Schizophrenia

ARTICLE OPEN

Multivariable prediction of functional outcome after first-episode psychosis: a crossover validation approach in EUFEST and PSYSCAN

Margot I. E. Slot[1,54]✉, Maria F. Urquijo Castro[2,54], Inge Winter - van Rossum[1,3,4], Hendrika H. van Hell[1], Dominic Dwyer[5,6], Paola Dazzan[7], Arija Maat[1], Lieuwe De Haan[8], Benedicto Crespo-Facorro[9,10], Birte Y. Glenthøj[11,12], Stephen M. Lawrie[13], Colm McDonald[14], Oliver Gruber[15], Thérèse van Amelsvoort[16], Celso Arango[17], Tilo Kircher[18], Barnaby Nelson[5,6], Silvana Galderisi[19], Mark Weiser[20,21], Gabriele Sachs[22], Matthias Kirschner[23,24], the PSYSCAN Consortium*, W. Wolfgang Fleischhacker[25], Philip McGuire[4], Nikolaos Koutsouleris[2,26,27,55] and René S. Kahn[1,3,55]✉

Several multivariate prognostic models have been published to predict outcomes in patients with first episode psychosis (FEP), but it remains unclear whether those predictions generalize to independent populations. Using a subset of demographic and clinical baseline predictors, we aimed to develop and externally validate different models predicting functional outcome after a FEP in the context of a schizophrenia-spectrum disorder (FES), based on a previously published cross-validation and machine learning pipeline. A crossover validation approach was adopted in two large, international cohorts (EUFEST, $n = 338$, and the PSYSCAN FES cohort, $n = 226$). Scores on the Global Assessment of Functioning scale (GAF) at 12 month follow-up were dichotomized to differentiate between poor (GAF current < 65) and good outcome (GAF current ≥ 65). Pooled non-linear support vector machine (SVM) classifiers trained on the separate cohorts identified patients with a poor outcome with cross-validated balanced accuracies (BAC) of 65-66%, but BAC dropped substantially when the models were applied to patients from a different FES cohort (BAC = 50–56%). A leave-site-out analysis on the merged sample yielded better performance (BAC = 72%), highlighting the effect of combining data from different study designs to overcome calibration issues and improve model transportability. In conclusion, our results indicate that validation of prediction models in an independent sample is essential in assessing the true value of the model. Future external validation studies, as well as attempts to harmonize data collection across studies, are recommended.

Schizophrenia (2024)10:89 ; https://doi.org/10.1038/s41537-024-00505-w

INTRODUCTION

Given the large variability in disease trajectories and functional outcomes after experiencing a first episode of psychosis (FEP)[1–4], research has focused on developing tools to predict functional outcomes in order to guide clinical decision-making. Machine learning techniques are increasingly being used in psychiatric research[1–14] and can capture patient heterogeneity to make individual outcome predictions[15], by learning complex associations from multivariate data[16,17].

Functional outcome measures, such as the Global Assessment of Functioning (GAF) scale[18], offer a holistic view of a patient's ability to function in daily life, encompassing social, occupational, and psychological domains. Unlike more narrowly defined criteria like symptom remission, functional assessments prioritize the impact of the disorder on the patient's life, aligning with a patient-centered approach. The GAF is a quick and easily administered measure which requires minimal training[18]. Additionally, GAF scores may serve as proxies for estimating other meaningful

[1]Department of Psychiatry, UMC Utrecht Brain Center, University Medical Center Utrecht, Utrecht, The Netherlands. [2]Department of Psychiatry and Psychotherapy, Ludwig-Maximilian-University, Munich, Germany. [3]Department of Psychiatry, Icahn School of Medicine, Mount Sinai, New York, USA. [4]Department of Psychiatry, University of Oxford, Oxford, UK. [5]Centre for Youth Mental Health, University of Melbourne, Melbourne, VIC, Australia. [6]Orygen, Melbourne, VIC, Australia. [7]Department of Psychological Medicine, Institute of Psychiatry, Psychology & Neuroscience, King's College London, De Crespigny Park, Denmark 458 Hill, SE5 8AF London, UK. [8]Amsterdam UMC, University of Amsterdam, Psychiatry, Department Early Psychosis, Meibergdreef 9, Amsterdam, The Netherlands. [9]Department of Psychiatry, Marqués de Valdecilla University Hospital, IDIVAL. School of Medicine, University of Cantabria, Santander, Spain. [10]CIBERSAM, Centro Investigación Biomédica en Red Salud Mental, Madrid, Spain. [11]Centre for Neuropsychiatric Schizophrenia Research (CNSR) & Centre for Clinical Intervention and Neuropsychiatric Schizophrenia Research (CINS), Mental Health Centre Glostrup, Glostrup, Denmark. [12]University of Copenhagen, Faculty of Health and Medical Sciences, Department of Clinical Medicine, Copenhagen, Denmark. [13]Division of Psychiatry, University of Edinburgh, Royal Edinburgh Hospital, Edinburgh EH10 5HF, UK. [14]Centre for Neuroimaging & Cognitive Genomics (NICOG), NCBES Galway Neuroscience Centre, National University of Ireland Galway, H91 TK33 Galway, Ireland. [15]Section for Experimental Psychopathology and Neuroimaging, Department of General Psychiatry, Heidelberg University, Heidelberg, Germany. [16]Department of Psychiatry and Neuropsychology, Maastricht University, Maastricht, The Netherlands. [17]Department of Child and Adolescent Psychiatry, Institute of Psychiatry and Mental Health, Hospital General Universitario Gregorio Marañón, IiSGM, CIBERSAM, ISCIII, School of Medicine, Universidad Complutense, Madrid, Spain. [18]Department of Psychiatry, University of Marburg, Rudolf-Bultmann-Straße 8, D-35039 Marburg, Germany. [19]Department of Mental and Physical Health and Preventive Medicine, University of Campania Luigi Vanvitelli, Largo Madonna delle Grazie, 80138 Naples, Italy. [20]Zachai Department of Psychiatry, Sheba Medical Center, Tel Hashomer 52621, Israel. [21]Tel Aviv University School of Medicine, Ramat Aviv, Israel. [22]Department of Psychiatry and Psychotherapy, 1090 Vienna, Austria. [23]Division of Adult Psychiatry, Department of Psychiatry, University Hospitals of Geneva, Geneva, Switzerland. [24]Department of Psychiatry, Psychotherapy and Psychosomatics, Psychiatric Hospital, University of Zurich, Zurich, Switzerland. [25]Medical University of Innsbruck, Innsbruck, Austria. [26]Department of Psychosis Studies, Institute of Psychiatry, Psychology & Neuroscience, King's College London, De Crespigny Park, Denmark 458 Hill, London SE5 8AF, UK. [27]Max Planck Institute of Psychiatry, Munich, Germany. [54]These authors contributed equally: Margot I. E. Slot, Maria F. Urquijo Castro. [55]These authors jointly supervised this work: Nikolaos Koutsouleris, René S. Kahn. *A list of authors and their affiliations appears at the end of the paper. ✉email: I.E.Slot-3@umcutrecht.nl; rene.kahn@mssm.edu

patient outcomes, including hospital readmission rates, treatment adherence, quality of life and occupational ability[11,19]. The integration of the GAF with other routinely collected variables such as sociodemographic characteristics, symptom severity, and comorbid disorders has demonstrated robust clinical relevance and predictive power for individual functional outcomes[1,11,20–23].

Several multivariate prognostic models have been published to predict functional outcomes in patients with a FEP, using demographic and clinical variables at baseline. Koutsouleris et al.[11] developed a Support Vector Machine (SVM) model to predict 1-year functional treatment outcome in a large cohort of FEP patients (in the context of a schizophrenia spectrum disorder, broadly defined as first-episode schizophrenia spectrum disorder; FES) participating in a pragmatic randomized controlled trial[11]. Cross-validated results were able to predict functional outcome with 73.8% balanced accuracy (BAC). Geographical generalizability of the models was tested using a leave-site-out approach, i.e. by iteratively validating the models in one study site which was completely left out from the training sample used to develop the models (the remaining sites). This yielded a BAC of 71.1% (including all selected variables) to 67.7% (model including the 10% of top-ranked variables only). Additional studies from Leighton et al.[5] and De Nijs et al.[10] followed, reporting models to predict individual symptomatic and functional outcomes at one-, three- and six-year follow-up in FEP patients and patients with a schizophrenia-spectrum disorder respectively. To date, only few studies have attempted to validate such models in large, independent samples; external validation of these models is essential, since a high-performing model in one sample may have limited predictive value when applied to other patients[24].

