Towards clinical application of prediction models for transition to psychosis

PRONIA Consortium

DOI:
10.1016/j.neubiorev.2021.02.032

Document Version
Peer reviewed version

Citation for published version (Harvard):
PRONIA Consortium 2021, 'Towards clinical application of prediction models for transition to psychosis: a systematic review and external validation study in the PRONIA Sample', *Neuroscience and biobehavioral reviews*, vol. 125, pp. 478-492. https://doi.org/10.1016/j.neubiorev.2021.02.032

Highlights:

- First comprehensive validation of prediction models for transition to psychosis
- In external PRONIA validation sample, two models show good discrimination performance
- Combining predictions from raters and transition models improves performance
- Prediction of transition to psychosis is feasible on global scale
- Yet transition models need additional research efforts before clinical implementation

A multitude of prediction models for a first psychotic episode in individuals at clinical high-risk (CHR) for psychosis have been proposed, but only rarely validated. We identified transition models based on clinical and neuropsychological data through a registered systematic literature search and evaluated their external validity in 173 CHRs from the Personalised Prognostic Tools for Early Psychosis Management (PRONIA) study. Discrimination performance was assessed with the area under the receiver operating characteristic curve (AUC), and compared to the prediction of clinical raters. External discrimination performance varied considerably across the 22 identified models (AUC 0.40-0.76), with two models showing good discrimination performance. None of the tested models significantly outperformed clinical raters (AUC = 0.75). Combining predictions of clinical raters and the best model descriptively improved discrimination performance (AUC = 0.84). Results show that personalized prediction of transition in CHR is potentially feasible on a global scale. For implementation in clinical practice, further rounds of external validation, impact studies, and development of an ethical framework is necessary.

Keywords: psychosis, clinical high-risk, prediction, model validation, early intervention precision medicine, translational psychiatry

Towards Clinical Application of Prediction Models for Transition to Psychosis: A Systematic Review and External Validation Study in the PRONIA Sample

Marlene Rosen[1]*, Linda T. Betz[1]*, Frauke Schultze-Lutter[2,3,4], Katharine Chisholm[5,6], Theresa K. Haidl[1] , Lana Kambeitz-Ilankovic[1,7], Alessandro Bertolino[8], Stefan Borgwardt[9,10], Paolo Brambilla[11,12] , Rebekka Lencer[9,13], Eva Meisenzahl[2], Stephan Ruhrmann[1], Raimo K. R. Salokangas[14], Rachel Upthegrove[5], Stephen J. Wood[5,15,16], Nikolaos Koutsouleris[7,17,18], Joseph Kambeitz[1], for the PRONIA Consortium†

[1] Department of Psychiatry and Psychotherapy, Faculty of Medicine and University Hospital of Cologne, Cologne, Germany
[2] Department of Psychiatry and Psychotherapy, Medical Faculty, Heinrich-Heine University, Düsseldorf, Germany
[3] Department of Psychology and Mental Health, Faculty of Psychology, Airlangga University, Surabaya, Indonesia
[4] University Hospital of Child and Adolescent Psychiatry and Psychotherapy, University of Bern, Bern, Switzerland
[5] Institute for Mental Health and Centre for Human Brain Health, University of Birmingham, Birmingham, UK
[6] Department of Psychology, Aston University, Birmingham, UK
[7] Department of Psychiatry and Psychotherapy, Ludwig-Maximilian-University, Munich, Germany
[8] Department of Neurological and Psychiatric Sciences, University of Bari, Bari, Italy
[9] Department of Psychiatry and Psychotherapy, University of Lübeck, Lübeck, Germany
[10] Department of Psychiatry, Psychiatric University Hospital, University of Basel, Switzerland
[11] Department of Neurosciences and Mental Health, Fondazione IRCCS Ca' Granda Ospedale Maggiore Policlinico, University of Milan, Milan, Italy
[12] Department of Pathophysiology and Transplantation, University of Milan, Milan, Italy
[13] Department of Psychiatry, University of Münster, Münster, Germany
[14] Department of Psychiatry, University of Turku, Turku, Finland
[15] Orygen, Melbourne, Australia
[16] Centre for Youth Mental Health, University of Melbourne, Melbourne, Australia
[17] Max-Planck Institute of Psychiatry, Munich, Germany
[18] Institute of Psychiatry, Psychology and Neuroscience, King's College London, London, UK

* The first two authors contributed equally to this work.

Word count (incl. abstract, text, and references): 5891
Number of figures: 3
Number of tables: 3

Send correspondence to Dr. Kambeitz (Department of Psychiatry and Psychotherapy, University Hospital of Cologne, Kerpener Str. 62, 50937 Cologne, Germany; joseph.kambeitz@uk-koeln.de).

The authors report no financial relationships with commercial interests.

Funding: PRONIA is a Collaboration Project funded by the European Union under the 7th Framework Programme under grant agreement n° 602152. J.K. has received funding from the German Research Foundation (DFG; grant agreement n° KA 4413/1-1). These funding sources had no role in the design and execution of this study, nor in analyses, interpretation of the data, or decision to submit results.

† **Group Information:** PRONIA consortium members listed here performed the screening, recruitment, rating, examination, and follow-up of the study participants and were involved in implementing the examination protocols of the study, setting up its information technological infrastructure, and organizing the flow and quality control of the data analyzed in this article between the local study sites and the central study database. **Department of Psychiatry and Psychotherapy, Ludwig-Maximilian-University, Munich, Bavaria, Germany:** Prof. Nikolaos Koutsouleris, MD, Lana Kambeitz-Ilankovic, PhD, Mark Sen Dong, MSc, Anne Erkens, Eva Gussmann, MSc, Shalaila Haas, PhD, Alkomiet Hasan, MD, Claudius Hoff, MD, Ifrah Khanyaree, BSc, Aylin Melo, MSc, Susanna Muckenhuber-Sternbauer, MD, Janis Kohler, Omer Faruk Ozturk, MD, David Popovic, MD, Nora Penzel, MSc, Adrian Rangnick, BSc, Sebastian von Saldern, MD, Rachele Sanfelici, MSc, Moritz Spangemacher, Ana Tupac, MSc, Maria Fernanda Urquijo, MSc, Johanna Weiske, MSc, Julian Wenzel, MSc, and Antonia Wosgien. **Department of Psychiatry and Psychotherapy, University of Cologne, North Rhineland–Westphalia, Germany:** Prof. Joseph Kambeitz, MD, Prof. Stephan Ruhrmann, MD, Marlene Rosen, PhD, Linda Betz, MSc, Theresa Haidl, MD, Karsten Blume, Mauro Seves, MSc, Nathalie Kaiser, PhD, Tanja Pilgram, MSc, Thorsten Lichtenstein, MD, and Christiane Woopen, MD. **Psychiatric University Hospital, University of Basel, Basel, Switzerland:** Prof. Stefan Borgwardt, MD, Christina Andreou, MD, PhD, Laura Egloff, PhD, Fabienne Harrisberger, PhD, Claudia Lenz, PhD, Letizia Leanza, MSc, Amatya Mackintosh, MSc, Renata Smieskova, PhD, Erich Studerus, PhD, Anna Walter, MD, and Sonja Widmayer, MSc. **Institute for Mental Health, University of Birmingham, Birmingham, United Kingdom:** Prof. Rachel Upthegrove, MBBS FRCPsych, PhD, Prof. Stephen J. Wood, PhD, Katharine Chisholm, PhD, Chris Day, BSc, Sian Lowri Griffiths, PhD, Mariam Iqbal, BSc, Mirabel Pelton, MSc, Pavan Mallikarjun, MBBS, DPM, MRCPsych, PhD, Alexandra Stainton, MSc, and Ashleigh Lin, PhD. **Department of Psychiatry, University of Turku, Turku, Finland:** Prof. Raimo K. R. Salokangas, MD, PhD, MSc, Alexander Denissoff, MD, Anu Ellila, RN, Tiina From, MSc, Markus Heinimaa, MD, PhD, Tuula Ilonen, PhD, Paivi Jalo, RN, Heikki Laurikainen, MD, Maarit Lehtinen, RN, Antti Luutonen, BA, Akseli Makela, BA, Janina Paju, MSc, Henri Pesonen, PhD, Reetta-Liina Armio (Saila), MD, Elina Sormunen, MD, Anna Toivonen, MSc, and Otto Turtonen, MD. **General Electric Global Research Inc, Munich, Germany:** Ana Beatriz Solana, PhD, Manuela Abraham, MBA, Nicolas Hehn, PhD, and Timo Schirmer, PhD. **Workgroup of Paolo Brambilla, MD, PhD,**

University of Milan, Milan, Italy: Department of Neuroscience and Mental Health, Fondazione IRCCS Ca' Granda Ospedale Maggiore Policlinico, University of Milan, Milan, Italy: Prof. Paolo Brambilla, MD, Carlo Altamura, MD, Marika Belleri, PsychD, Francesca Bottinelli, PsychD, Adele Ferro, PsychD, PhD, and Marta Re, PhD. Programma 2000, Niguarda Hospital, Milan: Emiliano Monzani, MD, Mauro Percudani, MD, and Maurizio Sberna, MD. San Paolo Hospital, Milan: Armando D'Agostino, MD, and Lorenzo Del Fabro, MD. Villa San Benedetto Menni, Albese con Cassano: Giampaolo Perna, MD, Maria Nobile MD, PhD, and Alessandra Alciati, MD. **Workgroup of Paolo Brambilla, MD, PhD, University of Udine, Udine, Italy:** Department of Medical Area, University of Udine: Matteo Balestrieri, MD, Carolina Bonivento, PsychD, PhD, Giuseppe Cabras, PhD, and Franco Fabbro, MD, PhD. IRCCS Scientific Institute "E. Medea", Polo FVG, Udine: Marco Garzitto, PsychD, PhD and Sara Piccin, PsychD, PhD. **Workgroup of Prof. Alessandro Bertolino, University of Bari Aldo Moro, Italy:** Prof. Alessandro Bertolino, Prof. Giuseppe Blasi; Prof. Linda A. Antonucci, Prof. Giulio Pergola, Grazia Caforio, PhD, Leonardo Faio, PhD, Tiziana Quarto, PhD, Barbara Gelao, PhD, Raffaella Romano, PhD, Ileana Andriola, MD, Andrea Falsetti, MD, Marina Barone, MD, Roberta Passatiore, M.Sc., Marina Sangiuliano, MD. **Department of Psychiatry and Psychotherapy, Westfaelische Wilhelms-University Muenster, Germany:** Prof. Rebekka Lencer, Marian Surman, M.Sc., Olga Bienek, MD, Georg Romer, MD, Udo Dannlowski, MD, PhD. **Department of Psychiatry and Psychotherapy of the University Düsseldorf, Germany,** Prof. Eva Meisenzahl, MD, Frauke Schultze-Lutter, PhD, Christian Schmidt-Kraepelin, MD, Susanne Neufang, PhD, Alexandra Korda, PhD, Henrik Rohner, MD.

1. Introduction

Within preventive psychiatry, a plethora of personalized models predicting the development of a first psychotic episode in patients with a clinical high-risk (CHR) for psychosis (henceforth,"transition models") have been proposed based on different modeling strategies and data modalities, including clinical, neurocognitive, and neurobiological data (Bodatsch et al., 2015; Sanfelici et al., 2020; Studerus et al., 2017).

Following an indicated preventive approach, CHR criteria enable the identification of a group of adolescents and young adults showing first signs of the potentially developing disorder who often are in need for treatment and experience serious functional impairments (Catalan et al., 2020; Fusar-Poli et al., 2013, 2020). However, a recent meta-analysis suggests transition risk in CHR individuals of 22% at 3-years (Fusar-Poli et al., 2020), reflecting only moderate capacity of currently established CHR criteria to rule in psychosis. One suggested reason behind the associated low specificity is risk enrichment in CHR samples, an inherent consequence of recruitment strategies aimed at indicated prevention in help-seeking samples (Fusar-Poli et al., 2012, 2016; F. Schultze-Lutter et al., 2015).

