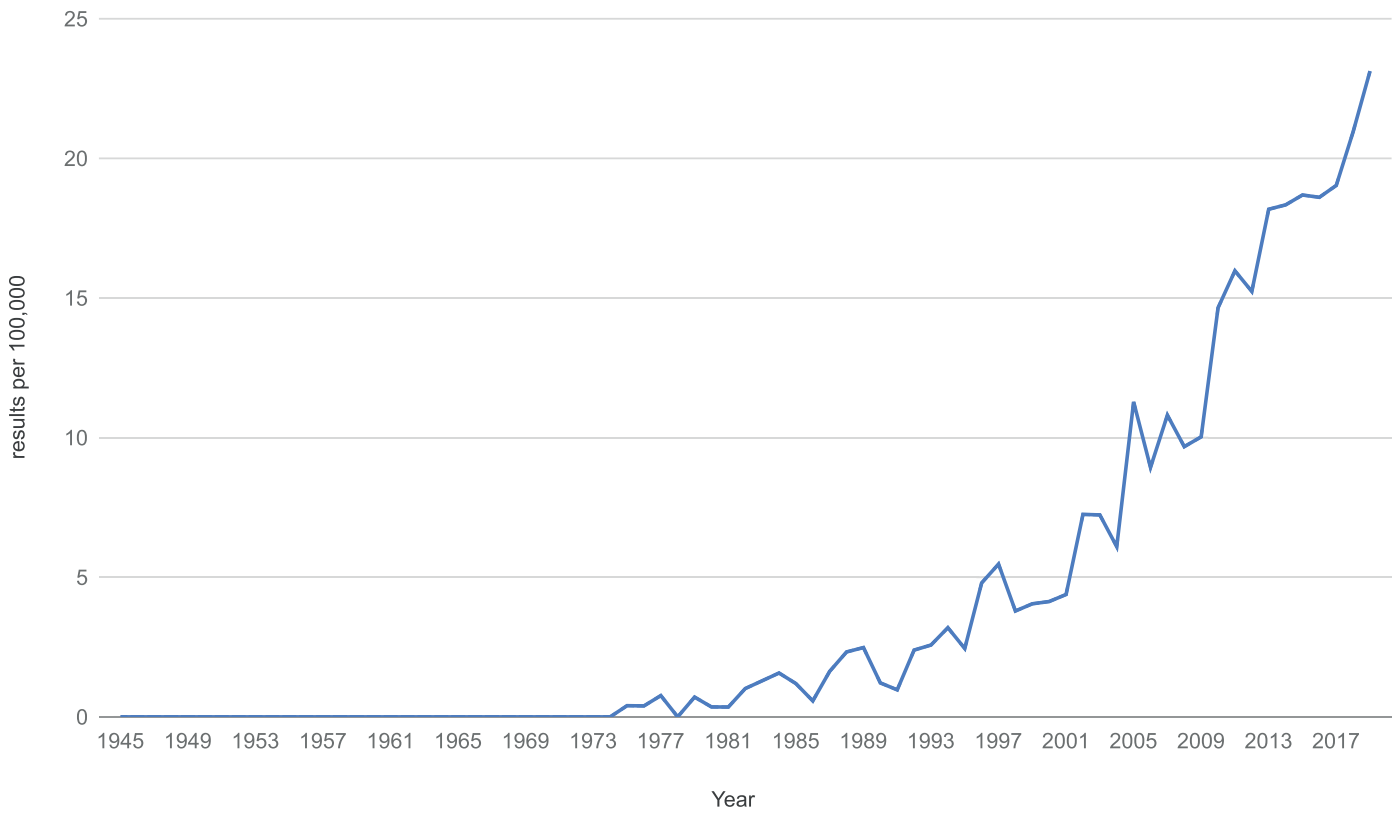
Correspondence and requests for materials should be addressed to Emanuele F. Osimo.

Peer review information *Nature Mental Health* thanks the anonymous reviewers for their contribution to the peer review of this work.

Emanuele F. Osimo[1,2,3,4,13] ✉, Benjamin I. Perry[2,3,13], Pavan Mallikarjun[5,6], Megan Pritchard[4], Jonathan Lewis[3], Asia Katunda[6], Graham K. Murray[2,3], Jesus Perez[2,3,7,8,9], Peter B. Jones[2,3,8], Rudolf N. Cardinal[2,3], Oliver D. Howes[1,4,10], Rachel Upthegrove[5,6,14] & Golam M. Khandaker[2,3,11,12,14]

