Psychological Medicine (2017), **47**, 1163–1178. © Cambridge University Press 2017
doi:10.1017/S0033291716003494

REVIEW ARTICLE

Prediction of transition to psychosis in patients with a clinical high risk for psychosis: a systematic review of methodology and reporting

E. Studerus[1], A. Ramyead[2] and A. Riecher-Rössler[1]*

[1] *University of Basel Psychiatric Hospital, Center for Gender Research and Early Detection, Basel, Switzerland*
[2] *Department of Psychiatry, Weill Institute for Neurosciences, University of California (UCSF), San Francisco, CA, USA*

Background. To enhance indicated prevention in patients with a clinical high risk (CHR) for psychosis, recent research efforts have been increasingly directed towards estimating the risk of developing psychosis on an individual level using multivariable clinical prediction models. The aim of this study was to systematically review the methodological quality and reporting of studies developing or validating such models.

Method. A systematic literature search was carried out (up to 14 March 2016) to find all studies that developed or validated a clinical prediction model predicting the transition to psychosis in CHR patients. Data were extracted using a comprehensive item list which was based on current methodological recommendations.

Results. A total of 91 studies met the inclusion criteria. None of the retrieved studies performed a true external validation of an existing model. Only three studies (3.5%) had an event per variable ratio of at least 10, which is the recommended minimum to avoid overfitting. Internal validation was performed in only 14 studies (15%) and seven of these used biased internal validation strategies. Other frequently observed modeling approaches not recommended by methodologists included univariable screening of candidate predictors, stepwise variable selection, categorization of continuous variables, and poor handling and reporting of missing data.

Conclusions. Our systematic review revealed that poor methods and reporting are widespread in prediction of psychosis research. Since most studies relied on small sample sizes, did not perform internal or external cross-validation, and used poor model development strategies, most published models are probably overfitted and their reported predictive accuracy is likely to be overoptimistic.

Received 30 May 2016; Revised 9 December 2016; Accepted 9 December 2016; First published online 16 January 2017

Key words: Clinical high risk, prediction, prognostic models, psychosis, schizophrenia.

Introduction

The early detection and treatment of psychoses already in their prodromal stage have become widely accepted goals in psychiatry during the last two decades (Fusar-Poli *et al.* 2013*b*). Consequently, a number of operational criteria aiming at identifying patients with a clinical high risk (CHR) for psychosis have been established internationally. However, meta-analyses suggest that – among help-seeking individuals – about one-third of those meeting internationally established CHR criteria will develop psychosis within 5 years (Fusar-Poli *et al.* 2012; Schultze-Lutter *et al.* 2015), with about 73% of these developing schizophrenic psychoses (Fusar-Poli *et al.* 2013*a*) and about one-third is having a clinical remission within 2 years (Simon *et al.* 2013). Hence, risk stratification of CHR patients offers great potential for enhancing clinical decision making and improving the cost–benefit ratio of preventive interventions (Ruhrmann *et al.* 2012). Accordingly, recent research efforts have been increasingly directed toward estimating the risk of developing psychosis on an individual level. The trend towards indicated prevention and personalized medicine in early stages of psychosis is exemplified by the fact that several large multicenter studies [i.e. Personalised Prognostic Tools for Early Psychosis Management (PRONIA), PSYSCAN and North American Prodrome Longitudinal Study (NAPLS) III] are currently underway aiming at developing prognostic tools in CHR patients. Furthermore, an ever-increasing number of studies are seeking to improve the prediction of psychosis in CHR patients by incorporating single risk factors and indicators into multivariable prediction models (e.g. Cannon *et al.* 2008; Riecher-Rössler *et al.* 2009; Ruhrmann *et al.* 2010). By

* Address for correspondence: A. Riecher-Rössler, M.D., University of Basel Psychiatric Clinics, Center for Gender Research and Early Detection, Kornhausgasse 7, CH-4051 Basel, Switzerland.
(Email: anita.riecher@upkbs.ch)

using the term 'multivariable models', we refer to models with multiple predictor variables (i.e. independent variables) and one outcome variable (i.e. dependent variable) as opposed to multivariate models, which have multiple outcome variables (Hidalgo & Goodman, 2013).

However, despite considerable research efforts, no psychosis risk prediction model has yet been adopted in clinical practice. The most likely explanation for this is that none of the published models has yet been convincingly demonstrated to have sufficient validity and clinical utility. While a lack of progress in this area could be partly attributed to the fact that psychoses are complex disorders with large phenomenological, pathophysiological and etiological heterogeneity (Keshavan *et al.* 2011) and that there are heterogeneous subgroups within CHR samples (Fusar-Poli *et al.* 2016), another important obstacle to consider is the widespread use of poor (i.e. biased and inefficient) modeling strategies, which can severely compromise the reliability and validity of the developed models. Examples of poor modeling strategies are relying on small event per variable (EPV) ratios (i.e. small number of patients with transition to psychosis relative to the number of considered predictor variables), using biased methods to select predictor variables for inclusion into the multivariable prediction model among a set of candidate predictor variables, not properly assessing the predictive accuracy of the model, using inappropriate model types, and not efficiently dealing with missing data (D'Amico *et al.* 2016; Wynants *et al.* 2016). Systematic reviews on the methodology of studies developing clinical prediction models for type 2 diabetes (Collins *et al.* 2011), cancer (Mallett *et al.* 2010), traumatic brain injury outcome (Mushkudiani *et al.* 2008), kidney disease (Collins *et al.* 2013) or medicine in general (Bouwmeester *et al.* 2012) all found that the use of such methods is widespread. Hence, it is reasonable to assume that poor methods are also a widespread problem in prediction of psychosis research.

Unfortunately, a systematic review on the methodology and reporting of studies developing or validating models predicting psychosis in CHR patients using rigorous quality criteria has not yet been conducted. Although one systematic review (Strobl *et al.* 2012) has focused on methods and performance of models predicting the onset of psychosis, several critical aspects, such as EPV ratios, selection of predictor variables, assessment of predictive performance and dealing with missing data, were not addressed. This might be because up until recently, no guidance existed to help form a well-defined review question and determine which details to extract and critically appraise from prediction modeling studies (Moons *et al.* 2014). Fortunately, such guidance has now become available with the publication of the Checklist for critical Appraisal and data extraction for systematic Reviews of Prediction Modeling Studies (CHARMS; Moons *et al.* 2014) which was developed by a panel of experts of the Cochrane Prognosis Methods Group.

The present systematic review therefore aims to critically appraise the methodology and reporting of studies developing or validating models predicting psychosis in CHR patients. We reviewed prediction modeling studies regardless of the domains that predictor variables were selected from. In accordance with the recently published CHARMS and other guidelines on clinical prediction modeling (e.g. Altman *et al.* 2012; Collins *et al.* 2015), all important methodological issues are addressed, including effective sample size, type of model used, selection and transformation of variables, assessment of predictive performance, internal and external validation, and treatment of missing data. The ultimate goal of this paper is to enhance the methodology and reporting of future studies not only by identifying frequent sources of bias but also by giving recommendations for improvement. To facilitate understanding, brief explanations of key statistical concepts in prognostic modeling are provided in Table 1 (see also Fusar-Poli & Schultze-Lutter, 2016).

Method

Search strategy

A literature search was carried out (up to 14 March 2016) in the databases of Medline, Embase, PsycINFO and Web of Science using the following search terms: (predict* OR 'vulnerability marker' OR 'risk factors for transition') AND psychosis AND ('clinically at high risk' OR 'clinically at risk' OR 'clinical high risk' OR 'ultra high risk' OR prodrom* OR 'at risk mental state' OR 'risk of psychosis'). The search was restricted to English-language papers published from 1998 onwards because this marks the time when the first prospective studies with patients meeting validated CHR criteria were published (Yung *et al.* 1998). The publication type was restricted to articles only, thus excluding meeting abstracts, editorials, letters, reviews and comments. In addition, the reference lists of the included studies were screened to identify further potentially relevant studies.

Study selection

Studies were included if they met the following criteria: (1) involved subjects with a CHR for psychosis that were prospectively followed up; (2) developed or validated a prognostic model that predicted later

https://doi.org/10.1017/S0033291716003494 Published online by Cambridge University Press

Table 1. *Definitions of key terms used in developing and validating prognostic models*[a]

Term	Definition
Model performance	The ability of the prognostic model to predict the outcome of interest. Depending on the data in which it is assessed, we can distinguish apparent, internally validated, and externally validated model performance. Two important aspects of performance to consider in models predicting binary or survival outcomes are discrimination and calibration (see below)
Apparent performance	The predictive performance that is achieved when the model is applied to the same data as it was derived from
Internally validated performance	The predictive performance that is achieved when the model is developed and evaluated within one study sample, but not the same cases are used for developing and testing the model. Frequent internal validation methods are: • Split sampling: The sample is randomly split into two parts. One part is used for developing and the other for testing the model • Bootstrapping: Multiple samples are drawn with replacement from the original sample. For each iteration, a model is developed on the selected subsample and tested on the non-selected subsample. Model performance is then estimated by averaging of all iterations • *k*-Fold cross-validation: The original sample is randomly split into *k* equal-sized subsamples. Of the *k* subsamples, a single subsample is retained for testing the model, and the remaining $k - 1$ subsamples are used for developing the model. The process is repeated *k* times so that each of the *k* subsamples is used exactly once as the validation data. Model performance is then again estimated by averaging of all iterations
Externally validated performance	The predictive performance that is achieved when the model is tested in plausibly similar samples of patients that did not contribute to the development data. The external validation sample can be obtained at the same center but another time (temporal validation) or at another center (geographical validation). External validation is preferably conducted using data from another center and by fully independent investigators
Discrimination	Discrimination measures how well a prediction model can discriminate those with the outcome from those without the outcome, that is, how high the probability is that a patient with the outcome will receive a higher predicted probability than a patient without the outcome. It is most frequently assessed using the area under the receiver operator characteristic curve or concordance index
Calibration	Calibration measures the agreement between observed outcomes and predictions. For example, if the probability of developing psychosis is estimated 30% for a specific patient, the development of psychosis should be observed in 30 of 100 patients having the same characteristics
Optimism	The difference between the measured performance and the true performance of the model in the underlying population
Overfitting	Overfitting occurs when the model fitting (including variable selection and transformation) is too adaptive and therefore capitalizes on specifics and idiosyncrasies of the sample that do not generalize to new subjects outside of the sample

[a] Descriptions are adapted from Steyerberg (2009).

transition to psychosis from variables obtained at baseline; (3) included at least two predictor variables in the prognostic model.