Given the only moderate clinical utility of the CHR status to rule in psychosis risk (Fusar-Poli et al., 2015), transition models aim at improving prognostic accuracy in CHR samples to inform patients about their risk more accurately, compliant with ethical principles (Starke et al., 2020; Woods et al., 2020). At the same time, improved transition prediction may prevent unnecessary treatment and self- and other stigma (Moritz et al., 2019). The hope is that transition models based on resource-efficient data, such as clinical and neuropsychological variables, can inform clinical practice in two ways. First, by refining prognosis and identifying decisive factors for transition to psychosis, these models can inform interventions in CHR populations in the context of precision medicine, preventing burden associated with psychotic disorders. Second, transition models could serve as a first

stage before costlier assessments, such as neuroimaging, in the context of adaptive sequential testing (Koutsouleris et al., 2018; Ruhrmann et al., 2010; Schmidt et al., 2017). However, the external validity and actual benefit for clinical practice of the published transition models remains unclear. Until today, predicting transition to psychosis in CHR populations has stalled in the stage of model development and models are rarely used in clinical practice (Salazar de Pablo et al. 2020). Crucial subsequent steps, namely external validation and impact evaluations of transition models on clinical practice and health outcomes (Riley et al., 2016; Steyerberg et al., 2013), have been largely neglected, with few exceptions (Addington et al., 2017; Carrión et al., 2016; Malda et al., 2019; Oliver et al., 2020; Wang et al., 2020; Zhang, Yang, et al., 2019). Overall, only 0.2% (n = 1) of published prediction modeling studies in psychiatry aimed to investigate the actual implementation of a prediction model in clinical practice (Salazar de Pablo et al., 2020; Wang et al., 2020). Hence, it is uncertain which of the plethora of proposed transition models are reliable and under which circumstances (Studerus et al., 2017). Moreover, methodological shortcomings, such as lack of adequate validation strategies and small sample sizes, might have led to an optimistic bias in the performance of the published transition models (Royston & Altman, 2013; Sanfelici et al., 2020; Studerus et al., 2017). As a result, it is an open question how well these transition models will generalize to new CHR populations, restricting their immediate clinical usefulness.

In order to evaluate the potential benefit of proposed prognostic models predicting transition to psychosis in CHR individuals and to pave their way into clinical application, we conducted a systematic external validation of transition models in the CHR population acquired within a large multicenter study, the Personalised Prognostic Tools for Early Psychosis Management (PRONIA) study (Koutsouleris et al., 2018). We focused on regression models developed in CHR samples with readily obtainable data, such as clinical

and demographic information as well as neuropsychological data, that can be applied in most clinical contexts. Beyond generalizability to unseen data, a transition model should also demonstrate some degree of clinical effectiveness, i.e. enhancing a clinician's decision-making (Altman & Royston, 2000; Wyatt & Altman, 1995). Thus, we evaluated the potential usefulness of transition models by determining their gain in predicting transition to psychosis compared to the individualised prediction of transition by clinical raters who had assessed patients extensively.

2. Methods

2.1 Identification of relevant models

Based on PRISMA guidelines (Moher et al., 2009), we identified relevant transition models (i.e. prognostic models developed in CHR individuals intended to predict transition to psychosis in CHR individuals), through systematic literature search in the databases *PubMed* and *Web of Science* from January 1, 1990 up to April 30, 2020, using the term (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 protocol was publicly registered at https://osf.io/bnq4s. To be eligible for the present validation study, articles had to: (1) be written in English; (2) report original data (as opposed to e.g., reviews and study protocols); (3) be published in a peer-reviewed journal; (4) involve a group of individuals (as opposed to e.g., single case or animal studies) (5) with a CHR for psychosis (internationally recognized criteria, i.e. ultra-high risk (UHR) or basic symptoms (BS)) (Schultze-Lutter et al., 2015; Studerus et al., 2017) that were prospectively followed up (6) and on the basis of which is a prognostic model that predicted later transition to psychosis from variables obtained at baseline developed (7) using demographic, clinical or neuropsychological data only (8) with a regression model (either Cox or logistic regression, including regularized

variants) (9) including at least two significant predictor variables in the final prognostic model ('multivariable model'); (10) use only variables (or variables measuring the same concept) in the final model that were available in our validation data set; (11) report sufficient details to allow application of the model in an external sample. We also checked recently published overview articles (Sanfelici et al., 2020; Studerus et al., 2017) for additional studies. Two researchers (MR, LB) performed the literature search. Discrepancies were solved in a consensus meeting with a third author (JK). Importantly, we did not exclude studies that used different questionnaires or assessments for a given predictor variable (e.g., different neurocognitive tests assessing auditory verbal learning), given that methodologic transportability is one formal aspect of generalizability, i.e., if models maintain accuracy when they are tested in data collected using alternative methods (Justice et al., 1999). However, we required available assessments in the PRONIA sample to be comparable or convertible to the scale of assessment used in an eligible prediction model (as is the case, for example, when z-scores are used in the prediction model). Similarly, we included studies with variable follow-up intervals, which allowed us to assess follow-up period transportability, another aspect of model generalizability that reflects whether prediction models maintain accuracy when predictions are tested over a longer or shorter follow-up period (Justice et al., 1999). Included studies were evaluated with respect to their methodological quality according to the Newcastle-Ottawa Scale (NOS (Wells, 2001)) to which we added three statistical criteria (dichotomization of continuous variables, event per predictor variable rate, validation strategy) to evaluate the quality of model development.

2.2 Prediction of clinical raters

We compared the identified data-driven models for prediction of psychosis to the individualised prediction of later transition (yes/no) provided by the health care professionals (all psychologists or psychiatrists) in the PRONIA consortium (henceforth, 'raters')

(Koutsouleris et al., 2018). Specifically, at the end of the extensive PRONIA assessment, the attending clinical rater had to provide a prediction as to whether the assessed patient would, in his or her opinion, transition to psychosis at some point or not. We also explored whether an integration of data-driven clinical models and the prediction of transition by clinical raters could improve prognostic accuracy ('combined model'). To this end, we averaged the prognostic score from the best-fitting model in our sample with prediction of transition by clinical raters. To assess effects of rater experience on prognostic accuracy, we stratified raters by experience with assessment of CHR participants into experienced (≥ 24 months of experience at the time of assessment) and less experienced raters (< 24 months of experience at the time of assessment) using a median split.

2.3 Validation Sample

We validated the eligible models in the CHR sample of the multisite, naturalistic PRONIA study (German Clinical Trials Register identifier DRKS00005042). Details on the study rationale and protocol are available elsewhere (Koutsouleris et al., 2018). In brief, PRONIA followed up individuals at CHR, patients with recent-onset depression, patients with recent-onset psychosis, and healthy control participants in 10 academic early-recognition services in 5 European countries. Participants were recruited between February 2014 and November 2017. The scheduled follow-up time was 18 months, which was extended given the patient's consent. The CHR state in PRONIA was defined by: (1) cognitive disturbances (COGDIS), as assessed by the Schizophrenia Proneness Instrument (SPI-A (Frauke Schultze-Lutter et al., 2007)); and/or (2) adapted PRONIA ultra-high-risk (UHR) criteria for psychosis, as measured by the Structured Interview for Psychosis-Risk Syndromes (SIPS (McGlashan et al., 2010)). For inclusion and exclusion criteria, see supplementary method 1. Transition to psychosis was defined by the presence of at least one

SIPS positive item with a severity score of 6 for more than seven days. CHR assessments and transitions to psychosis were supervised in monthly telephone conferences by an expert in CHR assessments (FSL).

All participants provided their written informed consent prior to study inclusion. For underage participants, guardians provided written informed assent in addition. Local research ethics committees at each site approved the study protocol.

2.4 Statistical analyses

We performed all analyses in the *R* language for statistical computing, version 3.6.3 (R Core Team, 2020). Throughout, we considered a significance level of $\alpha < .05$. For descriptive statistics, we computed Welch's two sample t-tests for continuous, Wilcoxon rank-sum tests for ordinal, and χ^2-tests for categorical data.

2.4.1 Model validation

We attempted to use variables from the PRONIA validation data set that corresponded exactly, or as closely as possible to those used in the identified prediction models (details in supplementary method 2 and supplementary table 1). We allowed up to 25% missing values per subject in each predictor model, and imputed missing values via the k-Nearest Neighbour (kNN) algorithm. For external validation of the identified models, we assessed discrimination and calibration in the PRONIA CHR sample. Discrimination refers to the ability of a prediction model to distinguish low-risk from high-risk individuals (Altman & Royston, 2000; Royston & Altman, 2013). We computed the prognostic index (PI) for each CHR participant (supplementary method 3). We used area under the receiver operator characteristic curve (AUC) to measure discrimination performance of the models in the PRONIA data. 95%-confidence intervals (CI) for the AUC were computed based on 1000 stratified bootstrap samples. To determine the optimal PI-cutpoint, we optimized the Youden index. Based on the optimal PI-cutpoint, we computed additional end-point related performance measures -

sensitivity, specificity, balanced accuracy (BAC), positive predictive value (PPV), negative predictive value (NPV) and positive and negative likelihood ratio (LR). For each model, we performed a permutation test (1000 permutations) to assess whether its performance (AUC) was significantly better than chance (Ojala & Garriga, 2010). Additionally, we quantified generalizability of each model in terms of the difference in AUC between the development and the validation sample within the models that report development performance, and following previous literature (Salazar de Pablo et al., 2020), provide the Pearson correlation between development and validation performance. For each model, we also tested if it performed significantly better than prediction of transition by clinical raters using a permutation test (1000 permutations).

Calibration refers to the accuracy of absolute risk estimates, i.e. the agreement between predicted and observed risks (Altman & Royston, 2000; Royston & Altman, 2013). A clinical prediction model is well-calibrated if, for example, it predicts a probability of 40% risk of transition to psychosis for one participant at CHR, and similar participants would transition to psychosis 4 out of 10 times. Calibration in the strict sense can only be assessed if the baseline hazard function is reported (Royston & Altman, 2013). If that was the case, we compared the predicted population-averaged cumulative hazard curves per proposed risk class of the model (typically based on the PI) with the observed Kaplan-Meier cumulative hazard curves of these risk classes (Royston, 2015) (supplementary methods 4). In all other cases, we used a simple form of calibration by estimating the regression slope on the PI with the PI as the single covariate. Regression slopes smaller than 1 indicate overfitting and need for recalibration (Royston & Altman, 2013). When available, we compared Kaplan-Meier cumulative hazard curves of risk classes proposed in the development data, which supports a rough evaluation of model calibration (Royston & Altman, 2013; Steyerberg et al., 2013).

2.4.2 Decision curve analysis

For the best performing model in our external validation data set, we additionally performed decision curve analysis (Vickers et al., 2019; Vickers & Elkin, 2006). In decision curve analysis, a clinical judgment of the benefit-harm ratio associated with the prediction by a model is plotted against threshold probability, defined as the minimum probability of transition to psychosis at which further intervention would be warranted. In the present case, benefits may refer to preventing psychosis in CHR individuals, and harms may involve, among other aspects, unnecessary treatment. Benefits are displayed as net benefits, i.e., the difference between true positive and false positives predictions from the model. The threshold probability reflects a clinician's weighting of differential outcomes in decision-making: the harm of missing a transition to psychosis on the one end and the harm of unnecessary treatment on the other; the lower the threshold probability, the lower the perceived harm of unnecessary treatment compared to the perceived benefits of predicting transition to psychosis (Vickers et al., 2019). For subsequent net benefit analysis, we used 7.7% as a reference threshold probability for recommending indicated interventions to prevent psychosis (Fusar-Poli et al., 2017). Converted to odds, this reflects a clinician's belief that missing a transition to psychosis is about 12 times worse than commencing unnecessary treatment (Fusar-Poli et al., 2017; Vickers et al., 2019).

2.4.3 Sensitivity analysis: meta-regression

We used multiple linear regression analysis to assess the influence of the following study characteristics on model performance (measured by AUC) in the PRONIA CHR sample: year, continent (Europe vs. not Europe), CHR assessment (SIPS vs. CAARMS), dichotomization of predictors in model development (yes vs. no), event per predictor rate, internal or external validation strategy used during model development (yes/no), difference in follow-up period to PRONIA 18-month-follow-up, predictors replaced in validation (yes/no)

3. Results

3.1 Validation Models

In the literature search, we identified 1806 unique records. Of these, $k = 22$ prediction models for transition to psychosis were eligible and included in the present validation study (Addington et al., 2017; Carrión et al., 2018; Cornblatt et al., 2015; Dragt et al., 2011; Haidl et al., 2018; Hengartner et al., 2017; Kotlicka-Antczak et al., 2019; Lemos-Giráldez et al., 2009; Lencz et al., 2006; Lin et al., 2011; Malda et al., 2019; Metzler et al., 2016; Michel et al., 2014; Niles et al., 2019; Rosen et al., 2019; Ruhrmann et al., 2010; Salokangas et al., 2012; Thompson et al., 2011; Walder et al., 2013; Yung et al., 2004; Zhang et al., 2020; Zhang, Xu, et al., 2019) (table 1). For a flow-chart of the literature search, see supplementary figure 1. Note that for studies that report separate models for complementary subgroups based on sex (Rosen et al., 2019; Walder et al., 2013), we referred to the weighted aggregate performance of the separate models in the results section, unless stated otherwise. One study (Zhang et al., 2020) reported age-specific prediction models for adolescents and adults. Here, formal external validation was only performed on the adult-model, given very limited prevalence of adolescent CHR participants (n = 10, including only 1 transition) in the PRONIA sample, which would result in uninformative measures of external discrimination performance.