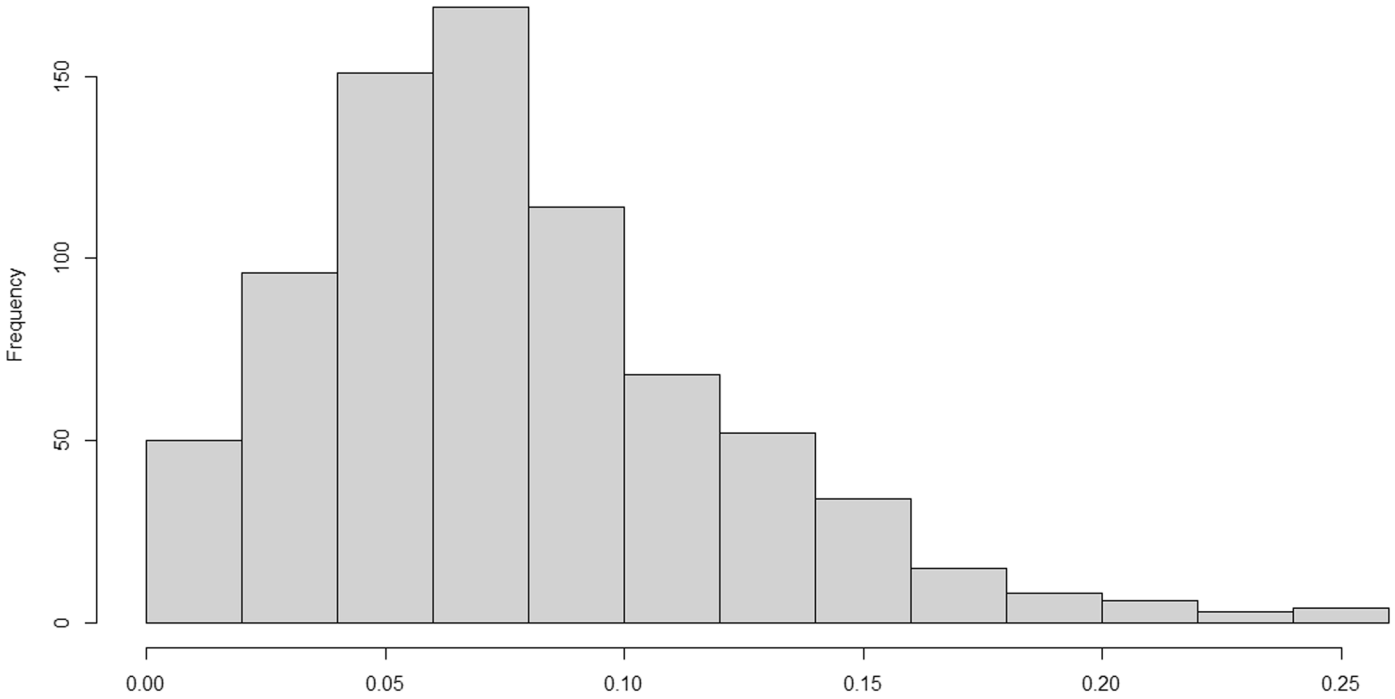
[1]MRC London Institute of Medical Sciences and Institute of Clinical Sciences, Hammersmith Campus, Imperial College London, London, UK. [2]Department of Psychiatry, University of Cambridge, Cambridge, UK. [3]Cambridgeshire and Peterborough NHS Foundation Trust, Cambridge, UK. [4]South London and Maudsley NHS Foundation Trust, London, UK. [5]Institute for Mental Health and Centre for Human Brain Health, University of Birmingham, Birmingham, UK. [6]Birmingham Early Intervention Service, Birmingham Women's and Children's NHS Foundation Trust, Birmingham, UK. [7]Norwich Medical School, University of East Anglia, Norwich, UK. [8]Applied Research Collaboration East of England, National Institute for Health Research, Cambridge, UK. [9]Department of Medicine, Institute of Biomedical Research (IBSAL), University of Salamanca, Salamanca, Spain. [10]Department of Psychosis Studies, Institute of Psychiatry, Psychology and Neuroscience, Kings College London, London, UK. [11]MRC Integrative Epidemiology Unit, Population Health Sciences, Bristol Medical School, University of Bristol, Bristol, England. [12]Centre for Academic Mental Health, Population Health Sciences, Bristol Medical School, University of Bristol, Bristol, England. [13]These authors contributed equally: Emanuele F. Osimo, Benjamin I. Perry. [14]These authors jointly supervised this work: Rachel Upthegrove, Golam M. Khandaker.
✉e-mail: efo22@cam.ac.uk