Leighton and colleagues assessed the prediction of 1-year outcome in terms of symptomatic remission and employment, education or training (EET) status in a naturalistic cohort of FEP patients[5] and externally validated their models in a different FEP cohort, both cohorts originating from Scotland. This yielded a Receiver Operating Characteristic Area Under the Curve (ROC-AUC) of 0.88 for EET status, and non-significant ROC-AUC values ranging from 0.63 to 0.65 for symptomatic remission, possibly due to a small sample size. In a second study from Leighton, Upthegrove et al.[6], prognostic models were developed to predict various outcomes, including symptom remission, social recovery, vocational recovery and quality of life at 1 year after a first episode of psychosis. The models were developed in a large, naturalistic cohort of FEP patients treated at Early Intervention Services in the UK (EDEN sample, $n = 1027$[25]) and validated in two external longitudinal FEP cohorts in Scotland (which formed the basis for the Leighton study described above; $n = 162$[5,26]), as well as in a randomized controlled study cohort of FEP patients receiving early intervention versus standard treatment in Denmark (OPUS sample (NCT00157313), $n = 578$[27]). The trained models identified patients with poor versus good symptomatic and functional outcomes significantly better than chance, with AUCs ranging from 0.703 to 0.736 (all $p < 0.0001$). External validation in the independent Scottish samples provided a mixed picture of the discriminative ability of the model with AUCs ranging from 0.679 to 0.867 (p-values ranging from <0.05 to <0.0001), while external validation in the Danish RCT yielded low AUCs ranging from 0.556 to .660 (three out of four AUCs reaching significance). More recently, Chekroud and colleagues [28], reported on an elaborate prediction model validation effort; the authors used five international, multisite randomized controlled treatment trials in patients with schizophrenia, resulting in heterogeneous patient samples ranging from pediatric to older adult patients, and chronically ill versus first episode patients. Machine learning methods were applied using baseline data, to predict clinically significant symptom improvement over a 4-week treatment period within each of the individual trials. These models were then cross-validated (within that same trial) as well as externally validated in the other study samples. The authors report that these models predicted patient outcomes with high accuracy within the trial in which the model was developed; however, they performed no better than chance when applied in the other trial samples. Aggregating data across trials to predict outcomes in the trial left out did not improve performance.

If we are to eventually use prediction tools in clinical practice, it is crucial to know how robust these models are when applied across different, potentially highly diverging patient populations. Given the relative importance and impact of treatment decisions in the early phase of schizophrenia, we focus specifically on the FES phase. In this subgroup of patients, external validation studies are scarce and the existing literature on external validation of prognostic models for 1-year treatment outcome in first episode patients has been mostly restricted to small samples or single-country study cohorts. With the few validation studies yielding low to moderate predictive strength, it is doubtful that this performance translates into clinical applicability. Hence, validation in large scale, naturalistic samples is required to further explore the generalizability of these models to more representative, real-world patient cohorts.

To respond to this need, the present work aimed to develop and externally validate prognostic models predicting functional outcome in FES, using a crossover approach on the EUFEST and PSYSCAN cohorts, two large scale samples from the European continent and Israel. Extending existing research in this field, our models are based on the machine learning models published by Koutsouleris et al.[11]. In view of its clinical applicability, easy to obtain demographic and clinical baseline predictors were used for model development. To evaluate geographical generalizability of the models, we applied additional strategies, such as data pooling and leave-site-out cross-validation.

METHODS

Participants and study design

The present work used data from EUFEST (ISRCTN68736636) as well as the FES cohort from the PSYSCAN study (HEALTH.2013.2.2.1-2-FEP). EUFEST is a multicenter, pragmatic, open randomized controlled trial comparing the effectiveness of second-generation antipsychotic drugs with that of a low dose of haloperidol in patients with first-episode schizophrenia spectrum disorder. Patients were randomized to treatment with haloperidol 1–4 mg, amisulpride 200–800 mg, olanzapine 5–20 mg, quetiapine 200–750 mg, or ziprasidone 40–160 mg daily, and followed for a period of 1 year. PSYSCAN is an international, multicenter, longitudinal study on the early stages of psychosis. Patients with first episode psychosis as defined by a DSM-IV diagnosis of schizophrenia, schizoaffective disorder (depressive type) or schizophreniform disorder were followed for a period of 1 year in a naturalistic, prospective design. All participants, or their legal representatives, provided written informed consent. Both studies were approved by the relevant ethics committees of the participating centers, and conducted in accordance with the Declaration of Helsinki (2013). A detailed description of the study design and inclusion and exclusion criteria has been provided elsewhere[39–41]. A summary of key differences between the studies is included in Supplementary Table 1. A subset of data was used; First, overlapping variables between the two studies were selected (see Table 1; a more elaborate description of the predictive features including an overview of the possible data values is included in Supplementary Table 2). After variable selection, only patients with equal to or less than 20% missing predictive variables and for whom a GAF score at month 12 was available, were included in the analyses. This led to a total of 338 subjects from the EUFEST cohort (slightly different from the 334 subjects in Koutsouleris et al.[11]) and 226 subjects from PSYSCAN. An overview of the sample size per site and study is provided in Supplementary Table 3.

Table 1. Baseline variables from the EUFEST and PSYSCAN databases selected for analysis.

Sociodemographic variables
1 Sex
2 Age
3 Years of education[a]
4 Weight
5 Body Mass Index (BMI)
6 Diastolic blood pressure
7 Systolic blood pressure
8 Marital status patient: married
9 Current occupation patient
10 Highest educational degree patient
11 Educational problems
12 Education father
13 Education mother
14 Living alone
15 Living environment
Antipsychotic medication
16 Haloperidol treatment
17 Olanzapine treatment
18 Quetiapine treatment
19 Amisulpride treatment
20 Ziprasidone treatment
Diagnostic interview
21 Disorganized schizophrenia
22 Catatonic schizophrenia
23 Paranoid schizophrenia
24 Schizophreniform disorder
25 Residual state
26 Schizoaffective disorder
27 Undifferentiated schizophrenia
28 MDE Current[b]
29 MDE Recurrent[b]
30 Substance induced mood disorder Past
31 MDE with melancholic features Current[b]
32 Dysthymia Current
33 Hypomanic episode Past[c]
34 Panic disorder Current past month
35 Panic disorder Lifetime
36 Agoraphobia Lifetime
37 Social phobia
38 Specific phobia
39 Obsessive-compulsive disorder
40 Schizophrenia Current
41 Schizophrenia Lifetime
Diagnostic interview
42 Schizoaffective disorder Current
43 Schizoaffective disorder Lifetime
44 Schizophreniform disorder Current
45 Schizophreniform disorder Lifetime
46 Substance induced psychotic disorder Lifetime
Clinician-rated scales
47 Clinical Global Impression (CGI)
48 Global Assessment of Functioning (GAF)
Positive and Negative Syndrome Scale (PANSS)
49 PANSS P1 Delusions
50 PANSS P2 Conceptual disorganization
51 PANSS P3 Hallucinations
52 PANSS P4 Hyperactivity
53 PANSS P5 Grandiosity
54 PANSS P6 Suspiciousness/persecution
55 PANSS P7 Hostility
56 PANSS N1 Blunted affect
57 PANSS N2 Emotional withdrawal
58 PANSS N3 Poor rapport
59 PANSS N4 Passive/apathetic social withdrawal
60 PANSS N5 Difficulty in abstract thinking
61 PANSS N6 Lack of spontaneity and flow of conversation
62 PANSS N7 Stereotyped thinking
63 PANSS G1 Somatic concern
64 PANSS G2 Anxiety
65 PANSS G3 Guilt feelings
66 PANSS G4 Tension
67 PANSS G5 Mannerisms and posturing
68 PANSS G6 Depression
69 PANSS G7 Motor retardation
70 PANSS G8 Uncooperativeness
71 PANSS G9 Unusual thought content
72 PANSS G10 Disorientation
73 PANSS G11 Poor attention
74 PANSS G12 Lack of judgment and insight
75 PANSS G13 Disturbance of volition
76 PANSS G14 Poor impulse control
77 PANSS G15 Preoccupation
78 PANSS G16 Active social avoidance
79 PANSS Positive score
80 PANSS Negative score
81 PANSS General score
82 PANSS Total score

Note. [a] Years of education = years in school and college/university (excluding kindergarten/nursery).
[b] MDE = Major Depressive Episode.
[c] Not included in the internally cross-validated model on PSYSCAN data, due to >20% missing values.