In terms of study quality, all studies achieved a score between 6 and 10 (median = 8) of 13 possible points on our adapted version of the NOS (supplementary table 2). Quality differences emerged in inclusion of control variables, drop-out rate, event per predictor ratio, dichotomization of continuous variables and application of model validation strategies.

3.2 External validation sample

Table 2 shows baseline demographic and clinical data for the PRONIA CHR sample used for external validation. In total, this sample comprised 277 participants with validated

CHR status. Of these, 173 had data available at least until the 18-month follow-up and were used for external validation (cumulative hazard function in supplementary figure 2). The validation sample comprised individuals transitioning to psychosis (n = 23) as well as those without transition during 18 months (n = 150). Due to different amounts of missing variables, the validation sample slightly varied across models (N between 148 and 173).

3.3 Validation performance

3.3.1 Discrimination

Discrimination performance in the external validation sample varied considerably across the tested studies (AUC = 0.40-0.76; table 3). For an overview of the contribution of different predictor domains to each model, see supplementary figure 3. All but eight models (Cornblatt et al., 2015; Dragt et al., 2011; Kotlicka-Antczak et al., 2019; Lemos-Giráldez et al., 2009; Rosen et al., 2019; Ruhrmann et al., 2010; Thompson et al., 2011; Yung et al., 2004; Zhang et al., 2020) predicted transition to psychosis significantly better than chance in the PRONIA data set. Only two of the 22 transition prediction models, the models by Hengartner (Hengartner et al., 2017) and Lencz (Lencz et al., 2006), both of which comprised a combination of positive symptoms and neurocognitive data, as well as the prediction of transition by clinical raters, achieved acceptable discrimination performance with an AUC ≥ 0.70 (Hosmer, 2000) (figure 1 for ROC curves). The combination of the best-performing transition prediction model by Lencz (Lencz et al., 2006) and prediction of transition by clinical raters yielded excellent discrimination performance (AUC ≥ 0.80 (Hosmer, 2000)). Performance of inexperienced raters (AUC = 0.68) rose markedly when combined with the Lencz model (ΔAUC = 0.13). Conversely, performance of experienced raters was already high (AUC = 0.81), with little improvement through the combination with the Lencz-model (ΔAUC = 0.02).

For 11 models, discrimination performance in the development sample was reported. Relative to the discrimination performance achieved in their respective development samples, performance of these models dropped when applied in the PRONIA external validation sample, except for the model by Malda (Malda et al., 2019). This finding might indicate varying degrees of overfitting in the large majority of tested models (supplementary figure 4). Development and validation discrimination performance was weakly negatively correlated (r = -0.13, 95% CI: -0.68-0.51, p = .693); yet influence analysis indicated that this result was driven by the study of Cornblatt et al. (2015) that showed substantially worse validation than development performance. After removal of this outlier study, there was a modest positive, albeit non-significant correlation between discrimination performance in development and validation sample (r = 0.42, 95% CI: -0.29-0.83, p = .227).

3.3.2 Calibration

Calibration in the strict sense was only possible for the Addington-model (2017). Results from the calibration analysis (supplementary figure 5) suggested that this model underestimated the individual transition probabilities in the PRONIA external validation sample across all proposed risk classes; i.e., outcomes tended to be worse than predicted (Royston & Altman, 2013). For the other models, only limited evaluation of calibration was possible. Calibration slopes smaller 1 indicated varying extents of overfitting in the development of the majority of the models tested. Kaplan-Meier hazard curves per risk class (available for five models, supplementary figures 6a-e) indicated relatively good agreement with the Kaplan-Meier hazard curves reported in the development sample for the model by Addington (Addington et al., 2017), but only poor agreement for the models by Haidl (Haidl et al., 2018), Michel (Michel et al., 2014), Ruhrmann (Ruhrmann et al., 2010), and Dragt (Dragt et al., 2011).

3.4 Decision curve analysis

Decision curve analysis (figure 3) showed that use of the transition model proposed by Lencz et al. (2006) was associated with relevant net benefits (≥ 0.01) across a wide range of threshold probabilities, starting at 4.6% compared to 'treat all' (figure 3). At the reference threshold of 7.7%, the net benefit compared to 'treat none' was 0.06, which translates to a strategy where treatment is commenced for 6 out of 100 CHR patients who would later develop psychosis, without conducting unnecessary treatment in any non-transitioners. Compared to 'treat all', use of the transition model by Lencz et al. (2006) would be equivalent to reducing the number of unnecessary treatments by about 22 per 100 CHR patients at a threshold probability of 7.7%, without missing treatment for any patients who would later transition to psychosis.

3.5 Sensitivity analysis: meta-regression

Explorative multiple regression analysis showed no significant impact of any of the model characteristics on external validation performance in the PRONIA sample (supplementary table 3). On statistical trend-level, there was evidence that those models that used SIPS for ascertainment of the CHR status performed better in the PRONIA sample than those that based inclusion on CAARMS (β = 0.10, 95% CI: -0.006-0.20, t = 2.00, p = .064). Looking at methodological characteristics, we found suggestive evidence that those models that featured dichotomized predictors showed worse external validation performance in the PRONIA sample (β = -0.09, 95% CI: -0.17-0.002, t = -2.10, p = .054). Finally, a higher event per predictor rate was, on trend-level, associated with better external discrimination performance in the PRONIA sample (β = 0.003, 95% CI: -0.001-0.007, t = 1.78, p = .096).

4. Discussion

We provide the first comprehensive external validation study of transition models conducted in the CHR sample of the multisite European PRONIA study. For maximal feasibility of clinical application, we focused our systematic literature search on Cox and

logistic regression models based on resource-efficient clinical and neuropsychological data. In total, we identified 22 transition models for validation that were developed on samples recruited from 1993 to 2016 in 14 countries with different health care systems across four continents with an average age range from 16 to 25 years.“

The external performance of the tested models in predicting transition to psychosis in terms of AUC ranged from 0.40 to 0.76, suggesting that the performance of the models varied from below chance level to candidate models that identified three-fourths of transitions correctly. Across all models, the PPV was low (but comparable to other medical fields), while the NPV was high, which is, in part, attributable to the overall low transition rate in our sample. Additionally, risk enrichment by opportunistic recruitment strategies identified as a crucial factor driving low transition risk in CHR samples, reflected by low PPV, needs to be better controlled for and deconstructed (Fusar-Poli et al., 2016). Positive LRs, which reflect the likelihood of obtaining a true transition in a patient with a predicted transition, independent of prevalence of transition, ranged from 1.13 to 4.95. This suggests that the validity of a positive transition prediction varied substantially across transition models, from providing virtually no or only very limited additional information about transition (LR+ < 2) to about a fivefold increase in transition likelihood given a positive test result. Arguably, a prediction model should demonstrate higher clinical utility than the CHR status alone, for which meta-analytic evidence suggests a LR+ of 1.82 (Fusar-Poli et al., 2015). Negative LRs of the validated transition models ranged from 0 to 0.91, suggesting that the meaning of a negative test result also differed considerably across models, from yielding no or only very limited additional indications for a non-transition (LR- > 0.5) in our external validation sample to effectively ruling out transition given a negative test result.

Thirteen of the 22 tested prediction models for transition to psychosis showed above chance performance in the PRONIA validation sample. Of these, the models by Hengartner

(Hengartner et al., 2017) and Lencz (Lencz et al., 2006) showed a generally acceptable prediction accuracy (AUC ≥ 0.70 (Hosmer, 2000)) in the PRONIA CHR sample. The Lencz-model (2006) performed best despite having been developed in a substantially younger, North American CHR sample, underlining an external validity promising for implementation. The model by Hengartner (2017) achieved a high sensitivity (95%), yet only limited specificity (45%). Reflecting this, LRs suggested that a negative prediction from this model was considerably more informative than a positive prediction. The model by Lencz (2006), by contrast, achieved a better balance between sensitivity (71%) and specificity (78%). LRs showed that positive and negative predictions from this model were approximately equally informative, increasing the likelihood of transition and non-transition about threefold, respectively. The increased LR+ represents a 1.8-fold increase in clinical utility above the CHR status alone (Fusar-Poli et al., 2015). Decision curve analysis corroborated the potential clinical utility of the Lencz-model across a wide range of threshold probabilities. In other words, use of the Lencz-model may be overall beneficial, largely independent of how much a clinician weighs the potential harms of a missed transition against those associated with a false positive prediction, such as unnecessary treatment.

When the Lencz-model was combined with rater prognoses, sensitivity increased substantially to 94%, while specificity dropped to 67%. Particularly a negative prediction from the combined model became more informative compared to the Lencz-model alone: the likelihood of a non-transition increased from factor 2.6 to 11.9.

Two studies proposed distinct models for complementary subgroups based on sex (Rosen et al., 2019; Walder et al., 2013). On aggregate, we found no evidence that these subgroup approaches outperformed generalized models.

Due to limited reporting, assessment of calibration in the strict sense was only possible for the Addington-model (Addington et al., 2017). This model predicted lower

transition probabilities for individual CHR participants than we observed in the PRONIA sample. PRONIA CHR participants were on average 4 years older than their counterparts in the PREDICT sample, reflecting well-known age-differences between European and North American CHR samples (Sanfelici et al., 2020), which may explain the overall higher transition risk in the PRONIA sample. Additionally, the PREDICT CHR sample was antipsychotic-naive at baseline, while there was (limited) exposure to antipsychotics in PRONIA CHR participants, associated with higher risk of transition (Cannon et al., 2008; Ruhrmann et al., 2010). Commencement of antipsychotic treatment may be more likely given severe psychosis-like symptoms on the verge of transition. For the other models, regression slopes on the PI provided a coarse indicator for the need for recalibration.

Overall, all the studies that reported development performance, except the Malda-transition model, performed worse in the validation than in the development sample. The Malda-model was built on the to-date largest amount of data collected in 15 studies, yielding a modest but probably realistic discrimination accuracy given that only basic demographic and clinical variables were used for prediction (Malda et al., 2019). A prognostic model built on such a substantial amount of data captures many aspects of transition, and therefore generalizes well to unseen data. Apart from being generalizable to new populations, prediction models should be clinically effective, i.e. enhance a clinician's decision beyond their baseline performance (Altman & Royston, 2000; Wyatt & Altman, 1995).

In the PRONIA data set, none of the tested transition models significantly outperformed the individualised prediction of transition by clinical raters. In line with a recently proposed multi-modal transition prediction model generated in the PRONIA sample (Koutsouleris et al., 2020), complementing prediction of transition by clinical raters with the best prediction model yielded the highest performance in the present data set. Transition models based on single specific information and clinical raters - having access to all

information of the extended assessments - may capture complementary aspects of prognostic information. Transition models may support clinical decision-making particularly in the presence of uncertainty, e.g. given an ambiguous clinical presentation or a clinician less experienced with at-risk and early stages of psychosis, which is likely the case for many practitioners working outside specialized early recognition centers. Less experienced raters in the PRONIA consortium had greater difficulties in correctly predicting transition to psychosis than experienced raters, and benefited particularly from transition models. Ultimately, if and under what circumstances the practical implementation of a transition model improves outcomes by informing clinical decision-making according to the model's predicted risk can only be determined by impact studies, ideally in a randomized controlled trial (Riley et al., 2016). Such impact studies should preferably only involve models that demonstrated generalizability in several external validation studies (Altman & Royston, 2000; Riley et al., 2016). Hence, the present study is only the first step towards clinical application of transition models (Addington et al., 2017; Carrión et al., 2016; Malda et al., 2019; Zhang, Yang, et al., 2019).

