


ARTICLE

Open Access

Clinical-learning versus machine-learning for transdiagnostic prediction of psychosis onset in individuals at-risk

Paolo Fusar-Poli^{1,2,3,4}, Dominic Stringer⁵, Alice M. S. Durieux⁵, Grazia Rutigliano¹, Ilaria Bonoldi¹, Andrea De Micheli¹ and Daniel Stahl⁵ 

Abstract

Predicting the onset of psychosis in individuals at-risk is based on robust prognostic model building methods including a priori clinical knowledge (also termed clinical-learning) to preselect predictors or machine-learning methods to select predictors automatically. To date, there is no empirical research comparing the prognostic accuracy of these two methods for the prediction of psychosis onset. In a first experiment, no improved performance was observed when machine-learning methods (LASSO and RIDGE) were applied—using the same predictors—to an individualised, transdiagnostic, clinically based, risk calculator previously developed on the basis of clinical-learning (predictors: age, gender, age by gender, ethnicity, ICD-10 diagnostic spectrum), and externally validated twice. In a second experiment, two refined versions of the published model which expanded the granularity of the ICD-10 diagnosis were introduced: ICD-10 diagnostic categories and ICD-10 diagnostic subdivisions. Although these refined versions showed an increase in apparent performance, their external performance was similar to the original model. In a third experiment, the three refined models were analysed under machine-learning and clinical-learning with a variable event per variable ratio (EPV). The best performing model under low EPVs was obtained through machine-learning approaches. The development of prognostic models on the basis of a priori clinical knowledge, large samples and adequate events per variable is a robust clinical prediction method to forecast psychosis onset in patients at-risk, and is comparable to machine-learning methods, which are more difficult to interpret and implement. Machine-learning methods should be preferred for high dimensional data when no a priori knowledge is available.

Introduction

Under standard care, outcomes of psychosis are poor¹. While early interventions at the time of a first psychotic episode are associated with some clinical benefits², they are not effective at preventing relapses² or reducing the duration of untreated psychosis (DUP)³; preventive interventions in individuals at clinical high risk for psychosis (CHR-P)⁴ may be an effective complementary

strategy. According to the World Health Organization, preventive strategies for mental disorders are based on the classification of the prevention of physical illness as universal, selective or indicated (targeted at the general public, those with risk factors, and those with minimal signs or symptoms of mental disorders respectively, as described by Gordon et al.) and on the classic public health classification as primary, secondary or tertiary (seeking to prevent the onset of a mental disorder, lower the rate of established disorder or reduce disability and relapses, respectively⁵). Universal, selective and indicated preventive interventions are “included within primary prevention in the public health classification” (page 17 in ref. ⁵). Since CHR-P individuals show attenuated

Correspondence: Paolo Fusar-Poli (paolo.fusar-poli@kcl.ac.uk)

¹Early Psychosis: Interventions and Clinical-detection (EPIC) lab, Department of Psychosis Studies, Institute of Psychiatry, Psychology & Neuroscience, King's College London, London, UK

²Department of Brain and Behavioural Sciences, University of Pavia, Pavia, Italy
Full list of author information is available at the end of the article.

© The Author(s) 2019



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

symptoms of psychosis coupled with help-seeking behaviour⁶ and functional impairments⁷, interventions in these individuals are defined as indicated primary prevention of psychosis. The conceptual and operational framework that characterises the CHR-P paradigm has been reviewed elsewhere^{8,9}. The empirical success of the CHR-P paradigm is determined by the concurrent integration of three core components: efficient detection of cases at-risk, accurate prognosis and effective preventive treatment^{10,11}. The underpinning methodology for each of these components is based on risk-prediction models¹². Unfortunately, a recent methodological review concluded that most of the CHR-P prediction modelling studies are of low quality, largely because they employ stepwise variable selection without proper internal and external validation¹³. These approaches overfit the data (i.e. the model learns the noise instead of accurately predicting unseen data¹⁴), inflate the estimated prediction performance on new cases and produce biased prognostic models that result in poor clinical utility¹⁴. Beyond stepwise model selection, overfitting can also occur when the number of events (e.g. number of at-risk patients who will develop psychosis over time) per variable (e.g. degree of freedoms of predictors of psychosis onset in at-risk patients) is low (event-per-variable, EPV <20^{14,15}). Low EPVs are frequently encountered in the CHR-P literature because the onset of psychosis in these samples is an infrequent, heterogeneous event (cumulating to 20% at 2-years, (eTable 4 in ref. 16; depending on the sampling strategies)^{17–20}.

A first approach to overcome these caveats is to use a priori clinical-learning or knowledge to identify a few robust predictors to be used in risk-prediction models¹³: it may be possible to use umbrella reviews (i.e. reviews of meta-analyses and systematic reviews that incorporate a stratification of the evidence²¹) on epidemiological risk/protective factors for psychosis²²). Because the selection of predictors would be limited in number (preserving the EPV¹⁴) and independent of the data on which the model is then tested, overfitting issues would be minimised¹³. For example, a recent risk estimation model has used a priori clinical-learning to select a few predictors of psychosis onset in CHR-P individuals²³. The prognostic model developed was robust and has already received several independent external replications²⁴. A second, increasingly popular approach is to bypass any clinical reasoning and instead use machine-learning procedures to select the predictors automatically²⁵: machine-learning studies have developed and internally validated models to stratify risk enrichment in individuals undergoing CHR-P assessment¹⁸ and functional outcomes in CHR-P samples²⁶. Machine-learning methods promise much to the CHR-P field because of their potential to assess a large number of predictors and to better capture non-linearities and

interactions in data; there is great confidence that they will outperform model-building based on clinical learning²⁵. Yet, modern machine-learning methods may not be a panacea²⁷, particularly because of the lack of empirical research comparing machine-learning vs clinical-learning theory-driven methods for the prediction of psychosis. The current manuscript advances knowledge by filling this gap.

Here we use a transdiagnostic, prognostic model that has been developed by our group using a priori meta-analytical clinical knowledge (hereafter clinical-learning)²⁸. The predictors used were collected as part of the clinical routine: age, gender, ethnicity, age by gender and ICD-10 index diagnostic spectrum. The model is cheap and “transdiagnostic”²⁹ because it can be applied at scale across several ICD-10 index diagnoses to automatically screen mental health trusts. This prognostic model has been externally validated twice^{28,30}, and is under pilot testing for real-world clinical use¹¹.

