Towards generalizable and transdiagnostic tools for psychosis prediction.

An independent validation and improvement of the NAPLS-2 risk calculator in the multi-site PRONIA cohort.

Nikolaos Koutsouleris, MD[1,2,3*,CA]; Michelle Worthington, MA[4,*]; Dominic B. Dwyer, PhD[1*]; Lana Kambeitz-Ilankovic, PhD[1,5]; Rachele Sanfelici, MSc[1]; Paolo Fusar-Poli, MD[2,6]; Marlene Rosen, PhD[5]; Stephan Ruhrmann, MD[5]; Alan Anticevic, PhD[4]; Jean Addington, PhD[7]; Diana O. Perkins, PhD, MPH[8]; Carrie E. Bearden, PhD[9]; Barbara A. Cornblatt, PhD, MBA[10]; Kristin S. Cadenhead, MD[11]; Daniel H. Mathalon, MD, PhD[12]; Thomas McGlashan, MD[13]; Larry Seidman, PhD[14]; Ming Tsuang, MD[11]; Elaine F. Walker, PhD[15]; Scott W. Woods, MD, PhD[13]; Peter Falkai, MD[1]; Rebekka Lencer, MD[16,17]; Alessandro Bertolino, MD[18], PhD; Joseph Kambeitz, MD[5]; Frauke Schultze-Lutter, PhD[19]; Eva Meisenzahl, MD[19]; Raimo K. R. Salokangas, MD, PhD[20]; Jarmo Hietala, MD, PhD[20]; Paolo Brambilla, MD, PhD[21,22]; Rachel Upthegrove, PhD[23,24]; Stefan Borgwardt, MD[17,25]; Stephen Wood, PhD[26,27]; Raquel E. Gur, MD, PhD[28]; Philip McGuire, MD, PhD[2]; Tyrone D. Cannon, PhD[4,13]

1 Department of Psychiatry and Psychotherapy, Ludwig-Maximilian-University, Munich, Germany
2 Institute of Psychiatry, Psychology and Neurosciences, King's College London, United Kingdom
3 Max-Planck Institute of Psychiatry, Munich, Germany
4 Department of Psychology, Yale University, USA
5 Department of Psychiatry and Psychotherapy, Faculty of Medicine and University Hospital, University of Cologne, Cologne, Germany
6 Department of Brain and Behavioral Sciences, University of Pavia, Italy
7 Hotchkiss Brain Institute, Department of Psychiatry, University of Calgary, Calgary, Alberta, Canada
8 Department of Psychiatry, University of North Carolina, USA
9 Departments of Psychiatry and Biobehavioral Sciences and Psychology, Semel Institute for Neuroscience and Human Behavior, UCLA, California, USA
10 The Zucker Hillside Hospital, Northwell Health, New York, USA
11 University of California, San Diego, California, USA
12 Department of Psychiatry, UCSF, and SFVA Medical Center, San Francisco, CA
13 Department of Psychiatry, Yale University, New Haven, CT
14 Department of Psychiatry, Harvard Medical School at Beth Israel Deaconess Medical Center, Boston, MA
15 Department of Psychology and Psychiatry, Emory University, Atlanta, GA
16 Department of Psychiatry and Psychotherapy, University of Münster, Germany
17 Department of Psychiatry and Psychotherapy, University of Lübeck, Germany
18 Department of Basic Medical Science, Neuroscience and Sense Organs, University of Bari Aldo Moro, Bari, Italy
19 Department of Psychiatry and Psychotherapy, Medical Faculty, Heinrich-Heine University, Düsseldorf, Germany
20 Department of Psychiatry, University of Turku, Finland
21 Department of Neurosciences and Mental Health, Fondazione IRCCS Ca' Granda Ospedale Maggiore Policlinico, Milano, Italy
22 Department of Pathophysiology and Transplantation, University of Milan, Milan, Italy
23 Institute of Mental Health, University of Birmingham, United Kingdom
24 School of Psychology, University of Birmingham, United Kingdom
25 Department of Psychiatry (Psychiatric University Hospital, UPK), University of Basel, Switzerland
26 Centre for Youth Mental Health, University of Melbourne, Australia
27 Orygen, the National Centre of Excellence for Youth Mental Health, Melbourne, Australia
28 Department of Psychiatry, Perelman School of Medicine, University of Pennsylvania, United States

* authors contributed equally

[CA] Corresponding author:
Nikolaos Koutsouleris
Professor for Neurodiagnostic Applications in Psychiatry
Department of Psychiatry and Psychotherapy, Ludwig-Maximilian-University Munich, Nussbaumstr. 7, D-80336 Munich, Germany; Tel.: 0049 (0) 89 4400 55885, Fax: 0049 (0) 89 4400 55776
Emails to: nikolaos.koutsouleris@med.uni-muenchen.de

Keywords:

Clinical high-risk states

First-episode depression

Psychosis prediction

Risk calculators

Reciprocal external validation

Machine learning

Word Count Manuscript

Abstract:	250
Main manuscript:	4000 (excluding abstract)

Tables: 3

Figures: 3

Abstract

Background: Transition to psychosis is among the most adverse outcomes of the clinical high-risk (CHR) syndromes encompassing ultra-high-risk (UHR) and basic symptoms states. Clinical risk calculators may facilitate an early and individualized interception of psychosis, but their real-world implementation requires thorough validation across diverse risk populations, including young patients with depressive syndromes.

Methods: We validated the previously described NAPLS-2 calculator in 334 patients (26 with psychosis transition) with CHR or recent-onset depression (ROD) drawn from the multisite European PRONIA study. Patients were categorized into three risk enrichment levels, ranging from UHR, over CHR, to a broad risk population comprising CHR or ROD patients (CHR|ROD). We assessed how risk enrichment and different predictive algorithms influenced prognostic performance using reciprocal external validation.

Results: After calibration, the NAPLS-2 model predicted psychosis with a balanced accuracy [BAC(sensitivity,specificity)] of 68%(73%,63%) in the PRONIA-UHR, 67%(74%,60%) in CHR, and 70%(73%,66%) in CHR|ROD patients. Multiple model derivation in PRONIA-CHR|ROD and validation in NAPLS-2-UHR patients confirmed that broader risk definitions produced more accurate risk calculators [CHR|ROD-based vs. UHR-based performance: 67%(68%,66%) vs. 58%(61%,56%)]. Support-vector machines (SVM) were superior in CHR|ROD (BAC=71%), while ridge logistic regression and SVM performed similarly in CHR (BAC=67%) and UHR cohorts (BAC=65%). Attenuated psychotic symptoms predicted psychosis across risk levels, while younger age and reduced processing speed became increasingly relevant for broader risk cohorts.

Conclusions: Clinical-neurocognitive machine-learning models operating in young patients with affective and CHR syndromes facilitate a more precise and generalizable prediction of psychosis. Future studies should investigate their therapeutic utility in large-scale clinical trials.

INTRODUCTION

Over the last 30 years, diverse research criteria emerged around the globe that defined the clinical high-risk (CHR) states for psychosis based on the ultra-high risk (UHR) or basic symptoms concepts (1–5). The purpose of these criteria has been to detect young persons with an increased risk for psychotic disorders, to study potential disease-modifying treatments in these persons and, ultimately to implement this application of preventive psychiatry in clinical care (6). Previous research showed that CHR ascertainment identifies vulnerable help-seeking populations with a substantially increased incidence for psychosis (7,8). Yet, their observed three-year transition risk steadily dropped from 36% to currently 22% as more sites adopted early recognition activities following the UHR concept (9). Hence, due to its modest prognostic value and dependency on sufficiently risk-enriched populations (10), the clinical utility and scalability of this CHR paradigm have been questioned (11).

Therefore, previous studies have proposed to augment the actual two-tier risk enrichment process—patient referral followed by CHR assessment—using algorithms that accurately measure psychosis risk in the individual patient (7,12–19). These studies demonstrated that individualized risk quantification could be achieved using Cox regression or machine learning models trained to estimate patients' likelihood of illness transition based on sociodemographic, clinical, neurocognitive, neurophysiological, neuroimaging, metabolic or genetic information. If these stratification models could operate across different risk-enriched cohorts and healthcare environments (20–25) a more personalized clinical management of vulnerable persons could be implemented: practitioners could flexibly tailor specific disease-interceptive strategies and combine them with treatments that target the array of psychiatric comorbidities and functional impairments present in these patients (26).