Predictions from transition models are most useful in clinical practice if they allow direct conclusions for an individual case. One convenient way is to implement the model in an online tool that automatically derives a risk estimate for an individual, as available for the model by Kotlicka-Antczak (Kotlicka-Antczak et al., 2019). Another way is to report risk classes that allow to stratify individuals based on the PI obtained from the model. Such risk classes were provided by five of the eleven tested models, the Ruhrmann-, Dragt-, Michel-, Haidl- and Addington-model (Addington et al., 2017; Dragt et al., 2011; Haidl et al., 2018; Michel et al., 2014; Ruhrmann et al., 2010). Risk classes allow to hold everyone at-risk or in need of treatment in focus, based on their predicted risk. At the same time, risk classes facilitate more accurate prognosis about transition to psychosis. Of the models with risk

classes examined in the present study, only the model by Addington (Addington et al., 2017) provided good agreement of transition risk between the development sample and our external validation sample. However, external discrimination performance of this model was not en par with the best models validated here, and its usefulness in clinical practice may therefore only be limited.

In the present external validation sample, the two best models, the Hengartner- (Hengartner et al., 2017) and the Lencz-model (Lencz et al., 2006), included a combination of clinical and neuropsychological predictors. Enrichment of core symptoms by neurocognitive data seems recommendable in the development of transition models. Overall, these results can be interpreted in the context of a meta-analysis that pointed to the benefits of combining different data types and modalities for predicting psychosis in a sequential testing approach (Schmidt et al., 2017). Only multiple data types and modalities may fully capture the complexity of the multifaceted architecture of psychosis risk (Koutsouleris et al., 2018, 2020; Sanfelici et al., 2020). In this context, there is also a growing call to enrich early recognition of psychosis by information beyond phenomenological aspects of psychosis (Schmidt et al., 2017). Evidently, models tested in the present study included relatively similar predictors. Hence, it remains to be tested if the addition of risk factors less specific to psychosis, such as depressive symptomatology, may improve prediction of outcomes in CHR populations and beyond (Fusar-Poli et al., 2017; McGorry et al., 2018; Sanfelici et al., 2020).

Mirroring findings from recent reviews (Christodoulou et al., 2019; Sanfelici et al., 2020; Studerus et al., 2017), the analysis of study quality highlights the need for conceptual and methodological guidelines to enhance model comparability and replicability in modern preventive psychiatry (Fusar-Poli et al., 2018; Sanfelici et al., 2020; Steyerberg et al., 2013): Few studies adopted rigorous internal model validation strategies, external validation in truly independent samples or adequate variable selection methods to evaluate and counteract

overfitting, respectively, and several models employed previously criticized dichotomization of model predictors that was indeed frequently associated with overall worse performance in our sample (Collins et al., 2016; Royston et al., 2006). Exploratory multiple regression analysis tentatively suggested that a higher event per predictor rate enhanced generalizability to unseen data in PRONIA, corroborating previous recommendations (Fusar-Poli et al., 2018; Studerus et al., 2017). Calibration of risk predictions, even though essential for clinical prediction models (Christodoulou et al., 2019; Royston & Altman, 2013), was rarely examined. In line with previous observations (Christodoulou et al., 2019; Studerus et al., 2017), we found that reporting of methodology and findings was sometimes ambiguous and incomplete such that application in an external sample was difficult or not possible.

Arguably, some of the models we examined were not explicitly proposed for clinical application. For a comprehensive picture, we included all prediction models. There are also several differences between the PRONIA validation sample and the development samples, i.e. inclusion criteria, transition rate, length of follow-up period, and operationalization of variables; none of which affected external validation performance in the PRONIA sample significantly. In particular, the role of different definitions of CHR criteria by different assessments remains unclear, and if a harmonization of these criteria will entail increased predictive performance (Malda et al., 2019). As clinical practice with variable assessment conditions is the intended target of the prediction models, extreme tests of generalizability in diverse samples are necessary to test if factors such as age, sex, cultural background, differences in follow-up interval, recruitment strategy and transition rate or the replacement of predictors with equivalent variables influence model generalizability to independent samples (Justice et al., 1999). However, modifications of proposed transition models by replacement of predictors might represent a limitation to our findings that needs to be investigated more in-depth in future studies. Finally, development as well as validation of

transition models was focused on the very specific context of CHR samples in our study. Comparison to models based on transdiagnostic samples and usefulness in more diverse and broader contexts such as secondary mental health care (Fusar-Poli et al., 2017) needs to be addressed in future work.

5. Conclusion

This first comprehensive external validation study showed that personalised prediction of transition in CHR is potentially feasible on a global scale. Complementing different prognostic assessments may refine prognostic accuracy in CHR samples. Two of 22 tested transition models based on clinical and neuropsychological data achieved satisfactory prediction performance in the European PRONIA CHR sample, which revealed promising potential for improvement of clinical decision-making by personalized risk prediction models, in particular for less experienced raters. To reach utility within preventive psychiatry, transition models need further rounds of external validation, as well as guidelines consolidating model comparability and replicability. Moreover, to finally close the translational gap in the field, impact studies developing a general and ethical framework for implementation are necessary (Fusar-Poli et al., 2018, 2020; Salazar de Pablo et al., 2020; Starke et al., 2020; Wang et al., 2020). Our study provides an important first step to pave the way for implementing transition models into clinical application to improve the personalized care of psychosis and prevent the burden of the disorder.

References

Addington, J., Liu, L., Perkins, D. O., Carrion, R. E., Keefe, R. S. E., & Woods, S. W. (2017). The Role of Cognition and Social Functioning as Predictors in the Transition to Psychosis for Youth With Attenuated Psychotic Symptoms. *Schizophrenia Bulletin*, *43*(1), 57–63. https://doi.org/10.1093/schbul/sbw152

Altman, D. G., & Royston, P. (2000). What do we mean by validating a prognostic model? *Statistics in Medicine*, *19*, 453. https://doi.org/10.1002/(SICI)1097-0258(20000229)19:4<453::AID-SIM350>3.0.CO;2-5

Bodatsch, M., Brockhaus-Dumke, A., Klosterkötter, J., & Ruhrmann, S. (2015). Forecasting psychosis by event-related potentials-systematic review and specific meta-analysis. *Biological Psychiatry*, *77*(11), 951–958. https://doi.org/10.1016/j.biopsych.2014.09.025

Cannon, T. D., Cadenhead, K., Cornblatt, B., Woods, S. W., Addington, J., Walker, E., Seidman, L. J., 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*(1), 28–37. https://doi.org/10.1001/archgenpsychiatry.2007.3

Carrión, R. E., Cornblatt, B. A., Burton, C. Z., Tso, I. F., Auther, A. M., Adelsheim, S., Calkins, R., Carter, C. S., Niendam, T., Sale, T. G., Taylor, S. F., & McFarlane, W. R. (2016). Personalized Prediction of Psychosis: External Validation of the NAPLS-2 Psychosis Risk Calculator With the EDIPPP Project. *The American Journal of Psychiatry*, *173*(10), 989–996. https://doi.org/10.1176/appi.ajp.2016.15121565

Carrión, R. E., Walder, D. J., Auther, A. M., McLaughlin, D., Zyla, H. O., Adelsheim, S., Calkins, R., Carter, C. S., McFarland, B., Melton, R., Niendam, T., Ragland, J. D., Sale, T. G., Taylor, S. F., McFarlane, W. R., & Cornblatt, B. A. (2018). From the psychosis prodrome to the first-episode of psychosis: No evidence of a cognitive decline. *Journal of Psychiatric Research*, *96*, 231–238. https://doi.org/10.1016/j.jpsychires.2017.10.014

Catalan, A., Salazar de Pablo, G., Vaquerizo Serrano, J., Mosillo, P., Baldwin, H., Fernández-Rivas, A., Moreno, C., Arango, C., Correll, C. U., Bonoldi, I., & Fusar-Poli, P. (2020). Annual Research Review: Prevention of psychosis in adolescents - systematic review and meta-analysis of advances in detection, prognosis and intervention. *Journal of Child*

Psychology and Psychiatry, and Allied Disciplines. https://doi.org/10.1111/jcpp.13322

Christodoulou, E., Ma, J., Collins, G. S., Steyerberg, E. W., Verbakel, J. Y., & Van Calster, B. (2019). A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *Journal of Clinical Epidemiology*, *110*, 12–22. https://doi.org/10.1016/j.jclinepi.2019.02.004

Collins, G. S., Ogundimu, E. O., Cook, J. A., Manach, Y. L., & Altman, D. G. (2016). Quantifying the impact of different approaches for handling continuous predictors on the performance of a prognostic model. *Statistics in Medicine*, *35*(23), 4124–4135. https://doi.org/10.1002/sim.6986

Cornblatt, B. A., Carrión, R. E., Auther, A., McLaughlin, D., Olsen, R. H., John, M., & Correll, C. U. (2015). Psychosis Prevention: A Modified Clinical High Risk Perspective From the Recognition and Prevention (RAP) Program. *The American Journal of Psychiatry*, *172*(10), 986–994. https://doi.org/10.1176/appi.ajp.2015.13121686

Dragt, S., Nieman, D. H., Veltman, D., Becker, H. E., van de Fliert, R., de Haan, L., & Linszen, D. H. (2011). Environmental factors and social adjustment as predictors of a first psychosis in subjects at ultra high risk. *Schizophrenia Research*, *125*(1), 69–76. https://doi.org/10.1016/j.schres.2010.09.007

Fusar-Poli, P., Bonoldi, I., Yung, A. R., Borgwardt, S., Kempton, M. J., 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*(3), 220–229. https://doi.org/10.1001/archgenpsychiatry.2011.1472

Fusar-Poli, P., Borgwardt, S., Bechdolf, A., Addington, J., Riecher-Rössler, A., Schultze-Lutter, F., Keshavan, M., Wood, S., Ruhrmann, S., Seidman, L. J., Valmaggia, L., Cannon, T., Velthorst, E., De Haan, L., Cornblatt, B., Bonoldi, I., Birchwood, M., McGlashan, T., Carpenter, W., … Yung, A. (2013). The Psychosis High-Risk State: A

Comprehensive State-of-the-Art Review. *JAMA Psychiatry* , *70*(1), 107–120. https://doi.org/10.1001/jamapsychiatry.2013.269

Fusar-Poli, P., Cappucciati, M., Rutigliano, G., Schultze-Lutter, F., Bonoldi, I., Borgwardt, S., Riecher-Rössler, A., Addington, J., Perkins, D., Woods, S. W., McGlashan, T. H., Lee, J., Klosterkötter, J., Yung, A. R., & McGuire, P. (2015). At risk or not at risk? A meta-analysis of the prognostic accuracy of psychometric interviews for psychosis prediction. *World Psychiatry: Official Journal of the World Psychiatric Association* , *14*(3), 322–332. https://doi.org/10.1002/wps.20250

Fusar-Poli, P., Hijazi, Z., Stahl, D., & Steyerberg, E. W. (2018). The Science of Prognosis in Psychiatry: A Review. *JAMA Psychiatry* , *75*(12), 1289–1297. https://doi.org/10.1001/jamapsychiatry.2018.2530

Fusar-Poli, P., Rutigliano, G., Stahl, D., Davies, C., Bonoldi, I., Reilly, T., & McGuire, P. (2017). Development and Validation of a Clinically Based Risk Calculator for the Transdiagnostic Prediction of Psychosis. *JAMA Psychiatry* , *74*(5), 493–500. https://doi.org/10.1001/jamapsychiatry.2017.0284

Fusar-Poli, P., Salazar de Pablo, G., Correll, C. U., Meyer-Lindenberg, A., Millan, M. J., Borgwardt, S., Galderisi, S., Bechdolf, A., Pfennig, A., Kessing, L. V., van Amelsvoort, T., Nieman, D. H., Domschke, K., Krebs, M.-O., Koutsouleris, N., McGuire, P., Do, K. Q., & Arango, C. (2020). Prevention of Psychosis: Advances in Detection, Prognosis, and Intervention. *JAMA Psychiatry* , *77*(7), 755–765. https://doi.org/10.1001/jamapsychiatry.2019.4779

Fusar-Poli, P., Schultze-Lutter, F., Cappucciati, M., Rutigliano, G., Bonoldi, I., Stahl, D., Borgwardt, S., Riecher-Rössler, A., Addington, J., Perkins, D. O., Woods, S. W., McGlashan, T., Lee, J., Klosterkötter, J., Yung, A. R., & McGuire, P. (2016). The Dark Side of the Moon: Meta-analytical Impact of Recruitment Strategies on Risk Enrichment

in the Clinical High Risk State for Psychosis. *Schizophrenia Bulletin*, *42*(3), 732–743. https://doi.org/10.1093/schbul/sbv162

Haidl, T., Rosen, M., Schultze-Lutter, F., Nieman, D., Eggers, S., Heinimaa, M., Juckel, G., Heinz, A., Morrison, A., Linszen, D., Salokangas, R., Klosterkötter, J., Birchwood, M., Patterson, P., Ruhrmann, S., & European Prediction of Psychosis Study (EPOS) Group. (2018). Expressed emotion as a predictor of the first psychotic episode - Results of the European prediction of psychosis study. *Schizophrenia Research*, *199*, 346–352. https://doi.org/10.1016/j.schres.2018.03.019

Hengartner, M. P., Heekeren, K., Dvorsky, D., Walitza, S., Rössler, W., & Theodoridou, A. (2017). Checking the predictive accuracy of basic symptoms against ultra high-risk criteria and testing of a multivariable prediction model: Evidence from a prospective three-year observational study of persons at clinical high-risk for psychosis. *European Psychiatry: The Journal of the Association of European Psychiatrists*, *45*, 27–35. https://doi.org/10.1016/j.eurpsy.2017.05.026

Hosmer, D. W. (2000). Lemeshow S. Applied logistic regression. *New York*.

