AUTHOR	Nikolaos Koutsouleris, Michelle Worthington, Dominic B. Dwyer, Lana Kambeitz-Ilankovic, Rachele Sanfelici, Paolo Fusar-Poli, Stephan Ruhrmann, Alan Anticevic, 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, Peter Falkai, Rebekka Lencer, Alessandro Bertolino, Joseph Kambeitz, Frauke Schultze-Lutter, Eva Meisenzahl, Raimo K. R. Salokangas, Paolo Brambilla, Rachel Upthegrove, Stefan Borgwardt, Stephen Wood, Philip McGuire, Raquel Gur, Tyrone D. Cannon
THE TITLE OF THE MANUCRIPT DIFFERES FROM THE PUBLISHED TITLE	Towards generalizable models for psychosis prediction An independent analysis of the NAPLS risk calculator in the multi-site PRONIA cohort
YEAR	2021
DOI	https://doi.org/10.1016/j.biopsych.2021.06.023
VERSION	Author's submitted manuscript
CITATION	Koutsouleris, Nikolaos, et al. "Toward Generalizable and Transdiagnostic Tools for Psychosis Prediction: An Independent Validation and Improvement of the NAPLS-2 Risk Calculator in the Multisite PRONIA Cohort." Biological psychiatry (2021). https://doi.org/10.1016/j.biopsych.2021.06.023

Towards generalizable models for psychosis prediction An independent analysis of the NAPLS risk calculator in the multi-site PRONIA cohort

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

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

Word Count Manuscript

Abstract:	249
Introduction:	589
Methods:	1360
Results:	631
Interpretation	903
Total:	3483 (excluding abstract)

Tables: 3

Figures: 3

Abstract

Background: Transition to psychosis (PT) is the most adverse outcome of the Clinical High-Risk (CHR) syndromes. The second phase of the North American Prodromal Longitudinal Study (NAPLS-2) proposed a psychosis risk calculator for patients with ultra-high risk (UHR) states operating on 8 clinical-neurocognitive variables.[1] The model's generalizability has not been sufficiently validated across diverse risk cohorts.

Methods: We validated the original model in the multi-site European PRONIA cohort (334 patients with CHR or recent-onset depression (ROD); 23/3 PT cases with CHR/ROD) by testing its performance in patients with UHR syndromes, UHR and basic symptoms (CHR), and a broader risk population encompassing patients with CHR states or ROD (CHR+). Using reciprocal external validation, we assessed how the choice of algorithm and the risk enrichment in different discovery populations moderated the prediction of PT in the validation samples.

Outcomes: After calibrating the PRONIA to the NAPLS-2 data, the original model predicted PT with a balanced accuracy [BAC(sensitivity,specificity)] of 64%(68%,59%) in the PRONIA-UHR, 70%(78%,61%) in CHR, and 70%(77%,62%) in CHR+ patients. Prognostic performance improved from UHR to CHR+-based models due to increased precision in NAPLS-2 UHR patients [UHR-based: 58%(61%,56%), CHR+-based: 67%(68%,66%)]. Attenuated psychotic symptoms predicted PT across risk levels, while age and processing speed were additionally predictive in the CHR+ cohort.

Interpretation: Multivariate models operating on the NAPLS-2 risk pattern reliably prognosticate PT in youth with diverse risk syndromes. Further studies should investigate the therapeutic utility of this risk signature, as well as the additional value of neurobiological information.

Research in context

Evidence before the study

We systematically searched PubMed to extract the relevant evidence from all papers published from inception to June 25th 2020 using the following search terms: ("psychosis" OR "schizophrenia") AND ("risk calculator") AND ("validation"). We identified 14 unique records, none of which involved the validation of a risk calculator model predicting likelihood of conversion to psychosis from an at-risk state in an external sample with a sufficiently clinically diverse risk cohort. Although multiple risk calculators consisting of clinical-neurocognitive measures, neuroimaging measures, or both, have been developed, cross-validation and robust external validation—which are essential for determining prognostic precision and ultimate clinical utility—has only been implemented with the individualized risk calculator developed in the second phase of the North American Prodrome Longitudinal Study (NAPLS-2). While this calculator has been validated in clinical-high risk cohorts in the US and Shanghai, it is still unknown how this model generalizes to diverse clinical risk cohorts, such as those in the European-based PRONIA study, and how Cox proportional hazard (Cox PH)-based regression models compare to different types of predictive algorithms.

Added value of the study

This is the first study to demonstrate generalizability of a risk prediction model for the development of psychosis in an international, multi-site framework within diverse risk samples including high-risk individuals with basic symptoms and affective disorders. We implemented strategies to account for cross-consortium differences between NAPLS-2 and PRONIA and showed that the model initially developed in the NAPLS-2 sample generalizes to a more transdiagnostic risk cohort. Further, not only did we show that the existing model generalizes to a more diverse risk cohort, we also demonstrated that the enrichment of the PRONIA validation sample with individuals experiencing basic symptoms and/or recent onset-depression resulted in the development of even higher performing prediction models when tested using both cross-validation and external validation in the NAPLS-2 sample.

Implications of all available evidence

Our study uses external validation strategies to demonstrate how prognostic models involving clinical-neurocognitive measures may be optimized to operate across diverse risk populations and healthcare settings across the globe, ultimately contributing to the increasing precision with which clinicians may be able to predict the development of psychosis in transdiagnostic populations.

INTRODUCTION

Over the last 30 years, research criteria defining the Clinical High-Risk (CHR) states for psychosis have been successfully established in academic sites around the globe.[2–4] The purpose of these criteria has been to early detect adolescents and young adults with an increased risk for developing psychotic disorders, study potential disease-modifying treatments in these risk cohorts and ultimately implement these strategies as a new approach of preventive psychiatry in real-world clinical care. Previous research showed that the pre-test patient referral process combined with the assessment of these CHR criteria identifies a risk population with a several hundred-fold increased incidence for psychosis.[5] Yet, the observed three-year transition rates have continuously dropped from 36% to currently 22% as more sites adopted and intensified early recognition activities.[6] Hence, due to this low prognostic value and the laborious, skill-dependent assessment of high-risk criteria, the clinical utility and scalability of the CHR paradigm have been questioned.[7]

To increase the prognostic value of the CHR designation, previous studies have proposed to augment the actual two-tier risk enrichment process—patient referral followed by CHR assessment—using algorithms that accurately measure the risk of psychosis in the individual CHR patient. These proof-of-principle studies demonstrated that individualized risk quantification could be achieved using Cox regression or machine learning models trained to estimate patients' disease transition likelihoods based on clinical, neurocognitive, neuroimaging, metabolic or genetic information.[1,8–12] If these stratification models could operate across risk cohorts and diverse healthcare environments, a more fine-grained and modular clinical management of CHR patients could be implemented: practitioners could tailor specific disease-interceptive strategies and flexibly combine them with treatments that target the varying array of psychiatric comorbidities and functional impairments present in these patients.[13]

However, for this vision to become reality major challenges must be addressed. One significant concern is that the generalizability of most risk calculators has not even been tested using cross-validation, let alone external validation approaches.[14] An exception is the clinical-neurocognitive risk calculator developed by the second phase of the North American Prodromal Longitudinal Study (NAPLS-2).[1] Based on 8 phenotypic variables, this Cox proportional-hazards (Cox-PH) model predicts PT with a sensitivity of 66% and specificity of 72% at a 20% predicted risk cutoff. So far, the NAPLS-2 model has been validated in independent CHR cohorts from the US[15] (sensitivity=58.3%, specificity=72.6%) and Shanghai (sensitivity=71.7%, specificity=45.8%).[16]

Due to these varying sensitivity and specificity levels, further validation is needed to test the model across risk populations and healthcare systems. Furthermore, as attenuated or brief limited intermittent psychotic symptoms may not mark the only pathway to psychosis,[17–20] the NAPLS-2 risk calculator should be probed across diverse risk cohorts, including youths with basic symptoms and affective disorders, as recently proposed by transdiagnostic studies of psychosis risk.[19–21] Finally, within this external validation framework, the original model should be compared with different predictive algorithms operating on the same NAPLS-2 variables to identify the model with the optimal prognostic precision across different risk populations.

The European PRONIA study (Personalised Prognostic Tools for Early Psychosis Management; www.pronia.eu)[22] recruited and followed such a diverse risk population, encompassing adolescents and young adults with different CHR states or recent-onset depression (ROD). In the current study, we first performed the external validation of the original NAPLS-2 model in PRONIA and then reciprocally trained, validated and compared different Cox-PH algorithms versus Support Vector Machine (SVM) models in the NAPLS-2 and PRONIA cohorts. In these analyses, we evaluated whether iteratively narrowing the discovery population from a cohort encompassing both CHR and ROD to patients experiencing only ultra-high-risk syndromes (UHR) moderated prognostic performance.

METHODS

Participants

Participants were drawn from the NAPLS-2[23] and the PRONIA studies.[22] NAPLS-2 is an 8-site observational consortium study examining the predictors and mechanisms related to transition to psychosis in the CHR population. Participants from NAPLS-2 were patients aged 12-35 who met criteria for an ultra-high (UHR) risk syndrome for psychosis as determined by the Criteria of Prodromal States (COPS)[24] and as measured by the Structured Interview for Psychosis-risk Syndromes (SIPS).[25,26] The PRONIA consortium is an observational consortium study across 7 sites located in 5 European countries aiming to implement personalized prognostic tools for the development of affective and non-affective psychoses. Participants from PRONIA were patients aged 15-40 who experienced (a) clinical high-risk syndromes for psychosis meeting UHR criteria and/or basic symptoms criteria, or (b) recent-onset depression (ROD). In PRONIA, CHR states were defined by 9 items of the Schizophrenia Proneness Instrument (SPI-A) which constitute the basic symptoms pattern termed cognitive disturbances (COGDIS),[27,28] or operationalized as a UHR syndrome using a modified version of the SIPS.[22,25,26] Individuals with ROD 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).[29]

The aim of our external validation strategy was to measure the performance of the NAPLS-2 risk calculator in predicting psychosis transition (PT)[30] in the PRONIA sample and vice versa. Disease transition was established when at least one of the 5 positive symptom items in the Structured Interview for Psychosis-Risk Syndromes[25] reached psychotic intensity daily for at least 7 days. Group-level differences in sociodemographic, clinical, and functional variables were compared between transition (PT) and non-transition (NT) patients in the NAPLS-2 and PRONIA cohorts (**Table 1**).

Risk Calculator Assessments

The original NAPLS2 risk calculator was developed with eight variables previously shown to be associated with PT. Of these variables, six were also assessed in the PRONIA study: age; positive symptom severity on the individual SIPS items measuring unusual thought content and suspiciousness; score on the Brief Assessment of Cognition in Schizophrenia (BACS) symbol coding test;[31] score on the Hopkins Verbal Learning Test-Revised (HVLT-R);[32] decline in social functioning over the past year as measured by the Global Functioning Social Scale (GFS);[33] and family history of psychotic disorders in a first-degree relative. The two variables that were omitted from the original risk calculator model, stressful life events as measured by the Research Interview Life Events Scale[34] and childhood traumas as measured by the Childhood Trauma and Abuse Scale,[35] did not have comparable measures in PRONIA and were also not significant at either the univariate or the multivariate level in the original risk calculator[1] and thus were excluded from the models trained and tested in this study.

HARMONY validation framework

A framework for reciprocal validation between the NAPLS-2 and PRONIA studies was made possible through 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 models predictive of psychosis development within and across independent datasets at the international scale. All analyses were performed using the

Virtual Pooling and Analysis of Research Data (ViPAR) portal.[36] This secure, web-based platform utilizes a centralized cloud server to securely and temporarily retrieve anonymized data from remote servers on demand. Once analyses are complete, output from the analyses can be accessed by the user and the data are removed from server random-access memory. The use of the ViPAR portal was approved by the ethics committees of the 15 participating study sites in the NAPLS-2 and PRONIA consortia.

External validation and evaluation of risk calculators

The original risk calculator described in Cannon et al. (2016) employed Cox-PH to determine the individual likelihood of PT in the NAPLS-2 sample. Prior to applying this model to the PRONIA data, we imputed missing values (26 out of 2004) using a standardized Euclidean distance-based nearest-neighbour approach.[22] Then, the regression coefficients of the NAPLS-2 risk calculator were applied to the full PRONIA sample (CHR+, n=334 total), the CHR-only sample (n =167) and the UHR-only sample (n=126). Patients were labelled with a future PT at a predicted risk of 0.2, as described in the original publication. Next, we evaluated the effect of adjusting the PRONIA data for consortium-level differences, by mean-centering each PRONIA sample to the NAPLS-2 data prior to risk estimation. Finally, we removed consortium-level differences in the entire PRONIA data using the PRONIA-UHR sample as reference and recomputed risk estimates and prognostic group assignments in each PRONIA sample. The performance of the NAPLS-2 risk calculator in these three external validation iterations was measured in terms of sensitivity, specificity, balanced accuracy (BAC), positive and negative predictive value (PPV, NPV), prognostic summary index (PSI),[37] positive and negative likelihood ratios (LR+, LR-), and area-under-the curve (AUC; see **Table 2** and **Figure 1**).