However, this vision is challenged by the unknown generalizability of most risk calculators which have not been vetted using cross-validation, let alone external validation (27,28). Exceptions are the clinical-neurocognitive models developed for UHR patients by the SHARP (18) and NAPLS-2 studies (16). The NAPLS-2 risk calculator predicted illness transition with sensitivity=66% and specificity=72% at a 20% two-year estimated risk cut-off. It has been validated in single-site UHR cohorts from the US (29,30) (sensitivity=58-71%, specificity=73-77%) and Shanghai (sensitivity=71.7%, specificity=45.8%) (31). Due to these varying sensitivity-specificity tradeoffs, the model's generalizability and calibration should be further tested across healthcare systems and at the international scale. Furthermore, attenuated or brief limited intermittent psychotic symptoms—key predictors in the NAPLS-2 model—may not mark the only pathway to psychosis (7,17,32–35). Thus, more diverse risk cohorts, including young persons with basic symptoms or affective disorders, as recently proposed by transdiagnostic studies of psychosis risk, are needed to test model generalizability (17,34,36,37). Finally, for clinical implementation, the NAPLS-2 Cox regression model should be compared with newer machine learning algorithms to identify the optimal predictive strategy for designated risk criteria.

By probing and further developing the NAPLS-2 risk calculator, we examined these intertwined challenges of individualized psychosis prediction, i.e., measuring model generalizability conditional to sociodemographic and clinical differences between model derivation and application cohorts, diverse levels of risk for subsequent psychosis development, and different algorithmic strategies for risk calculator development. To this end, we used the multisite European PRONIA study (www.pronia.eu) which recruited a diverse risk population, encompassing young people with CHR states according to UHR or basic symptoms criteria, or recent-onset depression (ROD) without CHR criteria (38). Originally, ROD patients were

enrolled as a clinical comparison group in PRONIA because of the high prevalence of depression in psychosis risk syndromes (39–41) and previous findings of shared neurobiology between depression and psychosis (42,43). Then, recently, we observed that 20% of the ROD patients developed psychosis-related outcomes during follow-up, either de-novo CHR states, or full psychosis (37). Hence, by including these patients into the current analysis, we aimed to evaluate how the three distinct levels of risk captured by the PRONIA design (**Table 1**)—ranging from UHR patients to more diverse risk samples encompassing patients with UHR, basic symptoms or ROD criteria—moderated algorithms' performance. After testing the NAPLS-2 model across these three risk levels, we employed reciprocal external validation and leave-site-out cross-validation to benchmark Cox regression, machine learning and combined algorithms (**Supplementary Table 4**) to enhance prognostic accuracy and model generalizability at each risk enrichment level.

METHODS AND MATERIALS

Participants

Nine-hundred-thirty participants (110 [11.8%] with psychosis transition) were drawn from the NAPLS-2 (44) and PRONIA databases (38). NAPLS-2 is an 8-site observational study examining the predictors and mechanisms related to psychosis transition in clinically defined at-risk populations. NAPLS-2 participants met CHR criteria based on UHR syndromes as determined by the Structured Interview for Psychosis-risk Syndromes (SIPS; **Table 1**). PRONIA (www.pronia.eu) is an observational study across 7 sites in 5 European countries aiming to develop personalized prognostic tools for affective and non-affective psychoses (37,38). PRONIA participants experienced (a) CHR syndromes based on UHR and/or cognitive basic symptoms criteria (COGDIS) (2,3), or (b) recent-onset depression (ROD). ROD patients met

criteria for an initial major depressive episode within 3 months of intake as determined by the Structured Clinical Interview for DSM-IV-TR (SCID) (45).

HARMONY validation framework

A framework for external validation between NAPLS-2 and PRONIA was established by the Harmonization of At Risk Multisite Observational Networks for Youth (HARMONY) collaboration, which also includes the PSYSCAN consortium (http://psyscan.eu/) and the Philadelphia Neurodevelopmental Cohort (PNC). This framework facilitates the development and validation of prognostic/predictive models across independent datasets at the international scale. All analyses were performed using the Virtual Pooling and Analysis of Research Data (ViPAR) portal (46). This web-based platform utilizes a centralized cloud server to retrieve anonymized data securely and temporarily from remote servers. Once analyses are complete, results can be accessed by the user and the data are removed from server's random-access memory. The use of ViPAR was approved by the ethics committees of the 15 study sites across NAPLS-2 and PRONIA.

External validation of the NAPLS-2 risk calculator

We followed the external validation guidelines by Royston and Altman (47) and first assessed the baseline group-level differences in sociodemographic, clinical, and functional variables between psychosis transition and non-transition patients in NAPLS-2 and PRONIA (**Table 2** and **Supplementary Table 2**). Then, we evaluated the effect of consortium-level differences in transition criteria on diagnostic outcomes: In PRONIA, a transition event was determined by ≥1 of the 5 SIPS positive symptom items reaching psychotic intensity daily for ≥7 days. SCID-based diagnoses were assessed at the follow-up visit after transition and corroborated in the ensuing visits. NAPLS-2 used the standard SIPS transition criteria (48), with diagnoses being

evaluated at the time of transition. We assessed the impact of these differences on diagnostic outcomes by comparing distributions of schizophrenia-spectrum, affective and other psychoses between NAPLS-2 and PRONIA (**Supplementary Table 1**). Then, we assessed follow-up and transition intervals (**Table 2, Supplementary Table 2**) and conducted a Kaplan-Meier analysis to compare the transition dynamics between PRONIA-UHR, PRONIA-CHR and NAPLS-2 cohorts (**Figure 1**). Finally, we used the PRONIA-CHR|ROD sample to assess the risk calculator's capacity to generalize to a broader, transdiagnostic risk population encompassing patients with basic symptoms and depressive syndromes (17,49,50).

The original NAPLS-2 risk calculator was developed with 8 preselected variables (16). Of these, 6 were also available in PRONIA: age; severity of the SIPS positive items 'unusual thought content' (P1) and 'suspiciousness' (P2); score on the Brief Assessment of Cognition in Schizophrenia (BACS) symbol coding test (51); score on the Hopkins Verbal Learning Test-Revised (HVLT-R) (52); decline in social functioning over the past year as measured by the Global Functioning Scale: Social (GFS) (53); and family history of psychotic disorders in a first-degree relative. The missing variables 'number of types of trauma endorsed' and 'undesirable life events score' were non-significant psychosis predictors in NAPLS-2 (16) and have not been regularly included in validation studies (31) or recent work (54,55).

In NAPLS-2, the 6-variable risk calculator performed with sensitivity=55%, specificity=79%, and balanced accuracy (BAC)=67% at an 0.2 estimated risk cutoff (BAC=68.5% in the 8-variable model). We used the Cox regression coefficients of these 6 variables to compute risk estimates for the PRONIA patients (**Figure 2A**). Before applying these coefficients to PRONIA-CHR|ROD (n=334), CHR (n =167) or UHR cohorts (n=126), we imputed missing values (26 [1.3%] out of 2004) using a standardized Euclidean distance-based nearest-neighbor approach. Due to differences in related sample characteristics (**Table 2**; **Supplementary Table 2**), we

evaluated a consortium-level calibration procedure. Specifically, we mean-centered each PRONIA predictor to the respective NAPLS-2 variable by computing the difference of means between variables and subtracting this difference from the respective PRONIA predictor. This procedure mitigated mean differences while preserving within-sample variance used by the regression model to determine transition (**Supplementary Table 2** and **Supplementary Figure 1**; i.e., calibration globally increased the SIPS-P1P2 scores and reduced HVLT, DSST scores and age of PRONIA patients). Then, we re-applied the risk calculator to the adjusted PRONIA data to recompute risk estimates and prognostic group assignments for the CHR|ROD, CHR, and UHR patients (**Figure 2B**). This procedure was repeated by using the PRONIA-UHR sample as reference for data calibration (**Figure 2C**) to determine the level of diagnostic specificity required for calibration. The performance of the NAPLS-2 model was measured in terms of sensitivity, specificity, balanced accuracy (BAC), positive and negative likelihood ratios, and area-under-the curve (**Table 3**). These metrics were also computed per PRONIA site (**Supplementary Table 3**). Finally, we evaluated the distribution and calibration of the model's estimates in the three PRONIA samples (**Supplementary Figures 2** and **3**). Model calibration was measured using the Expected Calibration Error (ECE) which is the weighted average difference between the fraction of correctly predicted outcomes and predicted probabilities across the binned probability range (56).