Justice, A. C., Covinsky, K. E., & Berlin, J. A. (1999). Assessing the generalizability of prognostic information. *Annals of Internal Medicine*, *130*(6), 515–524. https://doi.org/10.7326/0003-4819-130-6-199903160-00009

Kotlicka-Antczak, M., Karbownik, M. S., Stawiski, K., Pawełczyk, A., Żurner, N., Pawełczyk, T., Strzelecki, D., & Fusar-Poli, P. (2019). Short clinically-based prediction model to forecast transition to psychosis in individuals at clinical high risk state. *European Psychiatry: The Journal of the Association of European Psychiatrists*, *58*, 72–79. https://doi.org/10.1016/j.eurpsy.2019.02.007

Koutsouleris, N., Dwyer, D. B., Degenhardt, F., Maj, C., Urquijo-Castro, M. F., Sanfelici, R., Popovic, D., Oeztuerk, O., Haas, S. S., Weiske, J., Ruef, A., Kambeitz-Ilankovic, L.,

Antonucci, L. A., Neufang, S., Schmidt-Kraepelin, C., Ruhrmann, S., Penzel, N., Kambeitz, J., Haidl, T. K., … Writing Group for the PRONIA Consortium. (2020). Multimodal Machine Learning Workflows for Prediction of Psychosis in Patients With Clinical High-Risk Syndromes and Recent-Onset Depression. *JAMA Psychiatry* . https://doi.org/10.1001/jamapsychiatry.2020.3604

Koutsouleris, N., Kambeitz-Ilankovic, L., Ruhrmann, S., Rosen, M., Ruef, A., Dwyer, D. B., Paolini, M., Chisholm, K., Kambeitz, J., Haidl, T., & Others. (2018). 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*(11), 1156–1172. https://jamanetwork.com/journals/jamapsychiatry/article-abstract/2707956

Lemos-Giráldez, S., Vallina-Fernández, O., Fernández-Iglesias, P., Vallejo-Seco, G., Fonseca-Pedrero, E., Paíno-Piñeiro, M., Sierra-Baigrie, S., García-Pelayo, P., Pedrejón-Molino, C., Alonso-Bada, S., Gutiérrez-Pérez, A., & Ortega-Ferrández, J. A. (2009). Symptomatic and functional outcome in youth at ultra-high risk for psychosis: a longitudinal study. *Schizophrenia Research*, *115*(2-3), 121–129. https://doi.org/10.1016/j.schres.2009.09.011

Lencz, T., Smith, C. W., McLaughlin, D., Auther, A., Nakayama, E., Hovey, L., & Cornblatt, B. A. (2006). Generalized and specific neurocognitive deficits in prodromal schizophrenia. *Biological Psychiatry*, *59*(9), 863–871. https://doi.org/10.1016/j.biopsych.2005.09.005

Lin, A., Wood, S. J., Nelson, B., Brewer, W. J., Spiliotacopoulos, D., Bruxner, A., Broussard, C., Pantelis, C., & Yung, A. R. (2011). Neurocognitive predictors of functional outcome two to 13 years after identification as ultra-high risk for psychosis. *Schizophrenia Research*, *132*(1), 1–7. https://doi.org/10.1016/j.schres.2011.06.014

Malda, A., Boonstra, N., Barf, H., de Jong, S., Aleman, A., Addington, J., Pruessner, M., Nieman, D., de Haan, L., Morrison, A., Riecher-Rössler, A., Studerus, E., Ruhrmann, S., Schultze-Lutter, F., An, S. K., Koike, S., Kasai, K., Nelson, B., McGorry, P., … Pijnenborg, G. H. M. (2019). Individualized Prediction of Transition to Psychosis in 1,676 Individuals at Clinical High Risk: Development and Validation of a Multivariable Prediction Model Based on Individual Patient Data Meta-Analysis. *Frontiers in Psychiatry / Frontiers Research Foundation*, *10*, 345. https://doi.org/10.3389/fpsyt.2019.00345

McGlashan, T., Walsh, B., & Woods, S. (2010). *The Psychosis-Risk Syndrome: Handbook for Diagnosis and Follow-Up* (New edition). OXFORD UNIV PR. https://www.amazon.de/-/en/Thomas-McGlashan/dp/0199733317

McGorry, P. D., Hartmann, J. A., Spooner, R., & Nelson, B. (2018). Beyond the "at risk mental state" concept: transitioning to transdiagnostic psychiatry. *World Psychiatry: Official Journal of the World Psychiatric Association* , *17*(2), 133–142. https://doi.org/10.1002/wps.20514

Metzler, S., Dvorsky, D., Wyss, C., Nordt, C., Walitza, S., Heekeren, K., Rössler, W., & Theodoridou, A. (2016). Neurocognition in help-seeking individuals at risk for psychosis: Prediction of outcome after 24 months. *Psychiatry Research*, *246*, 188–194. https://doi.org/10.1016/j.psychres.2016.08.065

Michel, C., Ruhrmann, S., Schimmelmann, B. G., Klosterkötter, J., & Schultze-Lutter, F. (2014). A stratified model for psychosis prediction in clinical practice. *Schizophrenia Bulletin*, *40*(6), 1533–1542. https://doi.org/10.1093/schbul/sbu025

Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., & PRISMA Group. (2009). Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *PLoS Medicine*, *6*(7), e1000097. https://doi.org/10.1371/journal.pmed.1000097

Moritz, S., Gawęda, Ł., Heinz, A., & Gallinat, J. (2019). Four reasons why early detection centers for psychosis should be renamed and their treatment targets reconsidered: we should not catastrophize a future we can neither reliably predict nor change. *Psychological Medicine*, *49*(13), 2134–2140. https://doi.org/10.1017/S0033291719001740

Niles, H. F., Walsh, B. C., Woods, S. W., & Powers, A. R., 3rd. (2019). Does hallucination perceptual modality impact psychosis risk? *Acta Psychiatrica Scandinavica*, *140*(4), 360–370. https://doi.org/10.1111/acps.13078

Ojala, M., & Garriga, G. C. (2010). Permutation Tests for Studying Classifier Performance. *Journal of Machine Learning Research: JMLR*, *11*(Jun), 1833–1863. http://www.jmlr.org/papers/v11/ojala10a.html

Oliver, D., Spada, G., Colling, C., Broadbent, M., Baldwin, H., Patel, R., Stewart, R., Stahl, D., Dobson, R., McGuire, P., & Fusar-Poli, P. (2020). Real-world implementation of precision psychiatry: Transdiagnostic risk calculator for the automatic detection of individuals at-risk of psychosis. *Schizophrenia Research*. https://doi.org/10.1016/j.schres.2020.05.007

R Core Team. (2020). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Riley, R. D., Ensor, J., Snell, K. I. E., Debray, T. P. A., Altman, D. G., Moons, K. G. M., & Collins, G. S. (2016). External validation of clinical prediction models using big datasets from e-health records or IPD meta-analysis: opportunities and challenges. *BMJ* , *353*, i3140. https://doi.org/10.1136/bmj.i3140

Rosen, M., Haidl, T. K., Ruhrmann, S., Vogeley, K., & Schultze-Lutter, F. (2019). Sex differences in symptomatology of psychosis-risk patients and in prediction of psychosis. *Archives of Women's Mental Health*. https://doi.org/10.1007/s00737-019-01000-3

Royston, P. (2015). Tools for Checking Calibration of a Cox Model in External Validation: Prediction of Population-Averaged Survival Curves Based on Risk Groups. *The Stata Journal*, *15*(1), 275–291. https://doi.org/10.1177/1536867X1501500116

Royston, P., & Altman, D. G. (2013). External validation of a Cox prognostic model: principles and methods. *BMC Medical Research Methodology*, *13*, 33. https://doi.org/10.1186/1471-2288-13-33

Royston, P., Altman, D. G., & Sauerbrei, W. (2006). Dichotomizing continuous predictors in multiple regression: a bad idea. *Statistics in Medicine*, *25*(1), 127–141. https://doi.org/10.1002/sim.2331

Ruhrmann, S., Schultze-Lutter, F., Salokangas, R. K. R., Heinimaa, M., Linszen, D., Dingemans, P., Birchwood, M., Patterson, P., Juckel, G., Heinz, A., Morrison, A., Lewis, S., von Reventlow, H. G., & 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*(3), 241–251. https://doi.org/10.1001/archgenpsychiatry.2009.206

Salazar de Pablo, G., Studerus, E., Vaquerizo-Serrano, J., Irving, J., Catalan, A., Oliver, D., Baldwin, H., Danese, A., Fazel, S., Steyerberg, E. W., Stahl, D., & Fusar-Poli, P. (2020). Implementing Precision Psychiatry: A Systematic Review of Individualized Prediction Models for Clinical Practice. *Schizophrenia Bulletin*. https://doi.org/10.1093/schbul/sbaa120

Salokangas, R. K. R., Ruhrmann, S., von Reventlow, H. G., Heinimaa, M., Svirskis, T., From, T., Luutonen, S., Juckel, G., Linszen, D., Dingemans, P., Birchwood, M., Patterson, P., Schultze-Lutter, F., Klosterkötter, J., & EPOS group. (2012). Axis I diagnoses and transition to psychosis in clinical high-risk patients EPOS project: prospective follow-up of 245 clinical high-risk outpatients in four countries.

Schizophrenia Research, *138*(2-3), 192–197. https://doi.org/10.1016/j.schres.2012.03.008

Sanfelici, R., Dwyer, D. B., Antonucci, L. A., & Koutsouleris, N. (2020). Individualized diagnostic and prognostic models for patients with psychosis risk syndromes: A meta-analytic view on the state-of-the-art. *Biological Psychiatry*. https://doi.org/10.1016/j.biopsych.2020.02.009

Schmidt, A., Cappucciati, M., Radua, J., Rutigliano, G., Rocchetti, M., Dell'Osso, L., Politi, P., Borgwardt, S., Reilly, T., Valmaggia, L., McGuire, P., & Fusar-Poli, P. (2017). Improving Prognostic Accuracy in Subjects at Clinical High Risk for Psychosis: Systematic Review of Predictive Models and Meta-analytical Sequential Testing Simulation. *Schizophrenia Bulletin*, *43*(2), 375–388. https://doi.org/10.1093/schbul/sbw098

Schultze-Lutter, F., Addington, J., Ruhrmann, S., & Klosterkötter, J. (2007). Schizophrenia proneness instrument, adult version (SPI-A). *Rome: Giovanni Fioriti*. https://www.fioritieditore.com/wp-content/uploads/2016/06/summaryOfBasicSymptomCriteria.pdf

Schultze-Lutter, F., Michel, C., Schmidt, S. J., Schimmelmann, B. G., Maric, N. P., Salokangas, R. K. R., 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: The Journal of the Association of European Psychiatrists*, *30*(3), 405–416. https://doi.org/10.1016/j.eurpsy.2015.01.010

Starke, G., De Clercq, E., Borgwardt, S., & Elger, B. S. (2020). Computing schizophrenia: ethical challenges for machine learning in psychiatry. *Psychological Medicine*, 1–7. https://doi.org/10.1017/S0033291720001683

