Development and validation of multivariable prediction models of remission, recovery, and quality of life outcomes in people with first episode psychosis

Citation for published version:

Digital Object Identifier (DOI):
10.1016/S2589-7500(19)30121-9

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Publisher's PDF, also known as Version of record

Published in:
The Lancet Digital Health

General rights
Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact openaccess@ed.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.
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

Samuel P Leighton, Rachel Upthegrove, Rajeev Krishnadas, Michael E Benros, Matthew R Broome, Georgios V Gkoutos, Peter F Liddle, Swaran P Singh, Linda Everard, Peter B Jones, David Fowler, Vimal Sharma, Nicholas Freemantle, Rune H B Christensen, Nikolai Albert, Merete Nordentoft, Matthias Schwannauer, Jonathan Cavanagh, Andrew I Gumley, Max Birchwood*, Pavan K Mallikarjun*

Summary
Background Outcomes for people with first-episode psychosis are highly heterogeneous. Few reliable validated methods are available to predict the outcome for individual patients in the first clinical contact. In this study, we aimed to build multivariable prediction models of 1-year remission and recovery outcomes using baseline clinical variables in people with first-episode psychosis.

Methods In this machine learning approach, we applied supervised machine learning, using regularised regression and nested leave-one-site-out cross-validation, to baseline clinical data from the English Evaluating the Development and Impact of Early Intervention Services (EDEN) study (n=1027), to develop and internally validate prediction models at 1-year follow-up. We assessed four binary outcomes that were recorded at 1 year: symptom remission, social recovery, vocational recovery, and quality of life (QoL). We externally validated the prediction models by selecting from the top predictor variables identified in the internal validation models the variables shared with the external validation datasets comprised of two Scottish longitudinal cohort studies (n=162) and the OPUS trial, a randomised controlled trial of specialised assertive intervention versus standard treatment (n=578).

Findings The performance of prediction models was robust for the four 1-year outcomes of symptom remission (area under the receiver operating characteristic curve [AUC] 0·703, 95% CI 0·664–0·742), social recovery (0·731, 0·697–0·765), vocational recovery (0·736, 0·702–0·771), and QoL (0·704, 0·667–0·742; p<0·0001 for all outcomes), on internal validation. We externally validated the outcomes of symptom remission (AUC 0·680, 95% CI 0·587–0·773), vocational recovery (0·867, 0·805–0·930), and QoL (0·679, 0·522–0·836) in the Scottish datasets, and symptom remission (0·616, 0·553–0·679), social recovery (0·573, 0·504–0·643), vocational recovery (0·660, 0·610–0·710), and QoL (0·556, 0·481–0·631) in the OPUS dataset.

Interpretation In our machine learning analysis, we showed that prediction models can reliably and prospectively identify poor remission and recovery outcomes at 1 year for patients with first-episode psychosis using baseline clinical variables at first clinical contact.

Funding Lundbeck Foundation.

Copyright © 2019 The Author(s). Published by Elsevier Ltd. This is an Open Access article under the CC BY 4.0 license.

Introduction Psychosis is an illness with an early first onset, occurring usually in young people and with an incidence of 31 per 100,000 person-years.1 Patients with first-episode psychosis have heterogeneity of outcomes, with a 58% prevalence of remission and 38% of recovery.1 The identification of individual-patient outcomes at initial clinical contact might help to personalise treatment and lead to improved use of resources for those most in need or likely to respond to treatment.2 However, few validated tools are available for the accurate early identification of patients with good or poor outcomes.

Previous observational studies have identified predictors of outcomes at the group level, including sociodemographic factors, clinical and treatment response variables, comorbidity, and functional and cognitive deficits,3,4,5 with inconsistent reliability.6 More clarity is needed on how to apply group-level factors to an individual level of prediction. An approach that can be applied to stratify the individualised risk of a poor outcome at the initial clinical contact is required. One solution is the use of machine learning, in which algorithms can sift through a large array of predictor variables and detect complex high dimensional interactions that can reliably predict individual-patient outcomes.7

Lancet Digital Health 2019; 1: e261-70
Published Online September 12, 2019
http://dx.doi.org/10.1016/S2589-7500(19)30121-9
This online publication has been corrected. The corrected version first appeared at thelancet.com/digital-health on September 30, 2019
See Comment page e244
Joint senior authors
Institute of Health and Wellbeing (S P Leighton MBChB, Prof J Cavanagh MD, Prof A I Gumley PhD) and Institute of Neuroscience and Psychology (R Krishnadas PhD), University of Glasgow, Glasgow, UK; Institute for Mental Health (Prof R Upthegrove PhD, Prof M R Broome PhD, P K Mallikarjun PhD), Institute of Cancer and Genomics (Prof G V Gkoutos PhD), and Institute of Translational Medicine (Prof G V Gkoutos), University of Birmingham, Birmingham, UK; Copenhagen Research Center for Mental Health, Mental Health Centre Copenhagen, Copenhagen University Hospital, Copenhagen, Denmark (Prof M Nordentoft PhD, R H B Christensen PhD, N Albert PhD, Prof M Nordentoft PhD); Health Data Research UK Midlands, UK (Prof G V Gkoutos); University Hospitals Birmingham NHS Foundation Trust, Birmingham, UK (Prof G V Gkoutos); Institute of Mental Health, University of Nottingham, Nottingham, UK (Prof P F Liddle PhD); Mental Health and Wellbeing, Warwick Medical School, University of Warwick, Coventry, UK (Prof S P Singh MD),