https://doi.org/10.1038/s44220-022-00001-z



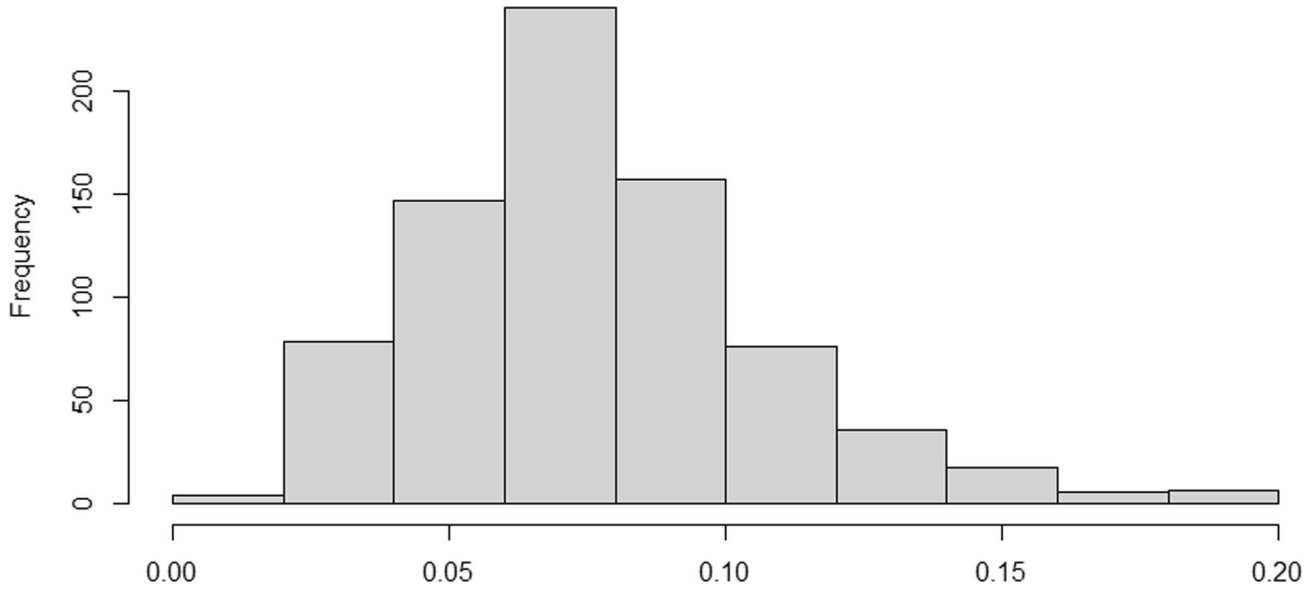
Extended Data Fig. 1 | PubMed results for ‘risk AND prediction AND psychosis’ by year. Generated with PubMed by Year. Available from http://esperr.github.io/pubmed-by-year/.

https://doi.org/10.1038/s44220-022-00001-z



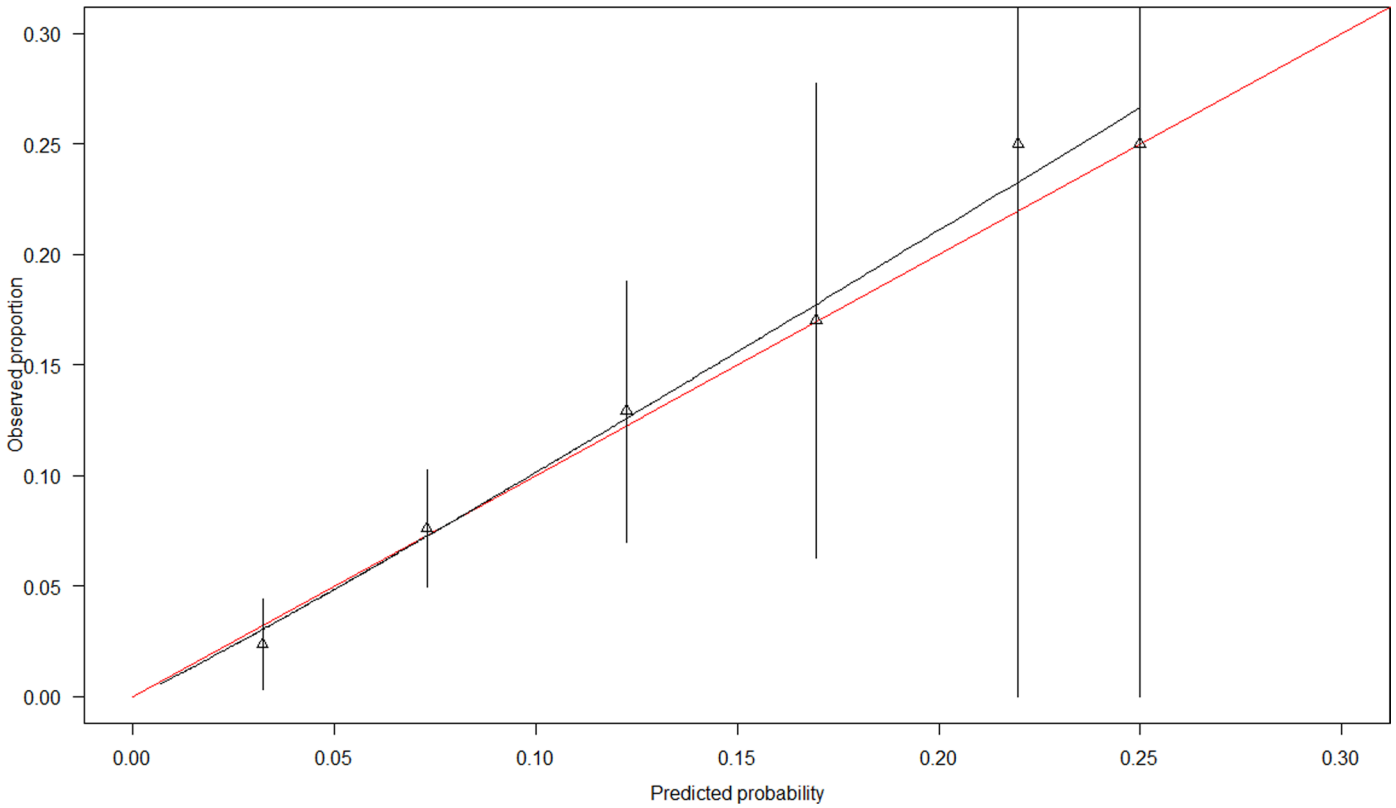
Extended Data Fig. 2 | Internal validation: distribution of predicted probabilities for MOZART. X axis: predicted probability.