Then, we integrated our open-source machine learning software NeuroMiner (version 1.05; https://github.com/neurominer-git) into ViPAR and performed a reciprocal external validation of the NAPLS-2-based clinical-neurocognitive risk signature. The rationale of this analysis was to evaluate the impact of different algorithms (univariate logistic regression [LR], Cox-PH, linear and non-linear Support Vector Machines [SVM]) and increasingly narrow risk definitions (CHR+, CHR, UHR) on prognostic performance. For all of the following experiments, we employed a repeated nested cross-validation design that used 10-fold cross-validation with 10 permutations at the inner cycle (CV_1) to determine optimal model parameters, and 10 repeats of reciprocal external validation at the outer cycle (CV_2) to estimate model generalizability from NAPLS-2 to PRONIA, and vice versa. Thus, in contrast to the single-model approach used by Cannon et al (2016),[1] we produced 10 CV_2-level prognostic ensembles for each consortium, each composed of 100 CV_1 models, which were applied to the respectively left-out data of the other consortium. All modelling and data pre-processing steps were entirely wrapped into this nested validation design, which involved scaling of the data to the range [0,1], nearest neighbour-based imputation of missing values using the Euclidean distance, as well as mean-centering and standardization of the CV_1 test and CV_2 validation data based on the parameters derived exclusively from the respective CV_1 training samples.

Due to the global risk estimate differences found between the unadjusted PRONIA data and the NAPLS-2 sample in the external validation analysis, we implemented a new, adaptive Cox-PH algorithm into NeuroMiner. Instead of using absolute risk cut-offs, this algorithm identifies a risk percentile that maximally separates PT from NT cases and applies this percentile to the test cases' distribution of risk estimates. Thus, the Cox-PH model learns to calibrate itself to risk samples with divergent absolute risks distributions but similar distribution shapes. Due to the highly unequal samples sizes of PT and non-transition (NT) cases in the two databases, we also tested whether combining this Cox-PH algorithm with an adaptive synthetic up-sampling method for the PT minority class would improve prediction performance (ADASYN algorithm[38] with pre-defined parameters: β=0.7, k_{SMOTE} = 5, and $k_{density}$=11). These multivariate algorithms were compared with machine learning strategies consisting of linear L_2-regularized, L_1-loss SVM[39] (regularization parameters $C = 2^{[-6 \xrightarrow[\in\mathbb{Z}]{} +4]}$) and non-linear SVM using Radial Basis Functions[40] ($C = 2^{[-6 \xrightarrow[\in\mathbb{Z}]{} +4]}$ and kernel parameters $\gamma = 2^{[-15 \xrightarrow[\in\mathbb{Z}]{} +2]}$). In addition, we compared these algorithms with simple univariate logistic regression. The models' risk estimates or decision scores were averaged across CV_2 repetitions and these out-of-training (OOT) predictions were evaluated using the performance metrics described above (**Table 3**). In addition, a supplementary leave-site-out cross-validation analysis across the pooled NAPLS-2 and PRONIA sites was conducted to measure the stability of PT prognostication within the multi-site context of the study.

We compared algorithms in terms of median BAC differences at the CV_2 level using Quade's non-parametric test[41,42] (**Figure 2**). The test was repeated for the classifier sets produced by the PRONIA-CHR+, CHR, and UHR samples, and hence the omnibus-level *P* values were corrected using the false-discovery rate (FDR),[43] followed by an FDR correction of pairwise classifier comparisons in each significant test. Statistical significance was determined at α=0.05. Finally, we evaluated how the increasingly narrow definition of psychosis risk across the three PRONIA discovery samples affected the algorithms' ability to predict PT in NAPLS-2 (**Figure 3, A**). We also compared the original NAPLS-2 model with the five ensemble-based NeuroMiner algorithms by evaluating their performance in the three PRONIA samples (**Figure 3, B**).

RESULTS

Group-level differences between samples

In the NAPLS-2 sample, 84 participants experienced a transition to psychosis during the follow-up period (transition rate: 14.1%). In the PRONIA sample, 26 (23 CHR and 3 ROD) participants developed psychosis during follow-up (transition rate: 12.2%). PRONIA and NAPLS-2 cohorts differed significantly on almost all examined sociodemographic, clinical, and neurocognitive variables, the including variables analyzed by the NAPLS-2 risk calculator (**Table 1, Suppl. Table 1**). Specifically, the PRONIA cohort was more than 5 years older, had more educational years, a higher percentage of female patients in the PT group, and a lower percentage of non-white participants. The PRONIA patients scored significantly lower on the SIPS-P1P2 summary item. In the BACS symbol coding and HVLT tests, the PRONIA PT cases scored between the NAPLS-2 non-transition and transition patients.

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

Though based on AUC measures the NAPLS2 model replicated in the PRONIA sample both overall and at the individual center level (see Supplement), the risk estimates based on the .2 cut-off in predicted risk produced by the original NAPLS-2 in the unadjusted PRONIA-CHR+, CHR, and UHR samples did not perform above chance levels due to highly unbalanced relation between sensitivity and specificity (BAC=49.4%-50.9%, sensitivity=3.8%-4.5%, specificity=94.3%-98.1%, **Table 2** and **Figure 1, A**). The removal of mean variable offsets between each PRONIA sample and the NAPLS-2 cohort, significantly increased performance across all PRONIA samples, with a broader risk definition being associated with higher prognostic precision (CHR+: BAC=69.6%, sensitivity=76.9%, specificity=62.3%; CHR: BAC=69.6%, sensitivity=78.3%, specificity=61.0%; UHR: BAC=63.6%, sensitivity=68.2%, specificity=59.0%; **Table 2** and **Figure 1, B**). When the PRONIA-UHR group served as reference sample for offset removal, the specificity of the NAPLS-2 model increased at the cost of sensitivity both in the

CHR+ sample (BAC=70.4%, sensitivity=57.7%, specificity=83.1%) and the CHR cohort (BAC=65.2%, sensitivity=68.1%, specificity=66.7%; **Table 2** and **Figure 1, C**).

Reciprocal external validation analyses

The reciprocal model discovery and validation of five different algorithms across the NAPLS-2 and PRONIA cohorts replicated the gains in prognostic precision when more broadly defined risk cohorts were included in the analysis (**Table 3**). This effect was particularly apparent in the NAPLS-2 UHR sample (**Figure 3, A**): When the five different PRONIA-derived models where derived from the PRONIA-CHR+ sample, the average performance measured BAC=67.0%, sensitivity=67.6%, specificity=66.4%. In contrast, when algorithms were developed using the PRONIA-UHR group, their average performance in the NAPLS-2 cohort was BAC=58.1%, sensitivity=60.7%, specificity=55.5%. This increase could be observed across all tested algorithms, except for the linear SVM, whose BAC ranged between 64.7% (CHR+) and 66.6% (CHR), and which performed best in the NAPLS-2 UHR (BAC=65.1%). Statistical classifier comparisons conducted across the full reciprocal external validation analysis (**Figure 2, A & B**) confirmed this finding by showing that the linear SVM outperformed all other algorithms when the PRONIA-training and validation sample was confined to the UHR or CHR subgroups. In contrast, our adaptive Cox-PH algorithm (with or without ADASYN) achieved superior prediction performance in the CHR+ sample, which included both the PRONIA-CHR and ROD patients. The supplementary leave-site-out analysis conducted across the UHR, CHR and CHR+ risk levels showed that the site-level variability of prognostic performance decreased from UHR to CHR+, as measured by the difference between the full sample and the mean (SD) performance metrics computed across sites (see **Suppl. Table 5**).

A qualitative comparison of predictive feature relevance between the Linear SVM and adaptive Cox-PH, as measured using the cross-validation ratio,[22] revealed commonalities and difference between algorithms, varying with the three risk enrichment levels (**Figure 2, C**): When the discovery sample was limited to UHR patients, feature profiles were similar between algorithms, except for the HVLT, which showed a high negative association with PT prediction in the linear SVM. Broadening the risk cohort to all PRONIA patients, increased the predictive value of both age and the BACS digit symbol coding test, while the HVLT importance was reduced in the linear SVM. Finally, differences between the five algorithms and the original NAPLS-2 risk calculator emerged when NAPLS-2 served for model discovery (**Figure 3, C**): In the PRONIA data, the highest prognostic performance was measured for the adaptive Cox-PH algorithm with (mean[SD] BAC=68.3%[4.4%]) or without ADASYN (68.6%[2.1%]), which was slightly increased compared to the original NAPLS-2 model combined with a priori mean-centering of each PRONIA sample to the NAPLS-2 data (67.6%[3.5%]).

INTERPETATION

The external validation of prognostic models has been identified as the major bottleneck and translational step for their implementation in clinical real-world settings.[44] In this regard, a reciprocal external validation environment that facilitates a standardized framework for model exchange and comparison between independent single- and multi-site projects may have the potential to mitigate multiple sources of bias caused by the idiosyncrasies of study purposes, patient recruitment strategies and predictive model designs.[45] To our knowledge, the HARMONY consortium, which authored the present study, is the first initiative to set up such a secure international forum for collaborative model discovery and validation in the field of psychosis prediction research and data analytics.

The HARMONY framework allowed us to test the generalizability and prognostic value of the NAPLS-2 psychosis risk signature[1] both at the international scale and across diverse risk samples provided by the European PRONIA project.[22] We encountered significant consortium-level differences between the NAPLS-2 and PRONIA cohorts, which were likely fuelled by systematic variation in participant referral, ascertainment, enrolment and retainment, resulting in two study cohorts that differed on sociodemographic (age, ethnicity) and clinical parameters (severity of attenuated psychotic symptoms). A key observation of the current work was that these differences considerably impaired the generalizability of the original risk calculator but could be overcome by mean-centering each predictor of the PRONIA sample to the respective variable of the NAPLS-2 cohort. This simple calibration procedure enabled the original model to predict PT in the PRONIA-UHR cohort with a 5.4% lower BAC compared to the NAPLS-2 discovery population (BAC=69%). Based on the observation of project-level differences between NAPLS-2 and PRONIA, we developed a new Cox-PH algorithm which learns an optimal *relative* risk cut-off compared to the fixed, absolute risk threshold (p=0.2) of the original risk calculator.[1] Based on this algorithm, we were able to show that the generalizability gap can be reduced to 2.8%. This finding is highly relevant for the successful clinical implementation of the risk calculator, as target populations will inevitably differ in their levels of absolute risk for the development of PT, as encountered in the NAPLS-2 sample (optimal probability cut-off for PT assignment: p=0.267) and the PRONIA-UHR cohort (p=0.184).

Furthermore, we observed that univariate logistic regression significantly trailed behind all multivariate methods, suggesting that generalizable prognostic precision can only be achieved when the relationships between psychosis risk variables are algorithmically modelled into a risk *pattern*. Among the pattern recognition algorithms, the linear SVM algorithm slightly but significantly outperformed Cox-PH in target populations encompassing only patients with CHR syndromes. The analysis of feature relevance indicated that the former algorithm learned a more complex clinical-neurocognitive pattern, while the Cox-PH model put less emphasis on neurocognitive information. As SVMs intrinsically learn decision boundaries between opposite classes by maximizing the distance between most similar cases, the higher complexity of the SVM pattern may have increased prognostic precision within a more homogenous risk population.[46] In contrast, the Cox-PH algorithm attained higher prognostic precision when the target population comprised both patients with CHR syndromes as well as patients with ROD. In this broader risk sample, the Cox-PH model identified attenuated psychotic symptoms, social functioning decline, and a positive family history as most reliable core predictors of subsequent PT. However, in summary, we did not find major differences between multivariate survival algorithms and SVM-based machine learning approaches. This finding was expected because the analysed risk space was spanned by just 6 variables, previously picked among many other potential sociodemographic, clinical, behavioural and neurocognitive predictors through a decade-long literature-driven and expert-based feature selection process.[1,47,48]

Strikingly, the present study revealed that a transdiagnostic risk designation which enriches the core group of CHR individuals with young patients experiencing their first episode of major depression, leads to risk calculators with superior, more generalizable, and stable prognostic performance. This finding is in line with previous studies,[19,20] suggesting that elevated risk for psychosis is not confined to CHR states but extends to other ‘neighbouring’ or comorbid conditions which typically co-occur with these syndromes. Of note, we observed that the increased prognostic performance of CHR+-based models was not driven by higher specificity due to the inclusion of ROD patients properly labelled as NT. Instead, we found that model performance increased particularly in the NAPLS-UHR sample, and, in addition, was more stable across study sites compared to UHR-only derived algorithms. This finding may point to an increased representational power of the CHR+-trained models due to the extension of the risk spectrum towards lower-risk individuals with early-onset affective disorders, who may share bio-behavioural features of psychosis.[49–52] Future studies should investigate whether this enrichment effect is specific to affective disorders or can also be traced in other conditions which evolve in adolescence and young adulthood.[53]

In summary, we found that the clinical-neurocognitive risk calculator previously described by the NAPLS-2 study provides an internationally scalable tool for individualised psychosis risk ascertainment in youth with diverse psychosis risk syndromes. The underlying risk signature may extend beyond the prevailing CHR-focused concepts of the current early recognition literature. This may have important ramifications for the design of future prognostic studies and the development of transdiagnostic precision medicine tools in the youth mental health field. The HARMONY initiative provided a useful resource for integrated model discovery and validation at the highest level of validity achievable with retrospective data. Future work should assess the prospective generalizability of the NAPLS-2 derived risk signature, its clinical utility for treatment stratification and the potential additive value of biological information.