www.thelancet.com/digital-health Vol 1 October 2019 e261

*Joint senior authors
Two models developed for outcome prediction in psychosis using baseline variables have been published.9,10 Koutsouleris and colleagues9 used machine learning to predict 4-week and 52-week functional outcomes in patients with first-episode psychosis to a 75·0% (for 4 weeks) and 73·8% (for 52 weeks) test-fold balanced accuracy (ie, average accuracy across the ten folds) on repeated nested internal cross-validation, with use of data from a randomised control study (n=334); however, this model was not externally validated. Leighton and colleagues1 developed 1-year remission and vocational recovery prediction models on 83 patients with first-episode psychosis and externally validated their models on 79 patients with the same condition; however, this study was limited by the small sample size.

To overcome the two major limitations of these previous studies (no external validation and small sample size), we aimed to apply a machine learning approach using one of the largest longitudinal cohort studies of patients with first-episode psychosis (n=1027), for model development and internal validation, and data from a large randomised control trial (n=578) and two longitudinal cohort studies (totalling 162 patients), for external validation. We developed prediction models for multiple outcomes, including symptom remission and functional recovery (social recovery, vocational recovery, and quality of life) at 1 year after first-episode psychosis.

Methods

Study design and sources of data

In this machine learning approach, we used data from several sources: the National EDEN studies,9,10 two Scottish validation datasets,11 and the OPUS trial.12 The National EDEN studies are a longitudinal naturalistic study of 1027 patients with first-episode psychosis recruited from 14 early intervention services across the National Health Service (NHS) in England (2005–10); the methods and baseline characteristics have been outlined previously.13 The Scottish validation datasets were two longitudinal cohort studies of patients with first-episode psychosis: the Compassionate Recovery: Individualised Support in early Psychosis (CR:ISP) study9 of 83 patients in NHS Greater Glasgow & Clyde (2011–14), and an earlier study10 of 79 patients in NHS Glasgow and NHS Edinburgh (2006–09). The methodologies and baseline characteristics of these studies have been outlined previously.13,14 The

Research in context

Evidence before this study

In patients with first-episode psychosis, prediction of remission and recovery outcomes is an important goal during initial clinical contact. These patients have heterogeneous outcomes, even with standardised interventions. Targeting extended or more intensive treatment to patients with poorer prognosis might lead to better outcomes. Previous studies have identified several group-level predictors, including poor premorbid adjustment, history of developmental disorder, symptom severity at baseline, and duration of untreated psychosis, as predictors of poor clinical, functional, and cognitive outcomes. Such group-level differences are not always replicated at the individual level, and how to combine the group-level factors for individualised prediction is unclear. We searched PubMed from inception to March 12, 2019, using the terms “psychosis” or “first episode psychosis” or “schizophrenia” AND “prediction” AND “outcome” in any field, with no language restrictions. We retrieved 470 articles, of which, after excluding articles not related to multivariable prediction of outcomes based on baseline clinical variables, we identified two articles that have published models for outcome prediction in psychosis using baseline variables. One study had developed an internally cross-validated model for prediction of functional outcomes in a large sample, but this model was not externally validated on an independent sample. The other study developed remission and recovery prediction models on a small sample from a longitudinal cohort study, and externally validated the models on patients from a different cohort study. Additionally, examples exist of outcome prediction models that have been internally cross-validated and externally validated for depression.

Added value of this study

To our knowledge, our study provides the first reliable evidence for the usefulness of machine learning to develop outcome prediction models, using baseline variables at first clinical contact, in a large sample of patients with first-episode psychosis. The models use baseline clinical and demographic data, rather than neuroimaging or other biomarkers, and, as such, are more accessible in a clinical setting for potential future applications. Our results were validated by the methods that we used, including internal–external validation of the outcome prediction models developed on data from a large multicentre cohort study and external validation on a small cohort study of patients with first-episode psychosis and a large randomised controlled trial of patients with first-episode psychosis. Our study attempted to develop outcome prediction models for multiple outcomes (clinical, recovery, and quality of life), although a single model might be useful to predict multiple outcomes, albeit with reduced accuracy.