Machine learning analyses

We integrated our machine learning software NeuroMiner (version 1.05; http://proniapredictors.eu/neurominer/index.html) into ViPAR to evaluate the interactions between different prognostic algorithms and the three risk enrichment levels. To improve the calibration of Cox Proportional Hazards (Cox-PH) regression to the different risk levels in NAPLS-2 and PRONIA, we extended NeuroMiner with an adaptive Cox-PH algorithm that identifies an optimal risk estimate cutoff for prediction based on the test cases' distribution of risk

estimates. Thus, the algorithm learns to calibrate itself to risk samples with divergent absolute risks distributions. Due to the unbalanced transition and non-transition samples, we also tested whether combining this algorithm with an adaptive synthetic up-sampling method for the transition minority class would improve prediction. To this end, we employed the Adaptive Synthetic (ADASYN) algorithm (57) which creates a weighted distribution of psychosis transition samples according to their difficulty of being correctly predicted and then produces a larger number of synthetic samples in the neighborhood of difficult-to-learn patients. We used ADASYN with pre-defined parameters β=0.7, k_{SMOTE}=5, and $k_{density}$=11, which respectively determine desired class balance, number of nearest neighbors used to create artificial samples for the minority class, and number of nearest neighbors irrespective of class membership.

We compared these Cox-PH algorithms with different forms of logistic regression and support vector machines (SVM), which are commonly used for predictive modelling (27) (**Supplementary Table 4**). Additionally, we tested whether meta-learning algorithms based on stacked generalization (58) outperformed single prediction strategies by optimally combining single algorithms' predictions (**Supplementary Figure 4**). The methods used for training, validating and comparing these algorithms are detailed in the **Supplementary Material**.

RESULTS

Group-level differences between samples

Eighty-four out of 596 NAPLS-2-UHR participants developed psychosis during follow-up (transition rate: 14.1%). In PRONIA, the transdiagnostic transition rate was significantly lower: 26 (22 UHR, 1 CHR and 3 ROD) out of 334 participants (167 CHR and 167 ROD) developed psychosis (transition rate: 7.8%; $\chi(1)^2$=4.83, *P*=.028). However, neither did CHR- or UHR-specific transition rates differed between PRONIA and NAPLS-2 (CHR: 23 out of 167 [13.8%,]

$\chi(1)^2$=0.25, *P*=.621; UHR: 22 out of 127 [17.3%], $\chi(1)^2$=2.32, *P*=.128) nor did transition dynamics distinguish both cohorts (**Figure 1**). The PRONIA and NAPLS-2 cohorts differed on almost all examined sociodemographic, clinical, and neurocognitive baseline variables, including the NAPLS-2 risk calculator features, as well as in the follow-up intervals (**Table 2, Supplementary Table 1**). Specifically, PRONIA patients were more than 5 years older, had more years of education, were more likely to be female in the transition group, and less likely to be of non-white ethnicity. PRONIA patients scored significantly lower on the SIPS-P1P2 summary item. In the BACS symbol coding and HVLT tests, the PRONIA transition cases scored between the NAPLS-2 transition and non-transition patients. Furthermore, comorbid DSM-IV diagnoses differed between both cohorts, with PRONIA patients showing a higher prevalence of psychiatric multi-morbidity and respective differences between transition and non-transition groups (**Table 2**). Besides more prevalent major depression due to the inclusion of ROD patients, the PRONIA samples had a higher frequency of other affective diagnoses and eating disorders, while the NAPLS-2 cohort showed an increased prevalence of substance abuse and anxiety disorders. Baseline psychometric depression scores did not differ between transition and non-transition groups in either cohort (**Table 2**). Finally, diagnostic outcomes of transition cases did not differ between PRONIA and NAPLS-2 (**Supplementary Table 1**).

External validation of the NAPLS-2 model in the PRONIA study

In the unadjusted PRONIA-CHR|ROD, CHR, and UHR samples, the performance of the NAPLS-2 model as determined by the original 0.2 cutoff in predicted risk ranged below the levels reported in the original publication due to unbalanced sensitivity-specificity relationships (BAC=58.4%-63.9%, sensitivity=38.5%-45.5%, specificity=71.4%-89.3%, **Table 3** and **Figure 2A**). The mean-centering of the PRONIA variables to the NAPLS-2 data significantly increased performance across all PRONIA samples, with the broadest risk definition being associated with

the highest prognostic accuracy (CHR|ROD: BAC=69.7%, sensitivity=73.1%, specificity=66.2%; CHR: BAC=67.1%, sensitivity=73.9%, specificity=60.3%; UHR: BAC=67.8%, sensitivity=72.7%, specificity=62.9%; **Table 3** and **Figure 2B**). When the PRONIA-UHR group served as reference sample for offset removal, the NAPLS-2 model performed best in the CHR|ROD sample (BAC=73.3%, sensitivity=61.5%, specificity=85.1%) followed by CHR (BAC=70.2%, sensitivity=69.6%, specificity=70.9%) and UHR samples (BAC=67.8%, sensitivity=72.7%, specificity=62.9%; **Table 3** and **Figure 2C**). The sensitivity-specificity relationships of the latter calibration approach were less balanced in the CHR|ROD and CHR samples than those of the former approach.

Reciprocal external validation analyses

The reciprocal model discovery and validation of 9 different algorithms replicated the prognostic accuracy gains seen in the CHR|ROD and CHR cohorts during the external validation of the NAPLS-2 model (**Supplementary Table 5, Supplementary Figure 4**). This effect was particularly apparent when the NAPLS-2 UHR sample served as external validation cohort (**Figure 3A**): When algorithms were derived from PRONIA-CHR|ROD patients, their average performance measured BAC=67.8% (sensitivity=69.6%, specificity=65.9%). Among these, the two stacking models achieved the highest BAC (69.2%-69.8%). In contrast, algorithms performed at BAC=57.1% (sensitivity=45.0%, specificity=69.4%) when trained on PRONIA-UHR patients, with the linear SVM producing the highest BAC (65.1%). Classifier comparisons confirmed these observations by showing that (1) algorithm derivation including PRONIA-CHR|ROD outperformed more confined risk enrichment levels ($BAC_{CHR|ROD}$=68.4%; BAC_{CHR}=64.9%; BAC_{UHR}=58.7%), and (2) machine learning models based on SVM or ridge logistic regression were superior to other approaches (**Supplementary Figure 4**). When derived from NAPLS-2 and tested in the three PRONIA samples, ridge logistic regression produced a

mean [SD] BAC of 70.4% [2.0%], outperforming the other strategies, including the NAPLS-2 risk calculator (mean [SD] BAC=68.2% [1.3%]; **Figure 3B**). The supplementary leave-site-out analysis corroborated these findings by showing that mean cross-site prognostic performances increased from UHR to CHR and CHR|ROD (**Supplementary Table 9**, **Supplementary Figure 6**).

The adaptive Cox-PH model showed better calibration compared to the NAPLS-2 risk calculator (**Supplementary Figure 3A** vs. **3D**): the lowest ECE was observed when the model was trained on the PRONIA-CHR|ROD sample and tested in NAPLS-2 (ECE=0.04; **Supplementary Figure 3C**). The event-per-variable simulation analysis performed for Cox-PH, ridge logistic regression and linear-kernel SVM in the CHR|ROD sample demonstrated that algorithms largely generated stable external validation performances (**Supplementary Figure 5**). Finally, non-regularized logistic regression produced inferior models (**Supplementary Figure 4**).