References

1 Cannon TD, Yu C, Addington J, *et al.* An individualized risk calculator for research in prodromal psychosis. *American Journal of Psychiatry* 2016; **173**: 980–988.

2 Yung AR, McGorry PD. The initial prodrome in psychosis: descriptive and qualitative aspects. *Australian and New Zealand Journal of Psychiatry* 1996; **30**: 587–599.

3 Yung AR, McGorry PD. The prodromal phase of first-episode psychosis: past and current conceptualizations. *Schizophrenia Bulletin* 1996; **22**: 353–370.

4 McGorry PD. Early intervention in psychosis: obvious, effective, overdue. *The Journal of nervous and mental disease* 2015; **203**: 310–318.

5 Fusar-Poli P, Schultze-Lutter F, Cappucciati M, *et al.* The dark side of the moon: meta-analytical impact of recruitment strategies on risk enrichment in the clinical high risk state for psychosis. *Schizophrenia Bulletin* 2015; **42**: 732–743.

6 Fusar-Poli P, Pablo GS de, Correll C, *et al.* Prevention of Psychosis: Advances in Detection, Prognosis and Intervention. *JAMA Psychiatry* 2020; **77**: 1–11.

7 Nelson B. Attenuated psychosis syndrome: don't jump the gun. *Psychopathology* 2014; **47**: 292–296.

8 Koutsouleris N, Davatzikos C, Bottlender R, *et al.* Early recognition and disease prediction in the at-risk mental states for psychosis using neurocognitive pattern classification. *Schizophr Bull* 2012; **38**: 1200–1215.

9 Koutsouleris N, Riecher-Rössler A, Meisenzahl EM, *et al.* Detecting the psychosis prodrome across high-risk populations using neuroanatomical biomarkers. *Schizophr Bull* 2015; **41**: 471–482.

10 Perkins DO, Jeffries CD, Addington J, *et al.* Towards a Psychosis Risk Blood Diagnostic for Persons Experiencing High-Risk Symptoms: Preliminary Results From the NAPLS Project. *Schizophrenia Bulletin* 2014; published online Aug. DOI:10.1093/schbul/sbu099.

11 Zhang T, Xu L, Tang Y, *et al.* Prediction of psychosis in prodrome: development and validation of a simple, personalized risk calculator. *Psychological medicine* 2018; : 1–9.

12 Perkins DO, Olde Loohuis L, Barbee J, *et al.* Polygenic Risk Score Contribution to Psychosis Prediction in a Target Population of Persons at Clinical High Risk. *American Journal of Psychiatry* 2019; **177**: 155–163.

13 Addington J, Piskulic D, Liu L, *et al.* Comorbid diagnoses for youth at clinical high risk of psychosis. *Schizophrenia Research* 2017; **190**: 90–95.

14 Sanfelici R, Dwyer D, Antonucci LA, Koutsouleris N. Individualized diagnostic and prognostic models for patients with psychosis risk syndromes: A meta-analytic view on the state-of-the-art. *Biological Psychiatry* 2020; **in press**.

15 Carrión RE, Cornblatt BA, Burton CZ, *et al.* Personalized prediction of psychosis: external validation of the NAPLS-2 psychosis risk calculator with the EDIPPP project. *American Journal of Psychiatry* 2016; **173**: 989–996.

16 Zhang T, Li H, Tang Y, *et al.* 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* 2018; **175**: 906–908.

17 Klosterkötter J, Hellmich M, Steinmeyer EM, Schultze-Lutter F. Diagnosing schizophrenia in the initial prodromal phase. *Archives of General Psychiatry* 2001; **58**: 158–164.

18 Schultze-Lutter F, Ruhrmann S, Berning J, Maier W, Klosterkötter J. Basic symptoms and ultrahigh risk criteria: symptom development in the initial prodromal state. *Schizophrenia Bulletin* 2010; **36**: 182–191.

19 Fusar-Poli P, Rutigliano G, Stahl D, *et al.* Development and validation of a clinically based risk calculator for the transdiagnostic prediction of psychosis. *JAMA Psychiatry* 2017; **74**: 493–500.

20 Lee TY, Lee J, Kim M, Choe E, Kwon JS. Can we predict psychosis outside the clinical high-risk state? A systematic review of non-psychotic risk syndromes for mental disorders. *Schizophrenia Bulletin* 2018; **44**: 276–285.

21 Fusar-Poli P, Stringer D, Durieux AM, *et al.* Clinical-learning versus machine-learning for transdiagnostic prediction of psychosis onset in individuals at-risk. *Translational Psychiatry* 2019; **9**: 1–11.

22 Koutsouleris N, Kambeitz-Ilankovic L, Ruhrmann S, *et al.* Prediction Models of Functional Outcomes for Individuals in the Clinical High-Risk State for Psychosis or With Recent-Onset Depression: A Multimodal, Multisite Machine Learning Analysis. *JAMA Psychiatry* 2018; **75**: 1156–1172.

23 Addington J, Cadenhead KS, Cornblatt BA, *et al.* North American Prodrome Longitudinal Study (NAPLS 2): overview and recruitment. *Schizophrenia Research* 2012; **142**: 77–82.

24 McGlashan T, Walsh B, Woods S. The psychosis-risk syndrome: handbook for diagnosis and follow-up. Oxford University Press, 2010.

25 McGlashan TH, Miller TJ, Woods SW, Hoffman RE, Davidson L. Instrument for the assessment of prodromal symptoms and states. In: Miller T, Madnick SA, McGlashan TH, Libiger J, Johannessen JO, eds. Early intervention in psychiatric disorders. Dordrecht: Kluwer Academic, 2001: 135–149.

26 Miller TJ, McGlashan TH, Rosen JL, *et al.* Prodromal assessment with the structured interview for prodromal syndromes and the scale of prodromal symptoms: predictive validity, interrater reliability, and training to reliability. *Schizophrenia Bulletin* 2003; **29**: 703–715.

27 Schultze-Lutter F, Addington J, Ruhrmann S, Klosterkötter J. Schizophrenia Proneness Instrument, Adult Version (SPI-A). 2007.

28 Schultze-Lutter F, Koch E. Schizophrenia Proneness Instrument, Children and Youth Version (SPI-CY). 2010.

29 First MB. Structured clinical interview for the DSM (SCID). *The encyclopedia of clinical psychology* 2014; : 1–6.

30 Yung AR, Nelson B, Thompson A, Wood SJ. The psychosis threshold in Ultra High Risk (prodromal) research: is it valid? *Schizophrenia Research* 2010; **120**: 1–6.

31 Keefe RS, Goldberg TE, Harvey PD, Gold JM, Poe MP, Coughenour L. The Brief Assessment of Cognition in Schizophrenia: reliability, sensitivity, and comparison with a standard neurocognitive battery. *Schizophrenia research* 2004; **68**: 283–297.

32 Benedict RH, Schretlen D, Groninger L, Brandt J. Hopkins Verbal Learning Test–Revised: Normative data and analysis of inter-form and test-retest reliability. *The Clinical Neuropsychologist* 1998; **12**: 43–55.

33 Cornblatt BA, Auther AM, Niendam T, *et al.* Preliminary findings for two new measures of social and role functioning in the prodromal phase of schizophrenia. *Schizophrenia Bulletin* 2007; **33**: 688–702.

34 Dohrenwend BS, Askenasy AR, Krasnoff L, Dohrenwend BP. Exemplification of a method for scaling life events: The PERI Life Events Scale. *Journal of health and social behavior* 1978; : 205–229.

35 Janssen I, Krabbendam L, Bak M, *et al.* Childhood abuse as a risk factor for psychotic experiences. *Acta Psychiatrica Scandinavica* 2004; **109**: 38–45.

36 Carter KW, Francis RW, Carter KW, *et al.* ViPAR: a software platform for the Virtual Pooling and Analysis of Research Data. *International journal of epidemiology* 2016; **45**: 408–416.

37 Linn S, Grunau PD. New patient-oriented summary measure of net total gain in certainty for dichotomous diagnostic tests. *Epidemiologic Perspectives & Innovations* 2006; **3**: 11.

38 Haibo H, Yang B, Edwardo GA, Shutao L. Adaptive Synthetic Sampling Approach for Imbalanced Learning. In: IEEE International Joint Conference on Neural Networks, IJCNN. 2016: 1322–1328.

39 Fan R, Chang K, Hsieh C, Wang X, Lin C. LIBLINEAR: A Library for Large Linear Classification. 2008.

40 Chang C-C, Lin C-J. LIBSVM: a library for support vector machines. 2001.

41 Quade D. Using Weighted Rankings in the Analysis of Complete Blocks with Additive Block Effects. *Journal of the American Statistical Association* 1979; **74**: 680–683.

42 Stkapor K. Evaluating and comparing classifiers: Review, some recommendations and limitations. In: International Conference on Computer Recognition Systems. Springer, 2017: 12–21.

43 Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the royal statistical society Series B (Methodological)* 1995; : 289–300.

44 Steyerberg EW, Moons KGM, Windt DA van der, *et al.* Prognosis Research Strategy (PROGRESS) 3: prognostic model research. *PLoS Medicine* 2013; **10**: e1001381.

45 Fusar-Poli P, Hijazi Z, Stahl D, Steyerberg EW. The Science of Prognosis in Psychiatry: A Review. *JAMA psychiatry* 2018; **75**: 1289–1297.

46 Vapnik V. The nature of statistical learning theory, 2nd edn. New York: Springer-Verlag, 2000.

47 Cannon TD, Cadenhead K, Cornblatt B, *et al.* Prediction of psychosis in youth at high clinical risk: a multisite longitudinal study in North America. *Archives of General Psychiatry* 2008; **65**: 28–37.

48 Fusar-Poli P, Sullivan SA, Shah JL, Uhlhaas PJ. 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* 2019; **10**: 774.

49 Koutsouleris N, Meisenzahl EM, Borgwardt S, *et al.* Individualized differential diagnosis of schizophrenia and mood disorders using neuroanatomical biomarkers. *Brain* 2015; **138**: 2059–2073.

50 Power RA, Tansey KE, ButtenschÃ¸n HNã¸, *et al.* Genome-wide Association for Major Depression Through Age at Onset Stratification: Major Depressive Disorder Working Group of the Psychiatric Genomics Consortium. *Biological Psychiatry* 2016; published online May. DOI:10.1016/j.biopsych.2016.05.010.

51 Byrne EM, Zhu Z, Qi T, *et al.* Conditional GWAS analysis to identify disorder-specific SNPs for psychiatric disorders. *Molecular Psychiatry* 2020; : 1–12.

52 Zhu Y, Womer FY, Leng H, *et al.* The Relationship Between Cognitive Dysfunction and Symptom Dimensions Across Schizophrenia, Bipolar Disorder, and Major Depressive Disorder. *Frontiers in psychiatry* 2019; **10**: 253.

53 Caspi A, Houts RM, Ambler A, *et al.* Longitudinal Assessment of Mental Health Disorders and Comorbidities Across 4 Decades Among Participants in the Dunedin Birth Cohort Study. *JAMA network open* 2020; **3**: e203221.

Table 1. Sociodemographic, clinical, and functional differences between non-transition and transition cases in the NAPLS-2 and PRONIA samples.

	NAPLS-2		PRONIA			
Variable	**NT**	**PT**	**NT**	**PT**	**Wald χ^2(df)**	***P***
Age, mean (SD)	18.6 (4.4)	18.1 (3.6)	24.7 (5.8)	23.5 (5.9)	$\chi^2(3) = 352.6$	**0.008**
Sex, % Female	43%	38.1%	49.1%	61.5%	$\chi^2(3) = 8.8$	**0.033**
Race, % non-white	41.8%	44%	13%	7.7%	$\chi^2(3) = 76.9$	**0.017**
Years of education, mean (SD)	11.3 (2.9)	11.0 (2.5)	14.3 (2.9)	13.3 (2.5)	$\chi^2(3) = 228.9$	**0.008**
Family history, % no history	84.4%	81%	90.6%	80.8%	$\chi^2(3) = 8.7$	**0.042**
Baseline positive symptoms (p1+p2), mean (SD)	5.9 (2.2)	7.1 (2.3)	2.6 (2.6)	5.5 (2.8)	$\chi^2(3) = 466.5$	**0.008**
HVLT, mean (SD)	25.8 (5.1)	24.2 (5.5)	28.5 (2.7)	26.5 (3.0)	$\chi^2(3) = 93.8$	**0.008**
BACS, mean (SD)	57.4 (13.2)	53.2 (11.6)	61.1 (11.8)	55.0 (13.0)	$\chi^2(3) = 33.1$	**0.025**
Change in GFS, mean (SD)	0.70 (1.0)	0.99 (1.2)	0.75 (0.9)	0.96 (1.1)	$\chi^2(3) = 7.6$	0.054

Abbreviations: *NT* non-transition cases, *PT* transition cases, *P* P value of comparison

Table 2. Results of the external model validation of the original NAPLS-2 risk calculator in the three PRONIA samples (CHR+, CHR, and UHR) with and without prior centering of the predictor variables to the means of the respective NAPLS variables. Offsets corrections were computed between the NAPLS-2 cohort and the PRONIA UHR sample.