CHR for psychosis was required to be diagnosed by internationally established criteria. That is, subjects had to fulfill either ultra-high-risk, basic symptom or unspecific prodromal symptom criteria (for a review, see Fusar-Poli *et al.* 2013*b*). Studies with overlapping samples were not excluded since the focus of our review was on methodology and reporting and not on the predictive performance of different models or the predictive potential of different predictor variables.

Studies were selected in a two-step procedure: First, all references retrieved from the databases were screened based on their titles and abstracts. Next, articles that were found to be potentially eligible were further evaluated based on their full texts. The study selection was performed by the first author (E.S.) and randomly checked by the second author (A.R.). Discrepancies in the final classification were discussed until consensus was reached.

Data extraction

We developed a comprehensive item list based on current methodological recommendations for developing and reporting clinical prediction models. To this end, we studied the item lists of previous systematic

https://doi.org/10.1017/S0033291716003494 Published online by Cambridge University Press

reviews evaluating prediction research in other medical fields (Mushkudiani *et al.* 2008; Mallett *et al.* 2010; Collins *et al.* 2011, 2013; Bouwmeester *et al.* 2012; van Oort *et al.* 2012), existing reporting statements and checklists [i.e. the CHARMS, Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD; Collins *et al.* 2015) and Reporting Recommendations for Tumor Marker Prognostic Studies (REMARK; Altman *et al.* 2012)], as well as current text books (Harrell, 2001; Steyerberg, 2009) and articles (Altman *et al.* 2009; Moons *et al.* 2009*a, b*; Royston *et al.* 2009; Steyerberg *et al.* 2010) on clinical prediction modeling. The first author (E.S.) extracted all data, which were randomly checked by the second author (A.R.). Discrepancies were resolved by mutual discussions.

Data analysis

In line with a recent systematic review on clinical prediction research (Bouwmeester *et al.* 2012), we distinguished between predictor finding studies, prediction model development studies and external validation studies. Predictor finding studies primarily aim to explore which predictors independently contribute to the prediction of the outcome, i.e. are associated with the outcome (Moons *et al.* 2009*b*; Bouwmeester *et al.* 2012). By contrast, model development studies aim to develop multivariable prediction models for clinical practice (i.e. for informed decision making) that predict the outcome as accurately as possible. While both types of studies make use of multivariable prediction models, the focus of the first is more on causal explanation and hypothesis testing whereas the latter is more concerned with accurate prediction. Although there are clear similarities in the design and analysis of etiological and prognostic studies, there are several aspects in which they differ. For example, calibration and discrimination are highly relevant to prognostic research but meaningless in etiological research (Moons *et al.* 2009*b*). Furthermore, establishing unbiased estimates of each individual predictor with the outcome is important in etiological research but not in prognostic research (for more details on the difference between prognostic and etiological research, see Moons *et al.* 2009*b*; Seel *et al.* 2012).

Studies were categorized as predictor model development studies if it was clearly stated in the paper that the aim was developing a model for clinical practice and not merely testing the predictive potential of certain predictor variables or domains. Studies were categorized as external validation studies if their aim was to assess the performance of a previously reported prediction model using new participant data that were not used in the development process. All other studies fulfilling inclusion criteria were termed predictor finding studies.

Since it would have been unfair to evaluate the different study types by exactly the same criteria, we grouped results by study type whenever necessary. Each extracted item was summarized in terms of absolute and relative frequencies and the results are reported according to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines (Moher *et al.* 2009).

Results

Literature search results

The literature search identified 91 articles eligible for full review (see Fig. 1). The included studies were published between November 2002 and February 2016. The number of studies published per year was increasing, with only a single study published in 2002 and 14 studies published in 2015. Three journals accounted for almost half of the publications: 28 articles (31%) appeared in *Schizophrenia Research*, nine (10%) in *Schizophrenia Bulletin* and eight (9%) in *Biological Psychiatry*. The full list of included studies is presented in online Supplementary Table S1.

Study aims

Only seven studies (Cannon *et al.* 2008; Ruhrmann *et al.* 2010; Michel *et al.* 2014; Nieman *et al.* 2014; Chan *et al.* 2015; Perkins *et al.* 2015*a, b*) (8%) aimed at developing a clinical prediction model for application in clinical practice and thus were categorized as model development studies. All other studies (92%) were considered predictor finding studies.

We did not identify any true external validation studies. Although Mason *et al.* (2004) aimed at replicating the results of Yung *et al.* (2004) and Thompson *et al.* (2011) aimed at replicating the results of Cannon *et al.* (2008), both studies did not evaluate an exact published model (i.e. applied a regression formula to new data) but re-estimated regression coefficients of previously identified predictors. As frequently pointed out in the literature (e.g. Royston & Altman, 2013; Moons *et al.* 2014), such studies are not model validation studies, but should be considered model re-development studies.

Study designs

All studies were cohort studies, except one (Thompson *et al.* 2011), which used a nested cohort design. Of the studies, 66 (73%) were single-center and 25 (27%) were multicenter studies. Data were collected at 27 different centers. Many studies had overlapping samples. For instance, more than one-quarter (25.3%) of the

https://doi.org/10.1017/S0033291716003494 Published online by Cambridge University Press

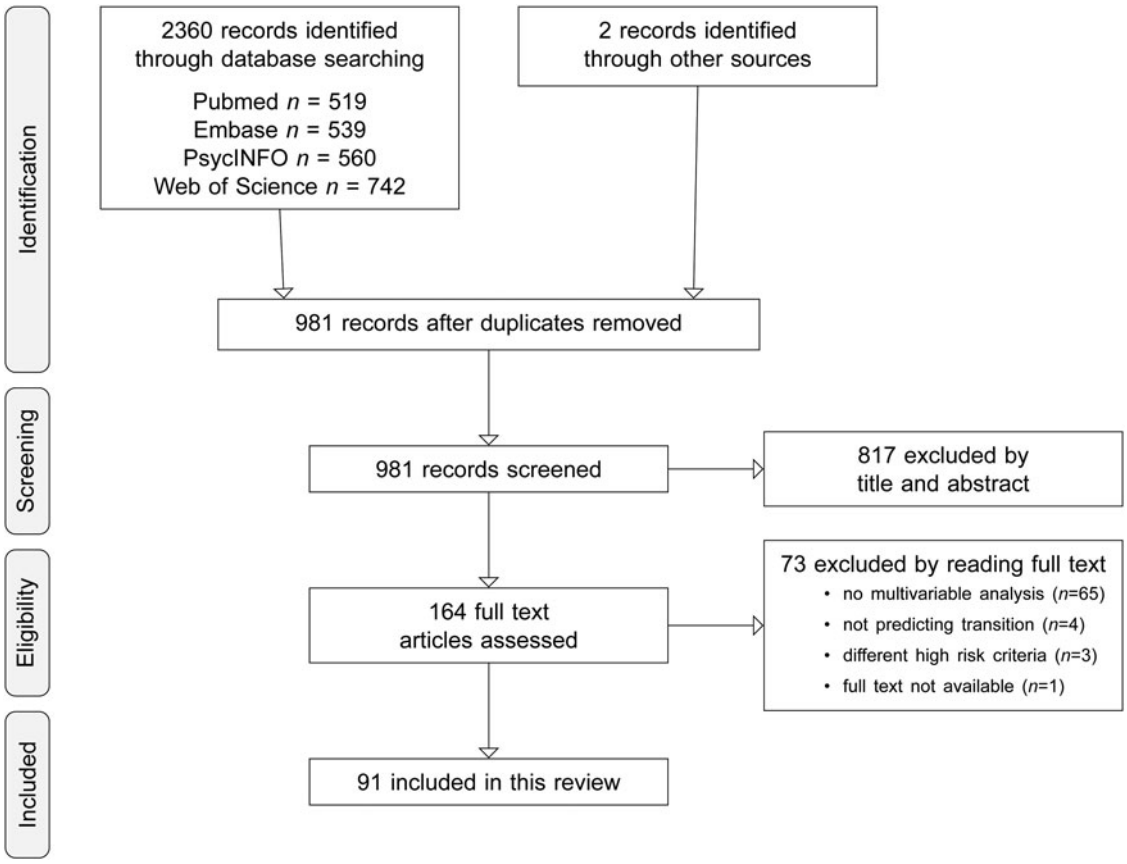

Fig. 1. Flowchart of the literature search.

published studies were based on data collected at the Personal Assessment and Crisis Evaluation (PACE) clinic in Melbourne, although not always from the same time periods. The criteria used for identifying CHR patients and assessing transition to psychosis are displayed in online Supplementary Table S2. The length and frequency of follow-up differed markedly between studies. Whereas some studies assessed transition to psychosis on a monthly basis in the first year (e.g. Riecher-Rössler *et al.* 2009), others conducted follow-up assessments only on a yearly or less frequent basis, which poses the risk of missing at least some transitions and might lead to a less accurate estimation of the time to transition. The average follow-up duration (from the 84 studies it could be determined) was 33.1 months (median 27.9 months, range 12–90 months). Of these studies, 17% had a follow-up duration of only 1 year, 33% of less than 2 years, and 60% of less than 3 years.