Implications of all the available evidence

Our study, and the two previous studies, showed that machine learning techniques applied to baseline clinical and demographic data can aid in the prediction of remission and recovery outcomes for patients with first-episode psychosis at first clinical contact. This approach can be extended to include other sources of data (neuroimaging data, immune biomarkers, and so on), which might enhance model performance. The next step before implementation into routine clinical practice would be to investigate the usefulness of the prediction models in prospective controlled trials.
OPUS trial (NCT00157313) was a randomised controlled trial of 578 patients with first-episode psychosis recruited from all inpatient and outpatient mental health services in Copenhagen (Copenhagen Hospital Corporation) and Aarhus County, Denmark. OPUS assessed standard (n=272) versus specialised assertive intervention integrated treatment (n=275; January, 1998, to December, 2000). The methods and baseline characteristics of OPUS have been outlined previously.18 Local ethics committees approved the studies and the trial. We have adhered to the transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD) statement.19

With data from the EDEN studies, we developed four predictive models for each of the outcomes assessed in our study and internally validated the models by nested leave-one-site-out cross-validation (LOSOCV). Subsequently, we identified shared variables (from the top predictor variables from the internally validated models) between the EDEN studies and the Scottish datasets and between the EDEN studies and OPUS trial. We used these shared variables to build separate prediction models for external validation on the Scottish datasets and OPUS trial.

The key differences between EDEN (development) and Scottish (validation) datasets were the setting (EDEN was done in NHS England, whereas the Scottish datasets were from studies done in NHS Scotland) and study period (2005–10 in EDEN, 2011–14 and 2006–09 in the Scottish datasets). Both NHS England and NHS Scotland are free at the point of delivery. The key differences between EDEN and the other validation dataset (OPUS) were the setting (England vs Denmark, but both free at the point of delivery), study period (2005–10 vs January, 1998 to December, 2000 in OPUS), study type (naturalistic in EDEN, for which everyone received early intervention, vs randomised clinical trial in OPUS, for which early intervention was compared with treatment as usual), and inclusion criteria (participants were aged 14–35 years with a presentation of psychotic symptoms in EDEN, whereas in OPUS, participants were aged 18–45 years with a diagnosis in the schizophrenia spectrum according to the International Classification of Diseases tenth edition codes in the F2 category, and participants in OPUS had not been given antipsychotic drugs for more than 12 weeks of continuous treatment). The inclusion and exclusion criteria for all three studies have been provided in the appendix (p 1).

Outcome variables
For EDEN, Scottish, and OPUS studies, assessments of predictors and outcomes were done by research assistants not directly involved in clinical care. We assessed four binary outcomes that were recorded at 1 year: symptom remission, meeting the Positive and Negative Syndrome Scale in Schizophrenia (PANSS) criteria at both 6 months and 1 year;26 social recovery, achieving a Global Assessment of Functioning (GAF) score (range 0–100) of 65 or higher in EDEN, and a mean GAF symptoms and GAF disability score of 65 or higher in OPUS;27 vocational recovery, assessing whether participants were in employment, education, or training;28 and quality of life (QoL), assessed with the 3-level European QoL 5 Dimensions Index (EQ-5D-3L) time trade-off index based on UK population norms and dichotomised to greater than median (0·848)29 in EDEN, the WHO QoL 26-item instrument with total score dichotomised to greater than median (88) in the 2006–09 Scottish study, and the Lancashire QoL score30 dichotomised at the median (43·5) in OPUS. We chose operationalised criteria for symptom remission26 and included the three outcome measures for recovery to cover a broader patient-centred experience of recovery. Social recovery was not measured in the Scottish studies.

Statistical analysis
The EDEN study was powered for duration of untreated psychosis. OPUS was powered for positive symptoms

(See Online for appendix)
Vocational recovery
Predictor (direction of effect)
<table>
<thead>
<tr>
<th>Predictor</th>
<th>Valid in OPUS, OR Scotinfo</th>
<th>Valid in EDEN, OR Scotinfo</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unemployment, education, or training at baseline</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>EIS=Early Intervention Service</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>LOSOCV=leave-one-site-out cross-validation</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Main income source is something other than family, salary, or benefits</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>In receipt of any state benefits</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Main income source is salary or wage</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Length of time since most recent self-harm</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>In education at baseline</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Male income source is state benefits</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>

Quality of life
Predictor (direction of effect)
<table>
<thead>
<tr>
<th>Predictor</th>
<th>Valid in OPUS, OR Scotinfo</th>
<th>Valid in EDEN, OR Scotinfo</th>
</tr>
</thead>
<tbody>
<tr>
<td>EQ-5D-3L=3-level European Quality of Life 5 Dimensions Index</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>EQ-5D-3L mobility</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>EQ-5D-3L health thermometer</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>

Figure 2: Top prediction variables for each outcome
Top predictor variables selected by elastic net regularisation across all 14 LOSOCV models for each outcome, ordered by their mean rank across the 14 models by absolute coefficient magnitude, along with their direction of effect (red is negative, grey is positive). PANSS=Positive And Negative Symptom Scale. GAF=Global Assessment of Functioning scale. DUP=Duration of Untreated Psychosis. PAS=Premorbid Adjustment Scale. LSD=Lysergic acid diethylamide. EQ-5D-3L=3-level European Quality of Life 5 Dimensions Index. UK TTO=time trade-off index based on UK population norms. EIS=Early Intervention Service. LOSOCV=leave-one-site-out cross-validation.

demographic, and clinical predictor variables were centred and scaled, variables with zero variance and near-zero variance were removed, and variables with more than 20% of missing data were excluded. For the remaining 163 (61%) of predictor variables (appendix pp 2–6), missing data were imputed by use of k-nearest neighbour imputation (k=5) to increase prediction performance.20 We did not complete any a-priori hypothesis-based feature selection.