PRONIA samples	TP	TN	FP	FN	Sens [%]	Spec [%]	BAC [%]	PPV [%]	NPV [%]	PSI [%]	LR+	LR-	AUC
PRONIA data not mean-centered to NAPLS-2													
CHR+	1	302	6	25	3.8	98.1	50.9	14.3	92.4	6.6	1.97	0.98	0.51
CHR	1	135	6	22	4.3	95.7	50.0	14.3	86.0	0.3	1.02	1.00	0.50
UHR	1	99	6	21	4.5	94.3	49.4	14.3	82.5	-3.2	0.80	1.01	0.49
PRONIA data mean-centered to NAPLS-2 using respective PRONIA sample as reference													
CHR+	20	192	116	6	76.9	62.3	69.6	14.7	97.0	11.7	2.04	0.37	0.70
CHR	18	86	55	5	78.3	61.0	69.6	24.7	94.5	19.2	2.01	0.36	0.70
UHR	15	62	43	7	68.2	59.0	63.6	25.9	89.9	15.7	1.66	0.54	0.64
PRONIA data mean-centered to NAPLS-2 using the PRONIA UHR sample as reference													
CHR+	15	256	52	11	57.7	83.1	70.4	22.4	95.9	18.3	3.42	0.51	0.70
CHR	15	96	45	8	65.2	68.1	66.7	25.0	92.3	17.3	2.04	0.51	0.67
UHR	15	62	43	7	68.2	59.0	63.6	25.9	89.9	15.7	1.66	0.54	0.64

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, *PPV* Positive Predictive Value, *NPV* Negative Predictive Value, *PSI* Prognostic Summary Index, *LR+* Positive Likelihood Ratio, *LR-* Negative Likelihood Ratio, *AUC* Area-under-the Curve.

Table 3. Algorithm comparisons in the reciprocal external validation (REV) experiments with performance measures computed separately for the risk calculators trained on the NAPLS-2 UHR cohort or on the three PRONIA samples (CHR+, CHR, and UHR).

	Confusion matrix				Performance measures								
Predictors	**TP**	**TN**	**FP**	**FN**	**Sens [%]**	**Spec [%]**	**BAC [%]**	**PPV [%]**	**NPV [%]**	**PSI [%]**	**LR+**	**LR-**	**AUC**
CHR+ enrichment level													
LR [Full Sample]	94	328	491	16	85.5	40.0	62.8	16.1	95.3	11.4	1.43	0.36	0.70
LR [NAPLS-2]	68	270	241	16	81.0	52.8	66.9	22.0	94.4	16.4	1.72	0.36	0.74
LR [PRONIA]	26	58	250	0	100	18.8	59.4	9.4	100	9.4	1.23	0.00	0.79
Cox-PH [Full Sample]	72	579	240	38	65.5	70.7	68.1	23.1	93.8	16.9	2.23	0.49	0.73
Cox-PH [NAPLS-2]	52	385	126	32	61.9	75.3	68.6	29.2	92.3	21.5	2.51	0.51	0.74
Cox-PH [PRONIA]	20	194	114	6	76.9	63.0	70.0	14.9	97.0	11.9	2.08	0.37	0.79
Cox-PH (+A) [Full Sample]	78	560	259	32	70.9	68.4	69.6	23.1	94.6	17.7	2.24	0.43	0.74
Cox-PH (+A) [NAPLS-2]	60	333	178	24	71.4	65.2	68.3	25.2	93.3	18.5	2.05	0.44	0.72
Cox-PH (+A) [PRONIA]	18	227	81	8	69.2	73.7	71.5	18.2	96.6	14.8	2.63	0.42	0.78
Linear SVM [Full Sample]	82	478	341	28	74.5	58.4	66.5	19.4	94.5	13.9	1.79	0.44	0.73
Linear SVM [NAPLS-2]	63	278	233	21	75.0	54.4	64.7	21.3	93.0	14.3	1.64	0.46	0.72
Linear SVM [PRONIA]	19	200	108	7	73.1	64.9	69.0	15.0	96.6	11.6	2.08	0.41	0.78
RBF-SVM [Full Sample]	58	680	139	52	52.7	83.0	67.9	29.4	92.9	22.3	3.11	0.57	0.74
RBF-SVM [NAPLS-2]	41	431	80	43	48.8	84.3	66.6	33.9	90.9	24.8	3.12	0.61	0.73
RBF-SVM [PRONIA]	17	249	59	9	65.4	80.8	73.1	22.4	96.5	18.9	3.41	0.43	0.78
CHR enrichment level													
LR [Full Sample]	106	71	584	1	99.1	10.8	55.0	15.4	98.6	14.0	1.11	0.09	0.73
LR [NAPLS-2]	83	34	477	1	98.8	6.7	52.7	14.8	97.1	12.0	1.06	0.18	0.73
LR [PRONIA]	23	37	107	0	100.0	25.7	62.8	17.7	100	17.7	1.35	0.00	0.74
Cox-PH [Full Sample]	81	371	284	26	75.7	56.6	66.2	22.2	93.5	15.6	1.75	0.43	0.71
Cox-PH [NAPLS-2]	64	277	234	20	76.2	54.2	65.2	21.5	93.3	14.7	1.66	0.44	0.71
Cox-PH [PRONIA]	17	94	50	6	73.9	65.3	69.6	25.4	94.0	19.4	2.13	0.40	0.75
Cox-PH (+A) [Full Sample]	79	385	270	28	73.8	58.8	66.3	22.6	93.2	15.9	1.79	0.45	0.68
Cox-PH (+A) [NAPLS-2]	64	277	234	20	76.2	54.2	65.2	21.5	93.3	14.7	1.66	0.44	0.68
Cox-PH (+A) [PRONIA]	15	108	36	8	65.2	75.0	70.1	29.4	93.1	22.5	2.61	0.46	0.74
Linear SVM [Full Sample]	80	381	274	27	74.8	58.2	66.5	22.6	93.4	16.0	1.79	0.43	0.72
Linear SVM [NAPLS-2]	64	291	220	20	76.2	56.9	66.6	22.5	93.6	16.1	1.77	0.42	0.72
Linear SVM [PRONIA]	16	90	54	7	69.6	62.5	66.0	22.9	92.8	15.6	1.86	0.49	0.74
RBF-SVM [Full Sample]	56	495	160	51	52.3	75.6	64.0	25.9	90.7	16.6	2.14	0.63	0.70
RBF-SVM [NAPLS-2]	44	384	127	40	52.4	75.1	63.8	25.7	90.6	16.3	2.11	0.63	0.70
RBF-SVM [PRONIA]	12	111	33	11	52.2	77.1	64.6	26.7	91.0	17.7	2.28	0.62	0.73
UHR enrichment level													
LR [Full Sample]	102	31	585	3	97.1	5.0	51.1	14.8	91.2	6.0	1.02	0.57	0.69
LR [NAPLS-2]	84	3	508	0	100	0.6	50.3	14.2	100	14.2	1.01	0.00	0.70
LR [PRONIA]	18	28	77	3	85.7	26.7	56.2	18.9	90.3	9.3	1.17	0.54	0.69
Cox-PH [Full Sample]	51	496	120	54	48.6	80.5	64.5	29.8	90.2	20.0	2.49	0.64	0.68
Cox-PH [NAPLS-2]	37	427	84	47	44.0	83.6	63.8	30.6	90.1	20.7	2.68	0.67	0.68
Cox-PH [PRONIA]	14	69	36	7	66.7	65.7	66.2	28.0	90.8	18.8	1.94	0.51	0.69

Cox-PH (+A) [Full Sample]	83	265	351	22	79.0	43.0	61.0	19.1	92.3	11.5	1.39	0.49	0.65
Cox-PH (+A) [NAPLS-2]	72	187	324	12	85.7	36.6	61.2	18.2	94.0	12.2	1.35	0.39	0.66
Cox-PH (+A) [PRONIA]	11	78	27	10	52.4	74.3	63.3	28.9	88.6	17.6	2.04	0.64	0.68
Linear SVM [Full Sample]	75	366	250	30	71.4	59.4	65.4	23.1	92.4	15.5	1.76	0.48	0.69
Linear SVM [NAPLS-2]	60	300	211	24	71.4	58.7	65.1	22.1	92.6	14.7	1.73	0.49	0.70
Linear SVM [PRONIA]	15	66	39	6	71.4	62.9	67.1	27.8	91.7	19.4	1.92	0.45	0.69
RBF-SVM [Full Sample]	12	583	33	93	11.4	94.6	53.0	26.7	86.2	12.9	2.13	0.94	0.70
RBF-SVM [NAPLS-2]	2	500	11	82	2.4	97.8	50.1	15.4	85.9	1.3	1.11	1.00	0.71
RBF-SVM [PRONIA]	10	83	22	11	47.6	79.0	63.3	31.3	88.3	19.5	2.27	0.66	0.68

Abbreviations: *LR* Logistic regression, *Cox-PH* Cox Proportional Hazard model, *SVM* Support Vector Machine, *RBF-SVM* Support Vector Machine with Radial Basis Kernel, *+A* Cox-PH (+ADASYN), *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, *PPV* Positive Predictive Value, *NPV* Negative Predictive Value, *PSI* Prognostic Summary Index, *LR+* Positive Likelihood Ratio, *LR-* Negative Likelihood Ratio, *AUC* Area-under-the Curve.

Figure 1. NAPLS-2 risk calculator estimates for the 2-year transition risk of PT (red) versus NT cases (blue) in three different risk cohorts of PRONIA (CHR+: 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).

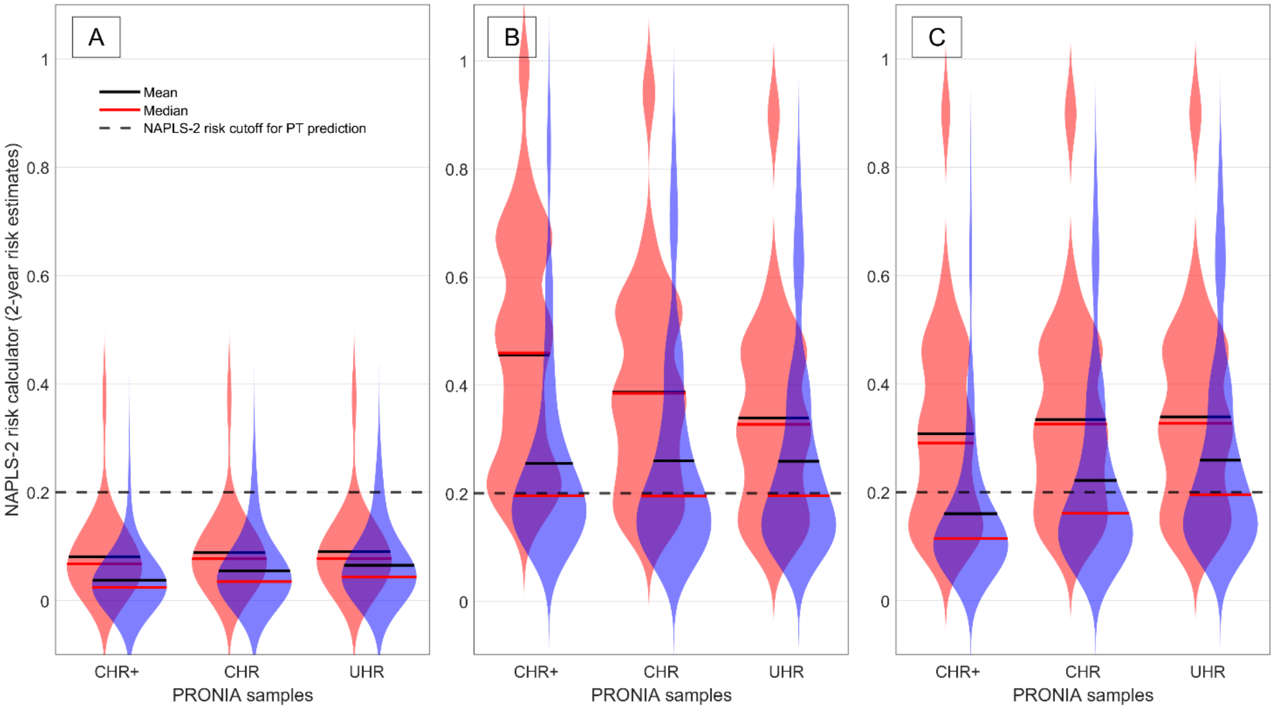


Figure 2. Analysis of classifier differences in the reciprocal external validation experiments. **A:** Balanced accuracy distributions of each classifier at the leave-project-out level described by the median, the 5% and 95% percentiles. Quade test omnibus analysis results were provided for each risk enrichment level (CHR+, CHR, UHR). **B:** Post hoc tests of pairwise BAC differences between risk calculators. **C:** Analysis of feature relevance for prediction of PT in the reciprocal external validation analysis (left: linear SVM, right: Cox-PH). Abbreviations: see **Table 3**.