Predictive feature relevance, as measured across 6 linear algorithms and the three risk enrichment levels (**Supplementary Figure 4C**) was highest for ridge logistic regression and linear-kernel SVM and lowest for the Cox-PH and linear SVM models. Particularly, for the former two algorithms, the broadening of risk from UHR to CHR|ROD patients increased the predictive value of younger age, a positive family history for psychosis, decline in global functioning, more pronounced attenuated psychotic symptoms, and lower performance in the BACS digital symbol test. Across the three risk levels, SVM or ridge logistic regression models showed a higher relevance compared to Cox-PH or non-regularized logistic regression when integrated using stacked generalization (**Supplementary Figure 7**).

DISCUSSION

The external validation of prognostic models has been identified as bottleneck, yet mandatory translational step for their clinical implementation (59). In this regard, a standardized framework for model comparison between independent projects may mitigate multiple sources of bias caused by the idiosyncrasies of study purposes, patient recruitment strategies and predictive model designs (60). To our knowledge, HARMONY is the first initiative to set up such a secure international forum for collaborative model discovery and validation in early recognition research.

HARMONY allowed us to test the generalizability and prognostic value of the NAPLS-2 psychosis risk signature (16) both at the international scale and across diverse risk samples provided by the European PRONIA project (38). We encountered significant consortium-level differences, which were likely fueled by systematic variation in participant referral, ascertainment, enrolment, and retainment, resulting in two cohorts that differed on sociodemographic, clinical parameters, treatments, and psychiatric comorbidities. A key observation was that these differences reduced the generalizability of the NAPLS-2 risk calculator but could be overcome by a simple data calibration procedure that replicated the derivation sample performance of the NAPLS-2 model (BAC=67%) in the corresponding PRONIA-UHR sample (BAC=68%). Importantly, we found evidence that calibration facilitated generalizability of the risk calculator beyond its original UHR scope, i.e., to risk samples that included patients with basic symptoms (BAC=67%) and even extended to patients with a first lifetime episode of major depression (BAC=70%; **Figure 1B**). Further BAC increases of +3.6% (CHR|ROD) or +3.1 (CHR) could be achieved in these broader risk samples by calibrating based on mean differences between UHR patients instead of adjusting for the mean differences between the NAPLS-2 UHR patients and respective target risk samples (**Figure 1C**). However, these increases came at the cost of lower prognostic sensitivity (**Table 3**) and would result in

transition cases being undetected. In this regard, our recent work indicated that prognostic models with high sensitivity would be preferable because they complement the high prognostic specificity of clinical raters, leading potentially to more accurate predictions in the clinical setting (37).

Based on the observation of project-level differences between NAPLS-2 and PRONIA and related literature on population-specific model re-calibration strategies (61), we developed a new Cox-PH algorithm which uses an optimal *relative* risk cut-off compared to the absolute risk threshold (*ER*=0.2) of the original model (16). The algorithm's precision and calibration performed well across the risk enrichment levels of the PRONIA sample (CHR|ROD: BAC=70%, ECE=13%; CHR: BAC=70%, ECE=9%; UHR: BAC=66%, ECE=8%). This finding is highly relevant for model implementation because target populations will inevitably differ in their absolute risk levels, as encountered in the NAPLS-2 sample (optimal estimated risk cut-off for transition prediction: *ER*=0.267) compared with the PRONIA-UHR cohort (*ER*=0.184).

An important caveat for the clinical implementation of these adaptive Cox-PH models should be considered: they require that youth mental health services have access to calibration data that represent their given help-seeking population. The following solution may facilitate model deployment to clinical sites without these legacy data: a public library of risk populations with algorithm-specific estimated risk distributions would be established through international collaboration. When new sites implement a risk prediction model from the library, they can extract the estimated risk distribution of the 'template population' that most closely approximates the sociodemographic and clinical characteristics of their (expected) target population. Then, following the principles of online machine learning (62), risk estimates generated during the clinical application of the model could be used to continuously update the template distribution, thus increasingly replacing the template with the optimal target risk distribution. Further

improvements could be achieved by developing pre-test risk decomposition methods that analyze the sociodemographic and clinical parameters of the given risk person and inform the adaptive Cox-PH model of the most appropriate risk distribution (63,64). Such methods would also address the limitation of mean centering, which requires a reference target sample for calibration.

However, our analyses also showed that the NAPLS-2 risk calculator could be further enhanced by replacing the underlying methodology with linear-kernel SVM (CHR|ROD, CHR) or ridge logistic regression (UHR). The analysis of feature relevance indicated that these regularized algorithms yielded more complex, yet stable clinical-neurocognitive patterns, while the Cox-PH methods put less emphasis on neurocognitive information. Higher pattern complexity may have facilitated increased prognostic accuracy in the CHR|ROD and CHR cohorts. Interestingly however, the use of kernelized SVM for predicting psychosis in UHR patients induced a drastic loss of sensitivity in the reciprocal external validation (**Supplementary Figure 4**). This indicates that kernel projections might have amplified residual between-consortia differences, and thus compounded the challenge of delineating the UHR minority class with a transition to psychosis (22 [17.3%]) from the majority of UHR patients, who did not develop psychosis but also did not recover from the risk syndrome (68 out of 105 [64.8%]). On the other hand, in the leave-site-out analyses we found that kernelized and stacking algorithms outperformed Cox-PH methods in terms of modestly higher prognostic performance (4-5% BAC) and lower cross-site variability (-5.3% to -7.4% BAC; **Supplementary Table 9**). These modest gains are expected because the analyzed risk space spanned just 6 variables, pre-selected among many other potential sociodemographic, clinical, behavioral and neurocognitive predictors through a decade-long literature-driven and expert-based process (9,16,65,66). Still, the observed performance differences may considerably impact on the success of model-informed early intervention strategies targeting at-risk populations at the international scale, and

they may further increase in high dimensional data, when no a priori knowledge about variable relevance exists (36).

Strikingly, the present study showed that model derivation from a heterogeneous help-seeking population comprising CHR states or recent-onset depression, leads to more precise and generalizable outputs. Of note, higher prognostic performance was not driven by the 'easy' prediction of non-transition in depressed patients, but by increased sensitivity for true transition events. This finding provides an independent replication of our recent work in the currently largest UHR sample provided by NAPLS-2 (37). These results also align with previous studies (17,34), suggesting that psychosis risk is not confined to UHR states but gradually increases from 'neighboring' affective conditions which typically co-occur with these syndromes, over cognitive basic symptoms to attenuated and brief limited intermittent psychosis. This finding may also point to an increased representational power of CHR|ROD-trained models due to the extension of the risk spectrum towards lower-risk individuals with early-onset affective disorders, who may share bio-behavioral features of psychosis (43,67–70). Because depressive and negative symptoms overlap (e.g., anhedonia or blunted affect), the results may alternatively suggest an unrecognized negative-symptom risk state or high-risk group within depression—although other research would suggest that risk will be generalized across help-seeking populations (17,49). Therefore, future studies should investigate whether this enrichment effect is specific to affective disorders or includes other conditions which evolve in adolescence and young adulthood (71,72).

In summary, we found that the NAPLS-2-derived risk pattern may generate internationally scalable, machine learning-based tools for psychosis risk ascertainment in youth with diverse psychosis risk syndromes. The clinical application scope of this risk signature extended beyond the prevailing UHR-focused concepts of the current early recognition literature. This may have

important consequences for the implementation of precision medicine tools in the youth mental health field, potentially facilitating more targeted and accessible preventive strategies in the near future (73). The HARMONY initiative provided a useful resource for integrated model discovery and validation at the highest level of validity achievable with retrospective data. Future studies should compare the generalizability of different clinical-neurocognitive risk signatures, their clinical utility for treatment stratification and the potential value of dynamic and biological information for further improving individualized predictions (37,74,75). Importantly, additional layers of biological data may not only increase prognostic accuracy, but also models' sensitivity to systematic covariate shifts between high-risk populations. Therefore, in-depth model validation based on international collaboration is pivotal to study these effects in greater detail and engineer sophisticated solutions (76) towards a more precise early detection and prevention of psychosis.