Steyerberg, E. W., Moons, K. G. M., van der Windt, D. A., Hayden, J. A., Perel, P., Schroter, S., Riley, R. D., Hemingway, H., Altman, D. G., & PROGRESS Group. (2013). Prognosis Research Strategy (PROGRESS) 3: prognostic model research. *PLoS Medicine*, *10*(2), e1001381. https://doi.org/10.1371/journal.pmed.1001381

Studerus, E., Ramyead, A., & Riecher-Rössler, A. (2017). Prediction of transition to psychosis in patients with a clinical high risk for psychosis: a systematic review of methodology and reporting. *Psychological Medicine*, *47*(7), 1163–1178. https://doi.org/10.1017/S0033291716003494

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*(1-3), 51–57. https://doi.org/10.1016/j.schres.2010.09.024

Vickers, A. J., & Elkin, E. B. (2006). Decision curve analysis: a novel method for evaluating prediction models. *Medical Decision Making: An International Journal of the Society for Medical Decision Making*, *26*(6), 565–574. https://doi.org/10.1177/0272989X06295361

Vickers, A. J., van Calster, B., & Steyerberg, E. W. (2019). A simple, step-by-step guide to interpreting decision curve analysis. *Diagnostic and Prognostic Research*, *3*, 18. https://doi.org/10.1186/s41512-019-0064-7

Walder, D. J., Holtzman, C. W., Addington, J., Cadenhead, K., Tsuang, M., Cornblatt, B., Cannon, T. D., McGlashan, T. H., Woods, S. W., Perkins, D. O., Seidman, L. J., Heinssen, R., & Walker, E. F. (2013). Sexual dimorphisms and prediction of conversion in the NAPLS psychosis prodrome. *Schizophrenia Research*, *144*(1-3), 43–50. https://doi.org/10.1016/j.schres.2012.11.039

Wang, T., Oliver, D., Msosa, Y., Colling, C., Spada, G., Roguski, Ł., Folarin, A., Stewart, R.,

Roberts, A., Dobson, R. J. B., & Fusar-Poli, P. (2020). Implementation of a Real-Time Psychosis Risk Detection and Alerting System Based on Electronic Health Records using CogStack. *Journal of Visualized Experiments: JoVE*, *159*. https://doi.org/10.3791/60794

Wells, G. (2001). The Newcastle-Ottawa Scale (NOS) for assessing the quality of non randomised studies in meta-analyses. *Http://www*. *Ohri*. *Ca/programs/clinical_epidemiology/oxford*. *Asp*. https://ci.nii.ac.jp/naid/20000796643/

Woods, S. W., Bearden, C. E., Sabb, F. W., Stone, W. S., Torous, J., Cornblatt, B. A., Perkins, D. O., Cadenhead, K. S., Addington, J., Powers, A. R., 3rd, Mathalon, D. H., Calkins, M. E., Wolf, D. H., Corcoran, C. M., Horton, L. E., Mittal, V. A., Schiffman, J., Ellman, L. M., Strauss, G. P., … Anticevic, A. (2020). Counterpoint. Early intervention for psychosis risk syndromes: Minimizing risk and maximizing benefit. *Schizophrenia Research*. https://doi.org/10.1016/j.schres.2020.04.020

Wyatt, J. C., & Altman, D. G. (1995). Commentary: Prognostic models: clinically useful or quickly forgotten? *BMJ* , *311*(7019), 1539–1541. https://doi.org/10.1136/bmj.311.7019.1539

Yung, A. R., Phillips, L. J., Yuen, H. P., & McGorry, P. D. (2004). Risk factors for psychosis in an ultra high-risk group: psychopathology and clinical features. *Schizophrenia Research*, *67*(2-3), 131–142. https://doi.org/10.1016/S0920-9964(03)00192-0

Zhang, T., Xu, L., Chen, Y., Wei, Y., Tang, X., Hu, Y., Li, Z., Gan, R., Wu, G., Cui, H., Tang, Y., Hui, L., Li, C., & Wang, J. (2020). Conversion to psychosis in adolescents and adults: similar proportions, different predictors. *Psychological Medicine*, 1–9. https://doi.org/10.1017/S0033291720000756

Zhang, T., Xu, L., Tang, Y., Li, H., Tang, X., Cui, H., Wei, Y., Wang, Y., Hu, Q., Liu, X., Li, C., Lu, Z., McCarley, R. W., Seidman, L. J., Wang, J., & SHARP (ShangHai At Risk for

Psychosis) Study Group. (2019). Prediction of psychosis in prodrome: development and validation of a simple, personalized risk calculator. *Psychological Medicine*, *49*(12), 1990–1998. https://doi.org/10.1017/S0033291718002738

Zhang, T., Yang, S., Xu, L., Tang, X., Wei, Y., Cui, H., Li, H., Tang, Y., Hui, L., Li, C., Chen, X., & Wang, J. (2019). Poor functional recovery is better predicted than conversion in studies of outcomes of clinical high risk of psychosis: insight from SHARP. *Psychological Medicine*, 1–7. https://doi.org/10.1017/S0033291719002174

Figure captions:

Figure 1. Discrimination illustrated with Receiver Operator Characteristic (ROC) curves. Models were grouped by area under the curve (AUC), from best (a) to worst (d). The ROC curve plots the true positive rate (sensitivity) against the false-positive rate (1-specificity) for different cut-points. Optimal cutpoints of the prognostic index (PI) for each model were derived via optimization of the Youden index.

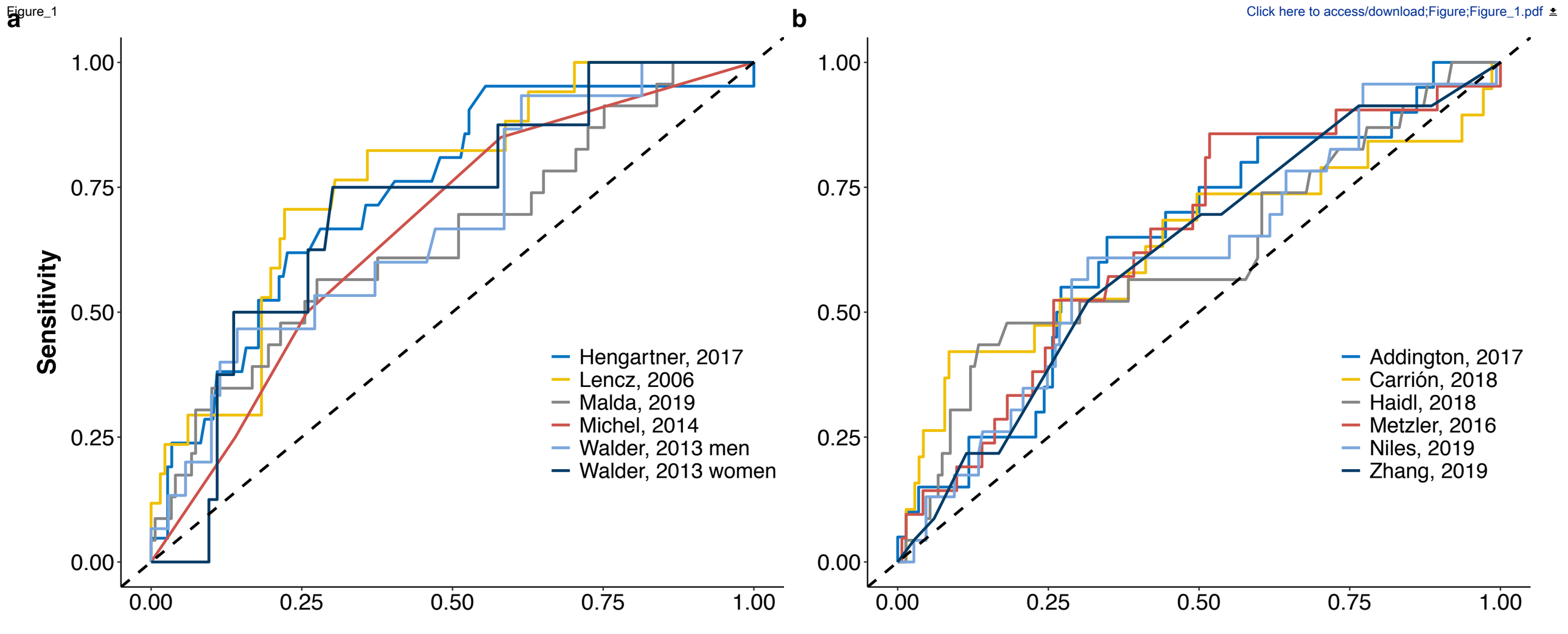
Figure 2. (a) Performance of prediction (in terms of area under the curve, AUC) of transition in the PRONIA sample per tested model. Error bars reflect the bootstrapped 95%-confidence interval. Dashed line reflects chance performance. Note that for studies reporting separate models for distinct subgroups (Rosen et al., 2019; Walder et al., 2013), we depict the weighted aggregate performance here. (b) Discrimination illustrated with Receiver Operator Characteristic (ROC) curves for the individualised prediction of transition by clinical raters, the best model, and the combination thereof. The ROC curve plots the true positive rate (sensitivity) against the false-positive rate (1-specificity) for different cut-points. Optimal

cutpoints of the prognostic index (PI) for each model were derived via optimization of the Youden index. Dashed line reflects chance performance.

Figure 3. Decision curve analysis estimated in the PRONIA validation data set, showing the potential clinical usefulness of the best-performing model by Lencz et al. (2006) at different threshold probabilities, compared with treating all patients ("treat all") or to treating no patients at all ("treat none").