We used the EDEN dataset for model development and undertook both internal and external validation with LOSOCV.20 We fit a logistic regression model by elastic net regularisation with variable selection in the caret package22 using the glmnet package.21 Glmnet fits a generalised linear model through penalised maximum likelihood (appendix pp 1–2). All the 163 predictor variables were used simultaneously with the elastic net regularisation model. Each of the 14 EDEN sites was left out once for the validation of a model based on the remaining 13 sites and trained by use of a ten-fold cross-validation (splits balanced by outcome class) over a 10×10 grid of α and λ hyperparameters, with Breiman’s 1 SE rule.24 We measured average performance across the resulting 14 best LOSOCV models using receiver operating characteristic (ROC) curve and area under the curve (AUC). AUCs, with 95% CIs, were established on the basis of U-statistic theory, and permutation testing confirmed significance. Representative model accuracy, sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), prognostic summary index (PSI), positive likelihood ratio (LR+) and negative likelihood ratio (LR−) are presented on the basis of the point on the ROC curve corresponding to Youden’s index. We assessed the stability (φ) of feature selection in the 14 best LOSOCV models using the approach described by Nogueira and colleagues,22 where φ lower than 0.4 shows poor agreement between the 14 models, 0.4 to 0.75 shows intermediate to good agreement, and higher than 0.75 shows excellent agreement. We did this model development procedure for each of our four binary outcomes.

We assessed the relatedness of the four models by computing the Yule φ correlation between the four outcomes, computing the Pearson correlation between probability outputs of the four logistic regression models, and assessing the prediction performance when using the probability outputs of one model as predictors of outcome for the other three models with LOSOCV (appendix pp 6–7). We used the shared predictor variables among the top variables for the four models to build generalised linear models for external validation.

For external (geographical and temporal) validation of the prediction models, we used the Scottish and OPUS datasets. For each outcome, we took the shared variables across both the EDEN and the external validation dataset from the top predictor variables determined during model development (those selected in all 14 LOSOCV models; figure 1). We standardised these variables according to the Scale for Assessment of Positive Symptoms (SAPS). The 2006–09 Scottish study was powered for the strength of association between duration of untreated psychosis and psychiatric symptomatology. The 2011–14 Scottish study was powered for positive and negative symptoms. Because our study is a post-hoc analysis, a sample size calculation is not applicable.

Studies with missing outcome data were removed from the analysis. Regarding predictor selection, during data pre-processing in EDEN, all 266 baseline social, psychiatric, and clinical variables were considered, and variables with zero variance and near-zero variance were removed, and variables with more than 20% of missing data were excluded. For the remaining 163 (61%) of predictor variables (appendix pp 2–6), missing data were imputed by use of k-nearest neighbour imputation (k=5) to increase prediction performance.20 We did not complete any a-priori hypothesis-based feature selection.

We used the EDEN dataset for model development and undertook both internal and external validation with LOSOCV.20 We fit a logistic regression model by elastic net regularisation with variable selection in the caret package22 using the glmnet package.21 Glmnet fits a generalised linear model through penalised maximum likelihood (appendix pp 1–2). All the 163 predictor variables were used simultaneously with the elastic net regularisation model. Each of the 14 EDEN sites was left out once for the validation of a model based on the remaining 13 sites and trained by use of a ten-fold cross-validation (splits balanced by outcome class) over a 10×10 grid of α and λ hyperparameters, with Breiman’s 1 SE rule.24 We measured average performance across the resulting 14 best LOSOCV models using receiver operating characteristic (ROC) curve and area under the curve (AUC). AUCs, with 95% CIs, were established on the basis of U-statistic theory, and permutation testing confirmed significance. Representative model accuracy, sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), prognostic summary index (PSI), positive likelihood ratio (LR+) and negative likelihood ratio (LR−) are presented on the basis of the point on the ROC curve corresponding to Youden’s index. We assessed the stability (φ) of feature selection in the 14 best LOSOCV models using the approach described by Nogueira and colleagues,22 where φ lower than 0.4 shows poor agreement between the 14 models, 0.4 to 0.75 shows intermediate to good agreement, and higher than 0.75 shows excellent agreement. We did this model development procedure for each of our four binary outcomes.