Acknowledgments

Author Contributions

Study design:
Nikolaos Koutsouleris, Lana Kambeitz-Ilankovic, Marlene Rosen, Stephan Ruhrmann, Jean Addington, Diana O. Perkins, Carrie E. Bearden, Barbara A. Cornblatt, Kristin S. Cadenhead, Daniel H. Mathalon, Thomas McGlashan, Larry Seidman, Ming Tsuang, Elaine F. Walker, Scott W. Woods, Eva Meisenzahl, Frauke Schultze-Lutter, Joseph Kambeitz, Raimo K. R. Salokangas, Paolo Brambilla, Stefan Borgwardt, Stephen Wood, Tyrone D. Cannon

Collection of data:
Nikolaos Koutsouleris, Michelle Worthington, Lana Kambeitz-Ilankovic, Rachele Sanfelici, Marlene Rosen, Stephan Ruhrmann, Jean Addington, Diana O. Perkins, Carrie E. Bearden, Barbara, A. Cornblatt, Kristin S. Cadenhead, Daniel H. Mathalon, Thomas McGlashan, Larry Seidman, Ming Tsuang, Elaine F. Walker, Scott W. Woods, Eva Meisenzahl, Frauke Schultze-Lutter, Joseph Kambeitz, Raimo K. R. Salokangas, Peter Falkai, Rebekka Lencer, Alessandro Bertolino, Paolo Brambilla, Rachel Upthegrove, Stefan Borgwardt, Stephen Wood, Tyrone D. Cannon

Data analysis:
Nikolaos Koutsouleris, Michelle Worthington, Dominic B. Dwyer

Data interpretation:
Nikolaos Koutsouleris, Michelle Worthington, Paolo Fusar-Poli, Kristin S. Cadenhead, Frauke Schultze-Lutter, Philip McGuire, Raquel Gur, Tyrone D. Cannon

Writing of the manuscript:
Nikolaos Koutsouleris, Michelle Worthington, Tyrone D. Cannon

Critical review of the manuscript:
Dominic B. Dwyer, Lana Kambeitz-Ilankovic, Rachele Sanfelici, Paolo Fusar-Poli, Marlene Rosen, Stephan Ruhrmann, Alan Anticevic, Jean Addington, Diana O. Perkins, Carrie E. Bearden, Barbara A. Cornblatt, Kristin S. Cadenhead, Daniel H. Mathalon, Thomas McGlashan, Ming Tsuang, Elaine F. Walker, Scott W. Woods, Eva Meisenzahl, Frauke Schultze-Lutter, Joseph Kambeitz, Raimo K. R. Salokangas, Peter Falkai, Rebekka Lencer, Alessandro Bertolino, Paolo Brambilla, Rachel Upthegrove, Stefan Borgwardt, Stephen Wood, Philip McGuire, Raquel Gur, Tyrone D. Cannon,

Role of funding sources

The following funders were not involved in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication:

PRONIA is a Collaboration Project funded by the European Union under the 7th Framework Programme under grant agreement n° 602152.

NAPLS-2 was further supported by NIH grants U01 MH081902 (to Dr. Cannon), P50 MH066286 (to Dr. Bearden), U01 MH081857 (to Dr. Cornblatt), U01 MH82022 (to Dr. Woods), U01 MH066134 (to Dr. Addington), U01 MH081944 (to Dr. Cadenhead), R01 U01 MH066069 (to Dr. Perkins), R01 MH076989 (to Dr. Mathalon), U01 MH081928 (to Dr. Seidman), and U01 MH081988 (to Dr. Walker).

The HARMONY collaboration was supported by the NIH administrative supplement 3U01MH081928-07S1 (Dr. Seidman).

Nikolaos Koutsouleris has not received any honoraria for the writing of this manuscript and had full access to the data in the study. He takes the final responsibility for the decision to submit this work for publication.

Financial Disclosures

Dr Koutsouleris and Dr Meisenzahl hold issued patent US20160192889A1 (‘Adaptive pattern recognition for psychosis risk modelling’). No other disclosures were reported with respect to the current work.

References

1. McGlashan TH, Miller TJ, Woods SW, Hoffman RE, Davidson L (2001): Instrument for the assessment of prodromal symptoms and states. In: Miller T, Madnick SA, McGlashan TH, Libiger J, Johannessen JO, editors. *Early Intervention in Psychiatric Disorders*. Dordrecht: Kluwer Academic, pp 135–149.
2. Schultze-Lutter F, Addington J, Ruhrmann S, Klosterkötter J (2007): *Schizophrenia Proneness Instrument, Adult Version (SPI-A)*. Rome.
3. Schultze-Lutter F, Koch E (2010): *Schizophrenia Proneness Instrument, Children and Youth Version (SPI-CY)*. Rome.
4. Yung AR, McGorry PD (1996): The prodromal phase of first-episode psychosis: past and current conceptualizations. *Schizophrenia Bulletin* 22: 353–370.
5. Yung AR, McGorry PD (1996): The initial prodrome in psychosis: descriptive and qualitative aspects. *Australian and New Zealand Journal of Psychiatry* 30: 587–599.
6. Woods SW, Bearden CE, Sabb FW, Stone WS, Torous J, Cornblatt BA, *et al.* (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
7. Ruhrmann S, Schultze-Lutter F, Salokangas RKR, Heinimaa M, Linszen D, Dingemans P, *et al.* (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.
8. Fusar-Poli P, Schultze-Lutter F, Cappucciati M, Rutigliano G, Bonoldi I, Stahl D, *et al.* (2015): 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: 732–743.
9. Fusar-Poli P, Pablo GS de, Correll C, Meyer-Lindenberg A, Millan MJ, Borgwardt S, *et al.* (2020): Prevention of Psychosis: Advances in Detection, Prognosis and Intervention. *JAMA Psychiatry* 77: 1–11.
10. Fusar-Poli P, Cappucciati M, Rutigliano G, Schultze-Lutter F, Bonoldi I, Borgwardt S, *et al.* (2015): At risk or not at risk? A meta-analysis of the prognostic accuracy of psychometric interviews for psychosis prediction. *World Psychiatry* 14: 322–332.
11. Nelson B (2014): Attenuated psychosis syndrome: don't jump the gun. *Psychopathology* 47: 292–296.
12. Bodatsch M, Ruhrmann S, Wagner M, Müller R, Schultze-Lutter F, Frommann I, *et al.* (2011): Prediction of psychosis by mismatch negativity. *Biological Psychiatry* 69: 959–966.
13. Koutsouleris N, Davatzikos C, Bottlender R, Patschurek-Kliche K, Scheuerecker J, Decker P, *et al.* (2012): Early recognition and disease prediction in the at-risk mental states for psychosis using neurocognitive pattern classification. *Schizophr Bull* 38: 1200–1215.
14. Koutsouleris N, Riecher-Rössler A, Meisenzahl EM, Smieskova R, Studerus E, Kambeitz-Ilankovic L, *et al.* (2015): Detecting the psychosis prodrome across high-risk populations using neuroanatomical biomarkers. *Schizophr Bull* 41: 471–482.
15. Perkins DO, Jeffries CD, Addington J, Bearden CE, Cadenhead KS, Cannon TD, *et al.* (2014): Towards a Psychosis Risk Blood Diagnostic for Persons Experiencing High-Risk Symptoms: Preliminary Results From the NAPLS Project. *Schizophrenia Bulletin*. https://doi.org/10.1093/schbul/sbu099