https://doi.org/10.1038/s44220-022-00001-z



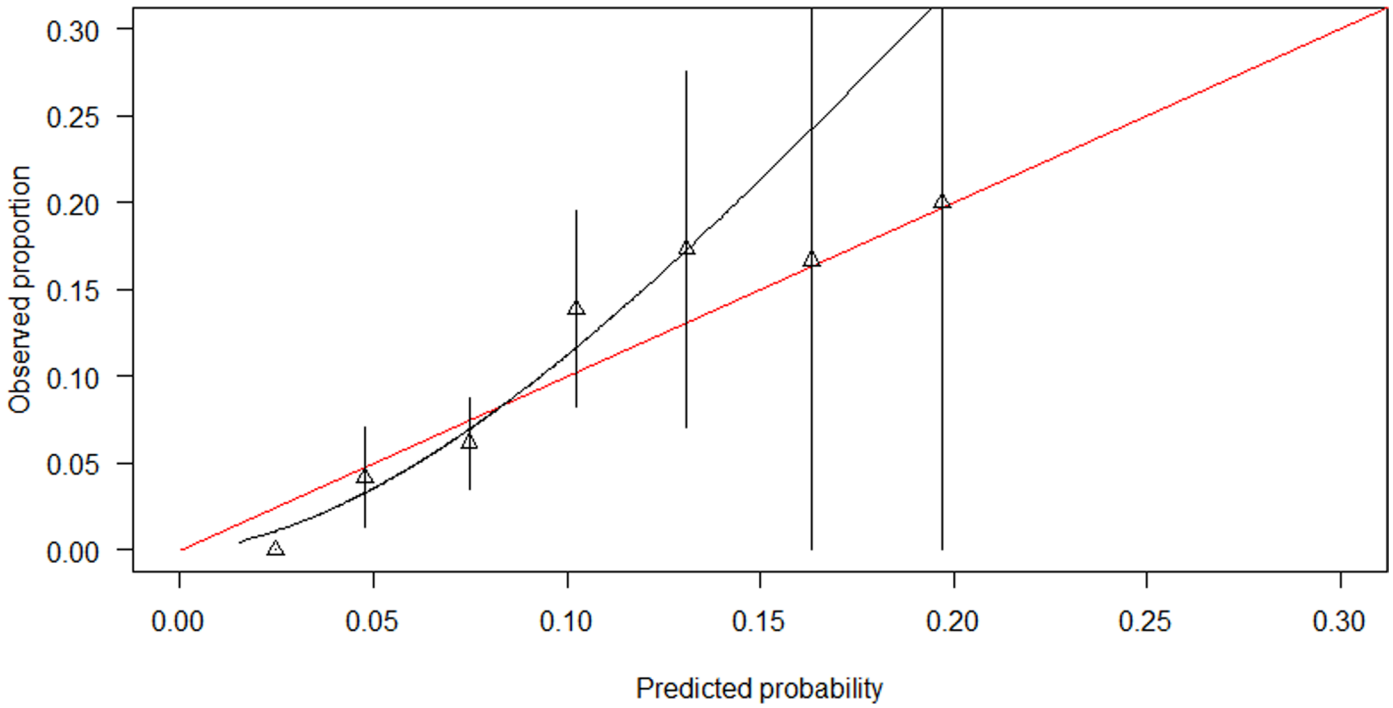
Extended Data Fig. 3 | Internal validation: distribution of predicted probabilities for the LASSO model. X axis: predicted probability.