Figure_1

Click here to access/download;Figure;Figure_1.pdf



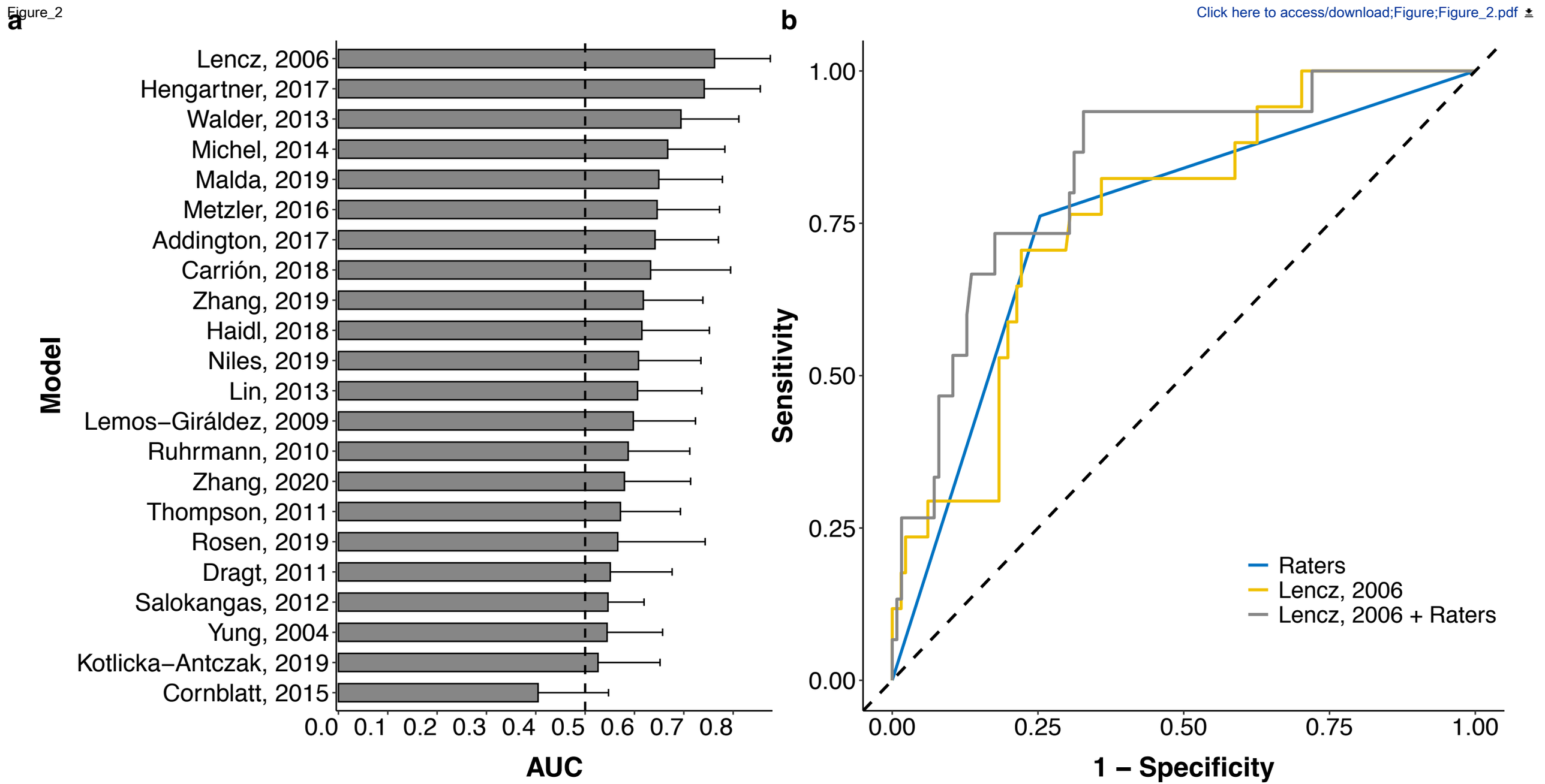


Figure_3

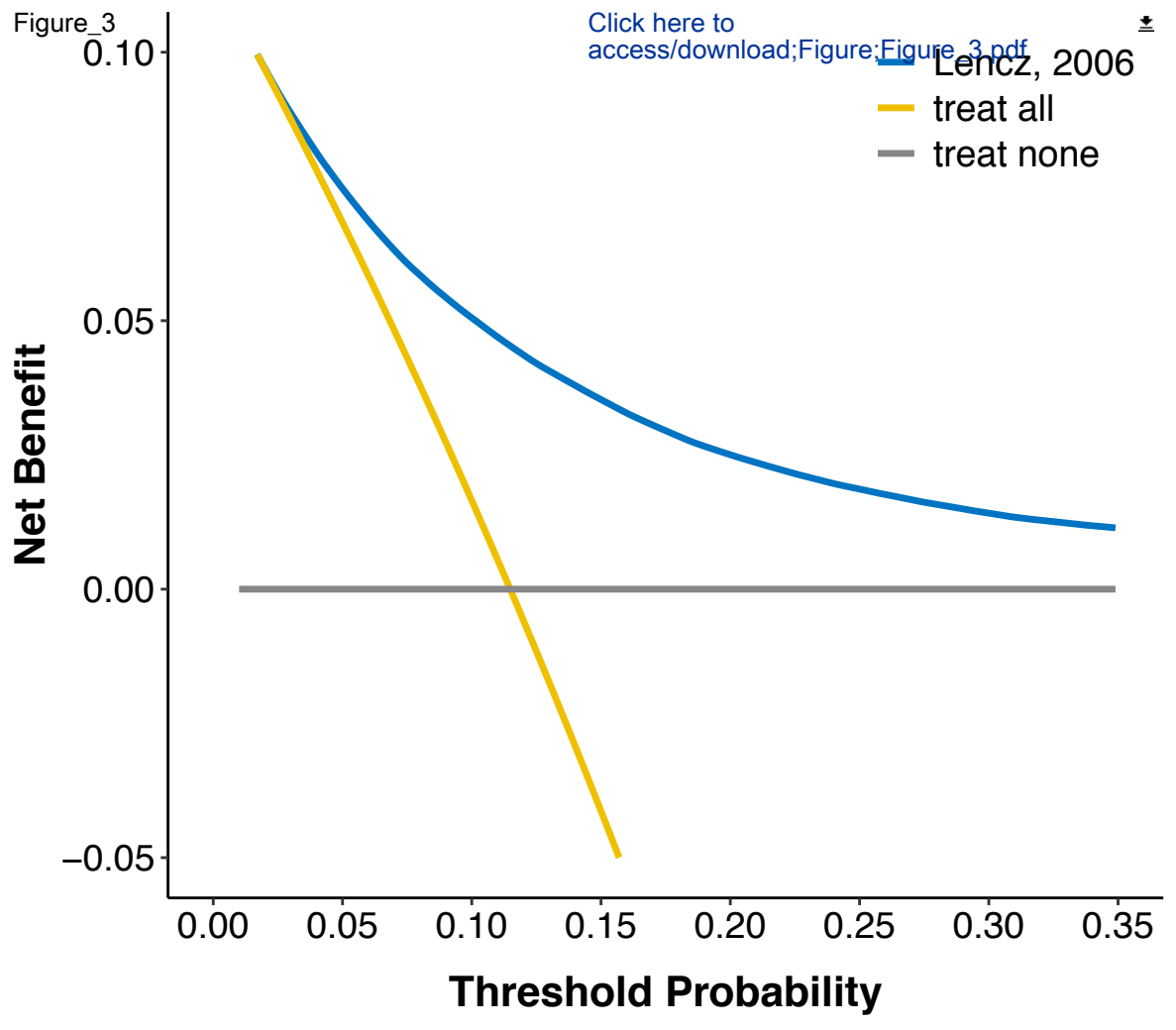


Table 1. Inclusion criteria, transition rate and description of predictor variables of models derived from the systematic literature search used in external validation with the PRONIA study.

First Author, Year	Countries	Sample	Inclusion Criteria (according to)	Sample (N)	Tran-sition n (%)	Age M (SD)	Female n (%)	Edu-cation	Follow-up (months)[a]	Recruiting years	Antipsychotic treatment at baseline	Predictors
Zhang, 2020 ("adult" model)	China	SHARP	UHR (SIPS)	219	55 (25.1)	24.3 (5.4)	106 (48.4)	13.1 years	36	n.a.	none	Drop in functioning (GAF), Positive symptoms (SIPS), General symptoms (SIPS)
Kotlicka-Antczak, 2019	Poland	PORT	UHR (CAARMS)	105	24 (22.9)	18.8 (3.5)	56 (53.3)	11 years	35.4	2010-2015	22.9%	Speech disorganization (CAARMS), Unusual thought content (CAARMS)
Malda, 2019	Canada, Netherlands, UK, Switzerland, Germany, Korea, Japan, Australia, Poland, Italy	ADAPT, CAYR, DUPS-A, EDIE-NL, EDIE-UK, FePsy, FETZ, GRAPE, IN-STEP, OASIS, PACE, PORT, ROME, SAFE, DUPS-U	UHR (CAARMS) and BS	1676	386 (23.0)	21.1 (4.4)	786 (46.9)	n.a.	32	1998-2016	exclusion criterion for n= 902 (53.8%)	Sex, Age, GRD, APS, BLIPS, GAF, SIPS positive subscale, SIPS negative subscale
Niles,	USA	PRIME	UHR (SIPS)	164	45	16.7	66	n.a.	6	2001-2016	n.a.	Unusual thought content

First Author, Year	Countries	Sample	Inclusion Criteria (according to)	Sample (N)	Tran-sition n (%)	Age M (SD)	Female n (%)	Edu-cation	Follow-up (months)[a]	Recruiting years	Antipsychotic treatment at baseline	Predictors
2019					(27.4)	(4.1)	(40.2)					(SIPS), Suspiciousness (SIPS), Grandiose ideas (SIPS), Auditory distortions (SIPS), Visual distortions (SIPS), Somatic distortions (SIPS), Disorganized communication (SIPS)
Rosen, 2019	Germany	FETZ	UHR (SIPS), BS, clinical impression	242	84 (34.7)	24.9 (6.0)	90 (37.2)	14.4 yea rs	40.8	1998-2003	n.a.	Suspiciousness (SIPS), Disorganized communication (SIPS), Avolition (SIPS), Ideational richness (SIPS), Impairment in personal hygiene (SIPS), Perceptual abnormalities (SIPS), Expression of emotions (SIPS), Trouble with focus and attention (SIPS), Occupational functioning (SIPS), Odd behavior or appearance (SIPS), Bizarre thinking (SIPS)
Zhang, 2019a	China	SHARP	UHR (SIPS)	417	83 (19.9)	20.9 (6.4)	217 (52.0)	11.4 years	42.4	n.a.	none	Functional decline (GAF), Positive symptoms (SIPS), Negative symptoms (SIPS), Disorganized symptoms (SIPS), General symptoms (SIPS)
Carrión, 2018	USA	EDIPPP	UHR (SIPS)	205	12 (5.9)	16.5 (3.3)	85 (41.5)	9.75 years	24.8	2007-2011	26.3%	Positive symptoms (SIPS), Age, Gender, Educational level, Anti-psychotics, Processing Speed (MATRICS),

First Author, Year	Countries	Sample	Inclusion Criteria (according to)	Sample (N)	Tran-sition n (%)	Age M (SD)	Female n (%)	Edu-cation	Follow-up (months)[a]	Recruiting years	Antipsychotic treatment at baseline	Predictors
												Working Memory (MATRICS), Attention/Vigilance (MATRICS), Verbal Learning (MATRICS), Visual Learning (MATRICS), Reasoning and Problem Solving (MATRICS)
Haidl & Rosen, 2018	Germany, Finland, Netherlands, UK	EPOS	UHR (SIPS) and BS	235	36 (15.3)	23.0 (5.3)	105 (44.7)	13.5 years	14.4	2002-2006	22.1%; no valid information: 12.8%	Positive symptoms (SIPS), Bizarre thinking (SIPS), Sleep disturbances (SIPS), Years of education, Irritability score (LEE)
Addington, 2017	USA, Canada	PREDICT	UHR (SIPS)	145	29 (20.0)	19.8 (4.7)	63 (43.4)	68.6% high school, 32.1% with some form of degree or professional training	24	n.a.	none	Functioning (SOFAS), Verbal fluency, Unusual thought content (SIPS), Verbal memory, Disorganized communication (SIPS), Processing speed, Baseline age
Hengartner, 2017	Switzerland	ZInEP	UHR (SIPS) and BS	188	24 (12.7)	20.5 (5.8)	88 (39.8)	n.a.	36	2010-2012	20.7%	Positive symptoms (SIPS), Verbal IQ

First Author, Year	Countries	Sample	Inclusion Criteria (according to)	Sample (N)	Tran-sition n (%)	Age M (SD)	Female n (%)	Edu-cation	Follow-up (months)[a]	Recruiting years	Antipsychotic treatment at baseline	Predictors
Metzler, 2016	Switzerland	ZInEP	UHR (SIPS) and BS	118	25 (21.1)	20.5 (5.9)	46 (39.0)	n.a.	24	2010-2012	30%	Positive symptoms (PANSS), Negative symptoms (PANSS), Verbal IQ
Cornblatt, 2015	USA	RAP	APS (SIPS)	92	15 (16.3)	16.0 (2.2)	27 (29.3)	9.5 years	36	2000-2006	20%	Disorganized communication (SIPS), Suspiciousness (SIPS), Verbal Memory, Decline in social functioning (baseline to last follow-up; GF:S), Baseline age
Michel, 2014	Germany	FETZ	UHR (SIPS) and BS	97	44 (45.4)	24.8 (5.5)	34 (35.1)	23%: 10 years, 10.3% 12 years, 48.8% 13 years, 14.6% still in school, 3.2% none	24	1998-2003	none	APS criteria, COGDIS criteria, Processing speed
Lin, 2013	Australia	PACE	UHR (CAARMS)	325	81 (24.9)	19.1 (3.3)	172 (52.9)	52.9% complet	86.2	1993-2006	n.a.	Unusual thought content (CAARMS), Matrix reasoning

First Author, Year	Countries	Sample	Inclusion Criteria (according to)	Sample (N)	Tran-sition n (%)	Age M (SD)	Female n (%)	Edu-cation	Follow-up (months)[a]	Recruiting years	Antipsychotic treatment at baseline	Predictors
								ed high school				
Walder, 2013	USA, Canada	NAPLS	UHR (SIPS)	276	70 (25.4)	18.3 (4.6)	113 (40.9)	n.a.	30	n.a.	41.7% psychotropic medications; no valid information for: 4.7%	Social adjustment in childhood (PAS), Academic adjustment in childhood (PAS), Functioning (GF), Positive symptoms (SIPS), Negative symptoms (SIPS), Disorganized symptoms (SIPS)
Salokangas, 2012	Germany, Finland, Netherlands, UK	EPOS	UHR (SIPS) and BS	245	37 (15.1)	22.4 (5.2)	108 (44.1)	13.5 years	14.2	2002-2006	n.a.	BLIPS criteria (SIPS), Bipolar disorder diagnosis (SCID), Somatoform disorder diagnosis (SCID)
Dragt, 2011 (41)	Netherlands	DUP (Amsterdam site)	UHR (SIPS) and BS	72	19 (26.4)	19.3 (4.0)	25 (34.7)	n.a.	36	2001-2009	23.6%	Social-sexual aspects (PAS), Social-personal adjustment (PAS), Urbanicity
Thompson, 2011	Australia	PACE	UHR (CAARMS)	104	41 (39.4)	19.4 (3.5)	53 (51.0)	41.6% secondary education qualification or above; 58.4% no seconda	28	n.a.	none	GRFD criteria (CAARMS), Unusual thought content (CAARMS), Suspicion/paranoia (BPRS), Functioning (GAF)