16. Cannon TD, Yu C, Addington J, Bearden CE, Cadenhead KS, Cornblatt BA, *et al.* (2016): An individualized risk calculator for research in prodromal psychosis. *American Journal of Psychiatry* 173: 980–988.
17. 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: 493–500.
18. Zhang T, Xu L, Tang Y, Li H, Tang X, Cui H, *et al.* (2018): Prediction of psychosis in prodrome: development and validation of a simple, personalized risk calculator. *Psychological medicine* 1–9.
19. Perkins DO, Olde Loohuis L, Barbee J, Ford J, Jeffries CD, Addington J, *et al.* (2019): Polygenic Risk Score Contribution to Psychosis Prediction in a Target Population of Persons at Clinical High Risk. *American Journal of Psychiatry* 177: 155–163.
20. Schultze-Lutter F, Michel C, Schimmelmann B, Ruhrmann S (2017): Clinical high risk symptoms and criteria in the community: Prevalence, clinical significance and risk factors for their occurrence. *European Psychiatry* 41: S226.
21. Schimmelmann BG, Michel C, Martz-Irngartinger A, Linder C, Schultze-Lutter F (2015): Age matters in the prevalence and clinical significance of ultra-high-risk for psychosis symptoms and criteria in the general population: Findings from the BEAR and BEARS-kid studies. *World Psychiatry* 14: 189–197.
22. Lindgren M, Jonninen M, Jokela M, Therman S (2019): Adolescent psychosis risk symptoms predicting persistent psychiatric service use: A 7-year follow-up study. *European Psychiatry* 55: 102–108.
23. Rickwood DJ, Telford NR, Parker AG, Tanti CJ, McGorry PD (2014): headspace - Australia's innovation in youth mental health: who are the clients and why are they presenting? *The Medical journal of Australia* 200: 108–111.
24. Pelizza L, Azzali S, Garlassi S, Paterlini F, Scazza I, Chiri LR, *et al.* (2018): Adolescents at ultra-high risk of psychosis in Italian neuropsychiatry services: prevalence, psychopathology and transition rate. *European child & adolescent psychiatry* 27: 725–737.
25. Zwicker A, MacKenzie LE, Drobinin V, Howes Vallis E, Patterson VC, Stephens M, *et al.* (2019): Basic symptoms in offspring of parents with mood and psychotic disorders. *BJPsych open* 5: e54.
26. Addington J, Piskulic D, Liu L, Lockwood J, Cadenhead KS, Cannon TD, *et al.* (2017): Comorbid diagnoses for youth at clinical high risk of psychosis. *Schizophrenia Research* 190: 90–95.
27. Sanfelici R, Dwyer D, Antonucci LA, 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* in press.
28. Salazar de Pablo G, Studerus E, Vaquerizo-Serrano J, Irving J, Catalan A, Oliver D, *et al.* (2021): Implementing Precision Psychiatry: A Systematic Review of Individualized Prediction Models for Clinical Practice. *Schizophrenia Bulletin* 47: 284–297.
29. Carrión RE, Cornblatt BA, Burton CZ, Tso IF, Auther AM, Adelsheim S, *et al.* (2016): Personalized prediction of psychosis: external validation of the NAPLS-2 psychosis risk calculator with the EDIPPP project. *American Journal of Psychiatry* 173: 989–996.
30. Osborne KJ, Mittal VA (2019): External validation and extension of the NAPLS-2 and SIPS-RC personalized risk calculators in an independent clinical high-risk sample. *Psychiatry Research* 279: 9–14.

31. Zhang T, Li H, Tang Y, Niznikiewicz MA, Shenton ME, Keshavan M, *et al.* (2018): Validating the Predictive Accuracy of the NAPLS-2 Psychosis Risk Calculator in a Clinical High-Risk Sample From the SHARP (Shanghai At Risk for Psychosis) Program. *American Journal of Psychiatry* 175: 906–908.
32. Klosterkötter J, Hellmich M, Steinmeyer EM, Schultze-Lutter F (2001): Diagnosing schizophrenia in the initial prodromal phase. *Archives of General Psychiatry* 58: 158–164.
33. Schultze-Lutter F, Ruhrmann S, Berning J, Maier W, Klosterkötter J (2010): Basic symptoms and ultrahigh risk criteria: symptom development in the initial prodromal state. *Schizophrenia Bulletin* 36: 182–191.
34. Lee TY, Lee J, Kim M, Choe E, Kwon JS (2018): Can we predict psychosis outside the clinical high-risk state? A systematic review of non-psychotic risk syndromes for mental disorders. *Schizophrenia Bulletin* 44: 276–285.
35. Schultze-Lutter F, Klosterkötter J, Ruhrmann S (2014): Improving the clinical prediction of psychosis by combining ultra-high risk criteria and cognitive basic symptoms. *Schizophrenia Research* 154: 100–106.
36. Fusar-Poli P, Stringer D, Durieux AM, Rutigliano G, Bonoldi I, De Micheli A, Stahl D (2019): Clinical-learning versus machine-learning for transdiagnostic prediction of psychosis onset in individuals at-risk. *Translational Psychiatry* 9: 1–11.
37. Koutsouleris N, Dwyer DB, Degenhardt F, Maj C, Urquijo-Castro MF, Sanfelici R, *et al.* (2021): Multimodal Machine Learning Workflows for Prediction of Psychosis in Patients With Clinical High-Risk Syndromes and Recent-Onset Depression. *JAMA Psychiatry* 78: 195–209.
38. Koutsouleris N, Kambeitz-Ilankovic L, Ruhrmann S, Rosen M, Ruef A, Dwyer DB, *et al.* (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: 1156–1172.
39. Fusar-Poli P, Nelson B, Valmaggia L, Yung AR, McGuire PK (2014): Comorbid Depressive and Anxiety Disorders in 509 Individuals With an At-Risk Mental State: Impact on Psychopathology and Transition to Psychosis. *Schizophrenia Bulletin* 40 (1). https://doi.org/10.1093/schbul/sbs136
40. Häfner H, Maurer K, Trendler G, Heiden W an der, Schmidt M, Könnecke R (2005): Schizophrenia and depression: challenging the paradigm of two separate diseases–a controlled study of schizophrenia, depression and healthy controls. *Schizophrenia Research* 77: 11–24.
41. Wigman JTW, van Nierop M, Vollebergh WAM, Lieb R, Beesdo-Baum K, Wittchen H-U, van Os J (2012): Evidence that psychotic symptoms are prevalent in disorders of anxiety and depression, impacting on illness onset, risk, and severity–implications for diagnosis and ultra-high risk research. *Schizophrenia Bulletin* 38: 247–257.
42. Musliner KL, Mortensen PB, McGrath JJ, Suppli NP, Hougaard DM, Bybjerg-Grauholm J, *et al.* (2019): Association of Polygenic Liabilities for Major Depression, Bipolar Disorder, and Schizophrenia With Risk for Depression in the Danish Population. *JAMA Psychiatry* 76: 516–525.
43. Koutsouleris N, Meisenzahl EM, Borgwardt S, Riecher-Rössler A, Frodl T, Kambeitz J, *et al.* (2015): Individualized differential diagnosis of schizophrenia and mood disorders using neuroanatomical biomarkers. *Brain* 138: 2059–2073.