Predictors and outcome measure

The primary outcome measure was functional outcome after 12 months of follow-up, assessed with the current score of the Global Assessment of Functioning (GAF) scale[42]. Early changes in GAF scores have been observed as a simple but effective predictor of various long-term symptomatic and functional outcomes[43]. GAF scores were dichotomized to differentiate patients with a so-called poor outcome (GAF score <65) from patients with a good outcome (GAF score ≥ 65), as previously defined by Koutsouleris et al.[11]. This cutoff was chosen since it is the cutoff used by the original study; scores between 51 and 70 have been suggested to distinguish between mild functional problems and moderate to severe functional impairment[44], as thresholds between at-risk and disease states[45], and cutoffs of 60 and 65 specifically have been used to define functional remission or recovery in previous naturalistic, longitudinal FEP studies[6,46–48]. To approach the models from the original paper as closely as possible, features available in both studies were identified, resulting in a total number of 82 overlapping demographic and clinical predictors selected for analysis (see Table 1). Schizophrenia spectrum disorders and comorbid DSM-IV diagnoses were confirmed using the MINI international neuropsychiatric interview plus (MINI plus; EUFEST)[49] or the Structured Clinical Interview for DSM-IV Disorders (SCID-I; PSYSCAN)[50].

Model development

Model development and validation followed the Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD) guidelines[51]. The open-source pattern recognition tool NeuroMiner version 1.1[52] in Matlab (release: R2021b; https://nl.mathworks.com/products/matlab.html) was used to rebuild the original prognostic classification models, i.e. based on the cross-validation and machine learning pipeline from Koutsouleris et al.[11] but only including those variables available in both studies. We used a repeated nested cross-validation (CV) framework with 20 folds at the outer level (CV$_2$) and 5 folds and 4 permutations at the inner cycle (CV$_1$). In the preprocessing phase, variables were first scaled to a range of [0–1], and missing values were imputed using a 5-nearest neighbor approach based on the Euclidean distance. Wrapper-based, greedy forward feature selection, using a non-linear support vector machine algorithm with a Radial Basis Function (RBF or Gaussian) kernel, identified a subset of most predictive variables. A number of steps similar to those used in Koutsouleris et al.[11] were implemented to prevent overfitting and increase generalizability and clinical utility of the model (see 'Machine Learning pipeline' in the Supplements). Subjects with a positive or negative mean decision score were classified as patients having a poor or good outcome, respectively. Higher absolute values of decision scores indicate a higher certainty of the patient belonging to either group. Significance of the prognostic model was determined using permutation analysis at $\alpha = 0.01$[53]; the classification performance of the model in terms of Balanced Accuracy (BAC) was compared with a null distribution of the out-of-training classification performance (BAC) of 500 random permutations of the outcome labels.

Cross-over validation

To determine the prognostic performance of the models beyond the discovery sample, a cross-over validation approach was adopted; a model was trained and cross-validated using data from one study (EUFEST) and then applied on to the external dataset (PSYSCAN), and vice versa. Following Steyerberg and Harrell's (2016) recommendations[54], we evaluated the geographical transportability of the models based on a leave-site-out cross-validation approach: we first pooled the data in the inner cycle (CV1) while data in the outer cycle (CV2) were split by site (see 'Machine learning pipeline' in the Supplements). Given the systematic decision score differences between the two study cohorts, all analyses were repeated after calibrating the data using the correction method described in Koutsouleris et al.[12], where group differences between study cohorts were corrected by (1) centering the variables to their global mean and (2) subtracting the difference between cohort-specific means and the global mean. Supplementary Table 4 lists results without mean offset correction.

Other analyses

Descriptive statistics and follow-up analyses were performed in SPSS version 29.0[55]. All statistical tests were two-sided. The significance level was set at $\alpha = 0.05$, unless otherwise indicated. Between-group comparisons of sociodemographic and clinical characteristics at baseline were performed using *t*-tests and Chi-Square tests. To assess the importance of each predictive feature for classification performance, we used the sign-based consistency as a measure of feature significance and the cross-validation ratio as a measure of feature reliability (see 'Importance of predictive features' in the Supplements)[38,56]. Predictive features were ranked for each classifier based on their sign-based consistency resulting from the inner cross-validation cycles. Features consistently selected as the most important predictors across the inner cross-validation cycles (i.e., those with a significant sign-based consistency value) were compared between the different classification models (EUFEST classifier, PSYSCAN classifier and leave-site-out classifier). To determine whether the overlapping features contributed to the prediction of the poor versus good outcome label, the cross-validation ratio was used. In a final step, Spearman and point-biserial correlation coefficients were calculated between the overlapping, significant predictive baseline features and the mean predicted decision scores resulting from the validation of the classifiers. The type of correlation analysis (Spearman's rho or point-biserial correlation coefficient) was selected according to the scale of the variables concerned. These correlation analyses provided insight into the direction of the associations, i.e. on whether a feature (e.g. PANSS total) was positively or negatively associated with the *predicted* outcome (regardless of whether this corresponds to the *actual* outcome).

RESULTS

Sociodemographic data and clinical characteristics at baseline, including results of between-sample comparisons, are presented in Table 2. After a year of follow-up, 78 patients from the EUFEST sample (23.1%) presented with a poor outcome (defined as GAF < 65) as compared to 113 patients (50.0%) in PSYSCAN, $\chi^2(1) = 44.15$, $p < 0.001$. The pooled non-linear SVM classifier trained on EUFEST data correctly identified patients with a poor outcome with a cross-validated BAC of 66.1%, $p < 0.002$ (Table 3). The decision scores of this adjusted model correlated strongly with the original model from Koutsouleris ($\rho = 0.751$, $p < 0.001$)[11,29]. The prognostic model trained on PSYSCAN participants achieved a slightly lower BAC of 64.6%, $p < 0.002$. When applying the models onto the study cohort not included in the discovery and cross-validation phase, classification performance substantially decreased, as reflected in BAC losses of 9.9% and 14.6%, respectively.

The drop in classification performance is also reflected in the imbalance between sensitivity and specificity emerging from the external validation analyses. Results of the internal cross-validation analyses indicate that 60.3% of the EUFEST and 69.9% of the PSYSCAN patients with a poor functional outcome were correctly classified as patients having a poor outcome at month 12 (sensitivity). A total of 71.9% of the good outcome patients in the EUFEST cohort and 59.3% of the good outcome patients in the PSYSCAN cohort were correctly classified as patients having a good outcome (specificity). Validation of the models in the external cohort resulted in a shift in sensitivity/specificity balance, as evidenced in sensitivity levels of 82.3% and 100.0%, and specificity levels of 30.1% and 0.0%, respectively. In other words, a 41.8–59.3% drop in specificity was observed when applying the models to the external test cohort. Despite these decreases in BAC and specificity performance when validating the EUFEST classifier in the PSYSCAN dataset and vice versa, the area under the receiver operating curve (AUC) was still within a range of 0.62-0.64.

The leave-site-out, inner pooled cross-validation analysis (LSO) on the combined dataset produced a significant BAC of 72.4%, $p < 0.002$, with an AUC of 0.79. The sensitivity (70.1%) and specificity (74.7%) of the leave-site-out classifier were within a similar range. Further inspection reveals a difference between the positive and negative likelihood ratios; the positive likelihood ratio (the probability that a poor outcome label is expected in a poor outcome patient, divided by the probability that a poor outcome label is expected in a patient with a good outcome[30]) is higher than the negative likelihood ratio (the probability of a patient with

Table 2. Sociodemographics and baseline clinical characteristics of the EUFEST sample and the PSYSCAN FES cohort.