Number of patients and transitions

The average number of included CHR patients per study was 128 (S.D. 134) and the average number of transitions was 29.8 (S.D. 25.2). Although model development studies tended to have a higher number of included patients and transitions than predictor finding studies (252 *v.* 118 and 56.1 *v.* 27.6, respectively), these differences did not reach statistical significance. The average proportion of patients with later transition to psychosis was 27% (median 26%, range 5–53%). Since for a binary or a time-to-event outcome the effective sample size is the smaller of the two outcome frequencies (Moons *et al.* 2015), the effective sample size in the included studies almost always corresponded to the number of cases with later transition to psychosis and thus on average was only about one-quarter of the number of included CHR patients.

Number and type of considered predictor variables

The number of considered predictor variables could be determined in 85 studies (93%) and was 23.7 on average (median 12, S.D. 36.9, range 2–225). Model development studies considered significantly more predictor variables than predictor finding studies (97 *v.* 17.1 predictors, $p = 0.040$). The most frequently covered domains were positive symptoms, followed by negative symptoms, sociodemographic characteristics, and general, social and occupational functioning (see Fig. 2).

https://doi.org/10.1017/S0033291716003494 Published online by Cambridge University Press

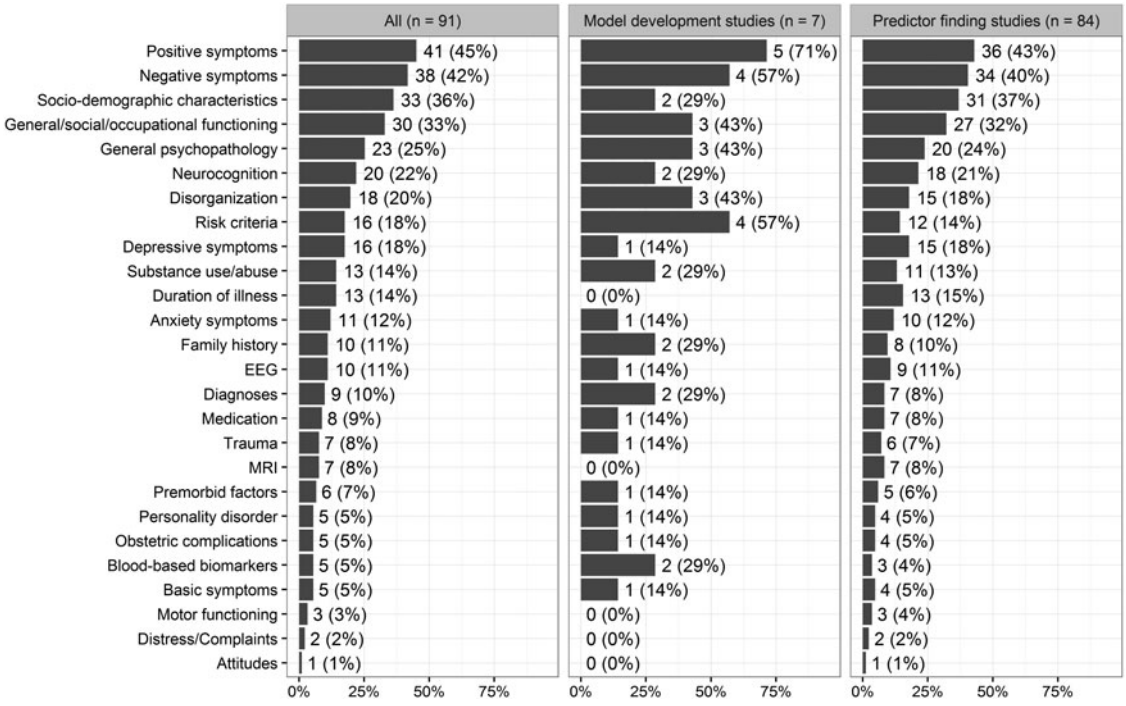

Fig. 2. Frequency of domains covered by candidate predictors. EEG, Electroencephalogram; MRI, magnetic resonance imaging.

EPV ratio

The average number of events per considered predictor variable (EPV) was 3 (median 1.8, S.D. 3, range 0.1–14.3). Although model development studies tended to have smaller average EPV than predictor finding studies (1.8 *v.* 3.1), this difference did not reach statistical significance. Only three studies (Velthorst *et al.* 2013*a*; Walder *et al.* 2013; Stowkowy *et al.* 2016) (3.5%), all of which were predictor finding studies, had an EPV of at least 10.

Missing data

Missing data at baseline were only explicitly mentioned in 28 studies (31%). The number of subjects with missing data was reported in 24 studies (26%), the number of missing values for each predictor in 11 studies (12%), and the number of subjects lost to follow-up in 35 studies (38%). Of the studies, nine (10%) reported to have omitted at least one predictor with missing values. The vast majority of studies handled missing data by performing complete case analyses, although this was only made explicit in 26 studies (29%) and must be assumed for those studies that did not mention missing data (Moons *et al.* 2015). Multiple imputation was only used in four studies (4%) (Seidman *et al.* 2010; Nieman *et al.* 2013, 2014; Rüsch *et al.* 2015) while single imputation was applied in two studies (2%) (Demjaha *et al.* 2012; Cornblatt *et al.* 2015).

Model types

The most frequently used model types were Cox proportional hazard and logistic regression models, which were used in 51 (56%) and 23 (25%) studies, respectively. Of the studies, five (5.5%) had fitted both of these models. A small number of studies applied more modern statistical learning methods, such as support vector machines (Koutsouleris *et al.* 2009, 2012*a*, *b*, 2015) (4%), least absolute shrinkage and selection operator (LASSO) (Chan *et al.* 2015; Ramyead *et al.* 2016) (2%), greedy algorithm (Perkins *et al.* 2015*a*, *b*) (2%), partial least squares discriminant analysis (Huang *et al.* 2007) (1%) and convex hull classification (Bedi *et al.* 2015) (1%). Linear discriminant analysis was used in one study (Mittal *et al.* 2010) (1%). One study (Healey *et al.* 2013) (1%) appeared to have used an ordinary least square regression model with a binary outcome, which clearly violates modeling assumptions.

Selection of predictor variables and dimensionality reduction

Pre-selection of candidate predictors for inclusion in the multivariable analyses based on univariable predictor–outcome associations was performed in 32 studies (35%). Of the studies, six reduced the number of predictors before inclusion to the final models by applying dimensionality reduction methods, such as principal component analysis (Huang *et al.* 2007;

https://doi.org/10.1017/S0033291716003494 Published online by Cambridge University Press

Koutsouleris *et al.* 2009, 2012*a*, 2015; Raballo *et al.* 2011), exploratory factor analysis (Demjaha *et al.* 2012) and latent class factor analysis (Velthorst *et al.* 2013*a*).

For selecting predictors within multivariable models, 34 studies (37%) used stepwise methods. Most of these used backward elimination methods, but six studies also used forward and backward stepwise, five used forward stepwise and two did not describe the specific stepwise method. Of the studies, nine (10%) applied stepwise variable selections in multiple steps, that is, first to each of several domains, and then to the variables retained in each domain. The most frequently used significance threshold for stepwise variable selection was $p = 0.05$. Automated variable selection within multivariable models using non-stepwise-methods was conducted in only four studies. Two of these (Chan *et al.* 2015; Ramyead *et al.* 2016) used the LASSO and two (Perkins *et al.* 2015*a, b*) a greedy algorithm.

Transformation of predictor variables

Three of the model development studies (Cannon *et al.* 2008; Ruhrmann *et al.* 2010; Perkins *et al.* 2015*b*) (43%) and 10 of the predictor finding studies (Yung *et al.* 2003, 2004; Mason *et al.* 2004; Amminger *et al.* 2006; Thompson *et al.* 2011; Nelson *et al.* 2013; Velthorst *et al.* 2013*b*; DeVylder *et al.* 2014; Cornblatt *et al.* 2015; O'Donoghue *et al.* 2015) (12%) fitted prediction models based on categorized or dichotomized continuous variables. Of these, six (Yung *et al.* 2003, 2004; Mason *et al.* 2004; Cannon *et al.* 2008; Thompson *et al.* 2011; Nelson *et al.* 2013) chose categorization cut-points based on the lowest p value, one (DeVylder *et al.* 2014) based on the maximal area under the receiver operating characteristic curve (AUC), one (O'Donoghue *et al.* 2015) based on quartiles, and five studies (Amminger *et al.* 2006; Cannon *et al.* 2008; Ruhrmann *et al.* 2010; Velthorst *et al.* 2013*b*; Perkins *et al.* 2015*b*) did not provide explanations for the chosen cut-points. In at least four studies (Yung *et al.* 2003, 2004; Mason *et al.* 2004; Nelson *et al.* 2013) the reason of dichotomizing continuous predictor variables was to provide a simple scoring rule.