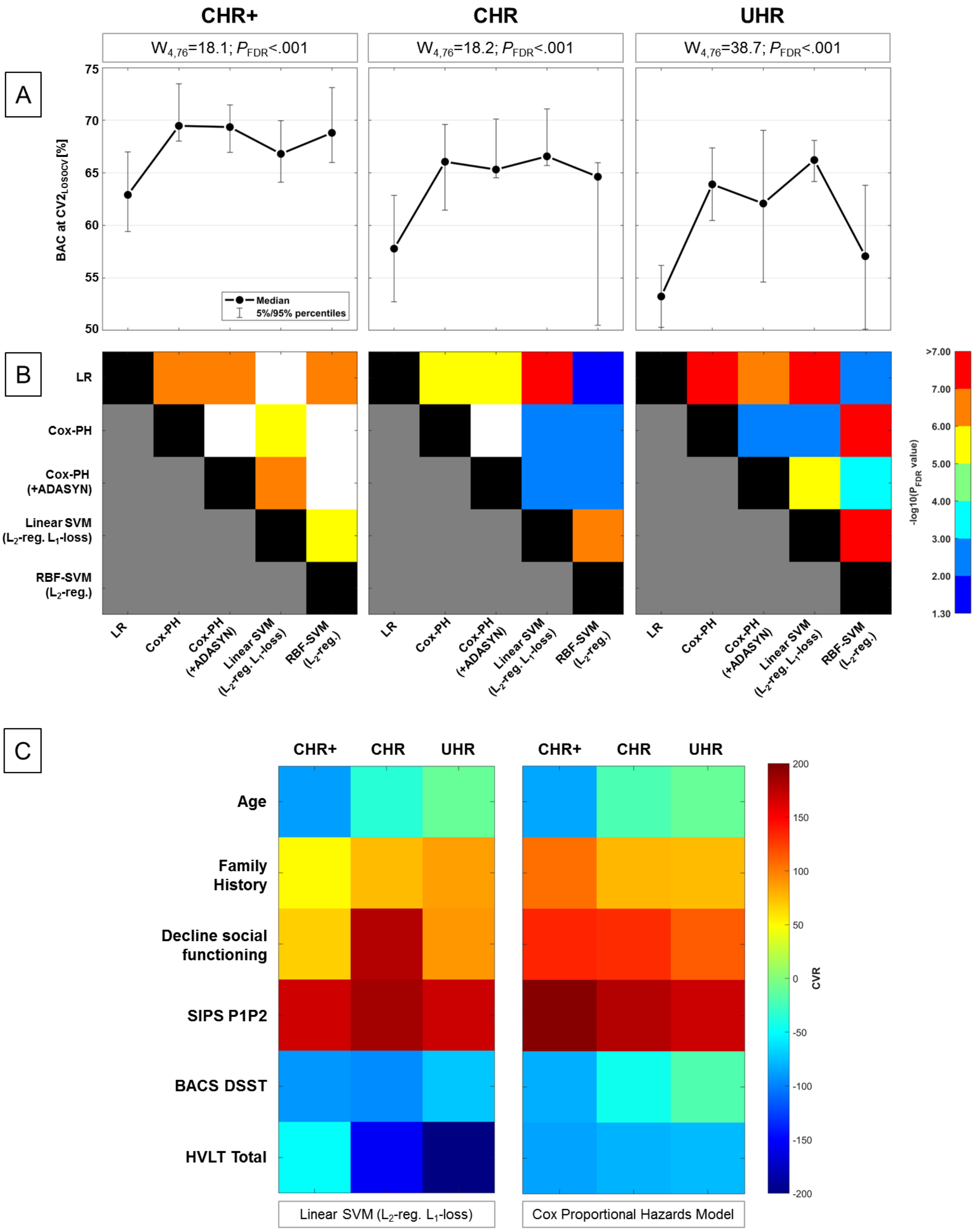
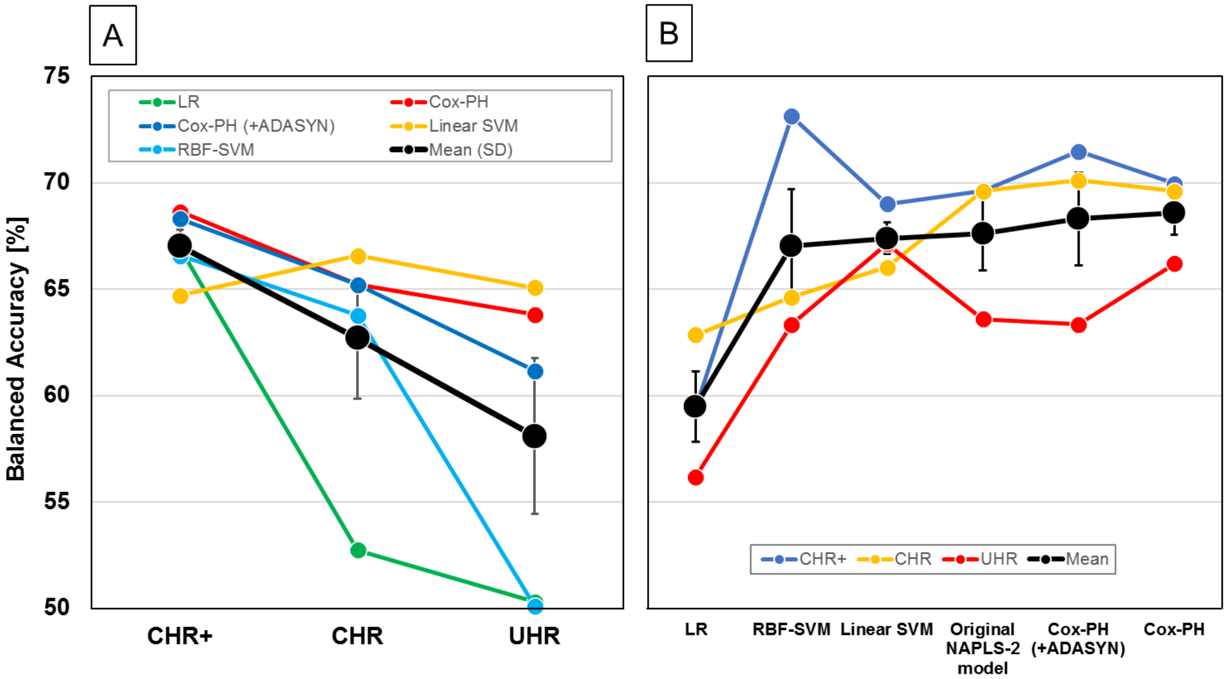


Figure 3. PRONIA risk enrichment effects on PT prediction in the NAPLS-2 cohort and algorithm effects on the prediction of PT in the PRONIA risk enrichment samples. **A:** Balanced accuracy of the 5 different prognostic algorithms in the NAPLS-2 cohort as a function of the PRONIA risk sample used to train these algorithms. **B:** Differences in balanced accuracy as a function of the type of algorithm applied to the three different PRONIA samples. Additionally, means and standard deviations are depicted for both A and B.



Supplementary Table 1. Descriptive analysis of means and standard deviations of the NAPLS-2 risk calculator variables shared between the two consortia.

	NAPLS-2			PRONIA-UHR			PRONIA-CHR			PRONIA-CHR+		
	N	**mean**	**SD**	**N**	**mean**	**SD**	**N**	**mean**	**SD**	**N**	**mean**	**SD**
Full sample												
Age [years]	596	18.51	4.27	126	23.11	4.97	167	23.79	5.37	334	24.71	5.76
BACS-DSST score	596	56.80	13.04	119	58.92	12.23	160	59.51	12.21	323	60.68	11.96
HVLT-Total score	596	25.61	5.15	121	27.89	3.04	162	27.94	3.07	322	28.32	2.77
GFS-Decline past year	596	0.74	1.04	126	0.74	0.92	167	0.71	0.92	334	0.79	0.95
Family history	596	0.16	0.37	126	0.20	0.40	167	0.17	0.37	334	0.10	0.30
SIPS-P1P2	596	2.61	1.57	126	2.13	1.87	167	1.72	1.84	334	0.91	1.56
Transition cases												
Age [years]	84	18.06	3.58	21	23.94	5.93	23	24.13	5.84	26	23.50	5.92
BACS-DSST score	84	53.20	11.60	18	55.50	13.78	20	54.70	13.55	23	55.00	12.97
HVLT-Total score	84	24.18	5.54	18	26.11	3.07	20	26.15	2.91	23	26.52	2.98
GFS-Decline past year	84	0.99	1.16	21	1.05	1.24	23	1.09	1.20	26	0.96	1.12
Family history	84	0.19	0.40	21	0.24	0.44	23	0.22	0.42	26	0.19	0.40
SIPS-P1P2	84	3.51	1.77	21	2.62	1.80	23	2.52	1.81	26	2.23	1.88
Months to transition	84	7.22	5.70	21	7.34	7.42	23	7.32	7.08	26	8.23	8.15
Non-transition cases												
Age [years]	512	18.59	4.37	105	22.95	4.77	144	23.74	5.31	308	24.72	5.77
BACS-DSST score	512	57.39	13.17	101	59.53	11.91	140	60.19	11.90	300	61.12	11.79
HVLT-Total score	512	25.84	5.05	103	28.20	2.95	142	28.20	3.02	299	28.45	2.71
GFS-Decline past year	512	0.70	1.01	105	0.68	0.84	144	0.65	0.86	308	0.75	0.93
Family history	512	0.16	0.36	105	0.19	0.39	144	0.16	0.37	308	0.09	0.29
SIPS-P1P2	512	2.46	1.49	105	2.03	1.88	144	1.60	1.82	308	0.80	1.48

Supplementary Table 2. Leave-site-out cross-validation analysis in the CHR+ sample comparing the five different PT prediction algorithms. The out-of-training performance of the given algorithm was broken down per site. In addition, the respective means and standards deviation of the given algorithm's performance measures were computed across sites. To avoid a biased estimate of the average leave-site-out performances, the PRONIA Udine site was excluded from this analysis because of no reported transition cases.