44. Addington J, Cadenhead KS, Cornblatt BA, Mathalon DH, McGlashan TH, Perkins DO, *et al.* (2012): North American Prodrome Longitudinal Study (NAPLS 2): overview and recruitment. *Schizophrenia Research* 142: 77–82.
45. First MB, Spitzer RL, Gibbon M, Williams JB, others (1994): Structured clinical interview for Axis I DSM-IV disorders. *New York: Biometrics Research*.
46. Carter KW, Francis RW, Carter KW, Francis RW, Bresnahan M, Gissler M, *et al.* (2016): ViPAR: a software platform for the Virtual Pooling and Analysis of Research Data. *International journal of epidemiology* 45: 408–416.
47. Royston P, Altman DG (2013): External validation of a Cox prognostic model: principles and methods. *BMC Medical Research Methodology* 13: 33.
48. McGlashan T, Walsh B, Woods S (2010): *The Psychosis-Risk Syndrome: Handbook for Diagnosis and Follow-Up*. Oxford University Press.
49. Fusar-Poli P, Werbeloff N, Rutigliano G, Oliver D, Davies C, Stahl D, *et al.* (2018): Transdiagnostic Risk Calculator for the Automatic Detection of Individuals at Risk and the Prediction of Psychosis: Second Replication in an Independent National Health Service Trust. *Schizophrenia Bulletin*. https://doi.org/10.1093/schbul/sby070
50. McGorry PD, Hartmann JA, Spooner R, Nelson B (2018): Beyond the "at risk mental state" concept: transitioning to transdiagnostic psychiatry. *World Psychiatry* 17: 133–142.
51. Keefe RS, Goldberg TE, Harvey PD, Gold JM, Poe MP, Coughenour L (2004): The Brief Assessment of Cognition in Schizophrenia: reliability, sensitivity, and comparison with a standard neurocognitive battery. *Schizophrenia research* 68: 283–297.
52. Benedict RH, Schretlen D, Groninger L, Brandt J (1998): Hopkins Verbal Learning Test–Revised: Normative data and analysis of inter-form and test-retest reliability. *The Clinical Neuropsychologist* 12: 43–55.
53. Cornblatt BA, Auther AM, Niendam T, Smith CW, Zinberg J, Bearden CE, Cannon TD (2007): Preliminary findings for two new measures of social and role functioning in the prodromal phase of schizophrenia. *Schizophrenia Bulletin* 33: 688–702.
54. Worthington MA, Walker EF, Addington J, Bearden CE, Cadenhead KS, Cornblatt BA, *et al.* (2020): Incorporating cortisol into the NAPLS2 individualized risk calculator for prediction of psychosis. *Schizophrenia research*. https://doi.org/10.1016/j.schres.2020.09.022
55. Chung Y, Addington J, Bearden CE, Cadenhead K, Cornblatt B, Mathalon DH, *et al.* (2019): Adding a neuroanatomical biomarker to an individualized risk calculator for psychosis: A proof-of-concept study. *Schizophrenia research* 208: 41–43.
56. Guo C, Pleiss G, Sun Y, Weinberger KQ (2017): *On Calibration of Modern Neural Networks*.
57. Haibo H, Yang B, Edwardo GA, Shutao L (2016): Adaptive Synthetic Sampling Approach for Imbalanced Learning. *IEEE International Joint Conference on Neural Networks, IJCNN*, vol. 8 8: 1322–1328.
58. Wolpert DH (1992): Stacked generalization. *Neural Networks* 5: 241–259.
59. Steyerberg EW, Moons KGM, van der Windt DA, Hayden JA, Perel P, Schroter S, *et al.* (2013): Prognosis Research Strategy (PROGRESS) 3: prognostic model research. *PLoS Medicine* 10: e1001381.
60. Fusar-Poli P, Hijazi Z, Stahl D, Steyerberg EW (2018): The Science of Prognosis in Psychiatry: A Review. *JAMA psychiatry* 75: 1289–1297.
61. Pennells L, Kaptoge S, Wood A, Sweeting M, Zhao X, White I, *et al.* (2019): Equalization of four cardiovascular risk algorithms after systematic recalibration: individual-participant meta-analysis of 86 prospective studies. *European Heart Journal* 40: 621–631.

62. Murphy KP (2012): *Machine Learning: A Probabilistic Perspective*, Illustrated edition. Cambridge, MA: The MIT Press.
63. Fusar-Poli P, Rutigliano G, Stahl D, Schmidt A, Ramella-Cravaro V, Hitesh S, McGuire P (2016): Deconstructing Pretest Risk Enrichment to Optimize Prediction of Psychosis in Individuals at Clinical High Risk. *JAMA Psychiatry* 73: 1260–1267.
64. Fusar-Poli P, Tantardini M, De Simone S, Ramella-Cravaro V, Oliver D, Kingdon J, *et al.* (2017): Deconstructing vulnerability for psychosis: Meta-analysis of environmental risk factors for psychosis in subjects at ultra high-risk. *European Psychiatry* 40: 65–75.
65. Cannon TD, Cadenhead K, Cornblatt B, Woods SW, Addington J, Walker E, *et al.* (2008): Prediction of psychosis in youth at high clinical risk: a multisite longitudinal study in North America. *Archives of General Psychiatry* 65: 28–37.
66. Fusar-Poli P, Sullivan SA, Shah JL, Uhlhaas PJ (2019): Improving the Detection of Individuals at Clinical Risk for Psychosis in the Community, Primary and Secondary Care: An Integrated Evidence-Based Approach. *Frontiers in psychiatry* 10: 774.
67. Power RA, Tansey KE, ButtenschÃ¸n HN, Cohen-Woods S, Bigdeli T, Hall LS, *et al.* (2016): Genome-wide Association for Major Depression Through Age at Onset Stratification: Major Depressive Disorder Working Group of the Psychiatric Genomics Consortium. *Biological Psychiatry*. https://doi.org/10.1016/j.biopsych.2016.05.010
68. Byrne EM, Zhu Z, Qi T, Skene NG, Bryois J, Pardinas AF, *et al.* (2020): Conditional GWAS analysis to identify disorder-specific SNPs for psychiatric disorders. *Molecular Psychiatry* 1–12.
69. Zhu Y, Womer FY, Leng H, Chang M, Yin Z, Wei Y, *et al.* (2019): The Relationship Between Cognitive Dysfunction and Symptom Dimensions Across Schizophrenia, Bipolar Disorder, and Major Depressive Disorder. *Frontiers in psychiatry* 10: 253.
70. Upthegrove R, Lalousis P, Mallikarjun P, Chisholm K, Griffiths SL, Iqbal M, *et al.* (2020): The Psychopathology and Neuroanatomical Markers of Depression in Early Psychosis. *Schizophrenia Bulletin*. https://doi.org/10.1093/schbul/sbaa094
71. Caspi A, Houts RM, Ambler A, Danese A, Elliott ML, Hariri A, *et al.* (2020): Longitudinal Assessment of Mental Health Disorders and Comorbidities Across 4 Decades Among Participants in the Dunedin Birth Cohort Study. *JAMA network open* 3: e203221.
72. Poletti M, Raballo A (2021): Early intervention in psychiatry through a developmental perspective [no. 1]. *npj Schizophrenia* 7: 1–2.
73. Oliver D, Spada G, Colling C, Broadbent M, Baldwin H, Patel R, *et al.* (2021): Real-world implementation of precision psychiatry: Transdiagnostic risk calculator for the automatic detection of individuals at-risk of psychosis. *Schizophr Res* 227: 52–60.
74. Yuen HP, Mackinnon A, Hartmann J, Amminger GP, Markulev C, Lavoie S, *et al.* (2018): Dynamic prediction of transition to psychosis using joint modelling. 202: 333–340.
75. Hauke DJ, Schmidt A, Studerus E, Andreou C, Riecher-Rössler A, Radua J, *et al.* (2021): Multimodal prognosis of negative symptom severity in individuals at increased risk of developing psychosis [no. 1]. *Translational Psychiatry* 11: 1–11.
76. Kouw WM, Loog M (2018): An introduction to domain adaptation and transfer learning. *arXiv e-prints* 1812: arXiv:1812.11806.

Table 1. Definition of the three risk enrichment levels (UHR, CHR, and CHR|ROD).

Risk enrichment levels	**Represented in**	**Definition**
UHR	PRONIA and NAPLS-2	Meeting Criteria of Prodromal States (COPS) based on the Structured Interview for Psychosis-risk Syndromes (SIPS) which defines three risk groups: (a) Attenuated Psychotic Symptoms, (b) Brief limited intermittent psychotic symptoms, (c) Genetic-risk functional deterioration syndrome
CHR	PRONIA	Meeting criteria for the UHR enrichment level *OR* Meeting Criteria for the COGnitive DISturbances (COGDIS) pattern based on the Schizophrenia Proneness Instrument for Adults (SPI-A) or Children and Youth (SPI-CY)
CHR|ROD	PRONIA	Meeting criteria for the CHR enrichment level *Or* Meeting Criteria for Recent-Onset Depression based on the Structured Clinical Interview for DSM-IV-TR and the following items: (a) First life-time depressive episode, and (b) duration of current depressive episode no longer than 24 months, and (c) diagnostic criteria fulfilled within past three months.

Table 2. Sociodemographic, clinical, and functional differences between non-transition and transition cases in the NAPLS-2 and PRONIA samples.