Model performance

Table 2 displays the frequency of reporting different performance measures stratified by study aim. Whereas all model development studies reported at least one model performance measure, this was only the case in 28 (33%) of the predictor finding studies. If model performance was assessed, this was mainly done using classification measures, such as sensitivity and specificity, and less frequently using overall performance and discrimination measures. Calibration was not assessed in any of the model development studies and only in five (5%) of the predictor finding studies. Four of these (Piskulic *et al.* 2012; Cornblatt *et al.* 2015; Rüsch *et al.* 2015; Xu *et al.* 2016) used the Hosmer–Lemeshow statistic and one (Perkins *et al.* 2015*b*) a calibration plot. From the 31 studies reporting at least one classification measure, 19 did not report the probability threshold for classification and whether it was chosen from the data or set *a priori*, three used model types that did not predict a probability, and eight chose the probability threshold from the data. From the 35 studies (38%) reporting at least one performance measure, 21 (60%) only reported the so-called apparent performance.

Model evaluation

Internal cross-validation was carried out in only four of the model development studies (57%) (Michel *et al.* 2014; Nieman *et al.* 2014; Perkins *et al.* 2015*a, b*) and 10 of the predictor finding studies (12%) (Schultze-Lutter *et al.* 2007; Koutsouleris *et al.* 2009, 2012*a, b*, 2015; Riecher-Rössler *et al.* 2009; Mittal *et al.* 2010; Bedi *et al.* 2015; Cornblatt *et al.* 2015; Ramyead *et al.* 2016). Of these, six (Koutsouleris *et al.* 2009, 2012*a, b*; 2015; Perkins *et al.* 2015*a*; Ramyead *et al.* 2016) used k-fold cross-validation, three (Michel *et al.* 2014; Nieman *et al.* 2014; Cornblatt *et al.* 2015) used bootstrapping, three (Riecher-Rössler *et al.* 2009; Mittal *et al.* 2010; Bedi *et al.* 2015) used leave-one-out cross-validation and two (Schultze-Lutter *et al.* 2007; Perkins *et al.* 2015*b*) used a split-sampling approach. However, five of these studies (Riecher-Rössler *et al.* 2009; Mittal *et al.* 2010; Michel *et al.* 2014; Nieman *et al.* 2014; Cornblatt *et al.* 2015) only cross-validated the final model and therefore did not take into account the uncertainty introduced by the variable selection and transformation. Only four studies (Koutsouleris *et al.* 2012*a, b*, 2015; Ramyead *et al.* 2016) used nested repeated cross-validation, which is considered the best approach for training and testing a prediction model in one sample (Krstajic *et al.* 2014).

Model presentation

Only four studies (Schultze-Lutter *et al.* 2007, 2012; Ziermans *et al.* 2014; Xu *et al.* 2016) (4%) provided the full model formula, seven (8%) used model types that cannot be easily described with a model formula (e.g. support vector machine), 20 (22%) only provided p values but not regression coefficients of predictors, and 60 studies (66%) only provided regression coefficients of the predictor variables but not the intercept or baseline survival function, which are required in logistic and Cox regression, respectively, to properly assess calibration (Royston & Altman, 2013; Moons *et al.* 2015). Three studies (Lencz *et al.* 2006;

https://doi.org/10.1017/S0033291716003494 Published online by Cambridge University Press

Table 2. *Model performance measures, stratified by type of prediction study*

	All studies ($n = 91$)	Model development studies ($n = 7$)	Predictor finding studies ($n = 84$)
Overall performance measures			
No overall performance measure	82 (90.1)	7 (100)	75 (89.3)
Cox–Snell R^2	2 (2.2)	0 (0.00)	2 (2.4)
Nagelkerke's R^2	7 (7.7)	0 (0.00)	7 (8.3)
R^2	1 (1.10)	0 (0.00)	1 (1.2)
Calibration measures			
No calibration measure	86 (94.5)	6 (85.7)	80 (95.2)
Calibration plot	1 (1.10)	1 (14.3)	0 (0.0)
Hosmer–Lemeshow statistic	4 (4.4)	0 (0.0)	4 (4.8)
Discrimination measures			
No discrimination measure	80 (87.9)	2 (28.6)	78 (92.9)
AUC	11 (12.1)	5 (71.4)	6 (7.14)
Classification measures			
No classification measure	60 (65.9)	3 (42.9)	57 (67.9)
Accuracy	8 (8.8)	1 (14.3)	7 (8.3)
BAC	4 (4.4)	0 (0.0)	4 (4.8)
DOR	2 (2.2)	0 (0.0)	2 (2.4)
FPR	2 (2.2)	1 (14.3)	1 (1.2)
LR	1 (1.1)	0 (0.0)	1 (1.2)
LR−	4 (4.4)	2 (28.6)	2 (2.4)
LR+	4 (4.4)	2 (28.6)	2 (2.4)
NPV	20 (22.0)	3 (42.9)	17 (20.2)
PPV	22 (24.2)	4 (57.1)	18 (21.4)
Sensitivity	31 (34.1)	4 (57.1)	27 (32.1)
Specificity	30 (33.0)	3 (42.9)	27 (32.1)
TNR	1 (1.1)	0 (0.0)	1 (1.2)
TPR	1 (1.1)	0 (0.0)	1 (1.2)
Method for testing performance			
Not testing performance	56 (61.5)	0 (0.0)	56 (66.7)
Apparent	21 (23.1)	3 (42.9)	18 (21.4)
Internal	14 (15.4)	4 (57.1)	10 (11.9)

Data are given as number of studies (percentage).

AUC, Area under the receiver operator characteristic curve; BAC, balanced accuracy; DOR, diagnostic odds ratio; FPR, false positive rate; LR, likelihood ratio; NPV, negative predictive value; PPV, positive predictive value; TNR, true negative rate; TPR, true positive rate.

Riecher-Rössler *et al.* 2009; Bang *et al.* 2015) also only provided regression coefficients for standardized or otherwise transformed variables without giving enough details to exactly replicate the variable transformation in a new dataset.

Discussion

Our systematic review identified 91 studies using a multivariable clinical prediction model for predicting the transition to psychosis in CHR patients. The vast majority of these studies ($n = 84$) were classified as predictor finding studies because they primarily aimed at hypothesis testing or evaluating the predictive potential of certain predictors or assessment domains. Only seven studies stated explicitly that they aimed at developing a prediction model for clinical practice and therefore were classified as model development studies. Thus, in prediction of psychosis research, studies seem to focus much more often on etiology/explanation than maximizing prognostic accuracy (for a more detailed explanation of the difference between prognostic and etiological research, see Moons *et al.* 2009*b*; Seel *et al.* 2012). However, it should be noted that this distinction was not always clear-cut as many authors did not clearly describe the aim of the study or possibly tried to achieve both accurate prognosis and a better understanding of causal relationships.

We found that poor conduct and reporting were widespread in both predictor finding and model

https://doi.org/10.1017/S0033291716003494 Published online by Cambridge University Press

developed studies and that almost all aspects of the modeling process were affected. The results of this review are therefore consistent with reviews of prediction modeling studies in other medical fields (Mushkudiani *et al.* 2008; Mallett *et al.* 2010; Collins *et al.* 2011; Bouwmeester *et al.* 2012; Collins *et al.* 2013).

One of the biggest concerns is that most studies relied on small effective sample sizes and number of events (i.e. patients with later transitions to psychosis) relative to the number of considered predictor variables (EPV). Small EPV ratios increase the risk of overfitting and overestimating the performance of the model, if it is developed and assessed in the same sample (Moons *et al.* 2015). Furthermore, it can lead to biased regression coefficients and unstable variable selection (Mushkudiani *et al.* 2008). Current guidelines and textbooks therefore recommend EPV ratios of at least 10 (Steyerberg, 2009; Moons *et al.* 2014; Collins *et al.* 2015). Unfortunately, in this review, an EPV of at least 10 was only achieved in three studies (Velthorst *et al.* 2013*a*; Walder *et al.* 2013; Stowkowy *et al.* 2016) and the median EPV was only 1.8. While low EPV ratios have also frequently been criticized in other fields of clinical prediction research (Collins *et al.* 2011; Bouwmeester *et al.* 2012), the problem seems to be particularly severe in prediction of psychosis as reviews on studies developing models predicting cancer (Mallett *et al.* 2010), kidney disease (Collins *et al.* 2013), type 2 diabetes (Collins *et al.* 2011) and cardiovascular disease (Wessler *et al.* 2015) have reported median EPV ratios of 10, 29, 19 and 11–34, respectively. The much lower sample sizes in prediction of psychosis research can be at least partially explained by the fact that CHR patients are difficult to recruit and follow-up durations of at least 2 years are needed to detect most later transitions to psychosis (Kempton *et al.* 2015).

However, although missing data are expected to be frequent in medical research in general (Sterne *et al.* 2009) and in early psychosis research in particular, only about one-third of the included studies mentioned any missing data. Furthermore, reporting on the type and frequency of missing data was often poor. Moreover, only four studies (Seidman *et al.* 2010; Nieman *et al.* 2013, 2014; Rüsch *et al.* 2015) (4%) performed multiple imputation, which is generally acknowledged as the preferred method for handling incomplete data (Sterne *et al.* 2009; Moons *et al.* 2014). Hence, it is likely that most studies had excluded subjects or variables with incomplete data, which not only leads to a waste of data and reduced power, but can also negatively affect the representativeness of the sample and consequently the generalizability of the resulting prediction model (Gorelick, 2006; Moons *et al.* 2015). Unfortunately, poor handling and reporting of missing data are widespread in any medical field (Bouwmeester *et al.* 2012). However, in prediction of psychosis the consequences might be particularly severe as samples are already quite small and a further loss of data can be less afforded.