Sociodemographics	EUFEST (*n* = 338)	PSYSCAN (*n* = 226)	Statistic	*p*-value
Age in years	25.6 (5.7)	24.7 (5.7)	$t(562) = 1.79$	$p = 0.073$
Male sex	190 (56.2%)	155 (68.6%)	χ^2 (1) = 8.73	$p = 0.003$
GAF current	40.6 (13.5)	55.6 (20.0)	$t(359.41) = -9.87$	$p < 0.001$
GAF < 65	320/338 (94.7%)	146/225 (64.9%)	χ^2 (1) = 84.04	$p < 0.001$
Living status: independently	40/336 (11.9%)	42 (18.6%)	χ^2 (1) = 4.84	$p = 0.028$
Relationship status: married	44 (13.0%)	8 (3.5%)	χ^2 (1) = 14.54	$p < 0.001$
Educational years[a]	12.6 (2.9)	14.1 (3.2)	t (561) = -5.79	$p < 0.001$
Education level				
University (finished)	28/337 (8.3%)	26 (11.5%)	χ^2 (6) = 28.88	$p < 0.001$
University (unfinished)	65/337 (19.3%)	47 (20.8%)		
Professional training (finished)	49/337 (14.5%)	25 (11.1%)		
Professional training (unfinished)	16/337 (4.7%)	29 (12.8%)		
High school (finished)	68/337 (20.2%)	56 (24.8%)		
High school (unfinished)	61/337 (18.1%)	32 (14.2%)		
Less than high school	50/337 (14.8%)	11 (4.9%)		
Employment status: employed	158 (46.7%)	65 (28.8%)	χ^2 (1) = 18.33	$p < 0.001$
Baseline clinical characteristics				
Duration of illness < 2 years[b]	337 (99.7%)	173 (76.9%)	χ^2 (1) = 82.45	$p < 0.001$
Schizophrenia spectrum diagnosis			χ^2 (2) = 9.78	$p = 0.008$
Schizophrenia	177/336 (52.7%)	145 (64.2%)		
Schizophreniform disorder	133/336 (39.6%)	74 (32.7%)		
Schizoaffective disorder	26/336 (7.7%)	7 (3.1%)		
Comorbid psychiatric diagnoses				
Major depressive episode current	29 (8.6%)	18 (8.1%)	χ^2 (1) = 0.04	$p = 0.844$
Major depressive disorder recurrent	26 (7.7%)	13 (5.8%)	χ^2 (1) = 0.77	$p = 0.373$
Antipsychotic medication				
Haloperidol	68 (20.1%)	8/223 (3.6%)	χ^2 (1) = 31.35	$p < 0.001$
Olanzapine	81 (24.0%)	42/223 (18.8%)	χ^2 (1) = 2.07	$p = 0.151$
Amisulpride	69 (20.4%)	10/223 (4.5%)	χ^2 (1) = 28.18	$p < 0.001$
Quetiapine	68 (20.1%)	23/223 (10.3%)	χ^2 (1) = 9.50	$p = 0.002$
Ziprasidone	52 (15.4%)	0/223 (0.0%)	χ^2 (1) = 37.81	$p < 0.001$
Clozapine	0 (0.0%)	23/226 (10.2%)	χ^2 (1) = 35.86	$p < 0.001$
Aripiprazole	0 (0.0%)	56/226 (24.8%)	χ^2 (1) = 92.99	$p < 0.001$
Paliperidone	0 (0.0%)	8/226 (3.5%)	χ^2 (1) = 12.14	$p < 0.001$
CGI	4.9 (0.8)	3.4 (1.4)	$t(316.67) = 14.43$	$p < 0.001$
PANSS total	89.1 (20.7)	55.7 (27.3)	$t(518.79) = 20.55$	$p < 0.001$
PANSS positive	23.4 (6.2)	13.2 (5.6)	$t(553) = 19.73$	$p < 0.001$
PANSS negative	21.2 (7.7)	14.6 (6.7)	$t(508.67) = 10.56$	$p < 0.001$
PANSS general	44.5 (10.7)	27.9 (8.4)	$t(533.87) = 20.46$	$p < 0.001$
HAM-D sum score (*Mdn, IQR*)		5.0 (7.0)		
CDSS sum score (*Mdn, IQR*)	4.0 (7.0)			
Symbol Substitution Test, raw score (WAIS)	50.7 (18.5)	62.9 (17.5)	$t(533) = -7.64$	$p < 0.001$

Note. Data are mean (SD), n (%), or n/N (%), unless otherwise indicated. Denominators change because of incomplete data. *GAF* Global Assessment of Functioning. *CGI* Clinical Global Impression. *PANSS* Positive and Negative Syndrome Scale. *WAIS* Wechsler Adult Intelligence Scale. *HAM-D* Hamilton Depression Rating Scale. Scores range from 0 to 50. Higher scores indicate more severe depressive symptoms. *CDSS* Calgary Depression Scale for Schizophrenia. Scores range from 0 to 27. Higher scores indicate more severe depressive symptoms. Some baseline variables were omitted from the analyses due to the violation of assumptions of minimum cell frequencies (e.g., PTSD).

[a] Years of education = years in school and college/university (excluding kindergarten/nursery).

[b] Duration of illness is defined as the time between the onset of frank psychosis and study entry.

a poor outcome receiving the good outcome label, divided by the probability of a patient with a good outcome being classified as such[30]). Repeating the leave-site-out analyses on the EUFEST and PSYSCAN sample separately resulted in a similar result in EUFEST (BAC = 70.6%, $p < 0.002$) but a performance decrease in PSYSCAN data (BAC = 58.6%, $p = 0.086$), suggesting that the enhanced performance of the leave-site-out model on the merged dataset is not just a consequence of an increase in power. Details on the prediction performance per site are provided in Supplementary Figs. 1 and 2. Results of additional analyses comparing the

Table 3. Prediction performance of the trained models on classifying 52 week outcome.

	N subjects	*N* sites	TP	TN	FP	FN	Sens (%)	Spec (%)	PPV	NPV	PSI	LR+	LR-	BAC	AUC	*p*-value
Pooled CV classifier EUFEST	338	46	47	187	73	31	60.3	71.9	39.2	85.8	24.9	2.1	0.6	**66.1**	0.75	<0.002
Validation EUFEST classifier in PSYSCAN sample	226	15	93	34	79	20	82.3	30.1	54.1	63.0	17.0	1.2	0.6	**56.2**	0.64	–
Pooled CV classifier PSYSCAN	226	15	79	67	46	34	69.9	59.3	63.2	66.3	29.5	1.7	0.5	**64.6**	0.70	<0.002
Validation PSYSCAN classifier in EUFEST sample	338	46	78	0	260	0	100.0	0.0	23.1	–	–	1.0	–	**50.0**	0.62	–
Leave-site-out / inner pooled CV classifier merged sample†	433	25	101	216	73	43	70.1	74.7	58.0	83.4	41.4	2.8	0.4	**72.4**	0.79	<0.002
Leave-site-out / inner pooled CV classifier EUFEST sample†	218	15	23	152	24	19	54.8	86.4	48.9	88.9	37.8	4.0	0.5	**70.6**	0.76	<0.002
Leave-site-out / inner pooled CV classifier PSYSCAN sample†	210	12	51	73	38	48	51.5	65.8	57.3	60.3	17.6	1.5	0.7	**58.6**	0.65	0.086

Note. Classified poor outcomes (GAF < 65) are labeled as positive predictions and good outcomes (GAF ≥ 65) as negative predictions, i.e. sensitivity measures the classifier's ability to correctly identify patients with poor outcomes as such. In all models, mean offset correction was applied as an extra preprocessing step. † Sites that included <10 participants were excluded from the analysis (see Suppl. Table 2). *CV* Cross-validated. *TP* True Positives. *TN* True Negatives. *FP* False Positives. *FN* False Negatives. *Sens* Sensitivity. *Spec* Specificity. *PPV* Positive Predicted Value. *NPV* Negative Predicted Value. *PSI* Prognostic Summary Index. *LR+* Positive Likelihood Ratio. *LR-* Negative Likelihood Ratio. *BAC* Balanced Accuracy. *AUC* Area Under the Receiver Operating Characteristic Curve.