We assessed the relatedness of the four models by computing the Yule φ correlation between the four outcomes, computing the Pearson correlation between probability outputs of the four logistic regression models, and assessing the prediction performance when using the probability outputs of one model as predictors of outcome for the other three models with LOSOCV (appendix pp 6–7). We used the shared predictor variables among the top variables for the four models to build generalised linear models for external validation.

For external (geographical and temporal) validation of the prediction models, we used the Scottish and OPUS datasets. For each outcome, we took the shared variables across both the EDEN and the external validation dataset from the top predictor variables determined during model development (those selected in all 14 LOSOCV models; figure 1). We standardised these variables

Figure 2: Top prediction variables for each outcome
Top predictor variables selected by elastic net regularisation across all 14 LOSOCV models for each outcome, ordered by their mean rank across the 14 models by absolute coefficient magnitude, along with their direction of effect (red is negative, grey is positive). PANSS=Positive And Negative Symptom Scale. GAF=Global Assessment of Functioning scale. DUP=Duration of Untreated Psychosis. PAS=Premorbid Adjustment Scale. LSD=Lysergic acid diethylamide. EQ-5D-3L=3-level European Quality of Life 5 Dimensions Index. UK TTO=time trade-off index based on UK population norms. EIS=Early Intervention Service. LOSOCV=leave-one-site-out cross-validation.
separately on each dataset before model fitting; therefore, we were able to assess EDEN model performance on the validation dataset even though some shared variables were measured on different scales. Afterwards, we used the entire EDEN dataset to fit a generalised linear model by maximum likelihood estimation (without regularisation) using these shared top predictor variables (having found no improvement in performance during initial scoping with more complex classifiers, including linear and radial support vector machines, elastic net, and random forest). We confirmed that the internal–external validation performance on the EDEN dataset remained robust with the new model using only the shared top predictor variables. The internally–externally validated EDEN model was then externally validated on the external dataset, with performance reported as already outlined. This process was repeated separately for the Scottish datasets and the OPUS dataset (figure 2).

All statistical analyses were done with R, and the code is available online. The comparison between EDEN, Scottish, and OPUS samples (demographic and social variables) is provided in the appendix (p 2).

Role of the funding source
The funder of the study had no role in study design, data collection, data analysis, data interpretation, or writing of the report. The corresponding author had full access to all the data in the study and had final responsibility for the decision to submit for publication.

Results
We included only participants for whom outcome data were available (table 1). In the EDEN studies, 673 (66%) of 1027 patients had complete symptom remission outcome data, 829 (81%) had complete social recovery outcome data, 807 (79%) had complete vocational recovery outcome data, and 729 (71%) had complete QoL outcome data. 15–39% of patients were missing outcomes data on model performance at 1 year for training cohorts.

Table 1: Outcome data for training and validation cohorts

<table>
<thead>
<tr>
<th></th>
<th>Training data (EDEN studies)</th>
<th>Validation data (Scottish studies)</th>
<th>Validation data (OPUS trial)</th>
<th>p value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Symptom remission</td>
<td>320/673 (48%)</td>
<td>66/131 (50%)</td>
<td>121/338 (36%)</td>
<td>0.0006*</td>
</tr>
<tr>
<td>Social recovery</td>
<td>388/829 (47%)</td>
<td>NA</td>
<td>73/518 (14%)</td>
<td>&lt;0.0001*</td>
</tr>
<tr>
<td>Vocational recovery</td>
<td>436/807 (54%)</td>
<td>59/142 (42%)</td>
<td>173/553 (31%)</td>
<td>&lt;0.0001*</td>
</tr>
<tr>
<td>Quality of life</td>
<td>328/729 (45%)</td>
<td>23/47 (49%)</td>
<td>113/226 (50%)</td>
<td>0.39</td>
</tr>
</tbody>
</table>

Data are n/N (%). NA=not applicable. *Significant differences (determined with Pearson’s χ² test) of rates of positive outcomes between the cohorts, after Holm-Bonferroni correction.

For the R code see https://github.com/sambleighton87/EDEN_R_Code

Figure 2: Analysis pipeline
Elastic net model development and internal–external validation using a leave-one-site-out cross-validation in the EDEN sample. Internally–externally validated generalised linear models were constructed with use of top predictors shared between the EDEN and Scottish datasets, and the EDEN and OPUS datasets. These were then externally validated on the Scottish datasets and the OPUS dataset.
and for validation cohorts these values were 4–61% for OPUS and 19–61% for Scottish studies.

During internal cross-validation with all the 163 predictor variables simultaneously, all of our four models had AUCs higher than 0·700, significantly better than chance (p<0·0001; figure 3, table 2). The accuracy achieved by the four models was higher than AUC 0·65, and the PSI of the four models was higher than 0·31, indicating a 31% additional gain in prediction certainty. The stability of feature selection in the 14 LOSOCV models was 0·54 for the remission model, 0·67 for the social recovery model, 0·71 for the vocational recovery model, and 0·70 for the QoL model (appendix p 7).