Predictors	**TP**	**TN**	**FP**	**FN**	**Sens**	**Spec**	**BAC**	**PPV**	**NPV**	**PSI**	**LR+**	**LR-**	**AUC**
LR [Full Sample]	*107*	*155*	*634*	*3*	*97.3*	*19.6*	*58.5*	*14.4*	*98.1*	*12.5*	*1.21*	*0.14*	*0.74*
LR [Basel]	2	10	22	0	100.0	31.3	65.6	8.3	100.0	8.3	1.45	0.00	0.86
LR [Birmingham]	2	3	31	0	100.0	8.8	54.4	6.1	100.0	6.1	1.10	0.00	0.51
LR [Calgary]	18	25	89	1	94.7	21.9	58.3	16.8	96.2	13.0	1.21	0.24	0.67
LR [Cologne]	4	8	51	0	100.0	13.6	56.8	7.3	100.0	7.3	1.16	0.00	0.75
LR [Emory]	8	15	41	0	100.0	26.8	63.4	16.3	100.0	16.3	1.37	0.00	0.66
LR [Harvard]	3	4	40	0	100.0	9.1	54.5	7.0	100.0	7.0	1.10	0.00	0.45
LR [Hillside]	6	12	53	0	100.0	18.5	59.2	10.2	100.0	10.2	1.23	0.00	0.94
LR [Milan]	2	6	17	0	100.0	26.1	63.0	10.5	100.0	10.5	1.35	0.00	0.70
LR [Munich]	10	12	87	0	100.0	12.1	56.1	10.3	100.0	10.3	1.14	0.00	0.80
LR [Turku]	5	16	15	1	83.3	51.6	67.5	25.0	94.1	19.1	1.72	0.32	0.81
LR [UCLA]	14	13	46	0	100.0	22.0	61.0	23.3	100.0	23.3	1.28	0.00	0.77
LR [UCSD]	9	15	35	0	100.0	30.0	65.0	20.5	100.0	20.5	1.43	0.00	0.68
LR [UNC]	15	10	54	0	100.0	15.6	57.8	21.7	100.0	21.7	1.19	0.00	0.84
LR [Yale]	9	6	53	1	90.0	10.2	50.1	14.5	85.7	0.2	1.00	0.98	0.70
LR [mean]					97.7	21.3	59.5	14.1	98.3	12.4	1.27	0.11	0.72
LR [SD]					5.1	11.6	5.0	6.5	4.0	6.8	0.19	0.27	0.13
Cox-PH [Full Sample]	*68*	*562*	*227*	*42*	*61.8*	*71.2*	*66.5*	*23.1*	*93.0*	*16.1*	*2.15*	*0.54*	*0.72*
Cox-PH [Basel]	2	20	12	0	100.0	62.5	81.3	14.3	100.0	14.3	2.67	0.00	0.86
Cox-PH [Birmingham]	0	23	11	2	0.0	67.6	33.8	0.0	92.0	-8.0	0.00	1.48	0.47
Cox-PH [Calgary]	9	84	30	10	47.4	73.7	60.5	23.1	89.4	12.4	1.80	0.71	0.68
Cox-PH [Cologne]	3	37	22	1	75.0	62.7	68.9	12.0	97.4	9.4	2.01	0.40	0.75
Cox-PH [Emory]	3	40	16	5	37.5	71.4	54.5	15.8	88.9	4.7	1.31	0.88	0.65
Cox-PH [Harvard]	1	25	19	2	33.3	56.8	45.1	5.0	92.6	-2.4	0.77	1.17	0.42
Cox-PH [Hillside]	6	50	15	0	100.0	76.9	88.5	28.6	100.0	28.6	4.33	0.00	0.94
Cox-PH [Milan]	2	16	7	0	100.0	69.6	84.8	22.2	100.0	22.2	3.29	0.00	0.70
Cox-PH [Munich]	8	74	25	2	80.0	74.7	77.4	24.2	97.4	21.6	3.17	0.27	0.80
Cox-PH [Turku]	4	24	7	2	66.7	77.4	72.0	36.4	92.3	28.7	2.95	0.43	0.81
Cox-PH [UCLA]	10	47	12	4	71.4	79.7	75.5	45.5	92.2	37.6	3.51	0.36	0.78
Cox-PH [UCSD]	3	35	15	6	33.3	70.0	51.7	16.7	85.4	2.0	1.11	0.95	0.68
Cox-PH [UNC]	11	44	20	4	73.3	68.8	71.0	35.5	91.7	27.2	2.35	0.39	0.84
Cox-PH [Yale]	6	43	16	4	60.0	72.9	66.4	27.3	91.5	18.8	2.21	0.55	0.68
Cox-PH [mean]					62.7	70.3	66.5	21.9	93.6	15.5	2.25	0.54	0.72
Cox-PH [SD]					29.6	6.4	15.7	12.4	4.6	13.2	1.18	0.45	0.14
Cox-PH (+A) [Full Sample]	*68*	*583*	*206*	*42*	*61.8*	*73.9*	*67.9*	*24.8*	*93.3*	*18.1*	*2.37*	*0.52*	*0.72*
Cox-PH (+A) [Basel]	2	24	8	0	100.0	75.0	87.5	20.0	100.0	20.0	4.00	0.00	0.84
Cox-PH (+A) [Birmingham]	1	24	10	1	50.0	70.6	60.3	9.1	96.0	5.1	1.70	0.71	0.56
Cox-PH (+A) [Calgary]	9	83	31	10	47.4	72.8	60.1	22.5	89.2	11.7	1.74	0.72	0.64
Cox-PH (+A) [Cologne]	2	42	17	2	50.0	71.2	60.6	10.5	95.5	6.0	1.74	0.70	0.74
Cox-PH (+A) [Emory]	2	39	17	6	25.0	69.6	47.3	10.5	86.7	-2.8	0.82	1.08	0.67
Cox-PH (+A) [Harvard]	1	27	17	2	33.3	61.4	47.3	5.6	93.1	-1.3	0.86	1.09	0.45
Cox-PH (+A) [Hillside]	6	50	15	0	100.0	76.9	88.5	28.6	100.0	28.6	4.33	0.00	0.94
Cox-PH (+A) [Milan]	1	16	7	1	50.0	69.6	59.8	12.5	94.1	6.6	1.64	0.72	0.70
Cox-PH (+A) [Munich]	8	75	24	2	80.0	75.8	77.9	25.0	97.4	22.4	3.30	0.26	0.80
Cox-PH (+A) [Turku]	4	24	7	2	66.7	77.4	72.0	36.4	92.3	28.7	2.95	0.43	0.81
Cox-PH (+A) [UCLA]	10	47	12	4	71.4	79.7	75.5	45.5	92.2	37.6	3.51	0.36	0.78
Cox-PH (+A) [UCSD]	5	38	12	4	55.6	76.0	65.8	29.4	90.5	19.9	2.31	0.58	0.71
Cox-PH (+A) [UNC]	10	51	13	5	66.7	79.7	73.2	43.5	91.1	34.5	3.28	0.42	0.84
Cox-PH (+A) [Yale]	7	43	16	3	70.0	72.9	71.4	30.4	93.5	23.9	2.58	0.41	0.70
Cox-PH (+A) [mean]					61.9	73.5	67.7	23.5	93.7	17.2	2.48	0.53	0.73

Predictors	**TP**	**TN**	**FP**	**FN**	**Sens**	**Spec**	**BAC**	**PPV**	**NPV**	**PSI**	**LR+**	**LR-**	**AUC**
Cox-PH (+A) [SD]					22.0	4.9	12.7	12.9	3.9	13.1	1.11	0.33	0.12
Linear SVM [Full Sample]	*86*	*476*	*313*	*24*	*78.2*	*60.3*	*69.3*	*21.6*	*95.2*	*16.8*	*1.97*	*0.36*	*0.74*
Linear SVM [Basel]	2	21	11	0	100.0	65.6	82.8	15.4	100.0	15.4	2.91	0.00	0.89
Linear SVM [Birmingham]	1	20	14	1	50.0	58.8	54.4	6.7	95.2	1.9	1.21	0.85	0.43
Linear SVM [Calgary]	13	65	49	6	68.4	57.0	62.7	21.0	91.5	12.5	1.59	0.55	0.69
Linear SVM [Cologne]	3	41	18	1	75.0	69.5	72.2	14.3	97.6	11.9	2.46	0.36	0.78
Linear SVM [Emory]	8	34	22	0	100.0	60.7	80.4	26.7	100.0	26.7	2.55	0.00	0.71
Linear SVM [Harvard]	1	23	21	2	33.3	52.3	42.8	4.5	92.0	-3.5	0.70	1.28	0.52
Linear SVM [Hillside]	6	36	29	0	100.0	55.4	77.7	17.1	100.0	17.1	2.24	0.00	0.92
Linear SVM [Milan]	2	16	7	0	100.0	69.6	84.8	22.2	100.0	22.2	3.29	0.00	0.70
Linear SVM [Munich]	8	65	34	2	80.0	65.7	72.8	19.0	97.0	16.1	2.33	0.30	0.83
Linear SVM [Turku]	5	22	9	1	83.3	71.0	77.2	35.7	95.7	31.4	2.87	0.23	0.80
Linear SVM [UCLA]	11	36	23	3	78.6	61.0	69.8	32.4	92.3	24.7	2.02	0.35	0.76
Linear SVM [UCSD]	7	27	23	2	77.8	54.0	65.9	23.3	93.1	16.4	1.69	0.41	0.65
Linear SVM [UNC]	12	41	23	3	80.0	64.1	72.0	34.3	93.2	27.5	2.23	0.31	0.83
Linear SVM [Yale]	7	29	30	3	70.0	49.2	59.6	18.9	90.6	9.5	1.38	0.61	0.65
Linear SVM [mean]					78.3	61.0	69.6	20.8	95.6	16.4	2.10	0.38	0.73
Linear SVM [SD]					19.4	6.9	11.7	9.3	3.5	9.8	0.72	0.36	0.14
RBF-SVM [Full Sample]	*53*	*634*	*155*	*57*	*48.2*	*80.4*	*64.3*	*25.5*	*91.8*	*17.2*	*2.45*	*0.64*	*0.74*
RBF-SVM [Basel]	2	25	7	0	100.0	78.1	89.1	22.2	100.0	22.2	4.57	0.00	0.83
RBF-SVM [Birmingham]	0	27	7	2	0.0	79.4	39.7	0.0	93.1	-6.9	0.00	1.26	0.43
RBF-SVM [Calgary]	9	91	23	10	47.4	79.8	63.6	28.1	90.1	18.2	2.35	0.66	0.70
RBF-SVM [Cologne]	2	48	11	2	50.0	81.4	65.7	15.4	96.0	11.4	2.68	0.61	0.78
RBF-SVM [Emory]	2	40	16	6	25.0	71.4	48.2	11.1	87.0	-1.9	0.88	1.05	0.69
RBF-SVM [Harvard]	1	34	10	2	33.3	77.3	55.3	9.1	94.4	3.5	1.47	0.86	0.47
RBF-SVM [Hillside]	6	55	10	0	100.0	84.6	92.3	37.5	100.0	37.5	6.50	0.00	0.93
RBF-SVM [Milan]	0	16	7	2	0.0	69.6	34.8	0.0	88.9	-11.1	0.00	1.44	0.70
RBF-SVM [Munich]	5	81	18	5	50.0	81.8	65.9	21.7	94.2	15.9	2.75	0.61	0.81
RBF-SVM [Turku]	3	24	7	3	50.0	77.4	63.7	30.0	88.9	18.9	2.21	0.65	0.77
RBF-SVM [UCLA]	7	48	11	7	50.0	81.4	65.7	38.9	87.3	26.2	2.68	0.61	0.76
RBF-SVM [UCSD]	3	38	12	6	33.3	76.0	54.7	20.0	86.4	6.4	1.39	0.88	0.65
RBF-SVM [UNC]	10	57	7	5	66.7	89.1	77.9	58.8	91.9	50.8	6.10	0.37	0.85
RBF-SVM [Yale]	3	50	9	7	30.0	84.7	57.4	25.0	87.7	12.7	1.97	0.83	0.66
RBF SVM [mean]					45.4	79.4	62.4	22.7	91.8	14.6	2.54	0.70	0.72
RBF SVM [SD]					29.8	5.1	16.4	15.9	4.6	16.8	1.98	0.41	0.14

Abbreviations. Algorithms: *LR* Logistic regression, *Cox-PH* Cox Proportional Hazard model, *SVM* Support Vector Machine, *RBF-SVM* Support Vector Machine with Radial Basis Kernel, *+A* Cox-PH (+ADASYN); **Performance Measures:** *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, *PPV* Positive Predictive Value, *NPV* Negative Predictive Value, *PSI* Prognostic Summary Index, *LR+* Positive Likelihood Ratio, *LR-* Negative Likelihood Ratio, *AUC* Area-under-the Curve; **Test sites:** *UCLA* University of California, Los Angeles, *UCSD* University of Californa, San Diego, *UNC* University of North Carolina.

Supplementary Table 3. Leave-site-out cross-validation analysis in the CHR sample comparing the five different PT prediction algorithms. The out-of-training performance of the given algorithm was broken down per site. In addition, the respective means and standards deviation of the given algorithm's performance measures were computed across sites. To avoid a biased estimate of the average leave-site-out performances, the PRONIA Udine site was excluded from this analysis because of no reported transition cases.