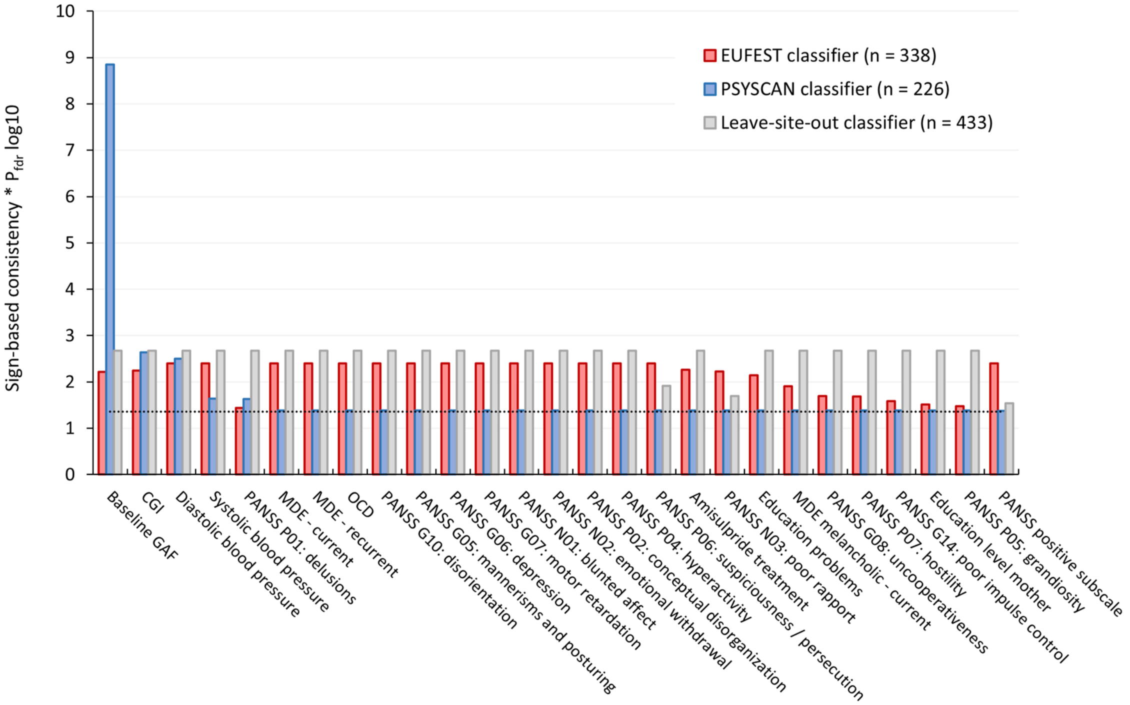

Fig. 1 A comparison of the most important predictive baseline variables per classifier based on the sign-based consistency resulting from the inner cross-validation cycles. Significant predictors overlapping across the models are presented only. Variables with a sign-based consistency * Pfdr log10 ≥ 1.36 are considered significant (dotted line reflects the significance threshold). EUFEST classifier = pooled non-linear cross-validated SVM model on EUFEST cohort. PSYSCAN classifier = pooled non-linear cross-validated SVM model on PSYSCAN cohort. Leave-site-out classifier = leave site-out inner pooled cross-validated SVM model on the merged dataset (sites that included < 10 participants were excluded).

accuracy versus error rates of the classification models for the different treatments arms in EUFEST are also provided in the Supplements (Supplementary Fig. 3).

Using the sign-based consistency method, a total of 27 overlapping, significant predictive baseline features were identified (Fig. 1). CV ratios indicated that illness severity (CGI), current and recurrent depressive episodes, obsessive-compulsive disorder and education problems consistently emerged as important predictors for classifying poor outcome (Fig. 2). The baseline GAF score and systolic blood pressure were observed as important predictors for classifying good outcome. For the leave-site-out classifier in particular, the baseline GAF score had a strong impact on the decision to assign a patient to the good outcome group. Treatment with amisulpride was also informative for the classification of good outcome, but for the EUFEST and leave-site-out classifier only. Correlations between the top predictive baseline features overlapping across the models and the mean predicted decision scores of the cross-validated models are presented in Supplementary Table 5.

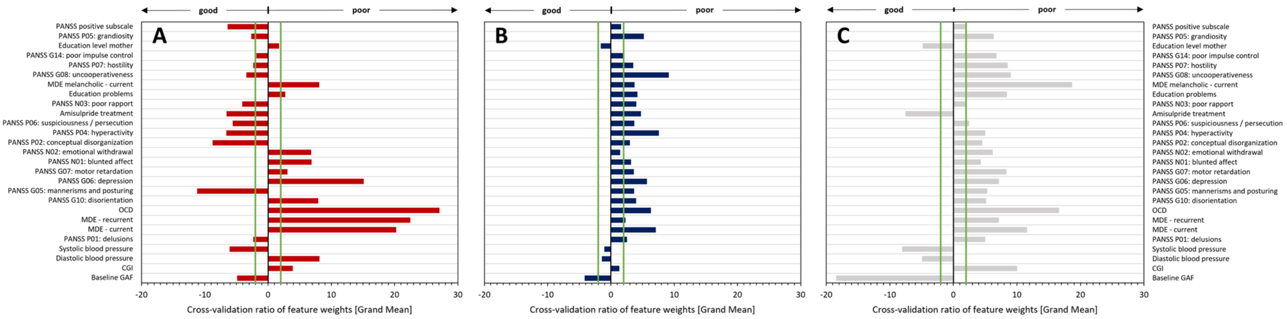

Fig. 2 Cross-validation ratios (CVR) of the significant predictive baseline variables overlapping across the models. The green lines reflect the 95% confidence threshold (CVR = ± 2). **a** EUFEST classifier = internally cross-validated SVM model on EUFEST data. **b** PSYSCAN classifier = internally cross-validated SVM model on PSYSCAN data. **c** Leave-site-out classifier = inner pooled/outer leave-site-out cross-validated SVM model on the merged dataset (sites that included < 10 participants were excluded). Positive CV ratios indicate that the variable contributes to the classification of the poor outcome label, whereas negative CV ratios indicate the opposite. The absolute CV ratio values indicate how strongly the variable affects the decision towards the outcome label (i.e., a variable with a higher absolute CV ratio drives the decision more strongly towards the classification of the outcome label than a variable with a lower absolute CV ratio).

DISCUSSION

Using clinical and sociodemographic data of one clinical trial (EUFEST) and one naturalistic study (PSYSCAN), the present work aimed to develop and externally validate different machine-learning based prediction models of functional outcome in FES, following the cross-validation and machine learning pipeline from Koutsouleris et al.[11]. For validation, a crossover approach was adopted, i.e. the model from one study was evaluated in the other and vice versa. This in turn allowed to test the generalizability of results stemming from inherently different study designs. Each cross-validated model discriminated between patients with good and poor functional outcome with balanced accuracies ranging from 65% to 66%, which reflects a low level of prediction accuracy (defined as 50–70% accuracy). In addition, when modelling outcome in a cross-over fashion, BAC dropped substantially while AUC was maintained, indicating potential residual calibration issues between the variable spaces of the two cohorts. Better performance (BAC = 72%) was achieved by the leave-site-out model on the merged sample, highlighting the effect of combining data from different study designs to overcome calibration issues and improve model transportability; in combination with results of the leave-site-out models on EUFEST and PSYSCAN data separately, it suggests that this model generalizes moderately well across sites (moderate prediction accuracy defined as 70–80%). In short, the prediction accuracy as yielded by the validation and cross-validation analyses of the two patient samples only reached a low level, which is far from sufficient when considering utilization of prediction models in a clinical setting.

Discrimination performance of the pooled cross-validated EUFEST classifier, the pooled cross-validated PSYSCAN classifier, and leave-site-out classifier in our study falls within the performance ranges of the pooled cross-validated (BAC = 74%) and leave-site-out cross-validated (BAC = 68%) models from Koutsouleris et al.[11]. Previous support vector machine models predicting the same endpoint at 3 year and 6 year follow-up yielded low internal cross-validated balanced accuracies between 64% and 68%, and low balanced accuracies of 57%–66% when using leave-site-out cross-validation[10]. Leighton, Upthegrove et al.[6] reported results of the development and external validation of a logistic regression model (with elastic net regularisation) for social recovery in a similar population. Social recovery was defined as a GAF score equal or above 65, which is identical to the operationalisation of good functional outcome in the current study. Our model performance in terms of AUC (0.70–0.79) matches results from the leave-site-out social recovery model reported by Leighton et al. (AUC = 0.73), both providing a moderate prediction quality. Discrimination performance of our external validation analysis was slightly higher (although still classified as low with an AUC ranging from 0.62 to 0.64) than the external validation performance of the social recovery model on the OPUS dataset (AUC = 0.57[27]) as conducted by Leighton and colleagues[6]. Although small, this difference may be partly explained by the international geographic coverage of the study cohorts used to develop the current models, or the stricter criteria for the diagnostic subgroups included in the present work; this may have contributed to the development of more robust models with a better chance of generalizability to new samples.