https://doi.org/10.1038/s44220-022-00001-z



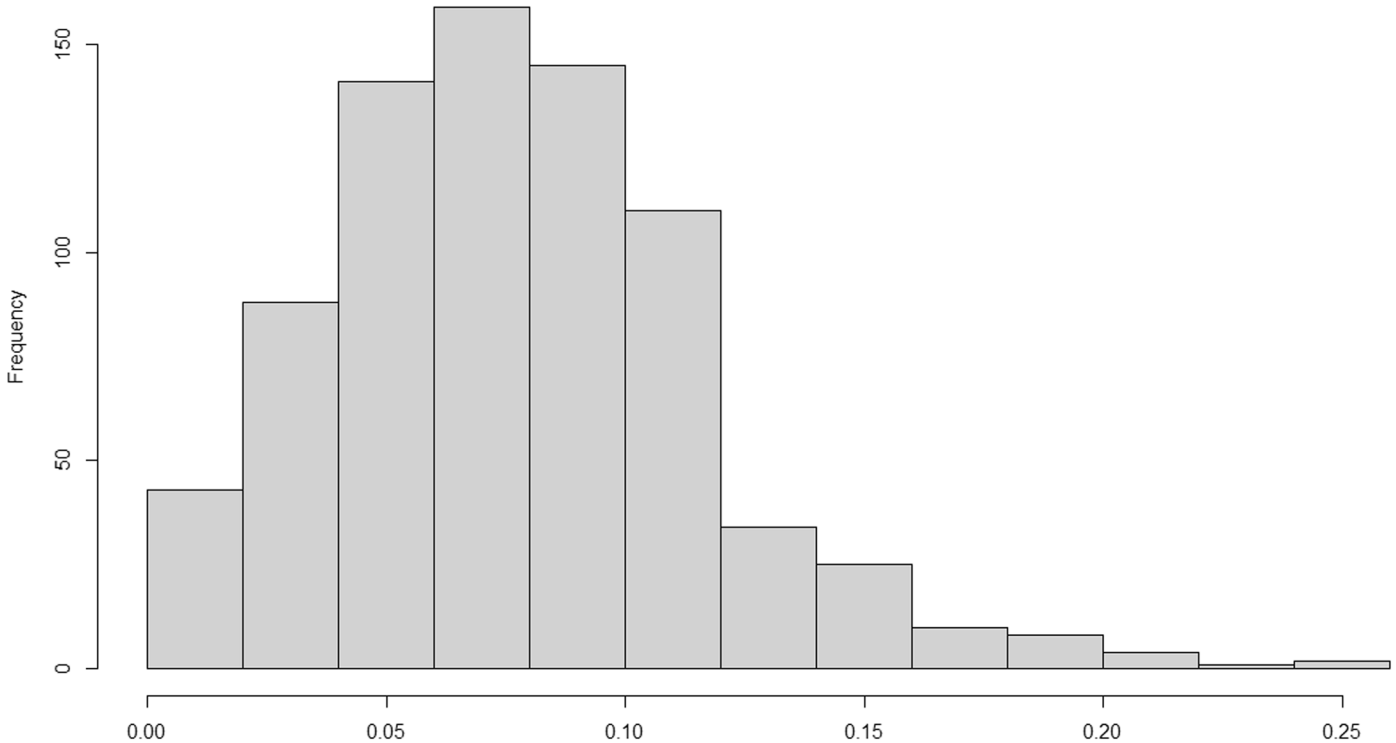
Extended Data Fig. 4 | Internal validation: calibration plot for MOZART. *Model calibration* is the extent to which outcomes predicted by the model are similar to those observed in the validation dataset. Calibration plots illustrate agreement between observed proportion (y axis) and predicted risk (x axis). Perfect agreement would trace the red line. Model calibration is shown by the continuous black line. Triangles denote grouped observations for participants at deciles of predicted risk, with 95% CIs indicated by the vertical black lines. Axes range between 0 and 0.3 since very few individuals received predicted probabilities greater than 0.3. N=785 participants in pooled development sample.

https://doi.org/10.1038/s44220-022-00001-z



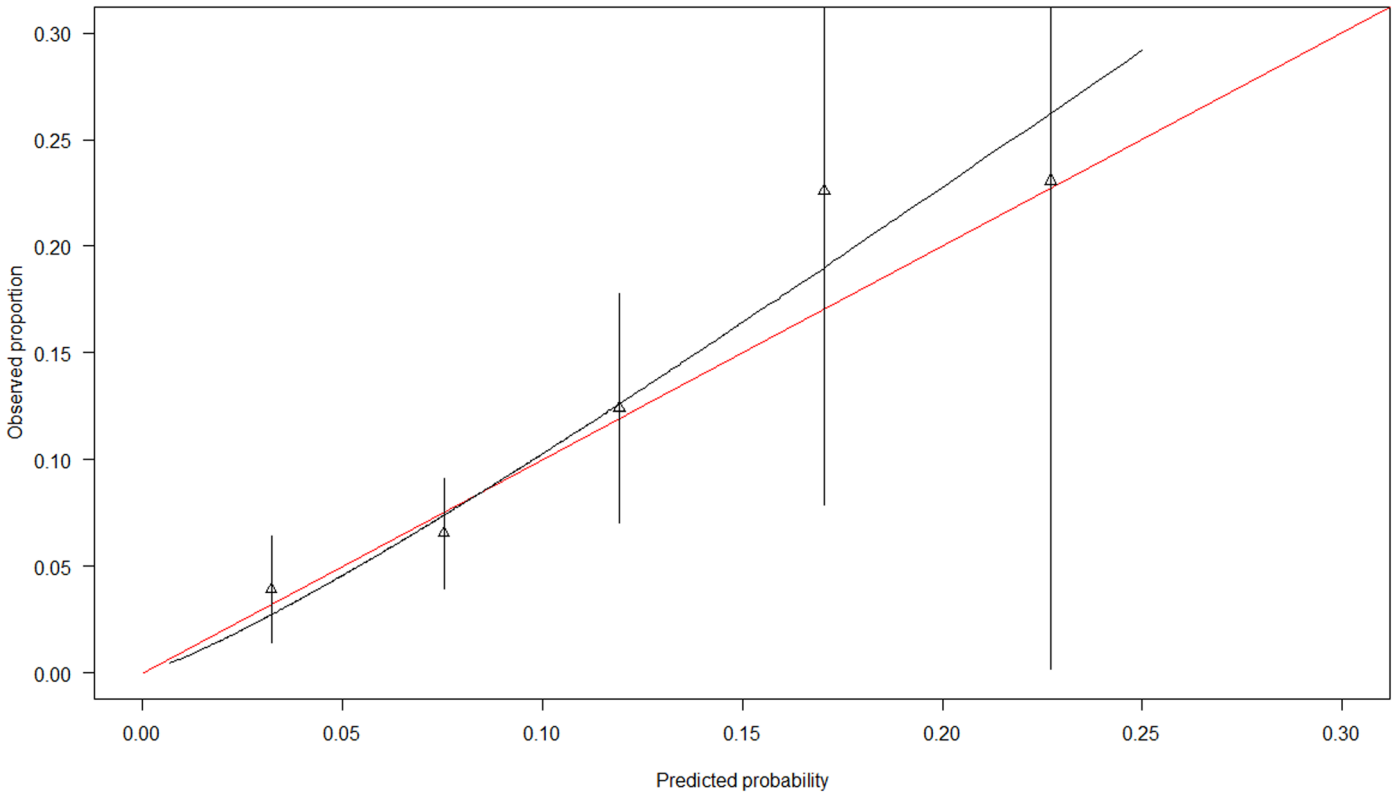
Extended Data Fig. 5 | Internal validation: calibration plot for the LASSO model. *Model calibration* is the extent to which outcomes predicted by the model are similar to those observed in the validation dataset. Calibration plots illustrate agreement between observed proportion (y axis) and predicted risk (x axis). Perfect agreement would trace the red line. Model calibration is shown by the continuous black line. Triangles denote grouped observations for participants at deciles of predicted risk, with 95% CIs indicated by the vertical black lines. Axes range between 0 and 0.3 since very few individuals received predicted probabilities greater than 0.3. N=785 participants in pooled development sample.