In the first experiment, we apply a machine-learning method to the same transdiagnostic individualised prognostic model and test the hypothesis that machine-learning methods produce models with better prediction accuracy than clinical-learning approach when the EPV is adequate. In the second experiment, we expand the granularity of the ICD-10 index diagnosis predictor and test the hypothesis that the use of more specific diagnostic specifications improves prognostic performance. In the third experiment, we test the hypothesis that machine-learning delivers better predicting prognostic models than clinical-learning under different models’ specifications, and in the specific scenario of low EPVs. Overall, this study provides much needed empirical research to guide prediction modelling strategies in early psychosis.

Materials and methods

Data source

Clinical register-based cohort selected through a Clinical Record Interactive Search (CRIS) tool³¹.

Study population

All individuals accessing South London and Maudsley (SLaM) services in the period 1 January 2008–31 December 2015, and who received a first ICD-10 index primary diagnosis of any non-organic and non-psychotic mental disorder (with the exception of Acute and Transient Psychotic Disorders, ATPDs) or a CHR-P designation (which is available in the whole SLaM through the Outreach And Support In South-London -OASIS- CHR-P service³²), were initially considered eligible. The ATPD group is diagnostically³³ and prognostically³⁴ similar to the Brief Limited Intermittent Psychotic Symptom (BLIPS) subgroup of the ARMS construct and to the Brief Limited Psychotic Symptoms (BIPS) subgroup of the

Structured Interview for the Psychosis-Risk Syndrome (SIPS; for details on these competing operationalisation see eTable 1 published in ref. ³⁴) and previous publications on the diagnostic and prognostic significance of short-lived psychotic disorders^{33,35,36}.

Those who developed psychosis in the three months immediately following the first index diagnosis were excluded. As previously detailed, this lag period was chosen to allow patients sufficient time after their index diagnosis to meet the ICD-10 duration criterion for ATPDs. Since we did not employ a structured assessment at baseline (see limitation), this lag period was also used to be conservative and exclude individuals who were underreporting psychotic symptoms at baseline (false transition to psychosis).

Ethical approval for the study was granted³¹.

Study measures

The outcome (risk of developing any ICD-10 non-organic psychotic disorder), predictors (index ICD-10 diagnostic spectrum, age, gender, ethnicity, and age by gender), and time to event were automatically extracted using CRIS³¹.

Statistical analyses

The original study was conducted according to the REporting of studies Conducted using Observational Routinely-collected health Data (RECORD) Statement³⁷.

Experiment 1: Machine-learning vs clinical-learning with adequate EPV for the prediction of psychosis

Development and validation of the original model (M1, diagnostic spectra) followed the guidelines of Royston et al.³⁸, Steyerberg et al.³⁹ and the Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD)⁴⁰. The details of model development and external validation have been presented previously²⁸. Briefly, predictors (ICD-10 diagnostic spectrum, age, gender, ethnicity, and age by gender interaction) were preselected on the basis of meta-analytical clinical knowledge⁴¹ as recommended¹³. The ICD-10 diagnostic spectrum was defined by all of the ten ICD-10 blocks (acute and transient psychotic disorders, substance abuse disorders, bipolar mood disorders, non-bipolar mood disorders, anxiety disorders, personality disorders, developmental disorders, childhood/adolescence onset disorders, physiological syndromes and mental retardation²⁸), with the exclusion of psychotic and organic mental disorders, and by CHR-P designation⁸. Accordingly, the diagnostic predictor of M1 encompassed 11 different levels. All other predictors together contributed 7 degrees of freedom, for a total of 18 degrees of freedom. Cox proportional hazards multivariable complete-case analyses were used to evaluate the effects of preselected

predictors on the development of non-organic ICD-10 psychotic disorders, and time to development of psychosis. Non-random split-sample by geographical location was used to create a development and external validation dataset⁴⁰. Performance diagnostics of individual predictor variables in the derivation dataset were explored with Harrell's C-index³⁸, which can be interpreted as a summary measure of the areas under the time-dependent ROC curve⁴². A value of $C = 0.5$ corresponds to a purely random prediction whereas $C = 1$ corresponds to perfect prediction. The model was then externally validated in the independent database from SLAM²⁸, and subsequently in another NHS Trust (Camden and Islington)³⁰. In the SLAM derivation database there were 1001 events (EPV 1001/18 = 55.61), and in the SLAM validation database there were 1010 events, both of which exceed the cut-off of 100 events required for reliable external validation studies⁴³.

In experiment 1, we tested the hypothesis that even when EPVs are above the recommended threshold and predictors are the same, machine-learning would outperform clinical-learning methods. Machine learning methods automate model building by learning from data with minimal human intervention⁴⁴; the best model is typically selected by assessing the prediction accuracy of unseen (hold-out) data for example using cross-validation methods⁴⁵. This is a key difference from classical statistical inferential methods, where the quality of a model is assessed by the sample used to estimate the model. Machine-learning methods typically introduce a regularisation term into the model to avoid overfitting, and this term usually imposes a penalty on complex models to reduce sample variance⁴⁵.

In our study we used regularised regression methods (also called penalised or shrinkage regression methods) as relatively simple, but often powerful machine learning methods which compare competitively to more complex machine learning methods like random forest or support vector machines^{46–48}. We chose regularised regression methods to enhance interpretability of the final model, in particular compared to models developed through clinical learning. It is important for clinicians to interpret prognostic models to gain knowledge and to detect their potential biases and limitations in real-world use⁴⁹. Regularised regression fits generalised linear models, for which the sizes of the coefficients are constrained to reduce overfitting. Two common regularised regression approaches to be considered in this study are RIDGE⁵⁰ and LASSO⁵¹. The primary difference between RIDGE and LASSO is that RIDGE regression constrains the sum of squares of the coefficients, whereas LASSO constrains the sum of absolute values of the coefficients⁴⁵. Unlike RIDGE, LASSO shrinks the coefficient to zero and thus performs an automatic selection of predictors. The degree

of constraint (or penalty) is determined by automated computer-intensive grid searches of tuning parameters. Because constraints depend on the magnitude of each variable, it is necessary to standardise variables. The final tuning parameter is chosen as the one which maximises a measure of prediction accuracy of unseen (hold-out) data using, for example, cross-validation methods⁴⁵.

Therefore, in experiment 1, we applied RIDGE and LASSO to the original unregularized Cox regression model in the same database to estimate their apparent and external performance (Harrell's C) in the derivation and validation datasets respectively. Their difference was then used to estimate the model's optimism.

Experiment 2: Diagnostic subdivisions vs diagnostic categories vs diagnostic spectra for the prediction of psychosis

We developed two refined prognostic models, M2 and M3, which differed from the original M1 model (diagnostic spectra, e.g. F30-F39 Mood [affective] disorders) by employing two expanded definitions of the predictor ICD-10 index diagnosis (the strongest predictor of the model^{28,30}). The model M2 (diagnostic categories) expanded the M1 model by adopting the 62 ICD-10 diagnostic categories—excluding psychotic and organic mental disorders—rather than the broader spectra (e.g. F30 manic episode, F31 Bipolar affective disorders etc.). The model M3 (diagnostic subdivisions) further expanded the M2 model by including all of the 383 specific ICD-10 diagnostic subdivisions of non-organic and non-psychotic mental disorder (e.g. F30.0 hypomania, F30.1 mania without psychotic symptoms, F30.2 mania with psychotic symptoms, F30.8 other manic episodes, F30.9 manic episode unspecified). From a clinical point of view, these refined models reflect the potential utility of specific vs block vs spectrum diagnostic formulations for the prediction of psychosis onset in at-risk individuals. The two previous independent replications of the original M1 model confirmed that the clinicians' pattern recognition of key diagnostic spectra is useful from a clinical prediction point of view. Thus, experiment 2a tested the clinical hypothesis that the use of more granular and specific ICD-10 index diagnoses would eventually improve the performance of the initial M1 model. The performance of the M1, M2 and M3 models was first reported in the derivation and validation dataset. In a subsequent stage, the model's performance (Harrell's C) was compared across each pair within the external validation dataset.

Experiment 3a and 3b. Machine-learning vs clinical-learning under variable EPVs

From a statistical point of view, increasing the number of levels of the ICD-10 diagnoses from M1 ($n = 10$) to M2 ($n = 62$) to M3 ($N = 383$) (plus the CHR designation), decreases the EPV from M1 to M2 to M3 respectively,

increasing the risk of overfitting in unregularised regression models (in particular when the EPV is lower than 20⁵²).

During experiment 3a, we tested the hypothesis that machine-learning would increasingly outperform clinician learning methods with decreasing EPVs. First, we compared the apparent performance of M1, M2, M3 in the whole dataset using RIDGE and LASSO versus unregularised Cox regression. Second, we compared the internal performance of M1, M2 and M3 in the whole dataset using ten-fold cross-validation repeated 100 times and taking the median Harrell's C across the 100 repetitions, again using RIDGE, LASSO versus unregularized Cox regression. We used the whole dataset because the refined M2 and M3 models have adopted different specifications of the ICD-10 diagnoses that were not always present in both derivation and validation datasets (in which case it would not have been possible to test the same model). In the light of the decreased EPVs we expected RIDGE and LASSO to perform better for M3 than for M2 than for M1, respectively⁴⁵.

In experiment 3b, we further assessed the impact of varied sample size and degree of EPV on the prognostic performance of the model M1 under machine-learning vs clinical-learning, without the confounding effect of including more potentially informative predictors. We randomly selected samples of different sizes from the derivation dataset and then fitted the machine-learning vs clinical-learning approaches to these samples. We then assessed the prediction accuracy in the external validation dataset. For each sample size, the results of ten repetitions with different random samples were averaged, and the median Harrell's C reported for both the derivation (apparent) and validation datasets. Samples sizes were 500, 1000, 2000 and 5000.

All analyses were conducted in STATA 14 and R 3.3.0. using the user-written R packages "Coxnet" for the regularised Cox regression models and "Hmisc" to calculate Harrell's C. The difference between two C's were calculated using the STATA package "Somersd" and the R package "Rms". Compute code is available from the authors (DS) upon request.

Results

Sociodemographic and clinical characteristics of the sample

91199 patients receiving a first index diagnosis of non-organic and non-psychotic mental disorder within SLaM in the period 2008–2015 fulfilled the study inclusion criteria and were included in the derivation (33820) or validation (54716) datasets. The baseline characteristics of the study population, as well as the derivation and validation datasets, are presented in Table 1²⁸. The mean follow-up was 1588 days (95% CI 1582–1595) with no

Table 1 Sociodemographic characteristics of the study population, including the derivation and validation dataset²⁸

	Derivation dataset		Validation dataset	
	Mean	SD	Mean	SD
Age (years)^c	34.4	18.92	31.98	18.54
	Count	%	Count	%
Gender				
Male	17303	48.81	27302	49.9
Female	16507	51.16	27398	50.07
Missing	10	0.03	16	0.03
Ethnicity				
Black	6879	20.34	7023	12.84
White	18627	55.08	35392	64.68
Asian	1129	3.34	2608	4.77
Mixed	1306	3.86	1957	3.58
Other	3466	10.25	2084	3.81
Missing	2413	7.13	5652	10.33
ICD-10 Index spectrum diagnosis				
CHR-P ^a	314	0.93	50	0.09
Acute and transient psychotic disorders	553	1.64	725	1.33
Substance use disorders	7149	21.14	6507	11.89
Bipolar mood disorders	950	2.81	1526	2.79
Non-bipolar mood disorders	6302	18.63	8841	16.16
Anxiety disorders	8235	24.35	15960	29.17
Personality disorders	1286	3.8	2116	3.87
Developmental disorders	1412	4.18	3706	6.77
Childhood/adolescence onset disorders	4200	12.42	9629	17.6
Physiological syndromes	2555	7.55	4424	8.09
Mental retardation	864	2.55	1232	2.25

^aLambeth and Southwark, $n = 33820$ ^bCroydon and Lewisham, $n = 54716$ ^cNot an ICD-10 Index spectrum diagnosis

significant differences between the derivation and validation datasets.

Experiment 1: Machine-learning vs clinical-learning and adequate EPV for the prediction of psychosis

The first analysis compared M1 model performance developed with clinician learning (a priori knowledge)

against RIDGE and LASSO in both the derivation and validation dataset. Harrell's C on derivation set was virtually the same for all three methods on both derivation (~ 0.8) and external validation data sets (~ 0.79 , Table 2).

Experiment 2: diagnostic subdivisions vs diagnostic categories vs diagnostic spectra for the prediction of psychosis

The database included the majority of the non-organic and non-psychotic ICD-10 diagnostic categories (57 out of 62, 92% in M2), and diagnostic subdivisions (353 out of 383, 92% in M2).

In the derivation dataset (apparent performance¹⁴), the M3 model (Harrell's C 0.833) seemed to perform better, than the M2 model (Harrell's C 0.811) and better than the original M1 model (Harrell's C 0.8). However, this was due to overfitting of the M3 to the derivation data, as confirmed by the external validation. In fact, in the validation dataset, using all of the ICD-10 diagnostic subdivisions (M3) yielded a comparable model performance (about 0.79) to M1 and comparable to the model with the diagnostic categories (M2). The latter model (M2) showed statistically significant, superior performance compared to M1. However, the magnitude of the improvement of the Harrell's C of 0.007 was too small to be associated with meaningful clinical benefits (see Table 3).

Experiment 3a and 3b. Prognostic performance using machine-learning vs clinical-learning under variable EPVs

The results from experiment 3a showed that the clinical-learning and machine-learning methods delivered similar apparent prognostic performance (Table 4). After internal validation, Harrell's C slightly decreased, and M1, M2 and M3 models were all similar (approximately 0.8). There were again small differences between clinical-learning and machine-learning methods, which were more marked as EPV decreased.

In experiment 3b, Harrell's C for M1 in the derivation dataset increased with decreasing sample size. The increase was larger for clinical-learning (unregularized regression: from 0.8 to 0.9), and smaller for machine-learning (RIDGE and LASSO: 0.79–0.83, Fig. 1). The opposite pattern was then seen in the external validation dataset, where Harrell's C for M1 decreased with decreasing sample size. Hence, optimism (the difference between Harrell's C in the apparent sample and with internal validation) increased with smaller sizes. As sample size decreased, Harrell's C decreased slightly more when using clinical-learning (unregularized regression: from 0.79 to 0.67 if $N = 500$) than when using machine-learning (RIDGE regression: from 0.79 to 0.70 and LASSO regression: from 0.79 to 0.69).

Table 2 Experiment 1: prognostic accuracy (Harrell's C) for the original model (M1, diagnostic spectra) developed through Clinical-learning (a priori clinical knowledge) vs machine learning (LASSO and RIDGE). The EPV is >20 (55.6)

Method	Derivation Data Set (N = 33,820)			Validation Data Set (N = 54,716)			Optimism
	Harrell's C	SE	95% C.I.	Harrell's C	SE	95% C.I.	
Unregularized	0.800	0.008	0.784–0.816	0.791	0.008	0.775–0.807	0.009
Lasso	0.798	0.008	0.782–0.814	0.789	0.008	0.773–0.805	0.009
Ridge	0.810	0.008	0.794–0.826	0.788	0.008	0.772–0.804	0.022

Table 3 Experiment 2: prognostic performance of the revised models in the derivation dataset and the validation dataset, and their comparative performance

Model	Type of clustering of ICD-10 index diagnoses	Harrell's C	SE	95% CI	
Derivation dataset					
M1	Diagnostic spectra	0.800	0.008	0.784	0.816
M2	Diagnostic categories	0.811	0.008	0.795	0.824
M3	Diagnostic subdivisions	0.833	0.008	0.821	0.847
Validation dataset					
M1	Diagnostic spectra	0.791	0.008	0.776	0.807
M2	Diagnostic categories	0.797	0.008	0.782	0.812
M3	Diagnostic subdivisions	0.792	0.008	0.776	0.808
M2-M1		0.006	0.003	0.001	0.012
M3-M1		0.001	0.005	−0.009	0.011
M3-M2		−0.005	0.005	−0.015	0.004

All models include age, gender, age by gender, ethnicity and ICD-10 index diagnosis (refined as specified in the methods)

Discussion

This study compared clinical-learning vs machine-learning methods for the prediction of individuals at-risk for psychosis. The first experiment indicated that clinical-learning methods with a priori selection of predictors and adequate EPV produce robust prognostic models that are comparable to those obtained through regularised regression machine-learning methods. The second experiment indicated that there is no improvement in prognostic accuracy when specific ICD-10 diagnoses are employed instead of broad diagnostic spectra. The third experiment indicated that machine learning methods can deliver more robust prognostic models that clinical-learning methods when the sample size is small and the EPV low, although the benefits are modest in magnitude.

The first hypothesis of the current study was that machine-learning methods would generally outperform clinical-learning methods using the same set of predictors.

This was not verified in our study, because when RIDGE and LASSO methods were applied to the previously published transdiagnostic individualised risk estimation model, there was no substantial difference in prognostic performance. This suggests that when a prognostic model is built on strong clinical knowledge, has a large sample and an adequate EPV (in this case it was 56), the model can perform very well without the use of machine-learning methods. Machine-learning methods are not always necessary to obtain an accurate prediction of psychosis onset and do not necessarily improve the performance of prognostic models developed on a priori clinical knowledge. For example, a recently published supervised machine-learning study failed to demonstrate improved prediction of transition to psychosis when using baseline clinical information with no a priori knowledge⁵³, suggesting that a priori clinical knowledge remains very important for developing good prognostic models. Given a comparable accuracy, models developed through clinical-learning tend to be more straightforward and thus more likely to be interpreted, assessed and accepted, and implemented in clinical care (see below).

Our second hypothesis was that adding more information to the model by expanding the granularity of the ICD-10 index diagnosis would improve prognostic performance. The results showed no prognostic benefit to using specific ICD-10 diagnoses compared to broad diagnostic spectra for the prediction of psychosis in secondary mental health care. The diagnostic spectra employed by the original version of the transdiagnostic individualised risk calculator²⁸ are robust because they originate in prototypical descriptions containing a core phenomenological structure (gestalt) of the disorder and its polysymptomatic manifestations²⁹. Examination of overlaps of etiological factors between disorders confirms that higher level broad diagnostic constructs may be more valid and clinically useful categories than specific diagnostic categories⁵⁴. The prognostic utility of the ICD-10 diagnostic spectra is also in line with recent meta-analytical findings indicating that diagnostic spectra (e.g. psychosis) are relatively stable at the time of a first episode of psychosis⁵⁵. These diagnostic spectra are certainly not optimal, yet they do not

Table 4 Experiment 3a. Prognostic performance using machine-learning vs clinical-learning under variable EPVs

	Unregularized								
	M1 (diagnostic spectra)			M2 (diagnostic categories)			M3 (diagnostic subdivisions)		
	Cox Regression	LASSO	RIDGE	Cox Regression	LASSO	RIDGE	Cox Regression	LASSO	RIDGE
Apparent performance									
C index	0.800	0.793	0.790	0.811	0.799	0.803	0.827	0.812	0.813
SE	0.005	0.005	0.006	0.005	0.005	0.005	0.005	0.005	0.005
Internal validation performance									
C index	0.799	0.794	0.790	0.804	0.795	0.795	0.805	0.793	0.797
SE	0.017	0.017	0.017	0.017	0.017	0.017	0.017	0.017	0.017
Events	2011			2011			2011		
Degrees of freedom of predictors	18			63			226		
EPV	111.7			31.9			8.9		

Upper part of the table: apparent performance of M1-M3 models in the whole dataset. Bottom part of the table: internal performance in the whole dataset using nested 10-fold CV and taking median values with 100 repetitions
 EPV events per variables, calculated as the number of transitions to psychosis over the degrees of freedom of predictors. Categorical predictors are counted as the number of indicator categories they consist of (i.e. number of categories-7)

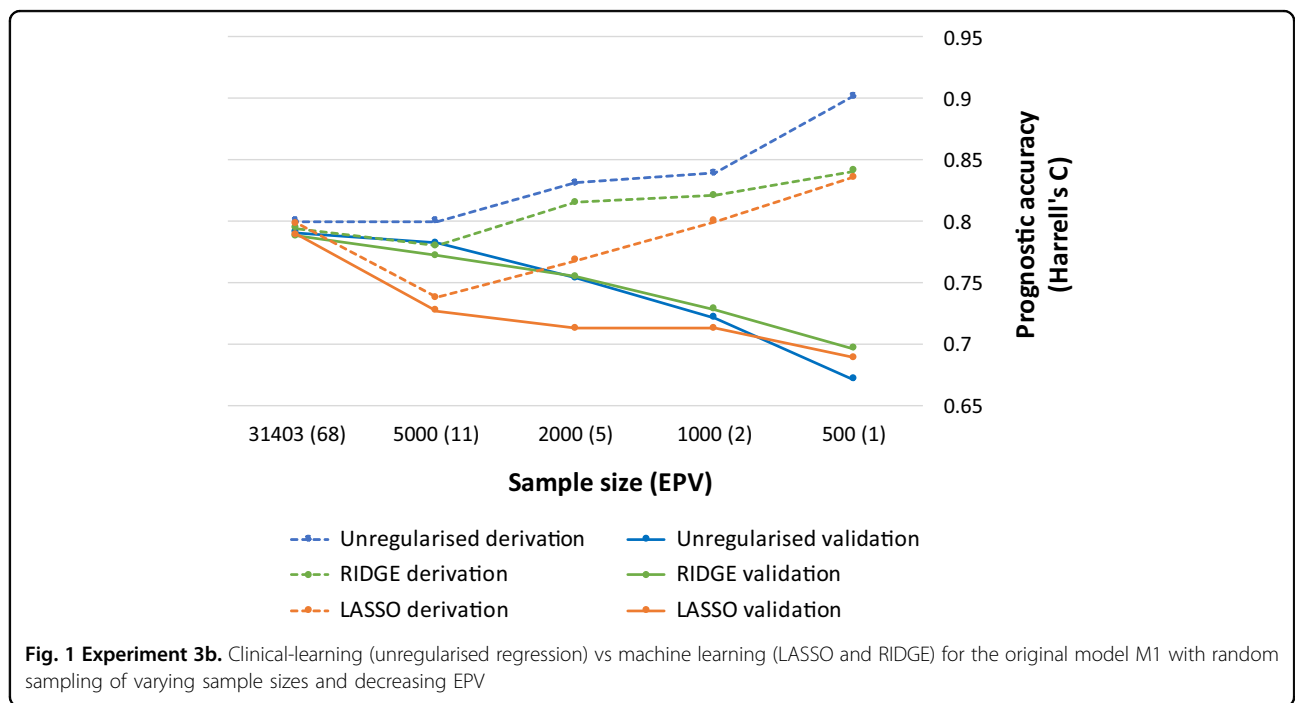


Fig. 1 Experiment 3b. Clinical-learning (unregularised regression) vs machine learning (LASSO and RIDGE) for the original model M1 with random sampling of varying sample sizes and decreasing EPV

present an insuperable barrier to scientific progress⁵⁶, and in terms of scalability in secondary mental health care⁵⁷ have yet to be beaten by other predictors of psychosis onset. Conversely, available clinical evidence indicates that the specific ICD-10 diagnoses are unreliable and unstable, and this may explain why their use is associated with overfitting problems and lack of

prognostic benefits⁵⁵. It is also possible that the small number of cases observed in some specific diagnostic categories may interfere with the efficacy of machine learning approaches.

The third hypothesis was that LASSO and RIDGE would perform better in the presence of either unstable (such as the specific ICD-10 diagnoses) or redundant

predictors, or infrequent events (low EPV); RIDGE is generally better with a small number of unstable predictors, and LASSO with a large number. This hypothesis was confirmed: the best performing model under low EPV and unstable predictors was obtained through machine-learning approaches¹³. However, the improvement in prognostic performance was modest, indicating that if strong predictors are known in advance through clinical-learning, it may be difficult to improve the model by adding many other variables which are more likely to be interpreted as noise, even when using penalized regression machine-learning methods. Notably, our study tested only two simple machine learning methods (RIDGE and LASSO), so we cannot exclude the possibility that prognostic improvements may have been larger if more complex machine learning methods (such as random forest or support vector machines for survival) have been used^{58,59}. However, Ploeg, Austin, and Steyerberg demonstrated that the development of robust models by machine-learning methods requires more cases-per-candidate predictors than traditional statistical methods when the dimensionality is not extremely high²⁷. Interestingly, even if large data sets are available, complex machine learning methods (i.e. random forests) only showed only minor improvement (at the expense of reduced interpretability and no automatic variable selection) over simple statistical models²⁷. This view was pragmatically supported by a recent systematic review which compared random forests, artificial neural networks, and support vector machines models to logistic regression. Across 282 comparisons, there was no evidence of superior performance of machine-over clinical-learning for clinical prediction modelling⁶⁰.

Not surprisingly, the prognostic tools used to date in the real world clinical routine of CHR-P services are still based on clinical-learning^{23,28}. However, in the current study, we could not test whether the addition of new multimodal predictors - beyond the clinical and socio-demographic ones—would improve the prognostic accuracy of psychosis onset. Some studies have suggested that the combination of clinical information with structural neuroimaging measures (such as gyrification and sub-cortical volumes) could improve prognostic accuracy⁶¹. However, available studies failed to provide convincing evidence that multimodal predictors under machine learning can substantially improve prognostic accuracy for predicting psychosis onset in patients at risk^{62,63}. Furthermore, complex models based on multimodal domains are constrained by logistical and financial challenges that can impede the ability to implement and scale these models in the real world. A potentially promising solution may be to adopt a sequential testing assessment to enrich the risk in a stepped framework, as demonstrated by our group with a simulation meta-analysis⁶⁴.

Interestingly, a recent machine-learning study on patients at-risk for psychosis confirmed that adding neuroimaging predictors to clinical predictors produced a 1.9-fold increase in prognostic certainty in uncertain cases of patients at-risk for psychosis²⁶.

Our study provides some conceptual and broad implications; although machine learning methods have attracted high expectations in the field^{25,65,66}, the enthusiasm may not be entirely substantiated in the field of psychosis. First, we have demonstrated that if robust a priori clinical knowledge is available, and if there are large sample sizes and EPVs, clinical-learning is a valid method to develop robust prognostic models. Clearly, a priori clinical knowledge may not always be available, and high dimensional databases with large sample sizes or strong signal to noise ratio may be needed to address the complexity of mental disorders. Under those circumstances, machine-learning methods can produce more robust prognostic models. Our study also provides support for this situation where detailed clinical information is not available; machine learning methods were able to identify models of similar prediction accuracy.

Second, the methodological, empirical and conceptual limitations of machine learning in psychiatry have not been completely addressed. Overoptimistic views, excessive faith in technology⁶⁷ and lack of knowledge of limitations of a specific methodology can lead to unrealisable promises⁶⁸. While machine learning methods can potentially achieve good predictive accuracy in high dimensional data when there is poor a priori knowledge, they tend to deliver “black-box” classifiers that provide very limited explanatory insights into psychosis onset⁶⁹. This is a fundamental limitation: without direct interpretability of a prognostic procedure, implementation in clinical practice may be limited⁶⁸. To have high impact and be adopted on a broader scale, a prognostic model must be accepted and understood by clinicians. Prediction models developed through clinical-learning are traditionally better understood by clinicians than machine learning models⁷⁰, while machine-learning models are challenging to evaluate and apply without a basic understanding of the underlying logic on which they are based⁷¹. A partial solution may be to incorporate a priori knowledge into machine-learning approaches⁷². Because of these issues, some authors argue that clinical-learning and reasoning will become even more critical to distil machine-learning and data-driven knowledge⁷³, and preliminary studies suggest that the combined use of theory-driven and machine learning approaches can be advantageous⁷⁴. There is a trend towards converting “big data” into “smart data” through contextual and personalised processing, allowing clinicians and stakeholders to make better decisions; our study supports such an approach⁷⁵.

Third, an additional pragmatic limitation is that for prediction models to ultimately prove useful, they must demonstrate impact⁷⁶—their use must generate better patient outcomes⁷⁰. Impact studies for machine-learning approaches in patients at-risk for psychosis are lacking. Rigorous tests on independent cohorts are critical requirements for the translation of machine-learning research to clinical applications⁷⁷. To our knowledge, the only study that has estimated the potential clinical benefit associated with the use of a prognostic model in secondary mental health care is our transdiagnostic individualised risk calculator analysis, which was based on clinical-learning²⁸. A recent review observed that although there are thousands of papers applying machine-learning algorithms to medical data, very few have contributed meaningfully to clinical care⁷⁸. Another recent empirical study focusing on the clinical impact of machine-learning in early psychosis concluded that the current evidence for the diagnostic value of these methods and structural neuroimaging should be reconsidered toward a more cautious interpretation⁷⁹.

Conclusions

Developing prognostic models on the basis of a priori clinical knowledge, large samples and adequate events per variable is a robust clinical prediction method for forecasting psychosis onset in patients at-risk. Under these circumstances, the prognostic accuracy is comparable to that obtained through machine-learning methods, which are more difficult to interpret and may present additional implementation challenges. The use of diagnostic spectra for transdiagnostic prediction of psychosis in secondary mental health care offers superior prognostic accuracy than the use of more specific diagnostic categories. Machine-learning methods should be considered in cases of high dimensional data when no a priori knowledge is available.

Acknowledgements

This study was supported by the King's College London Confidence in Concept award from the Medical Research Council (MRC) (MC_PC_16048) to PFP. This study also represents independent research part funded by the National Institute for Health Research (NIHR) Biomedical Research Centre at South London and Maudsley NHS Foundation Trust and King's College London. The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health and Social Care. The funders had no influence on the design, collection, analysis and interpretation of the data, writing of the report and decision to submit this article for publication.

Author details

¹Early Psychosis: Interventions and Clinical-detection (EPIC) lab, Department of Psychosis Studies, Institute of Psychiatry, Psychology & Neuroscience, King's College London, London, UK. ²Department of Brain and Behavioural Sciences, University of Pavia, Pavia, Italy. ³OASIS service, South London and Maudsley NHS Foundation Trust, London, UK. ⁴National Institute of Health Research – Mental Health – Translational Research Collaboration – Early Psychosis Workstream, London, UK. ⁵Department of Biostatistics and Health Informatics,

Institute of Psychiatry, Psychology & Neuroscience, King's College London, London, UK

Conflict of interest

The authors declare that they have no conflict of interest.

Publisher's note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Received: 9 February 2019 Revised: 3 May 2019 Accepted: 31 May 2019

Published online: 17 October 2019

References

1. Jaaskelainen, E. et al. A systematic review and meta-analysis of recovery in schizophrenia. *Schizophr. Bull.* **39**, 1296–1306 (2013).
2. Fusar-Poli, P., McGorry, P. & Kane, J. Improving outcomes of first episode psychosis. *World Psychiatry.* **16**, 251–265 (2017)
3. Oliver, D. et al. Can we reduce the duration of untreated psychosis? A meta-analysis of controlled interventional studies. *Schizophr. Bull.* **44**, 1362–1372 (2018).
4. Fusar-Poli, P. The clinical high-risk state for psychosis (CHR-P), Version II. *Schizophr. Bull.* **43**, 44–47 (2017).
5. WHO. Prevention of Mental Disorders. Effective Interventions and Policy Options. Geneva: Department of Mental Health and Substance Abuse; 2004. Contract No: ISBN 92 4 159215 X.
6. Falkenberg, I. et al. Why are help-seeking subjects at ultra-high risk for psychosis help-seeking? *Psychiatry Res.* **228**, 808–815 (2015).
7. Fusar-Poli, P. et al. Disorder, not just a state of risk: meta-analysis of functioning and quality of life in subjects at high clinical risk for psychosis. *Br. J. Psychiatry* **207**, 198–206 (2015).
8. Fusar-Poli, P. et al. Towards a standard psychometric diagnostic interview for subjects at ultra high risk of psychosis: CAARMS versus SIPS. *Psychiatry J.* **2016**, 7146341 (2016)
9. Fusar-Poli, P. et al. The psychosis high-risk state: a comprehensive state-of-the-art review. *JAMA Psychiatry* **70**, 107–120 (2013).
10. Fusar-Poli, P. Extending the benefits of indicated prevention to improve outcomes of first episode psychosis. *JAMA Psychiatry.* **74**, 667–668 (2017)
11. Fusar-Poli P. et al. Real-world implementation of a transdiagnostic risk calculator for the automatic detection of individuals at risk of psychosis in clinical routine: study protocol. *Front. Psychiatry.* **10**, 109 (2019)
12. Fusar-Poli, P. & Schultze-Lutter, F. Predicting the onset of psychosis in patients at clinical high risk: practical guide to probabilistic prognostic reasoning. *Evid. Based Ment. Health* **19**, 10–15 (2016).
13. Studerus, E., Rameyead, A. & Riecher-Rossler, A. Prediction of transition to psychosis in patients with a clinical high risk for psychosis: a systematic review of methodology and reporting. *Psychol. Med.* **47**, 1163–1178 (2017).
14. Fusar-Poli, P., Hijazi, Z., Stahl, D. & Steyerberg, E. W. The science of prognosis in psychiatry: a review. *JAMA Psychiatry.* **75**, 1289–1297(2018)
15. Austin, P. C. & Steyerberg, E. W. Events per variable (EPV) and the relative performance of different strategies for estimating the out-of-sample validity of logistic regression models. *Stat. Methods Med Res.* **26**, 796–808 (2017).
16. Fusar-Poli, P. et al. Heterogeneity of risk for psychosis within subjects at clinical high risk: meta-analytical stratification. *JAMA Psychiatry* **73**, 113–120 (2016).
17. Fusar-Poli, P. et al. The dark side of the moon: meta-analytical impact of recruitment strategies on risk enrichment in the clinical high risk state for psychosis. *Schizophr. Bull.* **42**, 732–743 (2016).
18. Fusar-Poli, P. et al. Deconstructing pretest risk enrichment to optimize prediction of psychosis in individuals at clinical high risk. *JAMA Psychiatry* **73**, 1260–1267 (2016).
19. Fusar-Poli, P. et al. Why transition risk to psychosis is not declining at the OASIS ultra high risk service: the hidden role of stable pretest risk enrichment. *Schizophr. Res.* **192**, 385–390 (2018). <https://doi.org/10.1016/j.schres.2017.06.015>. (Epub 20 Jul 2017).
20. Fusar-Poli, P. Why ultra high risk criteria for psychosis prediction do not work well outside clinical samples and what to do about it. *World Psychiatry* **16**, 212–213 (2017).

21. Fusar-Poli, P. & Radua, J. Ten simple rules for conducting umbrella reviews. *Evid. Based Ment. Health* **21**, 95–100 (2018).
22. Radua, J. et al. What causes psychosis? An umbrella review of risk and protective factors. *World Psychiatry* **17**, 49–66 (2018).
23. Cannon, T. D. et al. An individualized risk calculator for research in prodromal psychosis. *Am. J. Psychiatry*. **173**, 980–988 (2016)
24. Carrion, R. E. et al. Personalized prediction of psychosis: external validation of the NAPLS-2 psychosis risk calculator with the EDIPPP project. *Am. J. Psychiatry* **173**, 989–996 (2016).
25. Krystal, J. H. et al. Computational psychiatry and the challenge of schizophrenia. *Schizophr. Bull.* **43**, 473–475 (2017).
26. Koutsouleris, N. et al. Prediction models of functional outcomes for individuals in the clinical high-risk state for psychosis or with recent-onset depression: a multimodal, multisite machine learning analysis. *JAMA Psychiatry* **75**, 1156–1172 (2018).
27. van der Ploeg, T., Austin, P. C. & Steyerberg, E. W. Modern modelling techniques are data hungry: a simulation study for predicting dichotomous endpoints. *BMC Med. Res. Methodol.* **14**, 137 (2014).
28. Fusar-Poli, P. et al. Development and validation of a clinically based risk calculator for the transdiagnostic prediction of psychosis. *JAMA Psychiatry* **74**, 493–500 (2017).
29. Fusar-Poli, P. et al. Transdiagnostic psychiatry: a systematic review. *World Psychiatry*. **18**, 192–207 (2019)
30. Fusar-Poli, P. et al. Transdiagnostic risk calculator for the automatic detection of individuals at risk and the prediction of psychosis: second replication in an independent national health service trust. *Schizophr Bull.* **43**, 562–570 (2018). <https://doi.org/10.1093/schbul/sby070>.
31. Stewart, R. et al. The South London and Maudsley NHS Foundation Trust Biomedical Research Centre (SLAM BRC) case register: development and descriptive data. *BMC Psychiatry* **9**, 51 (2009).
32. Fusar-Poli, P., Byrne, M., Badger, S., Valmaggia, L. R. & McGuire, P. K. Outreach and support in south London (OASIS), 2001–2011: ten years of early diagnosis and treatment for young individuals at high clinical risk for psychosis. *Eur. Psychiatry* **28**, 315–326 (2013).
33. Fusar-Poli, P. et al. Diagnostic and prognostic significance of brief limited intermittent psychotic symptoms (BLIPS) in individuals at ultra high risk. *Schizophr. Bull.* **43**, 48–56 (2017).
34. Fusar-Poli, P. et al. Prognosis of brief psychotic episodes: a meta-analysis. *JAMA Psychiatry* **73**, 211–220 (2016).
35. Minichino, A. et al. Unmet needs in patients with brief psychotic disorders: too ill for clinical high risk services and not enough ill for first episode services. *Eur. Psychiatry*. **57**, 26–32 2018. <https://doi.org/10.1016/j.eurpsy.2018.12.006>. (Epub 15 Jan 2019).
36. Rutigliano, G. et al. Long term outcomes of acute and transient psychotic disorders: The missed opportunity of preventive interventions. *Eur. Psychiatry* **52**, 126–133 (2018).
37. Benchimol, E. I. et al. The REporting of studies Conducted using Observational Routinely-collected health Data (RECORD) statement. *PLoS Med.* **12**, e1001885 (2015).
38. Royston, P. & Altman, D. G. External validation of a Cox prognostic model: principles and methods. *BMC Med. Res. Methodol.* **13**, 33 (2013).
39. Steyerberg, E. W. et al. Assessing the performance of prediction models: a framework for traditional and novel measures. *Epidemiology* **21**, 128–138 (2010).
40. Collins, G. S., Reitsma, J. B., Altman, D. G. & Moons, K. G. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *Ann. Intern Med.* **162**, 55–63 (2015).
41. Kirkbride, J. B. et al. Incidence of schizophrenia and other psychoses in England, 1950–2009: a systematic review and meta-analyses. *PLoS ONE* **7**, e31660 (2012).
42. Schmid, M. & Potapov, S. A comparison of estimators to evaluate the discriminatory power of time-to-event models. *Stat. Med.* **31**, 2588–2609 (2012).
43. Collins, G. S., Ogundimu, E. O. & Altman, D. G. Sample size considerations for the external validation of a multivariable prognostic model: a resampling study. *Stat. Med.* **35**, 214–226 (2016).
44. Géron, A. Hands On Machine Learning With Scikit Learn And Tensorflow. (O'Reilly Media, Inc., 2017). <https://newbooksinpolitics.com/get/ebook.php?id=bRpYDgAAQBAJ>.
45. Hastie, T., Tibshirani, R. & Friedman, J. *Model assessment and selection. The elements of statistical learning: data mining, inference and prediction*. 2nd edn. (Springer, 2009)
46. Acharjee, A., Kloosterman, B., Visser, R. G. & Maliepaard, C. Integration of multi-omics data for prediction of phenotypic traits using random forest. *BMC Bioinforma.* **17**, 180 (2016).
47. Salvador, R. et al. Evaluation of machine learning algorithms and structural features for optimal MRI-based diagnostic prediction in psychosis. *PLoS ONE* **12**, e0175683 (2017).
48. Xie, R., Wen, J., Quitadamo, A., Cheng, J. & Shi, X. A deep auto-encoder model for gene expression prediction. *BMC Genomics.* **18**, 845 (2017).
49. Doshi-Velez, F. & Kim, B. Towards a rigorous science of interpretable machine learning. arXivorg [Internet]. (2017). <https://arxiv.org/abs/1702.08608#>.
50. Hoerl, A. & Kennard, R. Ridge regression: biased estimation for nonorthogonal problems. *Technometrics* **12**, 55–67 (1970).
51. Tibshirani, R. Regression shrinkage and selection via the LASSO. *J. R. Stat. Soc. Ser. B.* **58**, 267–288 (1996).
52. Ogundimu, E. O., Altman, D. G. & Collins, G. S. Adequate sample size for developing prediction models is not simply related to events per variable. *J. Clin. Epidemiol.* **76**, 175–182 (2016).
53. Mechelli, A. et al. Using clinical information to make individualized prognostic predictions in people at ultra high risk for psychosis. *Schizophr Res.* **184**, 32–38 2016. <https://doi.org/10.1016/j.schres.2016.11.047>. (Epub 4 Dec 2016).
54. Uher, R. & Zwickler, A. Etiology in psychiatry: embracing the reality of poly-gene-environmental causation of mental illness. *World Psychiatry* **16**, 121–129 (2017).
55. Fusar-Poli, P. et al. Diagnostic stability of ICD/DSM first episode psychosis diagnoses: Meta-analysis. *Schizophr. Bull.* **42**, 1395–1406 (2016).
56. Schizophrenia Working Group of the Psychiatric Genomics C. Biological insights from 108 schizophrenia-associated genetic loci. *Nature* **511**, 421–427 (2014).
57. Lawrie, S. M., O'Donovan, M. C., Saks, E., Burns, T. & Lieberman, J. A. Improving classification of psychoses. *Lancet Psychiatry* **3**, 367–374 (2016).
58. Ishwaran, H. & Lu, M. Standard errors and confidence intervals for variable importance in random forest regression, classification, and survival. *Stat. Med.* **38**, 558–582 (2019).
59. Van Belle, V., Pelckmans, K., Van Huffel, S. & Suykens, J. A. Support vector methods for survival analysis: a comparison between ranking and regression approaches. *Artif. Intell. Med.* **53**, 107–118 (2011).
60. Christodoulou, E. et al. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *J. Clin. Epidemiol.*, 110, 12–22 (2019)
61. 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).
62. Rameyad, A. et al. Prediction of psychosis using neural oscillations and machine learning in neuroleptic-naïve at-risk patients. *World J. Biol. Psychiatry* **17**, 285–295 (2016).
63. Pettersson-Yeo, W. et al. Using genetic, cognitive and multi-modal neuroimaging data to identify ultra-high-risk and first-episode psychosis at the individual level. *Psychol. Med.* **43**, 2547–2562 (2013).
64. Schmidt, A. et al. Improving prognostic accuracy in subjects at clinical high risk for psychosis: systematic review of predictive models and meta-analytical sequential testing simulation. *Schizophr Bull.* **43**, 375–388 (2017)
65. Vieira, S., Pinaya, W. H. & Mechelli, A. Using deep learning to investigate the neuroimaging correlates of psychiatric and neurological disorders: Methods and applications. *Neurosci. Biobehav. Rev.* **74**, 58–75 (2017).
66. Veronese, E., Castellani, U., Peruzzo, D., Bellani, M. & Brambilla, P. Machine learning approaches: from theory to application in schizophrenia. *Comput. Math. Methods Med.* **2013**, 867924 (2013).
67. Fusar-Poli, P., Broome, M., Barale, F. & Stanghellini, G. Why is psychiatric imaging clinically unreliable? Epistemological perspectives in clinical neuroscience. *Psychother. Psychosom.* **78**, 320–321 (2009).
68. Fusar-Poli, P. & Meyer-Lindenberg, A. Forty years of structural imaging in psychosis: promises and truth. *Acta Psychiatr. Scand.* **134**, 207–224 (2016).
69. Brodersen, K. H. et al. Dissecting psychiatric spectrum disorders by generative embedding. *Neuroimage Clin.* **4**, 98–111 (2014).
70. Cichosz, S. L., Johansen, M. D. & Hejlesen, O. Toward big data analytics: review of predictive models in management of diabetes and its complications. *J. Diabetes Sci. Technol.* **10**, 27–34 (2015).
71. Kubota, K. J., Chen, J. A. & Little, M. A. Machine learning for large-scale wearable sensor data in Parkinson's disease: concepts, promises, pitfalls, and futures. *Mov. Disord.* **31**, 1314–1326 (2016).

72. Libbrecht, M. W. & Noble, W. S. Machine learning applications in genetics and genomics. *Nat. Rev. Genet.* **16**, 321–332 (2015).
73. Bzdok, D. & Yeo, B. T. T. Inference in the age of big data: future perspectives on neuroscience. *Neuroimage.* **155**, 549–564 (2017)
74. Huys, Q. J., Maia, T. V. & Frank, M. J. Computational psychiatry as a bridge from neuroscience to clinical applications. *Nat. Neurosci.* **19**, 404–413 (2016).
75. Sheth, A., Jaimini, U., Thirunarayan, K. & Banerjee T. Augmented personalized health: how smart data with iots and ai is about to change healthcare. *RTSI.* **2017**, (2017). <https://doi.org/10.1109/RTSI.2017.8065963>. (Epub 12 Oct 2017).
76. Moons, K. G. et al. Risk prediction models: II. External validation, model updating, and impact assessment. *Heart* **98**, 691–698 (2012).
77. Yahata, N., Kasai, K. & Kawato, M. Computational neuroscience approach to biomarkers and treatments for mental disorders. *Psychiatry Clin. Neurosci.* **71**, 215–237 (2017).
78. Deo, R. C. Machine learning in medicine. *Circulation* **132**, 1920–1930 (2015).
79. Vieira S. et al. Using machine learning and structural neuroimaging to detect first episode psychosis: reconsidering the evidence. *Schizophr Bull.* pii: sby189 2019. <https://doi.org/10.1093/schbul/sby189>. [Epub ahead of print].