Nevertheless, the reduced model performance of our external validation analyses suggests that predictive models for functional outcomes in schizophrenia-spectrum disorder are highly context-dependent and have limited generalisability to new samples, even when applied to new patients in a similar disease stage (FES). This indicates the need to first focus on models that validate sufficiently to other, carefully controlled clinical environments before even considering the translation of a prediction model to daily clinical practice. Multiple differences between the two study samples were found at baseline. Although we tried to mitigate the impact of group differences using univariate mean-offset correction, residual multivariate difference patterns may exist which were not accounted for and could explain the drop in external validation performance caused by systematic shifts between decision scores, i.e. calibration problems of the respective models[31,32]. An important factor may be the shorter duration of illness in EUFEST compared to PSYSCAN, as a result of differences in the study eligibility criteria; EUFEST participants had a maximum duration of illness of 2 years (defined as the time interval between the onset of positive psychotic symptoms, and study entry), compared to a maximum illness duration of 3 years in PSYSCAN (defined as the time interval between the initiation of treatment for psychosis (i.e. date of hospital admission or acceptance at healthcare service for psychosis), and study entry). Second, due to the nature of the trial, none of the patients in EUFEST was treated with aripiprazole, clozapine or paliperidone at baseline, whereas respectively 24.8%, 10.2% and 3.5% of the PSYSCAN participants received this treatment at baseline. These differences in antipsychotic use may have affected the results, e.g. knowing that clozapine has been described to be superior to some other antipsychotics in ameliorating psychotic symptoms[33] and is more likely to be prescribed in treatment-resistant patients[34]. The variation in antipsychotic medication use is a good illustration of the heterogeneity in the PSYSCAN sample, compared to the better controlled RCT data (EUFEST), and may account for the superior performance of the EUFEST model over the PSYSCAN model, as observed in our leave-site-out approach. The pronounced difference in severity of psychotic illness at baseline, possibly related to variations in illness duration, is also worth mentioning; although this aspect was covered in the set of prognostic factors used to develop the models, in general, the EUFEST cohort was more severely ill than the naturalistic PSYSCAN cohort, as reflected in the severity of psychotic symptoms as well as the level of global functioning at baseline. We tried to match the two samples on the severity of psychotic symptoms and GAF scores at baseline to correct for these baseline group differences, but due to the large reduction in power as a result of excluding a large proportion of EUFEST participants, this analysis proved inadequate. As it is likely that the differences in the illness characteristics of the study samples reduced the prediction accuracy of the cross-over validation analyses, this will similarly hamper the translation of such machine learning models to the general patient population seen in clinical practice.

Another critical factor complicating the validation and generalizability of prediction models is the lack of consensus definitions for functional outcomes across studies. Differences in how functional outcomes are defined and measured may significantly impact the results[35,36]. We recommend establishing standardized definitions and outcome measures in the field, to facilitate accurate model comparisons. In this context, future studies with a large sample size could test a more nuanced categorization of the GAF (e.g. by including a middle group) without compromising statistical power. Alternatively, more objective or specific outcome measures (e.g. employment or educational status) could be implemented and evaluated, to explore whether this improves the robustness and generalizability of prediction models.

This study did not include clinicians' estimates of functional outcomes, which could potentially enhance the predictive power of the models without extending assessment times. Although algorithmic predictions are generally comparable to clinicians' estimations[37], the incorporation of clinician predictions could offer significant additional value in diagnostic and prognostic procedures[38]. Therefore, we strongly recommend integrating clinicians' prognostic evaluations into models of functional outcomes and other aspects of psychiatric illness.

Overall, our results are in line with the first external validation study of machine learning models in first episode cohorts, conducted by Leighton and colleagues[6], who also reported low prediction accuracy when models with international patient samples and different study designs (naturalistic versus intervention studies) were externally validated, although the leave-site-out cross-validation analyses resulted in moderate prediction accuracy in both the current study as well as Leighton's report.

The added value of the present work is the large geographical spread of the FES participants (the two samples include patients from 14 European countries, Israel and Australia). We showed that our models based on a previously published machine learning algorithm were able to classify patients from a new sample into good versus poor functional outcome groups when applied in patients from the same study and across sites. However, classification performance dropped significantly when applied in patients from a different FES cohort. In line with recent observations[28], our results indicate that models based on single data sets provide limited insight into performance in future patients; hence, external validation of prediction models in an unrelated and carefully controlled clinical environment is essential in assessing the true value of the model. Only then can the field move towards applications into daily clinical practice.

DATA AVAILABILITY

The data that support the findings of this study are available from the corresponding author upon reasonable request.

CODE AVAILABILITY

The code used in this manuscript is publicly available at http://www.proniapredictors.eu/ (NeuroMiner Model Library).

Received: 12 June 2024; Accepted: 4 September 2024;
Published online: 07 October 2024

REFERENCES

1. Soldatos, R. F. et al. Prediction of early symptom remission in two independent samples of first-episode psychosis patients using machine learning. *Schizophr. Bull.* **48**, 122–133 (2022).
2. de Wit, S. et al. Individual prediction of long-term outcome in adolescents at ultra-high risk for psychosis: Applying machine learning techniques to brain imaging data. *Hum. Brain Mapp* **38**, 704–714 (2017).
3. Nieuwenhuis, M. et al. Multi-center MRI prediction models: predicting sex and illness course in first episode psychosis patients. *Neuroimage* **145**, 246–253 (2017).
4. Rosen, M. et al. Towards clinical application of prediction models for transition to psychosis: a systematic review and external validation study in the PRONIA sample. *Neurosci. Biobehav. Rev.* **125**, 478–492 (2021).
5. Leighton, S. P. et al. Predicting one-year outcome in first episode psychosis using machine learning. *PLoS ONE* **14**, e0212846 (2019).
6. Leighton, S. P. et al. Development and validation of multivariable prediction models of remission, recovery, and quality of life outcomes in people with first episode psychosis: a machine learning approach. *Lancet Digit. Health* **1**, e261–e270 (2019).
7. Taylor, J. A., Larsen, K. M. & Garrido, M. I. Multi-dimensional predictions of psychotic symptoms via machine learning. *Hum. Brain Mapp.* **41**, 5151–5163 (2020).
8. Vieira, S. et al. Using machine learning and structural neuroimaging to detect first episode psychosis: reconsidering the evidence. *Schizophr. Bull.* **46**, 17–26 (2020).
9. Amoretti, S. et al. Identifying clinical clusters with distinct trajectories in first-episode psychosis through an unsupervised machine learning technique. *Eur. Neuropsychopharmacol.* **47**, 112–129 (2021).
10. de Nijs, J. et al. Individualized prediction of three- and six-year outcomes of psychosis in a longitudinal multicenter study: a machine learning approach. *NPJ Schizophr.* **7**, 34 (2021).
11. Koutsouleris, N. et al. Multisite prediction of 4 week and 52 week treatment outcomes in patients with first-episode psychosis: a machine learning approach. *Lancet Psychiatry* **3**, 935–946 (2016).
12. Koutsouleris, N. et al. Toward generalizable and transdiagnostic tools for psychosis prediction: an independent validation and improvement of the NAPLS-2 risk calculator in the multisite PRONIA cohort. *Biol. Psychiatry* **90**, 632–642 (2021).
13. Lalousis, P. A. et al. Heterogeneity and classification of recent onset psychosis and depression: a multimodal machine learning approach. *Schizophr. Bull* **.47**, 1130–1140 (2021).
14. Rosen, M. et al. Detailed clinical phenotyping and generalisability in prognostic models of functioning in at-risk populations. *Br. J. Psychiatry* **220**, 318–321 (2022).
15. Cearns, M., Hahn, T. & Baune, B. T. Recommendations and future directions for supervised machine learning in psychiatry. *Transl. Psychiatry* **9**, 271 (2019).
16. Dwyer, D. B., Falkai, P. & Koutsouleris, N. Machine learning approaches for clinical psychology and psychiatry. *Annu. Rev. Clin. Psychol*. **14**, 91–118 (2018).
17. Fusar-Poli, P., Hijazi, Z., Stahl, D. & Steyerberg, E. W. The science of prognosis in psychiatry: a review. *JAMA Psychiatry* **75**, 1280–1288 (2018).
18. Jones, S. H., Thornicroft, G., Coffey, M. & Dunn, G. A brief mental health outcome scale: reliability and validity of the global assessment of functioning (GAF). *Br. J. Psychiatry* **166**, 654–659 (1995).
19. Köhler, O., Horsdal, H. T., Baandrup, L., Mors, O. & Gasse, C. Association between global assessment of functioning scores and indicators of functioning, severity, and prognosis in first-time schizophrenia. *Clin. Epidemiol.* **8**, 323–332 (2016).
20. Del Fabro, L. et al. Machine learning methods to predict outcomes of pharmacological treatment in psychosis. *Transl. Psychiatry* **13**, 75 (2023).
21. Chang, W. C. et al. Patterns and predictors of trajectories for social and occupational functioning in patients presenting with first-episode non-affective psychosis: a three-year follow-up study. *Schizophr. Res.* **197**, 131–137 (2018).
22. Li, Y. et al. A random forest model for predicting social functional improvement in Chinese patients with schizophrenia after 3 months of atypical antipsychotic monopharmacy: a cohort study. *Neuropsychiatr. Dis. Treat.* **17**, 847–857 (2021).
23. Wu, C. S. et al. Development and validation of a machine learning individualized treatment rule in first-episode schizophrenia. *JAMA Netw. Open* **3**, e1921660 (2020).
24. Koutsouleris, N. Toward clinically useful models for individualised prognostication in psychosis. *Lancet Digit. Health* **1**, e244–e245 (2019).
25. Birchwood, M. 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* **8**, 59–67 (2014).
26. Gumley, A. I. et al. Insight, duration of untreated psychosis and attachment in first-episode psychosis: prospective study of psychiatric recovery over 12 month follow-up. *Br. J. Psychiatry* **205**, 60–67 (2014).