First Author, Year	Countries	Sample	Inclusion Criteria (according to)	Sample (N)	Tran-sition n (%)	Age M (SD)	Female n (%)	Edu-cation	Follow-up (months)[a]	Recruiting years	Antipsychotic treatment at baseline	Predictors
								ry education qualification				
Ruhrmann, 2010	Germany, Finland, Netherlands, England	EPOS	UHR (SIPS) and BS	245	37 (15.1)	23.0 (5.2)	108 (44.1)	13.5 years	14.2	2002-2006	22.1%; no valid information: 12.7%	Positive symptoms (SIPS), Bizarre thinking (SIPS), Sleep disturbances (SIPS), Schizotypal personality disorder (SIPS), Functioning (GAF), Years of education
Lemos-Giráldez, 2009	Spain	P3	UHR (SIPS)	61	14 (23.0)	21.7 (3.9)	21 (34.4)	10.8 years	36	n.a.	n.a.	Positive symptoms (SIPS), Negative symptoms (SIPS), Disorganized symptoms (SIPS), General symptoms (SIPS), Functioning (GAF), History of illegal drug use, Years of education, Gender, Family history of psychosis, Duration of untreated illness
Lencz, 2006	USA	RAP	APS (SIPS)	33	12 (36.4)	16.5 (2.2)	14 (42.4)	10.2 years	32	1998-2001	39.5%	Verbal Memory, Positive symptoms (SIPS)
Yung, 2004	Australia	PACE	UHR (CAARMS)	104	41 (39.4)	19.4 (3.5)	53 (51.0)	n.a.	12	1995-1999	none	GRFD and APS criteria, Functioning (GAF), Duration of attenuated symptoms, Attention (SANS)

Abbreviations: APS = Attenuated Positive Symptoms, BLIPS = Brief Limited Intermittent Symptoms, CAARMS = Comprehensive Assessment of At-Risk Mental States, BS = Basic Symptoms, GAF = Global Assessment of Functioning, GF:S = Global Functioning: Social Scale, GRD = Genetic Risk and Deterioration, LEE = Level of Expressed Emotion Scales, n.a. = not available, NOS = Newcastle-Ottawa Scale, PANSS = Positive and Negative Symptoms Scale, PAS = Premorbid Adjustment Scale, SIPS = Structured Interview of Prodromal Symptoms.

[a] mean or planned if mean is not stated

Table 2. Demographic and clinical characteristics of the PRONIA CHR sample used for external validation. Mean (SD) unless stated otherwise.

Variable	Transition	No Transition	Transition vs. No Transition	Known Outcome	Lost to Follow-Up	Known Outcome vs. Lost to Follow-Up
N	23	150	-	173	104	-
Age	23.7 (5.6)	23.6 (5.1)	t(27.9) = -0.07, p = .945	23.6 (5.1)	23.6 (5.3)	t(210.9) = 0.11, p = .916
Male, n (%)	15 (65)	70 (47)	$\chi^2(1) = 2.05$, p = .152	85 (49)	44 (42)	$\chi^2(1) = 0.96$, p = .328
Education (years)	13.1 (2.6)	13.8 (2.8)	t(30.5) = 1.16, p = .254	13.7 (2.7)	13.2 (2.5)	t(220.1) = -1.28, p = .203
SIPS score						
Positive	10.3 (4.3)	7.1 (4.2)	W = 1048, p = .003	7.6 (4.3)	7.7 (4.7)	W = 8844, p = .910
Negative	11.6 (8.6)	10.5 (6.6)	W = 1650, p = .814	10.6 (6.9)	9.9 (6.2)	W = 7929, p = .464
Disorganization	4.2 (4.0)	3.4 (2.8)	W = 1490, p = .356	3.5 (3.0)	2.9 (2.5)	W = 7288, p = .111
General	9.4 (4.4)	7.8 (3.7)	W = 1320, p = .083	8.0 (3.9)	7.6 (3.7)	W = 8002, p = .539
Total	35.5 (16.3)	28.8 (12.2)	W = 1291, p = .069	29.7 (13.0)	27.7 (12.9)	W = 7544, p = .307
GAF						
At baseline	47.8 (12.6)	50.0 (11.5)	W = 2036, p = .147	49.7 (11.6)	48.1 (10.9)	W = 8109, p = .240
Highest past year	61.9 (16.9)	61.3 (12.4)	W = 1780, p = .768	61.4 (13.0)	62.5 (14.2)	W = 9354, p = .437
CHR criteria, n (%)			$\chi^2(2) = 5.61$, p = .061			$\chi^2(2) = 2.04$, p = .365
UHR only	10 (43)	66 (44)	-	76 (44)	39 (38)	-
COGDIS only	1 (4)	35 (23)	-	36 (21)	29 (28)	-

Variable	Transition	No Transition	Transition vs. No Transition	Known Outcome	Lost to Follow-Up	Known Outcome vs. Lost to Follow-Up
UHR *and* COGDIS	12 (52)	49 (33)	-	61 (35)	36 (35)	-
Observation time (months)	20.7 (9.9)	26.0 (9.7)	t(24.0) = 2.25, p = .034	25.4 (9.8)	5.2 (5.5)	t(270.3) = -21.8, p < .001
Time to transition (months)	7.9 (6.4)	-	-	7.9 (6.4)	-	-

Abbreviations: UHR = Ultra High Risk criteria, COGDIS = Cognitive Disturbances, GAF = Global Assessment of Functioning, SIPS = Structured Interview for Prodromal Psychosis

Table 3. Results of external validation of prediction models for transition to psychosis in the PRONIA study.

	Development			Validation (PRONIA)										Cali-bration
Model	AUC	Sens	Spec	N (transition)	AUC (95%-CI)	p (vs. chance)	p (vs. rater)	BAC	Sens	Spec	PPV	NPV	LR+	LR-	m
Zhang et al., 2020 (adults)	na	na	na	161 (22)	0.58 (0.45, 0.71)	.068	.985	0.60	0.59	0.60	0.19	0.90	1.48	0.68	0.35
Kotlicka-Antczak et al., 2019	0.78	0.65	0.79	173 (23)	0.53 (0.40, 0.65)	.255	.999	0.55	0.87	0.23	0.15	0.92	1.13	0.56	0.13
Malda et al., 2019	0.63	na	na	172 (23)	0.65 (0.52, 0.78)	.002	.907	0.65	0.57	0.72	0.24	0.92	2.05	0.60	0.82
Niles et al., 2019	na	na	na	172 (23)	0.61 (0.48, 0.73)	.027	.955	0.65	0.61	0.68	0.23	0.92	1.93	0.57	0.32
Rosen et al., 2019 (women)	0.75	na	na	86 (8)	0.57 (0.38, 0.76)	.325	.943	0.64	1	0.28	0.12	1	1.39	0	0.31

Rosen et al., 2019 (men)	0.61	na	na	85 (15)	0.56 (0.40, 0.73)	.133	.969	0.63	0.73	0.53	0.25	0.9	1.55	0.51	0.53
Zhang wt al., 2019a	0.74	na	na	172 (23)	0.62 (0.50, 0.74)	.007	.966	0.60	0.52	0.68	0.20	0.90	1.65	0.70	0.35
Carrión et al., 2018	na	na	na	160 (19)	0.63 (0.47, 0.79)	.004	.953	0.67	0.42	0.91	0.40	0.92	4.95	0.63	0.12
Haidl et al., 2018	0.77	na	na	172 (23)	0.61 (0.48, 0.75)	.004	.960	0.65	0.43	0.87	0.33	0.91	3.24	0.65	0.32
Hengartner et al., 2017	0.85	0.86	0.85	167 (21)	0.74 (0.63, 0.85)	< .001	.564	0.70	0.95	0.45	0.20	0.98	1.72	0.11	0.12
Addington et al., 2017	0.73	na	na	164 (20)	0.64 (0.51, 0.77)	.012	.919	0.65	0.65	0.65	0.21	0.93	1.87	0.54	0.71
Metzler et al., 2016	na	na	na	164 (21)	0.65 (0.52, 0.77)	.013	.913	0.67	0.86	0.48	0.20	0.96	1.66	0.30	0.06

Cornblatt et al., 2015	0.92	0.60	0.97	170 (21)	0.40 (0.26, 0.55)	1	1	0.55	0.19	0.91	0.22	0.88	2.03	0.89	-0.07
Michel et al., 2014	na	na	na	163 (20)	0.67 (0.55, 0.78)	.003	.896	0.63	0.85	0.42	0.17	0.95	1.46	0.36	0.67
Lin et al., 2013	na	na	na	159 (19)	0.61 (0.48, 0.73)	.035	.965	0.60	0.53	0.68	0.18	0.91	1.64	0.70	0.25
Walder et al., 2013 (women)	na	na	na	81 (8)	0.71 (0.53, 0.89)	.010	.607	0.72	0.75	0.70	0.21	0.96	2.50	0.36	0.28
Walder et al., 2013 (men)	na	na	na	85 (15)	0.68 (0.53, 0.83)	< .001	.805	0.66	0.47	0.86	0.41	0.88	1.25	0.62	0.16
Salokan-gas et al., 2012	na	na	na	172 (23)	0.55 (0.47, 0.62)	.007	1	0.55	0.13	0.96	0.33	0.88	3.24	0.91	1.03
Dragt et al., 2011	na	1.0	0.69	168 (23)	0.55 (0.43, 0.68)	.154	.996	0.55	0.65	0.44	0.16	0.89	1.18	0.78	0.08

Thompson et al., 2011	na	na	na	172 (23)	0.57 (0.45, 0.69)	.128	.986	0.63	0.83	0.43	0.18	0.94	1.45	0.40	0.26
Ruhrmann et al., 2010	0.81	0.42	0.98	172 (23)	0.59 (0.46, 0.71)	.075	.976	0.61	0.52	0.70	0.21	0.91	1.77	0.68	0.27
Lemos-Giráldez et al., 2009	na	na	na	172 (23)	0.60 (0.47, 0.72)	.053	.980	0.60	0.48	0.72	0.21	0.90	1.70	0.73	0.07
Lencz et al., 2006	0.83*	0.82	0.79	148 (17)	0.76 (0.65, 0.88)	< .001	.453	0.74	0.71	0.78	0.29	0.95	3.19	0.38	1.07
Yung et al., 2004	na	na	na	170 (23)	0.54 (0.43, 0.66)	.134	.996	0.55	0.39	0.71	0.18	0.88	1.37	0.85	0.15
Raters (Over-all)	-	-	-	167 (21)	0.75 (0.65, 0.85)	< .001	-	0.75	0.76	0.75	0.30	0.96	3.01	0.32	-
Raters (Experien-ced)	-	-	-	73 (13)	0.81 (0.69, 0.92)	< .001	-	0.81	0.85	0.77	0.44	0.96	3.70	0.19	-

Raters (Less Experien-ced)	-	-	-	94 (8)	0.68 (0.49, 0.86)	< .001	-	0.68	0.63	0.73	0.18	0.95	2.33	0.51	-
Lencz et al., 2006 & Raters (Over-all)	-	-	-	143 (16)	0.84 (0.74, 0.94)	< .001	.163	0.80	0.94	0.67	0.26	0.99	2.83	0.09	-
Lencz et al., 2006 & Raters (Experien-ced)	-	-	-	60 (10)	0.83 (0.68, 0.95)	< .001	-	0.85	0.90	0.80	0.47	0.98	4.50	0.13	-
Lencz et al., 2006 & Raters (Less Experien-ced)	-	-	-	85 (7)	0.81 (0.69, 0.92)	< .001	-	0.84	1	0.68	0.22	1	3.13	0	-

Note: The reported p-values indicate whether the model performs better than chance or better than clinical raters (based on a permutation test with 1000 permutations (23)).

*AUC was calculated manually from the ROC curve provided in the publication.

Abbreviations: AUC = Area under the Curve; BAC = Balanced Accuracy; LR+ = Positive Likelihood Ratio; LR- = Negative Likelihood Ratio; m = calibration slope; NPV = Negative Predictive Value; PI = Prognostic Index; PPV = Positive Predictive Value; Sens = Sensitivity (True Positive Rate); Spec = Specificity (True Negative Rate).