Predictors	**TP**	**TN**	**FP**	**FN**	**Sens**	**Spec**	**BAC**	**PPV**	**NPV**	**PSI**	**LR+**	**LR-**	**AUC**
LR [Full Sample]	*104*	*80*	*560*	*3*	*97.2*	*12.5*	*54.8*	*15.7*	*96.4*	*12.0*	*1.10*	*0.20*	*0.72*
LR [Basel]	2	4	13	0	100.0	23.5	61.8	13.3	100.0	13.3	1.31	0.00	0.82
LR [Birmingham]	1	1	15	0	100.0	6.3	53.1	6.3	100.0	6.3	1.07	0.00	0.50
LR [Calgary]	18	10	104	1	94.7	8.8	51.8	14.8	90.9	5.7	1.04	0.60	0.68
LR [Cologne]	3	3	17	0	100.0	15.0	57.5	15.0	100.0	15.0	1.18	0.00	0.80
LR [Emory]	8	11	45	0	100.0	19.6	59.8	15.1	100.0	15.1	1.24	0.00	0.68
LR [Harvard]	3	3	41	0	100.0	6.8	53.4	6.8	100.0	6.8	1.07	0.00	0.48
LR [Hillside]	6	7	58	0	100.0	10.8	55.4	9.4	100.0	9.4	1.12	0.00	0.93
LR [Milan]	2	4	9	0	100.0	30.8	65.4	18.2	100.0	18.2	1.44	0.00	0.46
LR [Munich]	9	5	39	0	100.0	11.4	55.7	18.8	100.0	18.8	1.13	0.00	0.72
LR [Turku]	5	7	12	1	83.3	36.8	60.1	29.4	87.5	16.9	1.32	0.45	0.76
LR [UCLA]	14	8	51	0	100.0	13.6	56.8	21.5	100.0	21.5	1.16	0.00	0.77
LR [UCSD]	9	8	42	0	100.0	16.0	58.0	17.6	100.0	17.6	1.19	0.00	0.69
LR [UNC]	15	5	59	0	100.0	7.8	53.9	20.3	100.0	20.3	1.08	0.00	0.82
LR [Yale]	9	4	55	1	90.0	6.8	48.4	14.1	80.0	-5.9	0.97	1.48	0.68
LR [mean]					97.7	15.3	56.5	15.7	97.0	12.8	1.17	0.18	0.70
LR [SD]					5.1	9.4	4.4	6.1	6.3	7.5	0.13	0.42	0.14
Cox-PH [Full Sample]	*74*	*419*	*221*	*33*	*69.2*	*65.5*	*67.3*	*25.1*	*92.7*	*17.8*	*2.00*	*0.50*	*0.73*
Cox-PH [Basel]	2	11	6	0	100.0	64.7	82.4	25.0	100.0	25.0	2.83	0.00	0.82
Cox-PH [Birmingham]	0	9	7	1	0.0	56.3	28.1	0.0	90.0	-10.0	0.00	1.78	0.50
Cox-PH [Calgary]	13	73	41	6	68.4	64.0	66.2	24.1	92.4	16.5	1.90	0.49	0.69
Cox-PH [Cologne]	2	13	7	1	66.7	65.0	65.8	22.2	92.9	15.1	1.90	0.51	0.80
Cox-PH [Emory]	6	36	20	2	75.0	64.3	69.6	23.1	94.7	17.8	2.10	0.39	0.67
Cox-PH [Harvard]	1	27	17	2	33.3	61.4	47.3	5.6	93.1	-1.3	0.86	1.09	0.47
Cox-PH [Hillside]	6	43	22	0	100.0	66.2	83.1	21.4	100.0	21.4	2.95	0.00	0.92
Cox-PH [Milan]	0	7	6	2	0.0	53.8	26.9	0.0	77.8	-22.2	0.00	1.86	0.46
Cox-PH [Munich]	6	29	15	3	66.7	65.9	66.3	28.6	90.6	19.2	1.96	0.51	0.72
Cox-PH [Turku]	4	13	6	2	66.7	68.4	67.5	40.0	86.7	26.7	2.11	0.49	0.76
Cox-PH [UCLA]	10	40	19	4	71.4	67.8	69.6	34.5	90.9	25.4	2.22	0.42	0.76
Cox-PH [UCSD]	6	35	15	3	66.7	70.0	68.3	28.6	92.1	20.7	2.22	0.48	0.68
Cox-PH [UNC]	11	44	20	4	73.3	68.8	71.0	35.5	91.7	27.2	2.35	0.39	0.81
Cox-PH [Yale]	7	39	20	3	70.0	66.1	68.1	25.9	92.9	18.8	2.07	0.45	0.66
Cox-PH [mean]					61.3	64.5	62.9	22.5	91.8	14.3	1.82	0.63	0.70
Cox-PH [SD]					30.3	4.6	17.1	12.4	5.4	14.9	0.91	0.56	0.14
Cox-PH (+A) [Full Sample]	*72*	*415*	*225*	*35*	*67.3*	*64.8*	*66.1*	*24.2*	*92.2*	*16.5*	*1.90*	*0.50*	*0.72*
Cox-PH (+A) [Basel]	2	11	6	0	100.0	64.7	82.4	25.0	100.0	25.0	2.83	0.00	0.82
Cox-PH (+A) [Birmingham]	0	9	7	1	0.0	56.3	28.1	0.0	90.0	-10.0	0.00	1.78	0.56
Cox-PH (+A) [Calgary]	10	71	43	9	52.6	62.3	57.5	18.9	88.8	7.6	1.40	0.76	0.65
Cox-PH (+A) [Cologne]	2	13	7	1	66.7	65.0	65.8	22.2	92.9	15.1	1.90	0.51	0.77
Cox-PH (+A) [Emory]	6	37	19	2	75.0	66.1	70.5	24.0	94.9	18.9	2.21	0.38	0.67
Cox-PH (+A) [Harvard]	1	27	17	2	33.3	61.4	47.3	5.6	93.1	-1.3	0.86	1.09	0.45
Cox-PH (+A) [Hillside]	6	42	23	0	100.0	64.6	82.3	20.7	100.0	20.7	2.83	0.00	0.94
Cox-PH (+A) [Milan]	0	7	6	2	0.0	53.8	26.9	0.0	77.8	-22.2	0.00	1.86	0.46
Cox-PH (+A) [Munich]	6	28	16	3	66.7	63.6	65.2	27.3	90.3	17.6	1.83	0.52	0.71
Cox-PH (+A) [Turku]	4	13	6	2	66.7	68.4	67.5	40.0	86.7	26.7	2.11	0.49	0.78
Cox-PH (+A) [UCLA]	10	41	18	4	71.4	69.5	70.5	35.7	91.1	26.8	2.34	0.41	0.78
Cox-PH (+A) [UCSD]	7	34	16	2	77.8	68.0	72.9	30.4	94.4	24.9	2.43	0.33	0.69
Cox-PH (+A) [UNC]	11	44	20	4	73.3	68.8	71.0	35.5	91.7	27.2	2.35	0.39	0.83
Cox-PH (+A) [Yale]	7	38	21	3	70.0	64.4	67.2	25.0	92.7	17.7	1.97	0.47	0.69
Cox-PH (+A) [mean]					61.0	64.1	62.5	22.2	91.7	13.9	1.79	0.64	0.70

Predictors	**TP**	**TN**	**FP**	**FN**	**Sens**	**Spec**	**BAC**	**PPV**	**NPV**	**PSI**	**LR+**	**LR-**	**AUC**
Cox-PH (+A) [SD]					30.7	4.5	17.2	12.6	5.5	15.2	0.92	0.57	0.14
Linear SVM [Full Sample]	*78*	*378*	*262*	*29*	*72.9*	*59.1*	*66.0*	*22.9*	*92.9*	*15.8*	*1.80*	*0.50*	*0.73*
Linear SVM [Basel]	2	10	7	0	100.0	58.8	79.4	22.2	100.0	22.2	2.43	0.00	0.82
Linear SVM [Birmingham]	0	9	7	1	0.0	56.3	28.1	0.0	90.0	-10.0	0.00	1.78	0.44
Linear SVM [Calgary]	13	66	48	6	68.4	57.9	63.2	21.3	91.7	13.0	1.63	0.55	0.69
Linear SVM [Cologne]	2	14	6	1	66.7	70.0	68.3	25.0	93.3	18.3	2.22	0.48	0.82
Linear SVM [Emory]	7	34	22	1	87.5	60.7	74.1	24.1	97.1	21.3	2.23	0.21	0.71
Linear SVM [Harvard]	1	23	21	2	33.3	52.3	42.8	4.5	92.0	-3.5	0.70	1.28	0.52
Linear SVM [Hillside]	6	35	30	0	100.0	53.8	76.9	16.7	100.0	16.7	2.17	0.00	0.92
Linear SVM [Milan]	0	6	7	2	0.0	46.2	23.1	0.0	75.0	-25.0	0.00	2.17	0.46
Linear SVM [Munich]	7	27	17	2	77.8	61.4	69.6	29.2	93.1	22.3	2.01	0.36	0.73
Linear SVM [Turku]	4	13	6	2	66.7	68.4	67.5	40.0	86.7	26.7	2.11	0.49	0.76
Linear SVM [UCLA]	11	41	18	3	78.6	69.5	74.0	37.9	93.2	31.1	2.58	0.31	0.77
Linear SVM [UCSD]	7	28	22	2	77.8	56.0	66.9	24.1	93.3	17.5	1.77	0.40	0.68
Linear SVM [UNC]	11	42	22	4	73.3	65.6	69.5	33.3	91.3	24.6	2.13	0.41	0.84
Linear SVM [Yale]	7	30	29	3	70.0	50.8	60.4	19.4	90.9	10.4	1.42	0.59	0.66
Linear SVM [mean]					64.3	59.1	61.7	21.3	92.0	13.3	1.67	0.64	0.70
Linear SVM [SD]					31.7	7.3	17.7	12.6	6.1	15.7	0.85	0.64	0.14
RBF-SVM [Full Sample]	48	514	126	59	44.9	80.3	62.6	27.6	89.7	17.3	2.30	0.70	0.72
RBF-SVM [Basel]	1	12	5	1	50.0	70.6	60.3	16.7	92.3	9.0	1.70	0.71	0.82
RBF-SVM [Birmingham]	0	13	3	1	0.0	81.3	40.6	0.0	92.9	-7.1	0.00	1.23	0.56
RBF-SVM [Calgary]	8	90	24	11	42.1	78.9	60.5	25.0	89.1	14.1	2.00	0.73	0.68
RBF-SVM [Cologne]	2	17	3	1	66.7	85.0	75.8	40.0	94.4	34.4	4.44	0.39	0.80
RBF-SVM [Emory]	1	45	11	7	12.5	80.4	46.4	8.3	86.5	-5.1	0.64	1.09	0.68
RBF-SVM [Harvard]	1	33	11	2	33.3	75.0	54.2	8.3	94.3	2.6	1.33	0.89	0.52
RBF-SVM [Hillside]	6	52	13	0	100.0	80.0	90.0	31.6	100.0	31.6	5.00	0.00	0.93
RBF-SVM [Milan]	0	9	4	2	0.0	69.2	34.6	0.0	81.8	-18.2	0.00	1.44	0.46
RBF-SVM [Munich]	4	36	8	5	44.4	81.8	63.1	33.3	87.8	21.1	2.44	0.68	0.69
RBF-SVM [Turku]	3	16	3	3	50.0	84.2	67.1	50.0	84.2	34.2	3.17	0.59	0.77
RBF-SVM [UCLA]	7	46	13	7	50.0	78.0	64.0	35.0	86.8	21.8	2.27	0.64	0.75
RBF-SVM [UCSD]	2	39	11	7	22.2	78.0	50.1	15.4	84.8	0.2	1.01	1.00	0.66
RBF-SVM [UNC]	9	56	8	6	60.0	87.5	73.8	52.9	90.3	43.3	4.80	0.46	0.83
RBF-SVM [Yale]	4	50	9	6	40.0	84.7	62.4	30.8	89.3	20.1	2.62	0.71	0.66
RBF SVM [mean]					40.8	79.6	60.2	24.8	89.6	14.4	2.24	0.75	0.70
RBF SVM [SD]					26.8	5.3	14.6	17.2	4.8	18.3	1.65	0.36	0.13

Abbreviations. Algorithms: *LR* Logistic regression, *Cox-PH* Cox Proportional Hazard model, *SVM* Support Vector Machine, *RBF-SVM* Support Vector Machine with Radial Basis Kernel, *+A* Cox-PH (+ADASYN); **Performance Measures:** *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, *PPV* Positive Predictive Value, *NPV* Negative Predictive Value, *PSI* Prognostic Summary Index, *LR+* Positive Likelihood Ratio, *LR-* Negative Likelihood Ratio, *AUC* Area-under-the Curve; **Test sites:** *UCLA* University of California, Los Angeles, *UCSD* University of Californa, San Diego, *UNC* University of North Carolina.

Supplementary Table 4. Leave-site-out cross-validation analysis in the UHR sample comparing the five different PT prediction algorithms. The out-of-training performance of the given algorithm was broken down per site. In addition, the respective means and standards deviation of the given algorithm's performance measures were computed across sites. To avoid a biased estimate of the average leave-site-out performances, the PRONIA Udine site was excluded from this analysis because of no reported transition cases.