Approximately 60% of both predictor finding and model development studies used Cox regression and thus treated the outcome as a time-to-event variable, whereas the remaining studies used models with a binary outcome (i.e. transition *v.* non-transition). For prospective studies with longer-term diagnostic outcomes and regular follow-up assessments, as is the case in prediction of psychosis studies, time-to-event outcome models are more appropriate because they use more information, have more statistical power, and can deal with censoring (i.e. cases with incomplete follow-up) (van der Net *et al.* 2008; Moons *et al.* 2015). Since loss to follow-up is frequent in prediction of psychosis research and follow-up durations often too short to capture all transitions (Schultze-Lutter *et al.* 2015), the 40% of studies that have applied a binary outcome model are mainly faced with two shortcomings. First, they had to exclude non-transitioned cases with short follow-up durations, which again further aggravated the problem of already existing small sample sizes and might have hampered the representativeness of the sample. Second, patients with late transition to psychosis might have been misclassified as non-transitioned cases.

Several studies (Koutsouleris *et al.* 2009, 2012*a*, *b*, 2015) used so-called machine learning or pattern recognition methods, such as support vector machines. In line with Steyerberg *et al.* (2014), we herein use the term 'machine learning method' to refer to the more modern and flexible statistical learning methods originally developed in the field of computer science, such as random forest or neural networks, which can automatically capture highly complex non-linear relationships between predictor and response variables, and separate them from regression-based methods traditionally used in clinical prediction modeling, such as logistic and Cox regression or penalized versions thereof (i.e. models in which regression coefficients are shrunken towards zero, such as LASSO). Since the first results with machine learning methods have been encouraging, a more widespread use of these methods in the field of early detection of psychosis is now considered by many authors a promising strategy to improve the prediction of psychosis (Pettersson-Yeo *et al.* 2013; Koutsouleris & Kambeitz, 2016). However, many methodologists in the field of clinical prediction modeling (Steyerberg *et al.* 2014; Moons *et al.* 2015) do not share this enthusiasm for the following reasons: First, due to their higher flexibility, machine learning methods are more prone to

https://doi.org/10.1017/S0033291716003494 Published online by Cambridge University Press

overfitting than regression-based approaches, particularly in small datasets (van der Ploeg *et al.* 2014). Hence, when sample sizes are small, as is frequently the case in prediction of psychosis research, their performance advantage resulting from the increased ability to capture the true underlying relationship between predictors and response might not be high enough to compensate for the increased tendency to overfit (Steyerberg *et al.* 2014). Accordingly, van der Ploeg *et al.* (2016) have shown that logistic regression outperformed support vector machines, random forests and neural networks in external validation, when predicting 6-month mortality in traumatic brain injury patients from sociodemographic, computed tomography, and laboratory data. Similarly, logistic regression outperformed random forest and support vector machines, when predicting treatment resistance in major depressive disorder (Perlis, 2013). Of course, this does not mean that machine learning methods would also perform worse in every prediction of psychosis scenario (for example, they might still be superior when predicting psychosis from neuroimaging data). However, based on the above findings, it seems rather unlikely that they would be vastly superior in most scenarios. Second, machine learning methods are less interpretable and more difficult to communicate to clinicians (Steyerberg *et al.* 2014). For example, regression models can transparently be presented, with insight in relative effects of predictors by odds or hazard ratios, while many machine learning models are essentially black boxes with highly complex prediction equations. Third, while traditional methods can be easily adjusted to local settings (e.g. by changing the model intercept), this is more difficult for machine learning methods (Steyerberg *et al.* 2014). However, the ability to adjust the model, also called re-calibration (Steyerberg, 2009), is important in prediction of psychosis, as rates of transition to psychosis have been shown to vary considerably across time and location (Fusar-Poli *et al.* 2012).

With regard to variable selection strategies, we found that univariable screening of candidate predictors and/or stepwise variable selection were frequently conducted in both predictor finding and model development studies. However, these methods have long been criticized on multiple grounds (Harrell, 2001; Steyerberg, 2009; Núñez *et al.* 2011). Specifically, when the EPV ratio is low, the variable selection is unstable, the size and significance of the estimated regression coefficients are systematically overestimated, and the performance of the selected model is overoptimistic (Derksen & Keselman, 1992; Sun *et al.* 1996; Steyerberg *et al.* 1999; Steyerberg & Vergouwe, 2014). Since the bias introduced by these methods is more severe when EPV ratios are low, their use in prediction of psychosis research is particularly problematic. Unfortunately, we also found that most studies relied on high significance thresholds, such as $p < 0.05$, for variable selection, which leads to more bias and worse cross-validated predictive performance than higher thresholds, particularly in small datasets (Steyerberg *et al.* 1999, 2001). Furthermore, we found that several studies performed forward stepwise instead of the more recommended backward stepwise selection (Steyerberg, 2009; Núñez *et al.* 2011). Given that sample sizes in the field of early psychosis research are small, a more sensible approach for variable selection would be to rely more on external knowledge. For example, candidate predictors could be pre-selected by performing meta-analyses or based on theory. If external knowledge is not available, a more stable set of predictor variables and reduced overfitting can be achieved by applying shrinkage methods (Steyerberg *et al.* 2001; Núñez *et al.* 2011), such as the LASSO (Tibshirani, 1997), which have only been used in two (Chan *et al.* 2015; Ramyead *et al.* 2016) of the included studies.

We also found several studies that categorized or even dichotomized continuous predictor variables, which has been strongly discouraged by methodologists because it leads to a considerable loss of information, reduced statistical power, residual confounding, and decreased predictive accuracy (Royston *et al.* 2006; Altman *et al.* 2012; Collins *et al.* 2016). Furthermore, many of these studies chose cut-points by taking the value that produced the lowest p or highest AUC value, which can lead to a serious inflation of the type I error and to an overestimation of the prognostic effect (Hollander *et al.* 2004; Altman *et al.* 2012).

Another area that needs considerable improvement concerns model performance assessment and evaluation, although this is clearly more important for model development and less so for predictor finding studies. We found that none of the proposed models has been externally validated and internal cross-validation was carried out in only 57% of model development studies and 12% of predictor finding studies. Furthermore, half of these used poor internal cross-validation strategies, such as split-sampling, which wastes half of the data and leads to highly uncertain estimates of model performance (Austin & Steyerberg, 2014; Moons *et al.* 2015), or cross-validating only the final model after having conducted data-driven variable selection in the whole sample, which leads to highly overoptimistic performance estimates (Krstajic *et al.* 2014).

Since internal cross-validation was conducted infrequently, most studies only reported the so called 'apparent' performance, which tends to be strongly overoptimistic because it is calculated in the same

https://doi.org/10.1017/S0033291716003494 Published online by Cambridge University Press

Table 3. *Recommendations for improved methodology and reporting*

Sample size	The number of cases with transition to psychosis in the CHR sample should ideally be at least 10 times as high as the number of considered predictor variables. This ratio can be improved by: • Restricting the number of considered predictor variables by pre-selecting them based on theory or external knowledge • Performing multicenter studies • Avoiding loss of data through inefficient missing data methods
Missing data	• Avoid inefficient missing data methods, such as listwise deletion • Always report on the amount of missingness in the dataset
Model type	• If time to transition has been assessed, use survival instead of binary outcome models • Use more flexible models (i.e. machine learning methods) only if it can be demonstrated that compared with regression-based approaches they achieve an improved cross-validated predictive performance that is worth their extra complexity
Variable selection	• Pre-select variables blind to the outcome to avoid overfitting (e.g. based on theory or external knowledge) • Use the LASSO instead of univariate variable screening and/or stepwise variable selection methods
Variable transformation	• Avoid categorizing continuous predictor variables
Internal validation	• Always perform internal cross-validation using repeated *k*-fold cross-validation or bootstrapping in which variable selection and transformation are repeated at each iteration of the resampling procedure
External validation	• Apply the exact published model (formula) to the new data • If the model shows poor performance on new data, consider adjusting, updating or recalibrating the original model
Model performance	• Always report performance not only in terms of discrimination, but also calibration • In model development studies, always report the cross-validated predictive performance. The apparent predictive performance might additionally be provided to gain insight into the amount of over-optimism
Model presentation	• Always report the full model formula, including intercept or baseline survival function, which are required in logistic and Cox regression, respectively, to properly assess calibration • Consider publishing the model as an online risk calculator to ease applicability

CHR, Clinical high risk; LASSO, least absolute shrinkage and selection operator.

data as used for model building (Moons *et al.* 2015). Furthermore, most studies did not report the whole spectrum of recommended performance measures. For example, calibration, which is a key aspect of the model performance (Moons *et al.* 2015), was rarely assessed and mostly using the Hosmer–Lemeshow statistic instead of the more recommended calibration-in-the-large and calibration slope (Steyerberg *et al.* 2010; Collins *et al.* 2015). Moreover, many studies reporting classification measures (i.e. sensitivity and specificity) had searched for optimal probability thresholds for classification in the same sample as they used for testing, which again probably contributed to overoptimism (Leeflang *et al.* 2008).

We also found major deficiencies in the way that models were presented. Most importantly, most studies did not provide enough details to exactly apply the model in a new dataset, which might at least partially explain why none of these models has yet been externally validated. Furthermore, several studies only provided enough details to apply a simplified scoring rule but not the original model. However, as explained above, the perceived advantage of simplification/categorization comes at high costs. A much better way of facilitating the clinical application would be the creation of an online risk calculator (Steyerberg & Vergouwe, 2014). This would also allow the clinical use and external validation of more complex models (e.g. machine learning algorithms) that cannot be described with a simple model formula (Steyerberg *et al.* 2014).