27. Petersen, L. et al. Improving 1 year outcome in first-episode psychosis: OPUS trial. *Br. J. Psychiatry* **187**, s98–s103 (2005).
28. Chekroud, A. M. et al. Illusory generalizability of clinical prediction models. *Science (1979)* **383**, 164–167 (2024).
29. Schober, P., Boer, C. & Schwarte, L. A. Correlation coefficients: appropriate use and interpretation. *Anesth Analg.* **126**, 1763–1768 (2018).
30. Bolin, E. & Lam, W. A review of sensitivity, specificity, and likelihood ratios: evaluating the utility of the electrocardiogram as a screening tool in hypertrophic cardiomyopathy. *Congenit. Heart Dis.* **8**, 406–410 (2013).
31. Schnack, H. G. & Kahn, R. S. Detecting neuroimaging biomarkers for psychiatric disorders: sample size matters. *Front. Psychiatry* **7**, 50 (2016).
32. Altman, D. G., Vergouwe, Y., Royston, P. & Moons, K. G. M. Prognosis and prognostic research: validating a prognostic model. *BMJ (Online)* **338**, 1432–1435 (2009).
33. Leucht, S. et al. Comparative efficacy and tolerability of 15 antipsychotic drugs in schizophrenia: a multiple-treatments meta-analysis. *Lancet* **382**, 951–962 (2013).
34. Correll, C. U. & Howes, O. D. Treatment-resistant schizophrenia: definition, predictors, and therapy options. *J. Clin. Psychiatry* **82**, MY20096AH1C (2021).
35. Searle, A., Allen, L., Lowther, M., Cotter, J. & Barnett, J. H. Measuring functional outcomes in schizophrenia in an increasingly digital world. *Schizophr. Res. Cogn.* **29**, 100248 (2022).
36. Peuskens, J. & Gorwood, P. How are we assessing functioning in schizophrenia? a need for a consensus approach. *Eur. Psychiatry* **27**, 391–395 (2012).
37. Şahin, D. et al. Algorithmic fairness in precision psychiatry: analysis of prediction models in individuals at clinical high risk for psychosis. *Br. J. Psychiatry* **224**, 55–65 (2024).
38. Koutsouleris, N. et al. Multimodal machine learning workflows for prediction of psychosis in patients with clinical high-risk syndromes and recent-onset depression. *JAMA Psychiatry* **78**, 195–209 (2021).
39. Kahn, R. S. et al. Effectiveness of antipsychotic drugs in first-episode schizophrenia and schizophreniform disorder: an open randomised clinical trial. *Lancet* **371**, 1085–1097 (2008).
40. Tognin, S. et al. Towards precision medicine in psychosis: benefits and challenges of multimodal multicenter studies - PSYSCAN: translating neuroimaging findings from research into clinical practice. *Schizophr. Bull.* **46**, 432–441 (2020).
41. Slot, M. I. E. et al. A naturalistic cohort study of first-episode schizophrenia spectrum disorder: a description of the early phase of illness in the PSYSCAN cohort. *Schizophr. Res.* **266**, 237–248 (2024).
42. Hall, R. C. W. Global assessment of functioning: a modified scale. *Psychosomatics* **36**, 267–275 (1995).
43. Golay, P. et al. Six months functional response to early psychosis intervention program best predicts outcome after three years. *Schizophr. Res.* **238**, 62–69 (2021).
44. Amminger, G. P., Schäfer, M. R., Schlögelhofer, M., Klier, C. M. & McGorry, P. D. Longer-term outcome in the prevention of psychotic disorders by the Vienna omega-3 study. *Nat. Commun.* **6**, 6–12 (2015).
45. Scott, J. et al. Clinical staging in psychiatry: a cross-cutting model of diagnosis with heuristic and practical value. *Br. J. Psychiatry* **202**, 243–245 (2013).
46. Austin, S. F., Hjorthøj, C., Baagland, H., Simonsen, E. & Dam, J. Investigating personal and clinical recovery processes in people with first episode psychosis. *Early Interv. Psychiatry* **16**, 1102–1111 (2022).
47. Dazzan, P. et al. Symptom remission at 12 weeks strongly predicts long-term recovery from the first episode of psychosis. *Psychol Med.* **50**, 1452–1462 (2020).
48. Simonsen, C. et al. Early clinical recovery in first-episode psychosis: Symptomatic remission and its correlates at 1 year follow-up. *Psychiatry Res.* **254**, 118–125 (2017).
49. Sheehan, D. V. et al. The mini-international neuropsychiatric Interview (MINI): the development and validation of a structured diagnostic psychiatric interview for DSM-IV and ICD-10. *J. Clin. Psychiatry* **59**, 22–33 (1998).
50. First, M. B., Spitzer, R. L., Gibbon, M. & Williams, J. B. W. *Structured Clinical Interview for DSM-IV-TR Axis I Disorders, Research Version, Patient Edition (SCID-I/P)*. (Biometrics Research, New York State Psychiatric Institute, New York, 2002).
51. Collins, G. S. et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ* **385**, q902 (2024).
52. Koutsouleris, N., Vetter, C. & Wiegand, A. *Neurominer [Computer software]*. https://github.com/neurominer-git/NeuroMiner_1.2 (2023).
53. Golland, P. & Fischl, B. Permutation tests for classification: towards statistical significance in image-based studies. In *Biennial international conference on information processing in medical imaging* (eds. Taylor, C. & Noble, J. A.) 330–341 (Springer, 2003).
54. Steyerberg, E. W. & Harrell, F. E. Prediction models need appropriate internal, internal-external, and external validation. *J. Clin. Epidemiol.* **69**, 245–247 (2016).
55. IBM Corp. *IBM SPSS Statistics for Windows, Version 29.0.* https://www.ibm.com/spss. (2022).
56. Gómez-Verdejo, V., Parrado-Hernández, E. & Tohka, J. Sign-consistency based variable importance for machine learning in brain imaging. *Neuroinformatics* **17**, 593–609 (2019).

ACKNOWLEDGEMENTS

We acknowledge the contributions of the entire EUFEST and PSYSCAN consortia. EUFEST was funded by the European Group for Research in Schizophrenia (EGRIS) with grants from AstraZeneca, Pfizer and Sanofi Aventis. PSYSCAN was funded as part of the European Funding 7th Framework Programme (grant agreement no 603196).

AUTHOR CONTRIBUTIONS

M.S. and M.F.U.C. contributed equally to this work (co-first authorship). R.S.K. and N.K. contributed equally to this work (co-last authorship). R.S.K. and W.W.F. developed the overall study design of the EUFEST study, acquired funding and provided the clinical data. R.S.K. and P.McG. developed the overall study design of the PSYSCAN study, acquired funding and provided the clinical data. H.H.v.H. and I.W.v.R. contributed to the study management of PSYSCAN. The PSYSCAN consortium contributed to the data collection of PSYSCAN. N.K., M.S., M.F.U.C. H.H.H., I.W.R., D.D. and R.S.K. contributed to developing the study design and approach for data analysis of the present work. M.S. and M.F.U.C. contributed to the data analysis and writing of the paper. N.K., D.D., R.S.K., P.McG., H.H.v.H. and I.W.v.R. provided feedback on the paper and made revisions, as appropriate. All authors reviewed the manuscript and approved the final version.