Variable	NAPLS-2		PRONIA		Wald χ^2 / *t*(df)	*P*
	Non-transition	Transition	Non-transition	Transition		
Age[†], mean (SD)	18.6 (4.4)	18.1 (3.6)	24.7 (5.8)	23.5 (5.9)	$\chi^2(3) = 352.6$	**.008**
Sex, % Female	43%	38.1%	49.1%	61.5%	$\chi^2(3) = 8.8$	**.033**
Race, % non-white	41.8%	44%	13%	7.7%	$\chi^2(3) = 76.9$	**.017**
Years of education, mean (SD)	11.3 (2.9)	11.0 (2.5)	14.3 (2.9)	13.3 (2.5)	$\chi^2(3) = 228.9$	**.008**
Family history, % no history	84.4%	81%	90.6%	80.8%	$\chi^2(3) = 8.7$	**.042**
Baseline positive symptoms (p1+p2), mean (SD)	5.9 (2.2)	7.1 (2.3)	2.6 (2.6)	5.5 (2.8)	$\chi^2(3) = 466.5$	**.008**
HVLT, mean (SD)	25.8 (5.1)	24.2 (5.5)	28.5 (2.7)	26.5 (3.0)	$\chi^2(3) = 93.8$	**.008**
BACS, mean (SD)	57.4 (13.2)	53.2 (11.6)	61.1 (11.8)	55.0 (13.0)	$\chi^2(3) = 33.1$	**.025**
Change in GFS, mean (SD)	0.70 (1.0)	0.99 (1.2)	0.75 (0.9)	0.96 (1.1)	$\chi^2(3) = 7.6$	.054
BDI[‡], mean (SD)	--	--	23.9 (10.8)	25.6 (11.8)	$t(29.3) = -0.7$	.45
CDSS, mean (SD)	5.9 (4.6)	5.7 (4.8)	--	--	$t(110) = 0.3$	.73
Treatment with antipsychotics at baseline, % treated	17.1%	26.9%	17.2%	38.5%	$\chi^2(3) = 12.5$	**.006**
Treatment with antidepressants at baseline, % treated	28.4%	23.1%	56.8%	50.0%	$\chi^2(3) = 80.2$	**<.001**
Follow-up interval [days] (SD)	563.6 (199.3)	216.7 (171.1)	697.5 (293.7)	246.9 (244.5)	$\chi^2(3) = 430.5$	**<.001**
SCID-I Diagnosis at study inclusion						
Any comorbid affective, substance, anxiety, or eating disorder, No. (%)						
No diagnosis	127 (28.2)	16 (20.5)	28 (9.1)	5 (19.2)	$\chi^2(3) = 41.7$	**<.001**
1 diagnosis	168 (37.3)	32 (41.0)	131 (42.5)	5 (19.2)		
2 diagnoses	103 (22.8)	19 (24.4)	85 (27.6)	9 (34.6)		
3 or more diagnoses	53 (11.7)	11 (14.1)	62 (20.1)	7 (26.9)		
Comorbid major depressive disorder, No. (%)						
Yes	192 (42.6)	40 (51.3)	251 (81.5)	17 (65.4)	$\chi^2(3) = 137.6$	**<.001**
No	259 (57.4)	38 (48.7)	55 (17.9)	9 (34.6)		
Comorbid affective disorders (excluding major depressive disorder), No. (%)						
No diagnosis	419 (92.9)	73 (93.6)	200 (64.9)	12 (46.2)	$\chi^2(3) = 140.7$	**<.001**
1 diagnosis	32 (7.1)	5 (6.4)	93 (31.8)	14 (53.8)		
2 diagnoses	0 (0.0)	0 (0.0)	8 (2.6)	0 (0.0)		
Comorbid substance use disorders, No. (%)						
No diagnosis	424 (94.0)	70 (89.7)	302 (98.1)	25 (96.2)	$\chi^2(3) = 14.8$	**.002**
1 diagnosis	22 (4.9)	5 (6.4)	3 (1.0)	1 (3.8)		
2 diagnoses	5 (1.1)	0 (0.0)	0 (0.0)	0 (0.0)		
3 or more diagnoses	0 (0.0)	1 (1.3)	1 (0.3)	0 (0.0)		
Comorbid anxiety disorders, No. (%)						
No diagnosis	237 (52.5)	38 (48.7)	212 (68.8)	17 (65.4)	$\chi^2(3) = 11.1$	**.01**
1 diagnosis	157 (34.8)	29 (37.2)	59 (19.2)	7 (26.9)		
2 diagnoses	49 (10.9)	9 (11.5)	25 (8.1)	1 (3.8)		
3 or more diagnoses	8 (1.8)	2 (2.6)	10 (3.2)	1 (3.8)		
Comorbid eating disorders, No. (%)						
No diagnosis	444 (98.4)	78 (100.0)	290 (94.2)	24 (92.3)	$\chi^2(3) = 14.1$	**.003**
1 or more diagnosis	7 (1.6)	0 (0.0)	16 (5.2)	2 (7.7)		

Table 3. External validation of the original NAPLS-2 risk calculator in the three PRONIA risk enrichment samples (CHR|ROD, CHR, and UHR) with and without prior centering of the predictor variables to the means of the respective NAPLS variables. Mean centering parameters were either computed by calculating the mean variable differences between the NAPLS-2 and each PRONIA cohort and subtracting them from the respective PRONIA sample, or by using the PRONIA UHR group for mean difference derivation and applying these parameters to each PRONIA sample.

PRONIA samples	TP	TN	FP	FN	Sens [%]	Spec [%]	BAC [%]	LR+	LR-	AUC
PRONIA data not mean-centered to NAPLS-2										
CHR|ROD	10	275	33	16	38.5	89.3	63.9	3.59	0.69	0.64
CHR	10	110	31	13	43.5	78.0	60.7	1.98	0.72	0.61
UHR	10	75	30	12	45.5	71.4	58.4	1.59	0.76	0.58
PRONIA data mean-centered to NAPLS-2 using respective PRONIA sample as reference										
CHR|ROD	19	204	104	7	73.1	66.2	69.7	2.16	0.41	0.70
CHR	17	85	56	6	73.9	60.3	67.1	1.86	0.43	0.67
UHR	16	66	39	6	72.7	62.9	67.8	1.96	0.43	0.68
PRONIA data mean-centered to NAPLS-2 using the PRONIA UHR sample as reference										
CHR|ROD	16	262	46	10	61.5	85 .1	73.3	4.12	0.45	0.73
CHR	16	100	41	7	69.6	70.9	70.2	2.39	0.43	0.70
UHR	16	66	39	6	72.7	62.9	67.8	1.96	0.43	0.68

Abbreviations: *TP* number of true positives, *TN* number of true negatives, *FP* number of false positives, *FN* number of false negatives, *Sens* Sensitivity, *Spec* Specificity, *BAC* Balanced Accuracy, *LR+* Positive Likelihood Ratio, *LR-* Negative Likelihood Ratio, *AUC* Area-under-the Curve.

Figure Legends

Figure 1. Kaplan-Meier analysis of cumulative non-transition rates (survival curves) between the PRONIA-UHR (upper chart) and PRONIA-CHR (lower chart) risk enrichment cohort (blue curves) and the NAPLS-2 sample (red curve). Log-rank tests were performed to test survival curves for statistical differences. Statistical significance was determined at α=0.05.

Figure 2. NAPLS-2 risk calculator estimates for the 2-year transition risk in patients who developed (red) or did not develop psychosis (blue) in three different risk cohorts of PRONIA (CHR|ROD: Sample comprising both CHR and ROD patients, CHR: Sample consisting only of CHR patients, UHR: Sample consisting only of patients fulfilling UHR criteria). Predictor variables were either not adjusted for mean differences to the NAPLS-2 data (A), adjusted using the respective PRONIA sample (B), or adjusted using the PRONIA-UHR sample as reference group (C).

Figure 3. PRONIA risk enrichment effects on psychosis prediction in the NAPLS-2 cohort and algorithm effects on the balanced accuracy of psychosis prediction in the PRONIA risk enrichment samples. **A:** Balanced accuracy of the 9 different prognostic algorithms in the NAPLS-2 cohort as a function of the PRONIA risk enrichment level used to train these algorithms. **B:** Differences in balanced accuracy as a function of type of algorithm applied to the three different PRONIA samples. The NAPLS-2 model with mean-centering to each PRONIA target sample was included in the comparisons. Additionally, means and standard errors were depicted for both A and B.