Limitations

Our literature search was restricted to English-language journal articles only. Thus, it is possible that some relevant literature has been missed. A further limitation is that choosing an appropriate modeling strategy is complex and depends on many different factors, including research question, study design, sample size and number of variables. Although we grouped studies by their aim and relied on guidelines (i.e. the CHARMS) for critically appraising the methodology

https://doi.org/10.1017/S0033291716003494 Published online by Cambridge University Press

and reporting of the included studies as much as possible, some studies might have been treated unfairly due to not taking all specific factors into account.

Conclusion

Taken together, we found that most studies developing a model for predicting the transition to psychosis in CHR patients were poorly conducted and reported. Biased and inefficient methods, such as complete case analysis, modeling a time-to-event outcome as a binary outcome, data-driven univariable and stepwise selection of candidate variables, categorization of continuous predictors, and assessing only the apparent predictive performance, were widespread and often applied together and in datasets with small EPV ratios, which probably potentiated their harmful consequences. Consequently, most published predictive performance estimates in this field are likely to be considerably overoptimistic. Unfortunately, this was rarely acknowledged, since proper internal validation was infrequent and external validation not attempted. An essential requirement for future studies is therefore to improve model validation. While we acknowledge that – due to differences in measurement methods across centers – external validation is often difficult, internal validation can and should always be performed (Moons *et al*. 2012). To further enhance progress, future studies should more strictly adhere to current checklists and guidelines on clinical prediction models, such as the recently published TRIPOD statement (Collins *et al*. 2015; Moons *et al*. 2015). Since EPV ratios in prediction of psychosis research are small compared with other fields of prediction research, researchers in this field should take extra care to not waste valuable information and to avoid overfitting, for example, by more strongly relying on external information and applying models that are not too adaptive. In Table 3, we have summarized our recommendations for improved methodology and reporting in prediction of psychosis studies.

Supplementary material

The supplementary material for this article can be found at http://doi.org/10.1017/S0033291716003494

Acknowledgements

This work was supported by the Swiss National Science Foundation (P2BSP3-165392).

Declaration of Interest

None.

References

Altman DG, McShane LM, Sauerbrei W, Taube SE (2012). Reporting Recommendations for Tumor Marker Prognostic Studies (REMARK): explanation and elaboration. *PLoS Medicine* **9**, e1001216.

Altman DG, Vergouwe Y, Royston P, Moons KG (2009). Prognosis and prognostic research: validating a prognostic model. *BMJ* **338**, b605.

Amminger GP, Leicester S, Yung AR, Phillips LJ, Berger GE, Francey SM, Yuen HP, McGorry PD (2006). Early-onset of symptoms predicts conversion to non-affective psychosis in ultra-high risk individuals. *Schizophrenia Research* **84**, 67–76.

Austin PC, Steyerberg EW (2014). Events per variable (EPV) and the relative performance of different strategies for estimating the out-of-sample validity of logistic regression models. *Statistical Methods in Medical Research*. Published online 19 November 2014. doi:10.1177/0962280214558972.

Bang M, Kim KR, Song YY, Baek S, Lee E, An SK (2015). Neurocognitive impairments in individuals at ultra-high risk for psychosis: who will really convert? *Australian and New Zealand Journal of Psychiatry* **49**, 462–470.

Bedi G, Carrillo F, Cecchi GA, Slezak DF, Sigman M, Mota NB, Ribeiro S, Javitt DC, Copelli M, Corcoran CM (2015). Automated analysis of free speech predicts psychosis onset in high-risk youths. *NPJ Schizophrenia* **1**, 15030.

Bouwmeester W, Zuithoff NP, Mallett S, Geerlings MI, Vergouwe Y, Steyerberg EW, Altman DG, Moons KG (2012). Reporting and methods in clinical prediction research: a systematic review. *PLoS Medicine* **9**, 1–12.

Cannon TD, Cadenhead K, Cornblatt B, Woods SW, Addington J, Walker E, Seidman LJ, Perkins D, Tsuang M, McGlashan T, Heinssen R (2008). Prediction of psychosis in youth at high clinical risk: a multisite longitudinal study in North America. *Archives of General Psychiatry* **65**, 28–37.

Chan MK, Krebs MO, Cox D, Guest PC, Yolken RH, Rahmoune H, Rothermundt M, Steiner J, Leweke FM, van Beveren NJ, Niebuhr DW, Weber NS, Cowan DN, Suarez-Pinilla P, Crespo-Facorro B, Mam-Lam-Fook C, Bourgin J, Wenstrup RJ, Kaldate RR, Cooper JD, Bahn S (2015). Development of a blood-based molecular biomarker test for identification of schizophrenia before disease onset. *Translational Psychiatry* **5**, e601.

Collins GS, Mallett S, Omar O, Yu LM (2011). Developing risk prediction models for type 2 diabetes: a systematic review of methodology and reporting. *BMC Medicine* **9**, 103.

Collins GS, Ogundimu EO, Cook JA, Manach YL, Altman DG (2016). Quantifying the impact of different approaches for handling continuous predictors on the performance of a prognostic model. *Statistics in Medicine* **35**, 4124–4135.

Collins GS, Omar O, Shanyinde M, Yu LM (2013). A systematic review finds prediction models for chronic kidney disease were poorly reported and often developed using inappropriate methods. *Journal of Clinical Epidemiology* **66**, 268–277.

Collins GS, Reitsma JB, Altman DG, Moons KGM (2015). Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis Or Diagnosis (TRIPOD): the

TRIPOD statement. *Journal of Clinical Epidemiology* **68**, 112–121.

Cornblatt BA, Carrion RE, Auther A, McLaughlin D, Olsen RH, John M, Correll CU (2015). Psychosis prevention: a modified clinical high risk perspective from the Recognition and Prevention (RAP) Program. *American Journal of Psychiatry* **172**, 986–994.

D'Amico G, Malizia G, D'Amico M (2016). Prognosis research and risk of bias. *Internal and Emergency Medicine* **11**, 251–260.

Demjaha A, Valmaggia L, Stahl D, Byrne M, McGuire P (2012). Disorganization/cognitive and negative symptom dimensions in the at-risk mental state predict subsequent transition to psychosis. *Schizophrenia Bulletin* **38**, 351–359.

Derksen S, Keselman HJ (1992). Backward, forward and stepwise automated subset selection algorithms: frequency of obtaining authentic and noise variables. *British Journal of Mathematical and Statistical Psychology* **45**, 265–282.

DeVylder JE, Muchomba FM, Gill KE, Ben-David S, Walder DJ, Malaspina D, Corcoran CM (2014). Symptom trajectories and psychosis onset in a clinical high-risk cohort: the relevance of subthreshold thought disorder. *Schizophrenia Research* **159**, 278–283.

Fusar-Poli P, Bechdolf A, Taylor MJ, Bonoldi I, Carpenter WT, Yung AR, McGuire P (2013*a*). At risk for schizophrenic or affective psychoses? A meta-analysis of DSM/ICD diagnostic outcomes in individuals at high clinical risk. *Schizophrenia Bulletin* **39**, 923–932.

Fusar-Poli P, Bonoldi I, Yung AR, Borgwardt S, Kempton MJ, Valmaggia L, Barale F, Caverzasi E, McGuire P (2012). Predicting psychosis: meta-analysis of transition outcomes in individuals at high clinical risk. *Archives of General Psychiatry* **69**, 220–229.

Fusar-Poli P, Borgwardt S, Bechdolf A, Addington J, Riecher-Rössler A, Schultze-Lutter F, Keshavan M, Wood S, Ruhrmann S, Seidman LJ, Valmaggia L, Cannon T, Velthorst E, De Haan L, Cornblatt B, Bonoldi I, Birchwood M, McGlashan T, Carpenter W, McGorry P, Klosterkötter J, McGuire P, Yung A (2013*b*). The psychosis high-risk state: a comprehensive state-of-the-art review. *JAMA Psychiatry* **70**, 107–120.

Fusar-Poli P, Cappucciati M, Borgwardt S, Woods SW, Addington J, Nelson B, Nieman DH, Stahl DR, Rutigliano G, Riecher-Rössler A, Simon AE, Mizuno M, Lee TY, Kwon JS, Lam MM, Perez J, Keri S, Amminger P, Metzler S, Kawohl W, Rössler W, Lee J, Labad J, Ziermans T, An SK, Liu CC, Woodberry KA, Braham A, Corcoran C, McGorry P, Yung AR, McGuire PK (2016). Heterogeneity of psychosis risk within individuals at clinical high risk: a meta-analytical stratification. *JAMA Psychiatry* **73**, 113–120.

Fusar-Poli P, Schultze-Lutter F (2016). Predicting the onset of psychosis in patients at clinical high risk: practical guide to probabilistic prognostic reasoning. *Evidence-Based Mental Health* **19**, 10–15.

Gorelick MH (2006). Bias arising from missing data in predictive models. *Journal of Clinical Epidemiology* **59**, 1115–1123.

Harrell FE (2001). *Regression Modeling Strategies with Applications to Linear Models, Logistic Regression, and Survival Analysis*. Springer: New York.

Healey KM, Penn DL, Perkins D, Woods SW, Addington J (2013). Theory of mind and social judgments in people at clinical high risk of psychosis. *Schizophrenia Research* **150**, 498–504.

Hidalgo B, Goodman M (2013). Multivariate or multivariable regression? *American Journal of Public Health* **103**, 39–40.

Hollander N, Sauerbrei W, Schumacher M (2004). Confidence intervals for the effect of a prognostic factor after selection of an 'optimal' cutpoint. *Statistics in Medicine* **23**, 1701–1713.

Huang JT, Leweke FM, Tsang TM, Koethe D, Kranaster L, Gerth CW, Gross S, Schreiber D, Ruhrmann S, Schultze-Lutter F, Klosterkötter J, Holmes E, Bahn S (2007). CSF metabolic and proteomic profiles in patients prodromal for psychosis. *PLoS ONE* **2**, e756.