https://doi.org/10.1038/s44220-022-00001-z



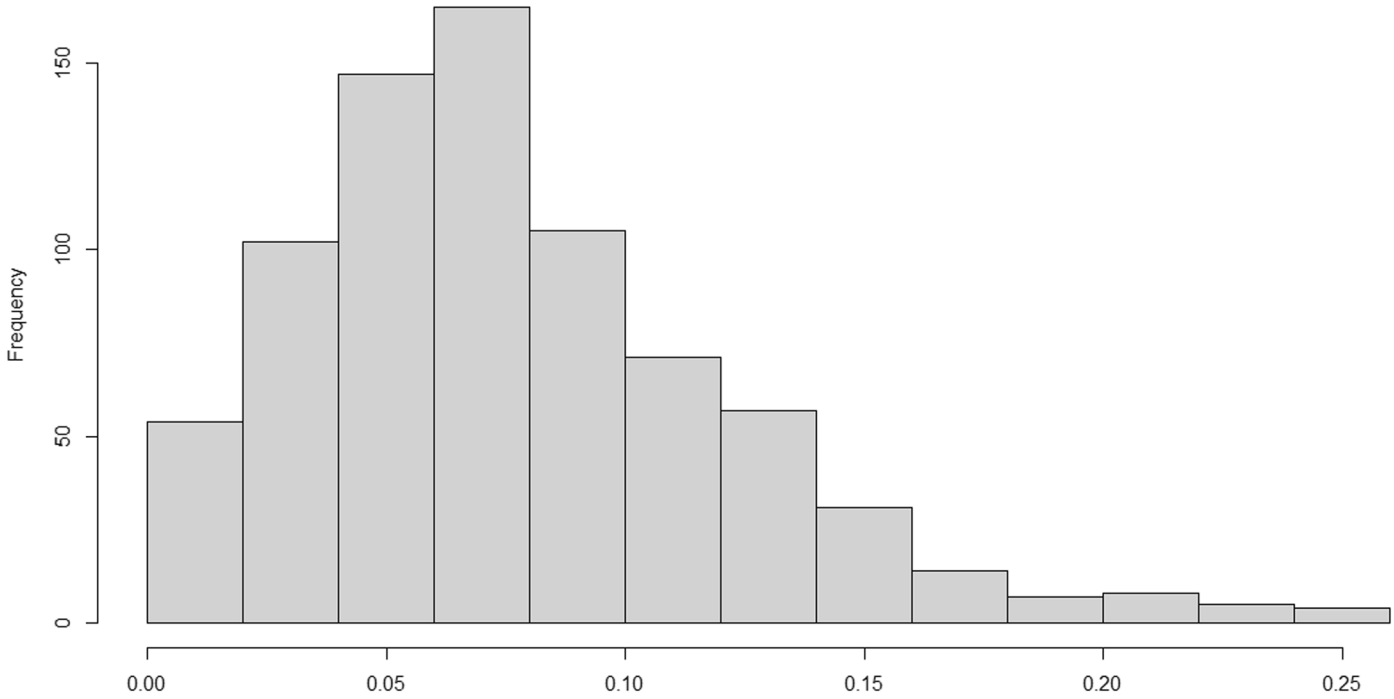
Extended Data Fig. 6 | Internal validation: distribution of predicted probabilities for M3. X axis: predicted probability.