The correlation of the probability outputs of the four models was higher than the correlation of the respective outcomes that they were trained to predict. Each model predicted its outcome best, but they also significantly predicted each of the other three outcomes, with a lower level of performance (appendix pp 6–7).

The top predictors for the four models selected by the elastic net model are provided in figure 1. The four models included predictor variables ranging from demographic

---

**Figure 3:** ROC curves showing internal–external LOSOCV model performance in the EDEN dataset for 1-year symptom recovery (A), social recovery (B), vocational recovery (C), and quality of life (D) models. ROC=receiver operating characteristic. LOSOCV=leave-one-site-out cross-validation. AUC=area under the curve.
In this study, we developed outcome prediction models for remission and recovery for people with first episode psychosis using baseline sociodemographic and clinical variables and internally cross-validated the models with external validation metrics for year 1 for binary symptomatic, social recovery, vocational recovery and quality of life outcomes. The models had good predictive performance, with AUCs ranging from 0.67 to 0.77. The models performed better than chance, with AUCs higher than 0.67 for symptom remission, 0.57 for social recovery, and 0.66 for vocational recovery than for the other outcomes, whereas in the standard treatment group for remission and social recovery, the models had AUCs of 0.54 and 0.57, respectively. The predictive performance of the models was significantly better than that of Koutsouleris and colleagues' model (0.66–0.72; p<0.0001). The PSIs indicated that our models had better performance than that of Koutsouleris and colleagues' model (0.65–0.70), whereas their model had a predictive performance of 0.57–0.67. The PSI indicated that our models had better performance than that of Koutsouleris and colleagues' model (0.66–0.72; p<0.0001). The PSIs indicated that our models had better performance than that of Koutsouleris and colleagues' model (0.66–0.72; p<0.0001).

### Table 2: Performance metrics for internal–external and external validation

<table>
<thead>
<tr>
<th>Symptom remission</th>
<th>Social recovery</th>
<th>Vocational recovery</th>
<th>Quality of life</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Internal-external validation (EDEN)</strong></td>
<td><strong>External validation (OPUS)</strong></td>
<td><strong>Internal-external validation (EDEN)</strong></td>
<td><strong>External validation (OPUS)</strong></td>
</tr>
<tr>
<td>Performance</td>
<td>0.703 (0.664–0.742; p&lt;0.0001)</td>
<td>0.680 (0.587–0.773; p&lt;0.0001)</td>
<td>0.616 (0.533–0.692; p&lt;0.0001)</td>
</tr>
<tr>
<td>Accuracy</td>
<td>0.670 (0.636–0.703)</td>
<td>0.685 (0.618–0.771)</td>
<td>0.554 (0.493–0.617)</td>
</tr>
<tr>
<td>Sensitivity</td>
<td>0.842 (0.808–0.878)</td>
<td>0.621 (0.512–0.732)</td>
<td>0.415 (0.306–0.526)</td>
</tr>
<tr>
<td>Specificity</td>
<td>0.751 (0.654–0.872)</td>
<td>0.769 (0.615–0.908)</td>
<td>0.796 (0.678–0.923)</td>
</tr>
<tr>
<td>PPV</td>
<td>0.679 (0.601–0.739)</td>
<td>0.729 (0.638–0.835)</td>
<td>0.462 (0.312–0.636)</td>
</tr>
<tr>
<td>NPV</td>
<td>0.667 (0.613–0.724)</td>
<td>0.667 (0.593–0.739)</td>
<td>0.472 (0.418–0.527)</td>
</tr>
<tr>
<td>PSI</td>
<td>0.346 (0.231–0.473)</td>
<td>0.396 (0.252–0.613)</td>
<td>0.217 (0.109–0.400)</td>
</tr>
<tr>
<td>LR-</td>
<td>0.554 (0.288–0.936)</td>
<td>0.493 (0.250–0.896)</td>
<td>0.617 (0.377–1.083)</td>
</tr>
</tbody>
</table>

Data are AUC (95% CI) or AUC (95% CI; p value). Leave-one-site-out internal–externally validated performance and external validation metrics for year 1 for binary symptomatic, social recovery, vocational recovery and internal-external validated performance for quality of life outcome metrics. 95% CIs of the AUC were established on the basis of U-statistic theory and significance level confirmed by permutation testing (n=10001). Representative accuracy, sensitivity, specificity, PPV, NPV, PSI, LR+, and LR- with bootstrap 95% CIs (n=9999) are based on the point on the receiver operating characteristic curve corresponding to Youden’s index. Different thresholds can be taken on the basis of the requirements of the diagnostic test. AUC area under the curve. PPV-positive predictive value. NPV-negative predictive value. PSI-prognostic summary index. LR+-positive likelihood ratio. LR--negative likelihood ratio.
Articles within each of the four prediction models.25 indicates an intermediate to good strength of agreement (vocational recovery model, and 0·70 for the QoL model) in the 14 LOSOCV models (0·54 for the remission model, 0·67 for the social recovery model, 0·71 for the QoL model) suggests an intermediate to good strength of agreement within each of the four prediction models.25

The external validation performance of the prediction models was similar to that of the training dataset for the Scottish datasets, although the performance was reduced in the OPUS dataset. Several possible explanations exist for this difference. The external validation models were necessarily built with use of shared variables alone, not with all the top identified predictor variables. However, a repeat internal validated LOSOCV performance with models using just the shared variables remained similar in the EDEN dataset. The way outcomes (and some predictors) were measured differed between the datasets: remission was defined with Andreasen criteria, but with use of PANSS for EDEN and Scottish datasets and SAPS–SANS (Scale for the Assessment of Negative Symptoms) for the OPUS dataset; social functioning was defined by GAF for EDEN, but by use of the mean of GAF symptoms and GAF disability for OPUS; and for QoL, EDEN used EQ-5D-3L, but the other three datasets used WHO QoL. The measurement of vocational recovery outcomes was similar across the datasets, and the fact that this model performed best in external validation could reflect this. Furthermore, we found significant differences in the balance of remission and recovery outcome rates between all datasets, although the OPUS dataset had much fewer remission and recovery outcomes than those of the other datasets. This finding might be explained by the differing timeframes of data collection and the fact that the EDEN and Scottish datasets were collected from patients in early-intervention services, whereas OPUS was a randomised controlled trial of intensive versus standard treatment. Contrary to our expectation, the validation performance was better for the remission and social recovery outcomes for the standard treatment group of the OPUS trial. The validation performance for the vocational recovery model was better for the intensive treatment group, which is similar to the performance in the training dataset. Taken together, these issues are unavoidable in the context of our analyses being opportunistic and post-hoc, with use of existing datasets. However, the fact that model performance was significantly better than chance on external validation (except for QoL in OPUS), despite these differences, is very promising for the ability of such methods to withstand heterogeneous data in real-world clinical settings.

Our analysis has several strengths. The data for the model development were derived from one of the largest naturalistic cohort studies in patients with first-episode psychosis treated in early-intervention services. We used LOSOCV for model development and internal validation. We found the stability of the feature selection with LOSOCV for 14 sites to have intermediate to good level of agreement. Furthermore, we externally validated the models in three independent datasets with different time periods, geographical regions, and recording methods. We used strict operationalised outcome criteria to define symptomatic outcomes and developed prediction models for multiple outcomes. Each of the individual prediction models predicted the other three outcomes better than chance, although with reduced performance. An argument exists for using one prediction model to predict multiple outcomes, although this would come with a trade-off of marginally reduced performance and needs further testing in prospective clinical trials.

Our study also has several limitations. About 49% of eligible patients consented to participate in the EDEN study, which might affect the generalisability of our prediction models to all patients with first-episode psychosis. However, participants who did not consent had characteristics largely similar at baseline to those of individuals who consented to participate.30 Despite this, we cannot assume that the models developed with data from the patients included in the EDEN study would have a better performance than chance in individuals not included in the EDEN study sample. The effect of missing outcomes data on model performance was not trivial in patients at 1 year for training cohorts (15–39%) and for validation cohorts (OPUS 4–61%; Scottish studies 19–61%). This effect might introduce bias and affect the generalisability of our results. Importantly, our models have not been validated for prediction after baseline as treatment progresses. Future studies could consider building models that account for change over time or in response to treatment (eg, dynamic Bayesian networks with continuous retraining). We did not collect cognitive and physical biomarkers of illness, including blood samples and neuroimaging, which previous studies have highlighted as potentially important for generating accurate predictions.32 The duration criteria for recovery has been proposed to be at least 2 years.33 However, the criteria used in our analysis for recovery outcomes were much narrower and, for three of four measures, were based on point outcomes (GAF; employment, education, or training status; and QoL) at 1 year. The prevalence of recovery in our training cohort was similar to that reported in a large meta-analysis,3 but higher than that reported in another meta-analysis,3 which might also affect the generalisability of our model.

The decision making process to determine which interventions to use and for how long in the treatment of patients with a first-episode psychosis is based on clinical intuition. We are not aware of evidence assessing how accurate clinicians using baseline information are at predicting 1-year outcomes for psychosis, although it has been shown that clinicians are poor at predicting outcomes in depression.34 Clinicians working with patients with first-episode psychosis might benefit from a reliable and methodologically robust tool to identify individuals with likelihood of a good or poor outcome at initial clinical contact, so that the information on

---

outcome prediction can be used alongside clinical judgment for stratification of treatment.