Kempton MJ, Bonoldi I, Valmaggia L, McGuire P, Fusar-Poli P (2015). Speed of psychosis progression in people at ultra-high clinical risk: a complementary meta-analysis. *JAMA Psychiatry* **72**, 622–623.

Keshavan MS, Nasrallah HA, Tandon R (2011). Schizophrenia, "Just the Facts" 6. Moving ahead with the schizophrenia concept: from the elephant to the mouse. *Schizophrenia Research* **127**, 3–13.

Koutsouleris N, Borgwardt S, Meisenzahl EM, Bottlender R, Moller HJ, Riecher-Rössler A (2012*a*). Disease prediction in the at-risk mental state for psychosis using neuroanatomical biomarkers: results from the FePsy study. *Schizophrenia Bulletin* **38**, 1234–1246.

Koutsouleris N, Davatzikos C, Bottlender R, Patschurek-Kliche K, Scheuerecker J, Decker P, Gaser C, Moller HJ, Meisenzahl EM (2012*b*). Early recognition and disease prediction in the at-risk mental states for psychosis using neurocognitive pattern classification. *Schizophrenia Bulletin* **38**, 1200–1215.

Koutsouleris N, Kambeitz J (2016). Pattern recognition methods in the prediction of psychosis. In *Early Detection and Intervention in Psychosis – State of the Art and Future Perspectives* (ed. A Riecher-Rössler and PD McGorry), pp. 95–102. Karger: Basel.

Koutsouleris N, Meisenzahl EM, Davatzikos C, Bottlender R, Frodl T, Scheuerecker J, Schmitt G, Zetzsche T, Decker P, Reiser M, Moller HJ, Gaser C (2009). Use of neuroanatomical pattern classification to identify subjects in at-risk mental states of psychosis and predict disease transition. *Archives of General Psychiatry* **66**, 700–712.

Koutsouleris N, Riecher-Rössler A, Meisenzahl EM, Smieskova R, Studerus E, Kambeitz-Ilankovic L, von Saldern S, Cabral C, Reiser M, Falkai P, Borgwardt S (2015). Detecting the psychosis prodrome across high-risk populations using neuroanatomical biomarkers. *Schizophrenia Bulletin* **41**, 471–482.

Krstajic D, Buturovic LJ, Leahy DE, Thomas S (2014). Cross-validation pitfalls when selecting and assessing regression and classification models. *Journal of Cheminformatics* **6**, 10.

Leeflang MM, Moons KG, Reitsma JB, Zwinderman AH (2008). Bias in sensitivity and specificity caused by data-driven selection of optimal cutoff values: mechanisms, magnitude, and solutions. *Clinical Chemistry* **54**, 729–737.

Lencz T, Smith CW, McLaughlin D, Auther A, Nakayama E, Hovey L, Cornblatt BA (2006). Generalized and specific neurocognitive deficits in prodromal schizophrenia. *Biological Psychiatry* **59**, 863–871.

Mallett S, Royston P, Dutton S, Waters R, Altman DG (2010). Reporting methods in studies developing prognostic models in cancer: a review. *BMC Medicine* **8**, 20.

Mason O, Startup M, Halpin S, Schall U, Conrad A, Carr V (2004). Risk factors for transition to first episode psychosis among individuals with 'at-risk mental states'. *Schizophrenia Research* **71**, 227–237.

Michel C, Ruhrmann S, Schimmelmann BG, Klosterkötter J, Schultze-Lutter F (2014). A stratified model for psychosis prediction in clinical practice. *Schizophrenia Bulletin* **40**, 1533–1542.

Mittal VA, Walker EF, Bearden CE, Walder D, Trottman H, Daley M, Simone A, Cannon TD (2010). Markers of basal ganglia dysfunction and conversion to psychosis: neurocognitive deficits and dyskinesias in the prodromal period. *Biological Psychiatry* **68**, 93–99.

Moher D, Liberati A, Tetzlaff J, Altman DG, Group P (2009). Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *PLoS Medicine* **6**, e1000097.

Moons KG, Altman DG, Reitsma JB, Ioannidis JP, Macaskill P, Steyerberg EW, Vickers AJ, Ransohoff DF, Collins GS (2015). Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration. *Annals of Internal Medicine* **162**, W1–W73.

Moons KG, Altman DG, Vergouwe Y, Royston P (2009*a*). Prognosis and prognostic research: application and impact of prognostic models in clinical practice. *BMJ* **338**, b606.

Moons KG, de Groot JA, Bouwmeester W, Vergouwe Y, Mallett S, Altman DG, Reitsma JB, Collins GS (2014). Critical appraisal and data extraction for systematic reviews of prediction modelling studies: the CHARMS checklist. *PLoS Medicine* **11**, e1001744.

Moons KG, Kengne AP, Woodward M, Royston P, Vergouwe Y, Altman DG, Grobbee DE (2012). Risk prediction models: I. Development, internal validation, and assessing the incremental value of a new (bio)marker. *Heart* **98**, 683–690.

Moons KG, Royston P, Vergouwe Y, Grobbee DE, Altman DG (2009*b*). Prognosis and prognostic research: what, why, and how? *BMJ* **338**, b375.

Mushkudiani NA, Hukkelhoven CW, Hernandez AV, Murray GD, Choi SC, Maas AI, Steyerberg EW (2008). A systematic review finds methodological improvements necessary for prognostic models in determining traumatic brain injury outcomes. *Journal of Clinical Epidemiology* **61**, 331–343.

Nelson B, Yuen HP, Wood SJ, Lin A, Spiliotacopoulos D, Bruxner A, Broussard C, Simmons M, Foley DL, Brewer WJ, Francey SM, Amminger GP, Thompson A, McGorry PD, Yung AR (2013). Long-term follow-up of a group at ultra high risk ("prodromal") for psychosis: the PACE 400 study. *JAMA Psychiatry* **70**, 793–802.

Nieman DH, Ruhrmann S, Dragt S, Soen F, van Tricht MJ, Koelman JH, Bour LJ, Velthorst E, Becker HE, Weiser M, Linszen DH, de Haan L (2014). Psychosis prediction: stratification of risk estimation with information-processing and premorbid functioning variables. *Schizophrenia Bulletin* **40**, 1482–1490.

Nieman DH, Velthorst E, Becker HE, de Haan L, Dingemans PM, Linszen DH, Birchwood M, Patterson P, Salokangas RK, Heinimaa M, Heinz A, Juckel G, von Reventlow HG, Morrison A, Schultze-Lutter F, Klosterkotter J, Ruhrmann S, group E (2013). The Strauss and Carpenter Prognostic Scale in subjects clinically at high risk of psychosis. *Acta Psychiatrica Scandinavica* **127**, 53–61.

Núñez E, Steyerberg EW, Núñez J (2011). [Regression modeling strategies]. *Revista Española de Cardiología* **64**, 501–507.

O'Donoghue B, Nelson B, Yuen HP, Lane A, Wood S, Thompson A, Lin A, McGorry P, Yung AR (2015). Social environmental risk factors for transition to psychosis in an ultra-high risk population. *Schizophrenia Research* **161**, 150–155.

Perkins DO, Jeffries CD, Addington J, Bearden CE, Cadenhead KS, Cannon TD, Cornblatt BA, Mathalon DH, McGlashan TH, Seidman LJ, Tsuang MT, Walker EF, Woods SW, Heinssen R (2015*a*). Towards a psychosis risk blood diagnostic for persons experiencing high-risk symptoms: preliminary results from the NAPLS project. *Schizophrenia Bulletin* **41**, 419–428.

Perkins DO, Jeffries CD, Cornblatt BA, Woods SW, Addington J, Bearden CE, Cadenhead KS, Cannon TD, Heinssen R, Mathalon DH, Seidman LJ, Tsuang MT, Walker EF, McGlashan TH (2015*b*). Severity of thought disorder predicts psychosis in persons at clinical high-risk. *Schizophrenia Research* **169**, 169–177.

Perlis RH (2013). A clinical risk stratification tool for predicting treatment resistance in major depressive disorder. *Biological Psychiatry* **74**, 7–14.

Pettersson-Yeo W, Benetti S, Marquand AF, Dell'acqua F, Williams SC, Allen P, Prata D, McGuire P, Mechelli A (2013). Using genetic, cognitive and multi-modal neuroimaging data to identify ultra-high-risk and first-episode psychosis at the individual level. *Psychological Medicine* **43**, 2547–2562.

Piskulic D, Addington J, Cadenhead KS, Cannon TD, Cornblatt BA, Heinssen R, Perkins DO, Seidman LJ, Tsuang MT, Walker EF, Woods SW, McGlashan TH (2012). Negative symptoms in individuals at clinical high risk of psychosis. *Psychiatry Research* **196**, 220–224.

Raballo A, Nelson B, Thompson A, Yung A (2011). The comprehensive assessment of at-risk mental states: from mapping the onset to mapping the structure. *Schizophrenia Research* **127**, 107–114.

Ramyead A, Studerus E, Kometer M, Uttinger M, Gschwandtner U, Fuhr P, Riecher-Rössler A (2016). Prediction of psychosis using neural oscillations and machine learning in neuroleptic-naive at-risk patients. *World Journal of Biological Psychiatry* **17**, 285–295.

Riecher-Rössler A, Pflueger MO, Aston J, Borgwardt SJ, Brewer WJ, Gschwandtner U, Stieglitz RD (2009). Efficacy of using cognitive status in predicting psychosis: a 7-year follow-up. *Biological Psychiatry* **66**, 1023–1030.

Royston P, Altman DG (2013). External validation of a Cox prognostic model: principles and methods. *BMC Medical Research Methodology* **13**, 33.

Royston P, Altman DG, Sauerbrei W (2006). Dichotomizing continuous predictors in multiple regression: a bad idea. *Statistics in Medicine* **25**, 127–141.

Royston P, Moons KG, Altman DG, Vergouwe Y (2009). Prognosis and prognostic research: developing a prognostic model. *BMJ* **338**, b604.

Ruhrmann S, Klosterkötter J, Bodatsch M, Nikolaides A, Julkowski D, Hilboll D, Schultz-Lutter F (2012). Chances and risks of predicting psychosis. *European Archives of Psychiatry and Clinical Neuroscience* **262** (Suppl. 2), S85–S90.

Ruhrmann S, Schultze-Lutter F, Salokangas RK, Heinimaa M, Linszen D, Dingemans P, Birchwood M, Patterson P, Juckel G, Heinz A, Morrison A, Lewis S, von Reventlow HG, Klosterkötter J (2010). Prediction of psychosis in adolescents and young adults at high risk: results from the prospective European Prediction of Psychosis Study. *Archives of General Psychiatry* **67**, 241–251.

Rüsch N, Heekeren K, Theodoridou A, Muller M, Corrigan PW, Mayer B, Metzler S, Dvorsky D, Walitza S, Rossler W (2015). Stigma as a stressor and transition to schizophrenia after 1 year among young people at risk of psychosis. *Schizophrenia Research* **166**, 43–48.

Schultze-Lutter F, Klosterkötter J, Michel C, Winkler K, Ruhrmann S (2012). Personality disorders and accentuations in at-risk persons with and without conversion to first-episode psychosis. *Early Intervention in Psychiatry* **6**, 389–398.

Schultze-Lutter F, Klosterkötter J, Picker H, Steinmeyer E-M, Ruhrmann S (2007). Predicting first-episode psychosis by basic symptom criteria. *Clinical Neuropsychiatry* **4**, 11–22.

Schultze-Lutter F, Michel C, Schmidt SJ, Schimmelmann BG, Maric NP, Salokangas RK, Riecher-Rössler A, van der Gaag M, Nordentoft M, Raballo A, Meneghelli A, Marshall M, Morrison A, Ruhrmann S, Klosterkötter J (2015). EPA guidance on the early detection of clinical high risk states of psychoses. *European Psychiatry* **30**, 405–416.

Seel RT, Steyerberg EW, Malec JF, Sherer M, Macciocchi SN (2012). Developing and evaluating prediction models in rehabilitation populations. *Archives of Physical Medicine and Rehabilitation* **93**, S138–S153.

Seidman LJ, Giuliano AJ, Meyer EC, Addington J, Cadenhead KS, Cannon TD, McGlashan TH, Perkins DO, Tsuang MT, Walker EF, Woods SW, Bearden CE, Christensen BK, Hawkins K, Heaton R, Keefe RS, Heinssen R, Cornblatt BA; North American Prodrome Longitudinal Study (NAPLS) Group (2010). Neuropsychology of the prodrome to psychosis in the NAPLS consortium: relationship to family history and conversion to psychosis. *Archives of General Psychiatry* **67**, 578–588.

Simon AE, Borgwardt S, Riecher-Rössler A, Velthorst E, de Haan L, Fusar-Poli P (2013). Moving beyond transition outcomes: meta-analysis of remission rates in individuals at high clinical risk for psychosis. *Psychiatry Research* **209**, 266–272.

Sterne JA, White IR, Carlin JB, Spratt M, Royston P, Kenward MG, Wood AM, Carpenter JR (2009). Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *BMJ* **338**, b2393.

Steyerberg EW (2009). *Clinical Prediction Models: a Practical Approach to Development, Validation, and Updating*. Springer: New York.

Steyerberg EW, Eijkemans MJ, Habbema JD (1999). Stepwise selection in small data sets: a simulation study of bias in logistic regression analysis. *Journal of Clinical Epidemiology* **52**, 935–942.

Steyerberg EW, Eijkemans MJ, Harrell Jr. FE, Habbema JD (2001). Prognostic modeling with logistic regression analysis: in search of a sensible strategy in small data sets. *Medical Decision Making* **21**, 45–56.

Steyerberg EW, van der Ploeg T, Van Calster B (2014). Risk prediction with machine learning and regression methods. *Biometrical Journal* **56**, 601–606.

Steyerberg EW, Vergouwe Y (2014). Towards better clinical prediction models: seven steps for development and an ABCD for validation. *European Heart Journal* **35**, 1925–1931.

Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N, Pencina MJ, Kattan MW (2010). Assessing the performance of prediction models: a framework for traditional and novel measures. *Epidemiology* **21**, 128–138.

Stowkowy J, Liu L, Cadenhead KS, Cannon TD, Cornblatt BA, McGlashan TH, Perkins DO, Seidman LJ, Tsuang MT, Walker EF, Woods SW, Bearden CE, Mathalon DH, Addington J (2016). Early traumatic experiences, perceived discrimination and conversion to psychosis in those at clinical high risk for psychosis. *Social Psychiatry and Psychiatric Epidemiology* **51**, 497–503.

Strobl EV, Eack SM, Swaminathan V, Visweswaran S (2012). Predicting the risk of psychosis onset: advances and prospects. *Early Intervention in Psychiatry* **6**, 368–379.

Sun GW, Shook TL, Kay GL (1996). Inappropriate use of bivariable analysis to screen risk factors for use in multivariable analysis. *Journal of Clinical Epidemiology* **49**, 907–916.

Thompson A, Nelson B, Yung A (2011). Predictive validity of clinical variables in the "at risk" for psychosis population: international comparison with results from the North American Prodrome Longitudinal Study. *Schizophrenia Research* **126**, 51–57.

Tibshirani R (1997). The lasso method for variable selection in the Cox model. *Statistics in Medicine* **16**, 385–395.

van der Net JB, Janssens ACJW, Eijkemans MJC, Kastelein JJP, Sijbrands EJG, Steyerberg EW (2008). Cox proportional hazards models have more statistical power than logistic regression models in cross-sectional genetic association studies. *European Journal of Human Genetics* **16**, 1111–1116.

van der Ploeg T, Austin PC, Steyerberg EW (2014). Modern modelling techniques are data hungry: a simulation study for predicting dichotomous endpoints. *BMC Medical Research Methodology* **14**, 137.

van der Ploeg T, Nieboer D, Steyerberg EW (2016). Modern modeling techniques had limited external validity in predicting mortality from traumatic brain injury. *Journal of Clinical Epidemiology* **78**, 83–89.

van Oort L, van den Berg T, Koes BW, de Vet RH, Anema HJ, Heymans MW, Verhagen AP (2012). Preliminary state of development of prediction models for primary care physical therapy: a systematic review. *Journal of Clinical Epidemiology* **65**, 1257–1266.

Velthorst E, Derks EM, Schothorst P, Becker H, Durston S, Ziermans T, Nieman DH, de Haan L (2013*a*). Quantitative and qualitative symptomatic differences in individuals at ultra-high risk for psychosis and healthy controls. *Psychiatry Research* **210**, 432–437.

Velthorst E, Nelson B, Wiltink S, de Haan L, Wood SJ, Lin A, Yung AR (2013*b*). Transition to first episode psychosis in ultra high risk populations: does baseline functioning hold the key? *Schizophrenia Research* **143**, 132–137.

Walder DJ, Holtzman CW, Addington J, Cadenhead K, Tsuang M, Cornblatt B, Cannon TD, McGlashan TH, Woods SW, Perkins DO, Seidman LJ, Heinssen R, Walker EF (2013). Sexual dimorphisms and prediction of conversion in the NAPLS psychosis prodrome. *Schizophrenia Research* **144**, 43–50.

Wessler BS, Lai Yh L, Kramer W, Cangelosi M, Raman G, Lutz JS, Kent DM (2015). Clinical prediction models for cardiovascular disease: tufts predictive analytics and comparative effectiveness clinical prediction model database. *Circulation. Cardiovascular Quality and Outcomes* **8**, 368–375.

Wynants L, Collins GS, Van Calster B (2016). Key steps and common pitfalls in developing and validating risk models. *BJOG: An International Journal of Obstetrics and Gynaecology*. Published online 30 June 2016. doi:10.1111/1471-0528.14170.

Xu L, Zhang T, Zheng L, Li H, Tang Y, Luo X, Sheng J, Wang J (2016). Psychometric Properties of Prodromal Questionnaire-Brief Version among Chinese help-seeking individuals. *PLOS ONE* **11**, e0148935.

Yung AR, Phillips LJ, McGorry PD, McFarlane CA, Francey S, Harrigan S, Patton GC, Jackson HJ (1998). Prediction of psychosis. A step towards indicated prevention of schizophrenia. *British Journal of Psychiatry. Supplement* **172**, 14–20.

Yung AR, Phillips LJ, Yuen HP, Francey SM, McFarlane CA, Hallgren M, McGorry PD (2003). Psychosis prediction: 12-month follow up of a high-risk ("prodromal") group. *Schizophrenia Research* **60**, 21–32.

Yung AR, Phillips LJ, Yuen HP, McGorry PD (2004). Risk factors for psychosis in an ultra high-risk group: psychopathology and clinical features. *Schizophrenia Research* **67**, 131–142.

Ziermans T, de Wit S, Schothorst P, Sprong M, van Engeland H, Kahn R, Durston S (2014). Neurocognitive and clinical predictors of long-term outcome in adolescents at ultra-high risk for psychosis: a 6-year follow-up. *PLOS ONE* **9**, e93994.