Predictors	**TP**	**TN**	**FP**	**FN**	**Sens**	**Spec**	**BAC**	**PPV**	**NPV**	**PSI**	**LR+**	**LR-**	**AUC**
LR [Full Sample]	*103*	*62*	*542*	*2*	*98.1*	*10.3*	*54.2*	*16.0*	*96.9*	*12.8*	*1.10*	*0.20*	*0.71*
LR [Basel]	2	3	8	0	100.0	27.3	63.6	20.0	100.0	20.0	1.38	0.00	0.68
LR [Birmingham]	1	0	8	0	100.0	0.0	50.0	11.1			1.00		0.25
LR [Calgary]	18	9	105	1	94.7	7.9	51.3	14.6	90.0	4.6	1.03	0.67	0.68
LR [Cologne]	3	1	14	0	100.0	6.7	53.3	17.6	100.0	17.6	1.07	0.00	0.73
LR [Emory]	8	10	46	0	100.0	17.9	58.9	14.8	100.0	14.8	1.22	0.00	0.67
LR [Harvard]	3	2	42	0	100.0	4.5	52.3	6.7	100.0	6.7	1.05	0.00	0.48
LR [Hillside]	6	7	58	0	100.0	10.8	55.4	9.4	100.0	9.4	1.12	0.00	0.92
LR [Milan]	1	3	7	0	100.0	30.0	65.0	12.5	100.0	12.5	1.43	0.00	0.40
LR [Munich]	8	3	34	0	100.0	8.1	54.1	19.0	100.0	19.0	1.09	0.00	0.71
LR [Turku]	5	4	8	1	83.3	33.3	58.3	38.5	80.0	18.5	1.25	0.50	0.69
LR [UCLA]	14	6	53	0	100.0	10.2	55.1	20.9	100.0	20.9	1.11	0.00	0.77
LR [UCSD]	9	7	43	0	100.0	14.0	57.0	17.3	100.0	17.3	1.16	0.00	0.69
LR [UNC]	15	4	60	0	100.0	6.3	53.1	20.0	100.0	20.0	1.07	0.00	0.82
LR [Yale]	10	3	56	0	100.0	5.1	52.5	15.2	100.0	15.2	1.05	0.00	0.67
LR [mean]					98.4	13.0	55.7	17.0	97.7	15.1	1.14	0.09	0.66
LR [SD]					4.6	10.3	4.4	7.5	6.0	5.3	0.13	0.22	0.17
Cox-PH [Full Sample]	*69*	*397*	*207*	*36*	*65.7*	*65.7*	*65.7*	*25.0*	*91.7*	*16.7*	*1.90*	*0.50*	*0.72*
Cox-PH [Basel]	1	7	4	1	50.0	63.6	56.8	20.0	87.5	7.5	1.38	0.79	0.73
Cox-PH [Birmingham]	0	4	4	1	0.0	50.0	25.0	0.0	80.0	-20.0	0.00	2.00	0.38
Cox-PH [Calgary]	12	73	41	7	63.2	64.0	63.6	22.6	91.3	13.9	1.76	0.58	0.69
Cox-PH [Cologne]	2	10	5	1	66.7	66.7	66.7	28.6	90.9	19.5	2.00	0.50	0.71
Cox-PH [Emory]	6	37	19	2	75.0	66.1	70.5	24.0	94.9	18.9	2.21	0.38	0.67
Cox-PH [Harvard]	1	26	18	2	33.3	59.1	46.2	5.3	92.9	-1.9	0.81	1.13	0.46
Cox-PH [Hillside]	6	43	22	0	100.0	66.2	83.1	21.4	100.0	21.4	2.95	0.00	0.93
Cox-PH [Milan]	0	6	4	1	0.0	60.0	30.0	0.0	85.7	-14.3	0.00	1.67	0.40
Cox-PH [Munich]	6	25	12	2	75.0	67.6	71.3	33.3	92.6	25.9	2.31	0.37	0.71
Cox-PH [Turku]	3	8	4	3	50.0	66.7	58.3	42.9	72.7	15.6	1.50	0.75	0.68
Cox-PH [UCLA]	10	40	19	4	71.4	67.8	69.6	34.5	90.9	25.4	2.22	0.42	0.76
Cox-PH [UCSD]	4	36	14	5	44.4	72.0	58.2	22.2	87.8	10.0	1.59	0.77	0.68
Cox-PH [UNC]	11	43	21	4	73.3	67.2	70.3	34.4	91.5	25.9	2.23	0.40	0.81
Cox-PH [Yale]	7	39	20	3	70.0	66.1	68.1	25.9	92.9	18.8	2.07	0.45	0.66
Cox-PH [mean]					55.2	64.5	59.8	22.5	89.4	11.9	1.64	0.73	0.66
Cox-PH [SD]					28.5	5.3	16.2	13.0	6.6	14.5	0.86	0.54	0.15
Cox-PH (+A] [Full Sample]	*63*	*401*	*203*	*42*	*60.0*	*66.4*	*63.2*	*23.7*	*90.5*	*14.2*	*1.80*	*0.60*	*0.70*
Cox-PH (+A) [Basel]	1	8	3	1	50.0	72.7	61.4	25.0	88.9	13.9	1.83	0.69	0.73
Cox-PH (+A) [Birmingham]	0	4	4	1	0.0	50.0	25.0	0.0	80.0	-20.0	0.00	2.00	0.25
Cox-PH (+A) [Calgary]	9	69	45	10	47.4	60.5	53.9	16.7	87.3	4.0	1.20	0.87	0.64
Cox-PH (+A) [Cologne]	2	11	4	1	66.7	73.3	70.0	33.3	91.7	25.0	2.50	0.45	0.71
Cox-PH (+A) [Emory]	3	39	17	5	37.5	69.6	53.6	15.0	88.6	3.6	1.24	0.90	0.66
Cox-PH (+A) [Harvard]	1	27	17	2	33.3	61.4	47.3	5.6	93.1	-1.3	0.86	1.09	0.45
Cox-PH (+A) [Hillside]	6	43	22	0	100.0	66.2	83.1	21.4	100.0	21.4	2.95	0.00	0.93
Cox-PH (+A) [Milan]	0	6	4	1	0.0	60.0	30.0	0.0	85.7	-14.3	0.00	1.67	0.40
Cox-PH (+A) [Munich]	5	28	9	3	62.5	75.7	69.1	35.7	90.3	26.0	2.57	0.50	0.71
Cox-PH (+A) [Turku]	4	9	3	2	66.7	75.0	70.8	57.1	81.8	39.0	2.67	0.44	0.63
Cox-PH (+A) [UCLA]	10	40	19	4	71.4	67.8	69.6	34.5	90.9	25.4	2.22	0.42	0.78
Cox-PH (+A) [UCSD]	3	35	15	6	33.3	70.0	51.7	16.7	85.4	2.0	1.11	0.95	0.69
Cox-PH (+A) [UNC]	12	44	20	3	80.0	68.8	74.4	37.5	93.6	31.1	2.56	0.29	0.82
Cox-PH (+A) [Yale]	7	38	21	3	70.0	64.4	67.2	25.0	92.7	17.7	1.97	0.47	0.70
Cox-PH (+A) [mean]					51.3	66.8	59.1	23.1	89.3	12.4	1.69	0.77	0.65

Predictors	**TP**	**TN**	**FP**	**FN**	**Sens**	**Spec**	**BAC**	**PPV**	**NPV**	**PSI**	**LR+**	**LR-**	**AUC**
Cox-PH (+A) [SD]					28.6	7.1	16.7	15.9	5.1	17.3	0.97	0.54	0.18
Linear SVM [Full Sample]	*74*	*346*	*258*	*31*	*70.5*	*57.3*	*63.9*	*22.3*	*91.8*	*14.1*	*1.60*	*0.50*	*0.71*
Linear SVM [Basel]	1	6	5	1	50.0	54.5	52.3	16.7	85.7	2.4	1.10	0.92	0.73
Linear SVM [Birmingham]	0	4	4	1	0.0	50.0	25.0	0.0	80.0	-20.0	0.00	2.00	0.38
Linear SVM [Calgary]	12	60	54	7	63.2	52.6	57.9	18.2	89.6	7.7	1.33	0.70	0.67
Linear SVM [Cologne]	2	12	3	1	66.7	80.0	73.3	40.0	92.3	32.3	3.33	0.42	0.71
Linear SVM [Emory]	7	34	22	1	87.5	60.7	74.1	24.1	97.1	21.3	2.23	0.21	0.71
Linear SVM [Harvard]	1	23	21	2	33.3	52.3	42.8	4.5	92.0	-3.5	0.70	1.28	0.49
Linear SVM [Hillside]	6	37	28	0	100.0	56.9	78.5	17.6	100.0	17.6	2.32	0.00	0.91
Linear SVM [Milan]	0	4	6	1	0.0	40.0	20.0	0.0	80.0	-20.0	0.00	2.50	0.40
Linear SVM [Munich]	6	21	16	2	75.0	56.8	65.9	27.3	91.3	18.6	1.73	0.44	0.71
Linear SVM [Turku]	3	9	3	3	50.0	75.0	62.5	50.0	75.0	25.0	2.00	0.67	0.68
Linear SVM [UCLA]	11	38	21	3	78.6	64.4	71.5	34.4	92.7	27.1	2.21	0.33	0.77
Linear SVM [UCSD]	7	28	22	2	77.8	56.0	66.9	24.1	93.3	17.5	1.77	0.40	0.68
Linear SVM [UNC]	11	42	22	4	73.3	65.6	69.5	33.3	91.3	24.6	2.13	0.41	0.82
Linear SVM [Yale]	7	28	31	3	70.0	47.5	58.7	18.4	90.3	8.7	1.33	0.63	0.67
Linear SVM [mean]					59.0	58.0	58.5	22.1	89.3	11.4	1.58	0.78	0.67
Linear SVM [SD]					30.0	10.6	17.9	14.6	6.9	16.6	0.92	0.70	0.15
RBF-SVM [Full Sample]	*50*	*480*	*124*	*55*	*47.6*	*79.5*	*63.5*	*28.7*	*89.7*	*18.5*	*2.30*	*0.70*	*0.71*
RBF-SVM [Basel]	1	8	3	1	50.0	72.7	61.4	25.0	88.9	13.9	1.83	0.69	0.73
RBF-SVM [Birmingham]	0	6	2	1	0.0	75.0	37.5	0.0	85.7	-14.3	0.00	1.33	0.25
RBF-SVM [Calgary]	8	90	24	11	42.1	78.9	60.5	25.0	89.1	14.1	2.00	0.73	0.68
RBF-SVM [Cologne]	2	13	2	1	66.7	86.7	76.7	50.0	92.9	42.9	5.00	0.38	0.71
RBF-SVM [Emory]	2	42	14	6	25.0	75.0	50.0	12.5	87.5	0.0	1.00	1.00	0.69
RBF-SVM [Harvard]	1	33	11	2	33.3	75.0	54.2	8.3	94.3	2.6	1.33	0.89	0.52
RBF-SVM [Hillside]	6	50	15	0	100.0	76.9	88.5	28.6	100.0	28.6	4.33	0.00	0.93
RBF-SVM [Milan]	0	6	4	1	0.0	60.0	30.0	0.0	85.7	-14.3	0.00	1.67	0.40
RBF-SVM [Munich]	4	32	5	4	50.0	86.5	68.2	44.4	88.9	33.3	3.70	0.58	0.70
RBF-SVM [Turku]	3	11	1	3	50.0	91.7	70.8	75.0	78.6	53.6	6.00	0.55	0.71
RBF-SVM [UCLA]	8	46	13	6	57.1	78.0	67.6	38.1	88.5	26.6	2.59	0.55	0.77
RBF-SVM [UCSD]	2	37	13	7	22.2	74.0	48.1	13.3	84.1	-2.6	0.85	1.05	0.67
RBF-SVM [UNC]	9	56	8	6	60.0	87.5	73.8	52.9	90.3	43.3	4.80	0.46	0.83
RBF-SVM [Yale]	4	50	9	6	40.0	84.7	62.4	30.8	89.3	20.1	2.62	0.71	0.67
RBF SVM [mean]					42.6	78.8	60.7	28.9	88.8	17.7	2.58	0.76	0.66
RBF SVM [SD]					26.3	8.1	15.7	21.7	5.0	21.6	1.92	0.42	0.17

Abbreviations. Algorithms: *LR* Logistic regression, *Cox-PH* Cox Proportional Hazard model, *SVM* Support Vector Machine, *RBF-SVM* Support Vector Machine with Radial Basis Kernel, *+A* Cox-PH (+ADASYN); **Performance Measures:** *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, *PPV* Positive Predictive Value, *NPV* Negative Predictive Value, *PSI* Prognostic Summary Index, *LR+* Positive Likelihood Ratio, *LR-* Negative Likelihood Ratio, *AUC* Area-under-the Curve; **Test sites:** *UCLA* University of California, Los Angeles, *UCSD* University of Californa, San Diego, *UNC* University of North Carolina.

Supplementary Table 5. Summary overview of prognostic performances in terms of sensitivity, specificity and balanced accuracy of the five different algorithms obtained at the three risk enrichment levels (CHR+, CHR, and UHR) in the reciprocal external validation analysis and the leave-site-out cross-validation experiment. For latter validation setup, performance metrics were provided at the full sample level as well as at the level of mean (SD) computed across participating sites.

Algorithms	**Reciprocal external validation**			**Leave-site-out cross-validation [Full Sample]**			**Leave-site-out cross-validation [Mean (SD)]**		
	Sens	**Spec**	**BAC**	**Sens**	**Spec**	**BAC**	**Sens**	**Spec**	**BAC**
PRONIA-CHR+									
LR	85.5	40.0	62.8	97.3	19.6	58.5	97.7 (5.1)	21.3 (11.6)	59.5 (5.0)
Cox-PH	65.5	70.7	68.1	61.8	71.2	66.5	62.7 (29.6)	70.3 (6.4)	66.5 (15.7)
Cox-PH (+A)	70.9	68.4	69.6	61.8	73.9	67.9	61.9 (22.0)	73.5 (4.9)	67.7 (12.7)
Linear SVM	74.5	58.4	66.5	78.2	60.3	69.3	78.3 (19.4)	61.0 (6.9)	69.6 (11.7)
RBF-SVM	52.7	83.0	67.9	48.2	80.4	64.3	45.4 (29.8)	79.4 (5.1)	62.4 (16.4)
PRONIA-CHR									
LR	99.1	10.8	55.0	97.2	12.5	54.8	97.7 (5.1)	15.3 (9.4)	56.5 (4.4)
Cox-PH	75.7	56.6	66.2	69.2	65.5	67.3	61.3 (30.3)	64.5 (4.6)	62.9 (17.1)
Cox-PH (+A)	73.8	58.8	66.3	67.3	64.8	66.1	61.0 (30.7)	64.1 (4.5)	62.5 (17.2)
Linear SVM	74.8	58.2	66.5	72.9	59.1	66.0	64.3 (31.7)	59.1 (7.3)	61.7 (17.7)
RBF-SVM	52.3	75.6	64.0	44.9	80.3	62.6	40.8 (26.8)	79.6 (5.3)	60.2 (14.6)
PRONIA-UHR									
LR	97.1	5.0	51.1	98.1	10.3	54.2	98.4 (4.6)	13 (10.3)	55.7 (4.4)
Cox-PH	48.6	80.5	64.5	65.7	65.7	65.7	55.2 (28.5)	64.5 (5.3)	59.8 (16.2)
Cox-PH (+A)	79.0	43.0	61.0	60.0	66.4	63.2	51.3 (28.6)	66.8 (7.1)	59.1 (16.7)
Linear SVM	71.4	59.4	65.4	70.5	57.3	63.9	59.0 (30.0)	58.0 (10.6)	58.5 (17.9)
RBF-SVM	11.4	94.6	53.0	47.6	79.5	63.5	42.6 (26.3)	78.8 (8.1)	60.7 (15.7)

Abbreviations. Algorithms: *LR* Logistic regression, *Cox-PH* Cox Proportional Hazard model, *SVM* Support Vector Machine, *RBF-SVM* Support Vector Machine with Radial Basis Kernel, *+A* Cox-PH (+ADASYN); **Performance Measures:** *Sens* Sensitivity, *Spec* Specificity, *BAC* Balanced Accuracy.