COMPETING INTERESTS

N.K. received honoraria for talks presented at education meetings organized by Otsuka/Lundbeck. W.W.F. has received grants from Lundbeck and Otsuka and lecture honoraria from Sumitomo-Pharma and Forum Medizinische Fortbildung. S.G. received advisory board/consultant fees from the following drug companies: Angelini, Boehringer Ingelheim Italia, Gedeon Richter-Recordati, Janssen Pharmaceutica NV and ROVI. S.G. received honoraria/expenses from the following drug companies: Angelini, Gedeon Richter-Recordati, Janssen Australia and New Zealand, Janssen Pharmaceutica NV, Janssen-Cilag, Lundbeck A/S, Lundbeck Italia, Otsuka, Recordati Pharmaceuticals, ROVI, Sunovion Pharmaceuticals. B.Y.G. has been the leader of a Lundbeck Foundation Centre of Excellence for Clinical Intervention and Neuropsychiatric Schizophrenia Research (CINS) (January 2009–December 2021), which was partially financed by an independent grant from the Lundbeck Foundation based on international review and partially financed by the Mental Health Services in the Capital Region of Denmark, the University of Copenhagen, and other foundations. All grants are the property of the Mental Health Services in the Capital Region of Denmark and administrated by them. The other authors declare no competing interests.

ADDITIONAL INFORMATION

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41537-024-00505-w.

Correspondence and requests for materials should be addressed to Margot I. E. Slot or René S. Kahn.

THE PSYSCAN CONSORTIUM

LONDON Philip McGuire[4,26], Stefania Tognin[26], Paolo Fusar-Poli[2,26,28,29], Matthew Kempton[26], Alexis E. Cullen[26,30], Gemma Modinos[7], Kate Merritt[26,31], Andrea Mechelli[26], Paola Dazzan[7], George Gifford[26], Natalia Petros[26], Mathilde Antoniades[26], Andrea De Micheli[26], Sandra Vieira[26], Tom Spencer[26], Zhaoying Yu[26], Dominic Oliver[26,32], Fiona Coutts[26], Emily Hird[26,33] and Helen Baldwin[26,34]

UTRECHT Rene Kahn[1,3], Arija Maat[1], Erika van Hell[1], Inge Winter[1,3,4] and Margot I. E. Slot[1,54]✉

AMSTERDAM Lieuwe de Haan[8] and Frederike Schirmbeck[8]

CANTABRIA Benedicto Crespo-Facorro[9,10], Diana Tordesillas-Gutierrez[9,10], Esther Setien-Suero[9,10], Rosa Ayesa-Arriola[9,10], Paula Suarez-Pinilla[9,10] and Victor Ortiz Garcia-de la foz[9,10]

COPENHAGEN Birte Glenthøj[11,12], Mikkel Erlang Sørensen[11], Bjørn H. Ebdrup[11,12], Jayachandra Mitta Raghava[11,12] and Egill Rostrup[11,35]

EDINBURGH Stephen M. Lawrie[13]

GALWAY Colm McDonald[14], Brian Hallahan[14], Dara M. Cannon[14], James McLoughlin[14] and Martha Finnegan[14]

HEIDELBERG Oliver Gruber[15], Anja Richter[15] and Bernd Krämer[15]

MAASTRICHT Thérèse van Amelsvoort[16], Bea Campforts[16], Machteld Marcelis[16] and Claudia Vingerhoets[16]

MADRID Celso Arango[17], Covadonga M. Díaz-Caneja[17], Miriam Ayora[17], Joost Janssen[17], Mara Parellada[17], Jessica Merchán-Naranjo[17], Roberto Rodríguez-Jiménez[36] and Marina Díaz-Marsá[37]

MARBURG Tilo Kircher[18], Irina Falkenberg[18], Florian Bitsch[18] and Jens Sommer[18]

MELBOURNE Barnaby Nelson[5,6], Patrick McGorry[5,6], Paul Amminger[5,6], Christos Pantelis[5,6], Meredith McHugh[5,6] and Jessica Spark[5,6]

NAPLES Silvana Galderisi[19], Armida Mucci[19], Paola Bucci[19], Giuseppe Piegari[19], Daria Pietrafesa[19], Alessia Nicita[19] and Sara Patriarca[19]

TEL HASHOMER Mark Weiser[20,21], Linda Levi[20] and Yoav Domany[20]

VIENNA Gabriele Sachs[22], Matthäus Willeit[22], Marcena Lenczowska[22], Ullrich Sauerzopf[22], Ana Weidenauer[22], Julia Furtner[38] and Daniela Prayer[22]

ZURICH Anke Maatz[38,39], Matthias Kirschner[23,24], Achim Burrer[24], Philipp Stämpfli[24], Naemi Huber[24], Stefan Kaiser[23] and Wolfram Kawohl[40]

SAO PAULO Rodrigo Bressan[41], André Zugman[41], Ary Gadelha[41] and Graccielle Rodrigues da Cunha[41]

SEOUL Jun Soo Kwon[42,43,44], Kang Ik Kevin Cho[44,45], Tae Young Lee[42,46], Minah Kim[42,43], Sun-Young Moon[42,47] and Silvia Kyungjin Lho[42,48]

TORONTO Romina Mizrahi[49], Michael Kiang[50,51], Cory Gerritsen[51,52], Margaret Maheandiran[51], Sarah Ahmed[51,53], Ivana Prce[51] and Jenny Lepock[51,53]

[28]Department of Brain and Behavioral Sciences, University of Pavia, Pavia, Italy. [29]Outreach and Support in South-London (OASIS) service, South London and Maudlsey (SLaM) NHS Foundation Trust, London, UK. [30]Division of Insurance Medicine, Department of Clinical Neuroscience, Karolinska Institutet, Stockholm, Sweden. [31]MRC Centre for Neurodevelopmental Disorders, King's College London, London, UK. [32]Department of Psychiatry, Division of Medical Sciences, University of Oxford, Warneford Hospital, OX3 7JX Oxford, UK. [33]Institute of Cognitive Neuroscience, Alexandra House, 17 Queen Square, London WC1N 3AZ, UK. [34]Health Service and Population Research (HSPR), Institute of Psychiatry, Psychology & Neuroscience, King's College London, De Crespigny Park, Denmark 458 Hill, London SE5 8AF, UK. [35]Functional Imaging Unit (FIUNIT), Rigshospitalet

Glostrup, University of Copenhagen, Glostrup, Denmark. [36]Department of Psychiatry, Instituto de Investigación Sanitaria Hospital 12 de Octubre (imas 12), CIBERSAM, ISCIII, School of Medicine, Universidad Complutense, Madrid, Spain. [37]Department of Psychiatry, Instituto de Investigación Hospital Clínico San Carlos (IdISSC), CIBERSAM, ISCIII, School of Medicine, Universidad Complutense, Madrid, Spain. [38]Medical University of Vienna, Department of Biomedical Imaging and Image-guided Therapy Währingergürtel 18-20, 1090 Vienna, Austria. [39]Department of Adult Psychiatry and Psychotherapy, Psychiatric University Clinic Zurich and University of Zurich, Zurich, Switzerland. [40]Department for Psychiatry and Psychotherapy, Psychiatric Services Aargau, Brugg, Switzerland. [41]Department of Psychiatry, Interdisciplinary Lab for Clinical Neurosciences (LiNC), Universidade Federal de Sao Paulo (UNIFESP), Sao Paulo, Brazil. [42]Department of Psychiatry, Seoul National University College of Medicine, 101 Daehakno, Jongno-gu, Seoul, Republic of Korea. [43]Department of Neuropsychiatry, Seoul National University Hospital, 101 Daehakno, Jongno-gu, Seoul, Korea. [44]Department of Brain and Cognitive Sciences, Seoul National University College of Natural Sciences, Gwanakro1, Gwanak-gu, Seoul, Republic of Korea. [45]Department of Psychiatry, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA. [46]Department of Psychiatry, Chonnam National University Hospital, Gwangju, Republic of Korea. [47]Department of Public Health Medical Services, Seoul National University Bundang Hospital, Seongnam, Republic of Korea. [48]Department of Psychiatry, Seoul Metropolitan Government-Seoul National University Boramae Medical Center, Seoul, Republic of Korea. [49]Department of Psychiatry, McGill University, Montreal, Canada. [50]Department of Psychiatry, University of Toronto, 250 College St 8th Floor, Toronto, M5T 1R8 Ontario, Canada. [51]Centre for Addiction and Mental Health, 250 College Street, Toronto M5T 1R8 Ontario, Canada. [52]Department of Psychology, University of Toronto, 100 St. George Street 4th Floor, Toronto, Ontario M5S 3G3, Canada. [53]Institute of Medical Science, University of Toronto, 1 King's College Circle Room 2374, Toronto, Ontario M5S 1A8, Canada.