https://doi.org/10.1038/s44220-022-00001-z



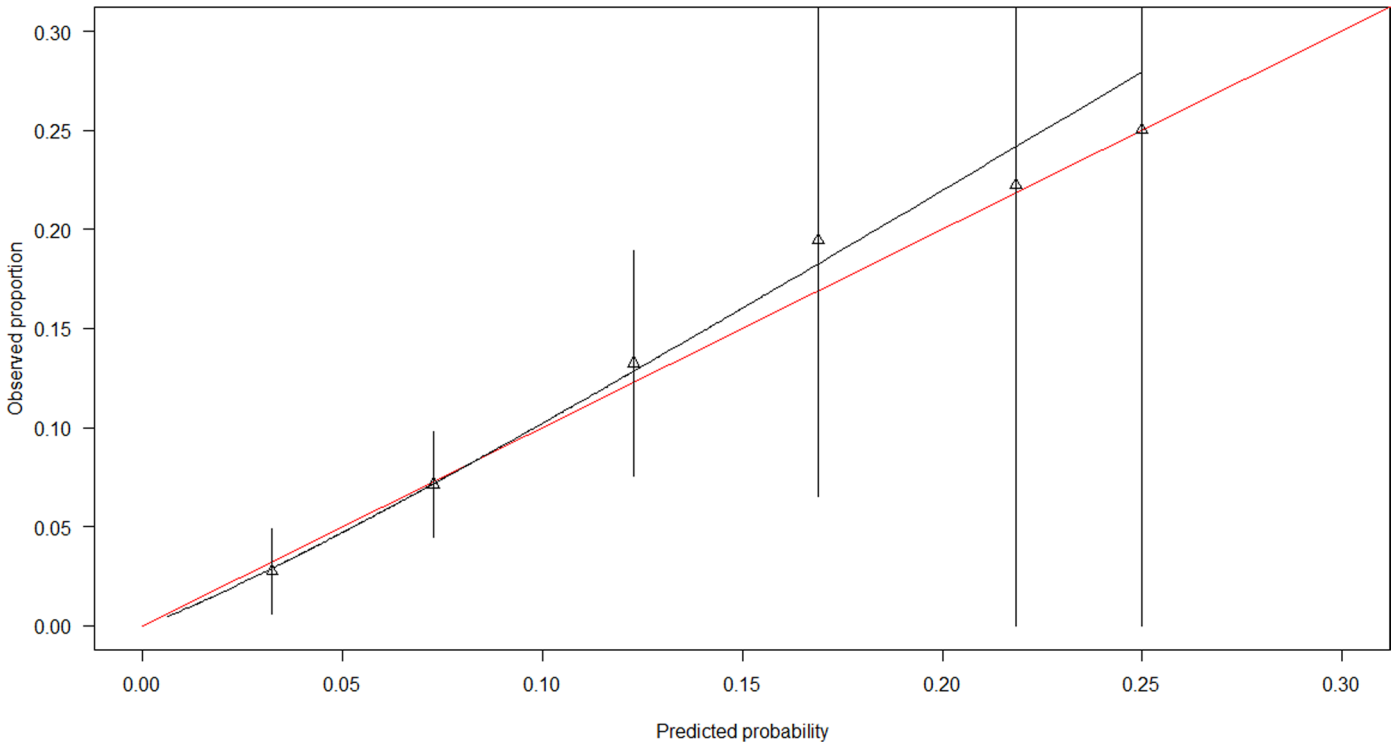
Extended Data Fig. 7 | Internal validation: calibration plot for M3. Calibration plots illustrate agreement between observed proportion (y axis) and predicted risk (x axis). Perfect agreement would trace the red line. Model calibration is shown by the continuous black line. Triangles denote grouped observations for participants at deciles of predicted risk, with 95% CIs indicated by the vertical black lines. Axes range between 0 and 0.3 since very few individuals received predicted probabilities greater than 0.3. N=785 participants in pooled development sample.

https://doi.org/10.1038/s44220-022-00001-z



Extended Data Fig. 8 | Internal validation: distribution of predicted probabilities for M4. X axis: predicted probability.

https://doi.org/10.1038/s44220-022-00001-z



Extended Data Fig. 9 | Internal validation: calibration plot for M4. Calibration plots illustrate agreement between observed proportion (y axis) and predicted risk (x axis). Perfect agreement would trace the red line. Model calibration is shown by the continuous black line. Triangles denote grouped observations for participants at deciles of predicted risk, with 95% CIs indicated by the vertical black lines. Axes range between 0 and 0.3 since very few individuals received predicted probabilities greater than 0.3. N=785 participants in pooled development sample.

nature portfolio

Corresponding author(s): Dr Emanuele F Osimo

Last updated by author(s): Oct 28, 2022

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒ A description of all covariates tested
☐	☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	All data was collected as part of clinical database systems; CRIS for SLAM (validation sample), and CRATE for CAMEO (development sample). CAMEO data were identified by anonymously searching for all EIS patients enrolled between 1st January 2013 and 31st May 2021 using the CPFT Research Database (UK National Health Service [NHS] Research Ethics Service references 12/EE/0407, 17/EE/0442). Anonymised data for all patients enrolled in the Birmingham EIS were collected between 1st January 2014 and 31st December 2018 as part of the National Clinical Audit of Psychosis Quality Improvement Programme, and were enhanced locally with biomarker data; the work conformed to the Health Research Authority definition of service evaluation (confirmed by Birmingham Women's and Children's Hospital NHS Foundation Trust). We used the Clinical Records Interactive Search (CRIS) resource to capture anonymised data from South London and Maudsley NHS Foundation Trust (SLaM) EIS (National Institute for Health Research [NIHR] Biomedical Research Centre [BRC] CRIS Oversight Committee reference 20-005). Data extraction and anonimisation tools, including NLP software in the development sample, are described in Cardinal et al, 2017. Data extraction and anonimisation tools, including NLP software in the external validation sample, are described in Jackson et al, 2016.
Data analysis	All data analyses were conducted in R 4.x. We performed sample size calculations using the R package pmsampsize v 1.1.2. For data imputation we used the MICE package v 3.14. For logistic modelling we used base R and the pROC package, v 1.18. For calibration plots we used the CalibrationCurves package, v 0.1.5. For LASSO model development, we used the package MAMI v 0.9.13. Finally, for coefficient shrinking we used the psfmi package, v 1.0.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