Patients with good outcomes are likely to need a different set of interventions and duration of treatment compared with patients with poorer outcomes. If outcome prediction models are developed into clinically applicable tools after further rigorous testing of their usefulness in a prospective clinical trial, they could assist in clinical decision making, leading to better use of clinical resources by providing targeted interventions based on individual predictions of patient outcomes. Guidelines could be developed in consultation with stakeholders on how to put such tools into practice to facilitate a stepped model of care. Future work should identify, in a prospective clinical trial, whether one prediction model might accurately predict multiple outcomes and whether it is possible to update the prediction models prospectively over time and in response to different interventions. Whether the addition of other predictors, including biomarkers, will improve prediction accuracy of the models remains to be tested.

In our machine learning analysis of a longitudinal cohort of patients with first-episode psychosis treated in early-intervention services, we were able to show that multiple outcomes can be reliably predicted for patients by use of baseline demographic and clinical variables at 1 year, with external generalisability. Our prediction models have similar discriminatory power to other available predictive models.45 Our models benefit from being developed with use of a naturalistic cohort study and externally validated in a cohort study and a randomised control trial, together with the use of readily available clinical data and, to our knowledge, the largest sample size used in a machine learning study of first-episode psychosis to date. Furthermore, to our knowledge, our study represents the first published evidence for the use of machine learning models of QoL outcome in patients with a first-episode psychosis.

Contributors
PKM conceptualised the analysis plan. PKM and SPL designed the analysis plan and drafted and revised the paper. SPL did the analysis. RU, MRB, RK, JC, GVG, MS, and PFL contributed to the interpretation of the analysis. RU, MRB, RK, JC, GVG, MS, PFL, SPS, LE, PBJ, DF, VS, NF, AIG, MS, MEB, and RHBC revised the draft. SPS, LE, PBJ, DF, VS, NF, AIG, MEB, and RHBC revised the draft. SPS, LE, PBJ, DF, VS, and NF designed the EDEN study. SPS, LE, PBJ, DF, VS, NF, AIG, MS, MEB, and RHBC contributed data. MEB, RHBC, and NF did the validation analysis on OPUS data. MB monitored data collection for the OPUS trial and supervised it. MB, AIG, and MN monitored the data collection for the Scottish studies and supervised them. MN monitored the data collection for the OPUS trial and supervised it. MB, AIG, and MN revised and approved the final version of the manuscript.

Declaration of interests
PKM has received honorariums from Sunovion and Sage. RU has received honorariums from Sunovion. JC has received grants from Wellcome Trust and Sackler Trust and honorariums from Johnson & Johnson. SPS and MB are part-funded by the National Institute for Health Research (NIHR) Collaboration for Leadership in Applied Health Research and Care WM (CLAHRC WM). GVG has received support from H2020-EINFRA, the NIHR Birmingham ECRC, NIHR Birmingham SRMRC, the NIHR Birmingham Biomolecular Research Centre, and the 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. All other authors declare no competing interests.

Data sharing
MB acts as custodian of the EDEN dataset and data sharing and secondary analyses are supported under the auspices of the University of Warwick (Coventry, UK); please contact MB for all requests. AIG acts as custodian of the Scottish datasets and data sharing and secondary analyses are supported under the auspices of the University of Glasgow (Glasgow, UK); please contact AIG for all requests. MN acts as custodian of the OPUS trial dataset and data sharing and secondary analyses are supported under the auspices of the University of Copenhagen (Copenhagen, Denmark); please contact MN for all requests.

Acknowledgments
EDEN was funded by the UK Department of Health (RDD/ARF2) and National Institute of Health Research under the Programme Grants for Applied Research Programme (RP-PG-0109–10074). The Scottish studies were funded by the National Health Service Research Scotland (NRS), through the Chief Scientist Office (CZJH/4/295 and CZJH/3/5), the Scottish Mental Health Research Network, and the Wellcome Trust (104025/Z/14/Z). The OPUS trial was funded by the Danish Ministry of Health (r.nr. 96-0790/71), Danish Ministry of Social Affairs, University of Copenhagen, Copenhagen Hospital Corporation, Danish Medical Research Council (r.nr. 9601612 and 9900734), and Slagtermester Wørzners Fond. This study and external validation were part funded by an unrestricted grant from The Lundbeck Foundation to PRECISE (R277–2018-1431). The views expressed in this publication are those of the authors and not necessarily those of the NHS, the NIHR, the Medical Research Council, the UK Department of Health, or the CLAHRC WM collaborative organisations. We thank all participants of the EDEN, OPUS, and Scottish studies and the EIP teams who supported this research. We thank the anonymous reviewers for providing very helpful reviews to improve the manuscript.

References
10 Birchwood M, Lester H, McCarthy L, et al. The UK national evaluation of the development and impact of Early Intervention Services [the National EDEN studies]: study rationale, design and baseline characteristics. Early Interv Psychiatry 2014; 8: 59–67.
33 Faerden A, Nesvåg R, Marder SR. Definitions of the term ‘recovered’ in schizophrenia and other disorders. Psychopathology 2008; 41: 271–78.