The source data for this work is anonymised patient records from three UK NHS Trusts: CPFT CRATE, SLAM CRIS and the Birmingham EIS. Data from these Trusts are only available to clinicians and clinical researchers with clinical contracts with the Trusts. The data are securely held on clinical systems and available following ethical approval in order to preserve patient confidentiality. Therefore, the raw data cannot be shared.
However, we developed an online data visualisation tool for both the original and recalibrated MOZART models, which allows to interactively explore the effect of each predictor and their combinations on the risk of clozapine use based on the predictors included in this study. See https://eosimo.shinyapps.io/trs_app/

Human research participants

Policy information about studies involving human research participants and Sex and Gender in Research.

Reporting on sex and gender	All participants were classified in clinical systems by sex. All analyses include sex as a co-variate. Gender was not available.
Population characteristics	We considered the following parameters, measured within 100 days of EIS start: sex (female or male); age (years); ethnicity (categorical: white European or not recorded [reference], Black or African- Caribbean, Asian, or other); triglyceride concentration (mmol/L); lymphocyte and neutrophil blood cell counts (billion/L); alkaline phosphatase levels (ALP, units/L), smoking status (binary, at least one cigarette on average daily); body mass index (BMI, kg/m^2); and random glucose levels (mmol/L). See Supplementary Methods for full rationale and details of data extraction.
Recruitment	Anonymised electronic health records from clinical systems were utilised, following Research Ethics approvals. Patients are automatically included in anonymised Trust records if they are treated by a team within the Trust. Enrolment into an EIS fosters confidence in the psychiatric phenotype of included participants, and into the naturalistic nature of the sample including many consecutive referrals with little possibility of selection bias from the sampling frame. Most EISs in the UK NHS, including all three in this analysis, are the only treatment providers for FEP in each geographical area, thus covering a large proportion of all incident cases of first-episode psychosis.
Ethics oversight	CAMEO data were identified by anonymously searching for all EIS patients enrolled between 2013-01-01 and 2021-05-31 using the CPFT Research Database (UK National Health Service [NHS] Research Ethics Service references 12/EE/0407, 17/EE/0442). Anonymised data for all patients enrolled in the Birmingham EIS were collected between 2014-01-01 and 2018-12-31 as part of the National Clinical Audit of Psychosis Quality Improvement Programme, and were enhanced locally with biomarker data; the work conformed to the Health Research Authority definition of service evaluation (confirmed by Birmingham Women's and Children's Hospital NHS Foundation Trust). We used the Clinical Records Interactive Search (CRIS) resource to capture anonymised data from South London and Maudsley NHS Foundation Trust (SLaM) EIS (National Institute for Health Research [NIHR] Biomedical Research Centre [BRC] CRIS Oversight Committee reference 20-005).

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	We performed sample size calculations using the R package pmsampsize, an open source tool that computes the minimum sample size required for the development of a new multivariable prediction model using the criteria proposed by Riley et al. (2018) <doi:10.1002/sim.7992>. pmsampsize calculations aim to minimise the overfitting and to ensure precise estimation of key parameters in the prediction model. In our case, the sample size required was estimated from the estimated outcome prevalence, the a priori estimated R^2 of the model, and the estimated required model shrinkage. For 11 predictors, the minimum sample required was 412. We did not consider non-linear terms or interactions to reduce the risk of overfitting. See Supplementary Methods for detailed sample size calculations.

Data exclusions	We excluded any participant who had missing data on >50% predictor variables, and non-cases (patients who did not use clozapine) who had less than 2 years of follow-up to reduce the probability of including future TRS cases as non-cases. All patients who developed TRS were included regardless of duration of follow-up. As predictors must pre-date outcomes, we also excluded all cases where the outcome start date (clozapine treatment start date, see below) pre-dated the earliest available baseline bloods in the CAMEO cohort (and SLaM cohort, see below), or participants who started taking clozapine within 100 days of baseline in the Birmingham cohort.
Replication	Both models were developed in CAMEO (Cambridge, UK), and externally validated in a separate early intervention service for psychosis based in South London (UK).
Randomization	We used retrospective analysis of all cases meeting criteria in an electronic health records population. No randomization was possible. Risk prediction took into account the following parameters: sex (female or male); age (years); ethnicity (categorical: white European or not recorded [reference], Black or African- Caribbean, Asian, or other); triglyceride concentration (mmol/L); lymphocyte and neutrophil blood cell counts (billion/L); alkaline phosphatase levels (ALP, units/L), smoking status (binary, at least one cigarette on average daily); body mass index (BMI, kg/m²); and random glucose levels (mmol/L). Risk prediction models do not make use of randomisation.
Blinding	As this is a risk prediction model, it requires the modeller to know the outcomes to be able to model them. Therefore, no blinding is possible.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging