Translating polygenic risk scores for clinical use by estimating the confidence bounds of risk prediction

Jiangming Sun, Yunpeng Wang, Lasse Folkesen, Yan Borne, Inge Amlien, Alfonso Buil, Marju Orho-Melander, Anders D. Barglum, David M. Hougaard, Regeneron Genetics Center, Olle Melander, Gunnar Engström, Thomas Werge & Kasper Lage

A promise of genomics in precision medicine is to provide individualized genetic risk predictions. Polygenic risk scores (PRS), computed by aggregating effects from many genomic variants, have been developed as a useful tool in complex disease research. However, the application of PRS as a tool for predicting an individual’s disease susceptibility in a clinical setting is challenging because PRS typically provide a relative measure of risk evaluated at the level of a group of people but not at individual level. Here, we introduce a machine-learning technique, Mondrian Cross-Conformal Prediction (MCCP), to estimate the confidence bounds of PRS-to-disease-risk prediction. MCCP can report disease status conditional probability value for each individual and give a prediction at a desired error level. Moreover, with a user-defined prediction error rate, MCCP can estimate the proportion of sample (coverage) with a correct prediction.

1 Institute of Biological Psychiatry, Mental Health Center Sct. Hans, Mental Health Services Copenhagen, Roskilde, Denmark. 2 The Lundbeck Foundation Initiative for Integrative Psychiatric Research (PSYCH), Copenhagen, Denmark. 3 Department of Clinical Sciences, Malmö, Lund University, Malmö, Sweden. 4 Lifespan Changes in Bain and Cognition (LCBC), Department of Psychology, University of Oslo, Oslo, Norway. 5 Department of Biomedicine, Human Genetics and Centre for Integrative Sequencing, Aarhus University, Aarhus, Denmark. 6 Department for Congenital Disorders, Center for Neonatal Screening, Statens Serum Institut, Copenhagen, Denmark. 7 Regeneron Genetics Center, Regeneron Pharmaceuticals, Tarrytown, NY, USA. 8 Department of Clinical Medicine, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark. 9 Stanley Center at Broad Institute of MIT and Harvard, Cambridge, MA, USA. 10 Department of Surgery, Massachusetts General Hospital, Boston, MA, USA. 11 These authors contributed equally: Jiangming Sun, Yunpeng Wang. *A list of authors and their affiliations appears at the end of the paper. **Email: sunjiangming@gmail.com; lage.kasper@mgh.harvard.edu
The last decade has witnessed the tremendous success of genome-wide association studies (GWAS), which have discovered tens of thousands of common variants robustly associated with a range of human complex traits and diseases, including cancers, cardiovascular diseases, neuropsychiatric, and neurodegenerative diseases. Although the identified variants individually have small to modest effect sizes, polygenic risk scores (PRS), which summarize effects from large numbers of variants, have proven to be a useful research tool. For example, PRS has been used to investigate the genetic overlaps of neuropsychiatric disorders, identify individuals of high risk for coronary artery disease, predict the age of onset for Alzheimer’s disease, and improve clinical diagnoses of cancers and type 2 diabetes mellitus.

Compared with environmental risk factors, PRS has many advantages. As an individual’s DNA is largely stable after conception, PRS for complex disorders is also stable. Therefore, it is unlikely that non-genetic factors can cause large numbers of changes to DNA, i.e., inverse causations. In addition, PRS is easy to compute. Thus, there is little doubt that the implementation of PRS will be an integral part of the clinical setting. In contrast to rare disease-causing mutations, which have large penetrance, PRS are continuous measures of the liability to disease as well as probabilistic measures of the risk of developing a condition. Thus, it is unclear what thresholds of PRS should be used by clinicians to assess an individual’s risk to develop a disease. To mitigate this, standard statistical models in the field divide the sample into different strata based on arbitrary PRS thresholds and evaluate the effect of PRS on disease risk within and across strata by several statistical metrics, for example, area under the receiver operating curve (AUC), the proportion of risk variation explained ($R^2$), odds ratio, and hazard ratio. The deciles, quintiles, top 10%, 5%, and even 1% versus bottom 10%, 5%, and 1%, respectively, are frequently used thresholds in the literature. As the performance of PRS is intimately related to genetic architectures of complex disorders, which vary for different disorders and in different populations, a systematic strategy for choosing these risk stratifying thresholds is imperative. In addition, the prediction accuracy per individual has rarely been investigated in PRS studies.

Here, we introduce a machine-learning technique, Mondrian Cross-Conformal predictor (MCCP), to complement the current state-of-the-art PRS methodology. In contrast to arbitrary PRS thresholds used in the literature, MCCP, functioning as a calibrator (Fig. 1) for PRS prediction in a test sample, is able to compute the proportion of the sample (termed coverage hereafter) for which the prediction of case-control status is reliable, i.e., below a pre-specified prediction error rate. For an individual with a predicted status, MCCP splits the sample by their respective classes and then estimates confidence levels for each class. Here, we implemented MCCP to estimate the confidence levels of risk prediction in a sample for which genetic and disease status information was available (Fig. 1; target sample). Our implementation first computed PRS for each individual in the target sample. We, then, divided the target sample into two subsets: the training and the testing set. The training set was further randomly partitioned into equal-sized subsets, one of which was retained as the calibration subset for calculating the MCCP probability value described by Eq. (1), and the remaining $n-1$ subset was used as the proper training set for model building. We fitted a logistic regression model on the proper training set and made predictions on both the calibration and the testing sets. A nonconformity measure (NCM; Fig. 1 and Methods) was calculated for every individual in the calibration and testing sets. Assuming that the training and the calibration sets were independent and identically distributed, we ranked the NCMS in the calibration set for both the case and control groups, respectively. Based on NCMS in the calibration set, probability values for assigning case or control labels to reach individuals in the testing set were then computed (equation [1]). We repeated this procedure $n$ times, using each of the $n$ subsamples exactly once as the calibration set. The $n$ probability values were averaged to produce a single estimation for the final prediction region of predicted subjects.

$$p_{ij} = \left[ \left\{ j = 1, \ldots, N_{\text{cal}} : y_j = y, \text{NCM}_j \geq \text{NCM}_{ij} \right\} \right] / (N_{\text{cal}} + 1 : y_j = y)$$

(1)

where $p_{ij}$ is the probability value of individual $i$ to be in class $y$ and $N_{\text{cal}}$ is the sample size of class $y$ in the calibration set.

In binary classifications, the probability values for assigning an individual as a case ($p_1$) and a control ($p_0$) are obtained from the MCCP, respectively. Given a prediction error rate of $\alpha$, a subject can be predicted by MCCP as a case ($p_1 > \alpha$ and $p_0 \leq \alpha$) or a control ($p_0 > \alpha$ and $p_1 \leq \alpha$), uncertain ($p_1 > \alpha$ and $p_0 > \alpha$) or unpredictable ($p_1 \leq \alpha$ and $p_0 \leq \alpha$) with a confidence level of $1-\alpha$.

The prediction coverage is defined as the proportion of samples predicted as case or control at the given error rate $\alpha$. To assess the clinical significance of the MCCP results versus standard methods, we computed the AUC, positive predictive value (PPV), and negative predictive value (NPV) restricted to predicted cases and controls at an error rate $\alpha$.

**Prediction error and coverage of MCCP.** We examined the calibration property of MCCP on prediction error for PRS across a range of simulated genetic architectures and case-control GWAS designs based on real genotypes (Methods and Supplementary Figure 1). For each simulated data set, MCCP and a simple logistic regression (LR) model were applied to the proper training and calibration sets. The predicted errors for the two models were compared in the hold-off testing set (Methods). In line with previous studies on CP, the prediction errors from MCCP were perfectly aligned with those expected across all simulated scenarios (Fig. 2 and Supplementary Figures 1–4). However, the naively implemented LR model underestimated the error rate when it was small; but, tended to overestimate when it was large (Fig. 2). Moreover, such biases varied with different genetic architectures. For example, for a fixed polygenicity of 0.01 and prevalence of 0.01, the LR models consistently underestimated the true error rate for heritability of 0.8, and, for a fixed heritability of 0.5 when the true error rate was < 0.5. A similar finding was found for cases with a low polygenicity (i.e., 0.001).
These results suggested that MCCP was superior in calibrating PRS prediction compared with the naive LR model.

Next, we studied the coverage of MCCP—how many subjects can be predicted at a given error rate—in our simulated data set (Supplementary Fig. 5). We observed that, as expected, the maximum coverage was achieved with a low error rate when the heritability was high. This was also true for data sets with low polygenicity for a fixed prevalence and heritability. Low prevalence tended to achieve high coverage at a lower error rate compared with high prevalence. This was mainly caused by the inherent imbalance of case versus control in the data. Moreover, as expected, larger GWAS discovery studies, which can generate more accurate effect estimates than smaller ones, always showed better prediction performance. Performance of MCCP using fivefold cross-validation at an error rate of 0.05 is given in Table 1.

Applications of MCCP to complex diseases. We used MCCP to evaluate the capacity of PRS in predicting risk for four common complex diseases (CAD, T2D, IBD, and BRCA; \( N = 276,299 \)) from the UK Biobank\(^{23,24} \), and SCZ from a Danish population study (iPSYCH, \( N = 24,072 \))\(^{25} \) using MCCP (Table 2). As for simulated data sets, MCCP predictions were well-calibrated across all studied diseases (Supplementary Fig. 6).

A commonly used approach for making decisions in trait prediction is contrasting the top \( x \)% with the bottom \( x \)% of PRS (termed empirical method). We compared the performance of MCCP to that of the empirical method in disease risk prediction with a fixed coverage, i.e., the number of predictable subjects in a testing sample. As the empirical method does not formally use the concept of coverage, we defined it as two times the extreme \( x \)%.

Using PRS as the only predictor, we found that MCCP performed marginally better than the empirical method in all studied diseases at error rates ranging from 0 to 0.2 (Fig. 3). A key advantage of using MCCP is that it can estimate confidence bounds for individual prediction. Based on these individualized predictions, MCCP can help clinicians make a decision at an error rate \( \alpha \).

We, next, evaluated the performance of MCCP when age, sex, and genetic ancestry information (computed by genetic principal components PC1–6, see Methods) were included in model building, calibration, and prediction steps. We found that MCCP outperformed the empirical method in studied data sets, which had varying case-control ratios from 1:4 to 1:111 (Fig. 4 and Supplementary Fig. 7). Such excellent performances were especially apparent for CAD, T2D, and SCZ. As expected, AUC, PPV, and NPV from MCCP decreased with increased coverage and with a decreasing preset confidence level. At an error rate of 0.05, MCCP can predict 35.2% of subjects for CAD, 22.7% for T2D, 15.5% for IBD, 19.0% for BRCA from the UK Biobank, and 31.4% of subjects of iPSYCH SCZ data set either as cases or controls. The prediction accuracies measured by AUCs were 0.865 (CAD, 95% CI 0.859–0.871), 0.788 (T2D, 95% CI 0.779–0.796), 0.705 (IBD, 95% CI 0.678–0.732), 0.754 (BRCA, 95% CI 0.742–0.766) and 0.842 (SCZ, 95% CI 0.828–0.856), respectively (Supplementary Table 1). In contrast to MCCP, the empirical method at a coverage of 10% (i.e., top 5% vs bottom 5% of PRS), gave 0.01–0.11 less in AUCs. Importantly, the empirical method cannot report confidence levels for risk prediction.
Validation of MCCP prediction by a follow-up study. Because complex diseases have different ages of onset between individuals, some healthy controls at the time of prediction may develop the disease later in life, we further examined the performance of MCCP in prediction using the MDC T2D follow-up data set. At baseline, median ages for men, women, and both sexes were 59.1 years (interquartile range [IQR]: 53.1–64.6), 57.4 (IQR: 50.1–63.7), and 57.8 (IQR: 51.3–64.2). T2D status of subjects in this data set was followed for 20 years and assessed at five time points, i.e., baseline, year 5, 10, 15, and 20. We estimated the prediction probability values at baseline using MCCP and the empirical method, and then, evaluated the performance of these predictions at each of the four follow-ups, using the corresponding T2D status or censored. In line with the results for the cross-sectional T2D data set from UK Biobank, MCCP slightly outperformed the empirical method when using PRS as the only predictor (Fig. 3 and Supplementary Fig. 8) but outperformed the empirical method strikingly at baseline using MDC data set (Fig. 4 and Supplementary Fig. 7) when age, sex, and genetic PC were considered in MCCP. When these covariates were considered in MCCP, the improved performances of MCCP also hold in the follow-up years 5, 10, 15, and 20 (Supplementary Fig. 9). Not surprisingly, MCCP also showed improved PPVs and decreasing NPVs along with follow-ups (Supplementary Figs. 10 and 11). Although a decreasing AUC with follow-up time was observed for MCCP, it still performed better than the empirical method at all studied time points (Supplementary Figs. 9a, 10a, and 11a). We also trained and applied MCCP and the empirical method at each time point of follow-ups. Results from these models consistently suggested a better performance of MCCP than that from the empirical method (Supplementary Figs. 9b, 10b, and 11b).

Discussion
In the present study, we introduce the machine-learning method MCCP as a calibration tool within the current polygenic risk prediction methodology. As a proof of concept, we show that MCCP is capable of estimating confidence levels for an individual’s predicted risk. The estimates obtained with MCCP enjoy the validity property, i.e., in long run, the prediction errors are guaranteed to fall below a preset error rate. The estimated probability values for hypothetically assigning an individual as a case or control will have more direct utility in clinics than group-wise estimates, which arbitrarily define the top 10%, 5%, or 1% of samples as the high-risk group, and, similarly, the bottom x% as a healthy group. In addition, with a preset error rate, e.g., α = 0.05, MCCP can systematically estimate the proportion of samples that can be predicted with an error rate below α. Such functionality is extremely valuable given that the performance of PRS depends on genetic architectures of complex diseases, which are typically unknown. For example, with a desired prediction error rate of 0.05, for one disease, 10% of samples can be reliably predicted by MCCP; but for others, only 5% may be predictable. In its current form, PRS is designed to predict risk at the population or group level, whereas ideally, one would want to know the individual’s susceptibility to disease. Thus, MCCP could aid clinicians in diagnosis by telling the prediction confidence for each individual.

Human diseases generally vary in prevalence, which in turn vary in time and across populations. Differences in prevalence between a reference GWAS sample and the tested population need to be considered before choosing the approach to use, such as defining the top x% of PRS in the tested population as the risk group. Our MCCP model can systematically overcome a situation where prevalence between GWAS sample and the tested population is different. Using a calibration set, MCCP estimates personalized confidences of prediction being case and control status, respectively. As such, MCCP can provide the users confidences on how much of their sample can be accurately predicted, no matter case or controls. Knowing the probability of being healthy (low risk of developing a disease) is equally important for the sake of public health management. In the present analysis, we observed a well-calibrated confidence level for predictions of common disorders in UK biobank data sets, SCZ in a Danish data set, and T2D in a Swedish data set. We also explored the proportion of samples that can be predicted within a given error rate. Thus, using confidence of prediction at various expected error rates to make decisions seems feasible.

Another advantage of MCCP is that it is indifferent to the bias of an imbalanced case-control study design. As the percentage of years lived with disability (% YLDs) for most complex diseases or conditions are <10% [53], data imbalance is common in population-based studies. Such study design typically results in a
majority class and a minority class in the sample. In extreme cases, even a simple predictor that treats all samples as the majority class can have good accuracy but fail to predict in the minority class. MCCP handles imbalanced data without the need to consider explicit balancing measures, such as over- or undersampling. Our simulations and real data applications for complex diseases showed that MCCP performs remarkably well and its validity was preserved even when the data were severely imbalanced, e.g., with prevalence ranging from 0.01 to 0.2, which reflected an imbalance level ranging from 1:99 to 1:4.

We demonstrated that predictions made by MCCP reflected the lifetime probability of being case or control using the MDC follow-up studies. As individuals typically show varying ages of onset for a specific disease, some healthy controls at the time of recruitment in a study may later develop such disease. Assigning probability values to an individual of being case or control at baseline using MCCP allows us to identify high-risk individuals. Meanwhile, such functionality automatically makes the MCCP prediction testable in the future. Our application to the 20 years follow-up data set from MDC showed that the prediction made by MCCP became more accurate as follow-up continues, confirming the applicability of MCCP. In addition, directly applying MCCP at each follow-up time point further improves the prediction performance both in prediction error rate and sample

Table 1 Performances of Mondrian cross-conformal prediction at an error of 0.05 on various genetic architecture by simulation (proportion of discovery set for GWAS is 0.5).

| Heritability | Polygenicity | Prevalence | AUC (95% CI) | PPV (95% CI) | NPV (95% CI) | Coverage |
|-------------|--------------|------------|--------------|--------------|--------------|-----------|
| 0.3         | 0.001        | 0.01       | 0.84 (0.80–0.88) | 0.48 (0.44–0.53) | 0.96 (0.94–0.98) | 0.18      |
| 0.3         | 0.05         | 0.05       | 0.76 (0.73–0.79) | 0.75 (0.72–0.81) | 0.67 (0.61–0.72) | 0.17      |
| 0.3         | 0.2          | 0.2        | 0.78 (0.71–0.84) | 0.25 (0.19–0.35) | 0.97 (0.95–0.99) | 0.16      |
| 0.3         | 0.01         | 0.01       | 0.75 (0.71–0.79) | 0.61 (0.57–0.65) | 0.82 (0.79–0.85) | 0.18      |
| 0.3         | 0.05         | 0.05       | 0.70 (0.70–0.77) | 0.69 (0.66–0.73) | 0.74 (0.71–0.78) | 0.17      |
| 0.3         | 0.2          | 0.2        | 0.67 (0.58–0.76) | 0.18 (0.10–0.33) | 0.95 (0.92–0.99) | 0.12      |
| 0.3         | 0.01         | 0.01       | 0.67 (0.62–0.71) | 0.57 (0.52–0.63) | 0.74 (0.70–0.78) | 0.14      |
| 0.3         | 0.05         | 0.05       | 0.69 (0.55–0.64) | 0.65 (0.59–0.73) | 0.56 (0.53–0.61) | 0.12      |
| 0.3         | 0.2          | 0.2        | 0.68 (0.60–0.76) | 0.24 (0.20–0.29) | 0.94 (0.91–0.97) | 0.11      |
| 0.5         | 0.001        | 0.01       | 0.81 (0.81–0.90) | 0.39 (0.34–0.44) | 0.98 (0.96–0.99) | 0.23      |
| 0.5         | 0.05         | 0.05       | 0.80 (0.77–0.82) | 0.77 (0.74–0.80) | 0.79 (0.76–0.82) | 0.21      |
| 0.5         | 0.2          | 0.2        | 0.70 (0.66–0.73) | 0.68 (0.64–0.72) | 0.71 (0.67–0.74) | 0.15      |
| 0.5         | 0.01         | 0.01       | 0.78 (0.73–0.85) | 0.72 (0.69–0.76) | 0.88 (0.86–0.90) | 0.26      |
| 0.5         | 0.05         | 0.05       | 0.75 (0.69–0.81) | 0.28 (0.24–0.33) | 0.96 (0.94–0.98) | 0.15      |
| 0.5         | 0.2          | 0.2        | 0.80 (0.77–0.83) | 0.69 (0.66–0.73) | 0.83 (0.81–0.86) | 0.21      |
| 0.8         | 0.001        | 0.01       | 0.81 (0.79–0.89) | 0.80 (0.60–0.89) | 0.75 (0.60–0.79) | 0.22      |
| 0.8         | 0.05         | 0.05       | 0.78 (0.67–0.85) | 0.65 (0.61–0.70) | 0.68 (0.64–0.72) | 0.14      |
| 0.8         | 0.2          | 0.2        | 0.82 (0.66–0.73) | 0.68 (0.64–0.72) | 0.71 (0.67–0.74) | 0.15      |

Table 2 Description of complex diseases.

| Disease | Discovery GWAS | Discovery GWAS sample size (#case/#total) | Prevalence in testing data set (#case/#total) | #Polymorphisms used in PRS construction |
|---------|----------------|-------------------------------------------|---------------------------------------------|----------------------------------------|
| CAD     | Nikpay et al. 43 | 60,801/123,504 | 13,689/276,299 (5.0%) | 9912 |
| T2D     | Scott et al. 44 | 26,676/132,532 | 15,006/276,299 (5.4%) | 19,054 |
| IBD     | Liu et al. 45 | 12,882/21,770 | 247/276,299 (0.9%) | 10,878 |
| BCCA    | Michaïlou et al. 46 | 122,977/105,974 | 455/147,317 (3.1%) | 28,945 |
| SCZ     | PGC 47 | 35,642/111,748 | 5125/24,072 (21.3%) | 31,755 |
| T2D (MDC) | Mahajan et al. 42 | 74,124/824,006 | 943/24,298 (3.9%) | 126,748 |

*Restricted to European unrelated participants from the UK biobank.
*Participants from the Danish sub-cohorts were removed.
*European unrelated participants from the Integrative Psychiatric Research (PSYCH) schizophrenia sample.
*European unrelated participants from the Malmo Diet and Cancer (MDC) study at baseline.

CAD coronary artery disease, T2D type 2 diabetes mellitus, IBD inflammatory bowel disease, BCCA breast cancer, restricted to women in testing data set, SCZ schizophrenia, T2D (MDC) T2D data set from the MDC study at the baseline.
Fig. 3 Comparison of the performance of MCCP and the empirical method on complex disease risk prediction using PRS as predictor. For MCCP, sample coverage (x axis) indicates the proportion of samples predicted as cases or controls, whereas, for the empirical method, it indicates extreme PRS, e.g., top and bottom x% of PRS. AUCs are computed from multivariate logistic regressions adjusted for age, sex, and PC1–6 on these samples stratified from MCCP and empirical method, respectively. The expected error rates for MCCP are indicated by the size of data points up to 0.20. Vertical lines correspond to an expected error of 0.05 from the MCCP. The solid lines and shades represent the median and 95% confidence intervals of AUCs. CAD coronary artery disease, T2D type 2 diabetes mellitus, IBD inflammatory bowel disease, BRCA breast cancer, SCZ schizophrenia, T2D (MDC) T2D data set from the MDC study at the baseline. Source data are provided as a Source Data file.

Fig. 4 Comparison of the performance of MCCP and the empirical method on complex disease risk prediction using PRS and additional information (age, sex, and PC1–6) in MCCP. For MCCP, sample coverage (x axis) indicates the proportion of samples predicted as cases or controls, whereas, for the empirical method, it indicates extreme PRS, e.g., top and bottom x% of PRS. AUCs are computed from multivariate logistic regressions adjusted for age, sex, and PC1–6 on these samples stratified from MCCP and empirical method, respectively. The expected error rates for MCCP are indicated by the size of data points up to 0.20. Vertical lines correspond to an expected error of 0.05 from the MCCP. The solid lines and shades represent the median and 95% confidence intervals of AUCs. CAD coronary artery disease, T2D type 2 diabetes mellitus, IBD inflammatory bowel disease, BRCA breast cancer, SCZ schizophrenia, T2D (MDC) T2D data set from the MDC study at the baseline. Source data are provided as a Source Data file.
coverage. Surprisingly, these crucial functions have not been formally developed by the empirical method.

We implemented MCCP onto PRS as an extension to the polygenic risk prediction paradigm such that advances in constructing PRS will also improve the performance of MCCP in personalized risk prediction. It has been shown that directly modeling linkage disequilibrium structure among SNPs and combining PRS constructed from different diseases or traits may improve the predictability of PRS34–36. Thus, we hypothesize that incorporating these advanced PRS models into our implementation will further improve the performance of MCCP. Moreover, the major improvement achieved by MCCP is the calibration step. Thus, the performance of MCCP could also be improved by constructing a disease risk score combining both genetic and non-genetic factors, such as lifestyle, socioeconomic status, and environmental exposures. We used logistic regression as our underlying model to construct a nonconformal measure (NCM). Other machine-learning models, such as k-nearest neighbors, random forest, support vector machine, or deep neural networks, can also be used to improve the overall power of MCCP in risk prediction.

In this work, we evaluated the performance of MCCP on samples of European ancestry. However, the application of our method in trans-ethnicity (e.g., training on European population and testing on Africa or Asian population) should be an interesting next step.

In conclusion, we described an approach for personalized genetic risk assessment of complex diseases. By estimating personalized confidences of risk prediction, it can help clinical professionals to assess the value of genetic data in disease risk prediction.

**Methods**

**Samples for UK Biobank, iPSYCH, and MDC.** The UK Biobank project is a prospective cohort study, comprising of ~500,000 individuals from the United Kingdom aged between 40 and 69 at recruitment23,24. Participants were genotyped using the Affymetrix UK BiLEVE Axiom array and the Affymetrix UK Biobank Axiom array, respectively. Quality control, ancestral origins, and cryptic relatedness were described elsewhere37. Phenotypes of CAD, T2D, and BRCA were retrieved using ICD9/10 codes, operation and procedure codes from hospital inpatient records (UK Biobank fields 41270, 41271, and 41272), as well as reported medical conditions and procedures (UK Biobank fields 20001, 20002, 20004). After standard quality, 276,299 unrelated participants with European ancestry remained for further analysis. The UK Biobank received approval from the National Information Governance Board for Health and Social Care and the National Health Service North West Center for Research Ethics Committee (Ref: 11/NW/0382). This research has been conducted using the UK Biobank Resource under application number 32048.

A detailed description of the iPSYCH cohort has been reported elsewhere38. In brief, the iPSYCH is a representative sample of the entire Danish population born between 1981 and 2005, including 1,472,762 subjects. Initial genotyping was performed at the Broad Institute with amplified DNA extracted from dried blood spots and assayed on the Infinium PsychChip v1.0 array. SNPs were phased into haplotypes using SHAPEIT3 and imputed using impute2 with European reference haplotypes from the 1000 genomes project phase 339. Individuals were censored to ensure no pair has closer than third-degree kinship. In total, 5125 cases of SCZ and 24,298, 23,211, 21,936, 20,247, and 16,146, respectively. Here, we implemented MCCP in two ways: (1) model building at baseline, and prediction at baseline and re-examined at each of the four follow-ups with respective T2D status. (2) model building and prediction at five time points, respectively.

MDC was approved by the Ethics Committee of Lund University (LU 51–90) and all patients provided written informed consent. The use of data in this study was approved by the MDC steering Committee.

**Simulation of different genetic architectures.** To examine the applicability of MCCP in various situations of GWAS design, we simulated case-control phenotypes based on real genotypes on chromosome 2 of 10,000 individuals from the iPSYCH project. The synthetic phenotypes were generated by combinations of varying polygenicity of 0.001, 0.01, 0.1, and 0.2, the heritability of 0.3, 0.5, and 0.8, and prevalence of 0.01, 0.05, 0.1, and 0.2 using the GCTA software38. The prevalence was chosen based on the reported prevalence of the considered diseases35. GWAS was performed using sample proportions of 0.2, 0.5, and 0.7 as discovery data sets. Consequently, PRS was obtained by a thresholding (p < 0.05) and pruning (LD r² < 0.1) approach. Logistic regressions within the MCCP setting were performed in the calibration and testing samples.

**PRS construction.** PRS was computed as the weighted sum of effect for the pruned SNPs (LD r² < 0.1) with MAF ≥ 0.01 in the target individuals. Indels and variants in the extended MHC region (build hg19, chromosome 6: 25–34 Mb) were removed. Effect sizes were taken from the reference GWAS or from the discovery set in simulation studies. In this study, summary statistics were from Psychiatric Genomics Consortium phase 2 without the Danish sub-cohorts3 and DIAGRAM42 (2018) for iPSYCH SCZ and MDC T2D data sets (MAF ≥ 0.05), respectively. For CAD, T2D, IBD, and BRCA from UK Biobank, we used summary statistics from respective GWAS studies43–46 where UK Biobank samples were not included, same to a recent study47. PLINK47 was used to construct PRS with the following parameters: p value < 0.05 and r² threshold of 0.1 within a window size of 100 kb and step size of 50 bp. And the LD structure from the 1000 Genomes Project phase 3 European subpopulation was used for LD pruning.

**Logistic regression.** As proof of principle, we built models based on simple LR to make a prediction. In simulation studies, LR was performed by simply using PRS alone as an independent variable. In the real-world clinical studies, LR was performed using PRS as an independent variable and age, sex, genotyping batches, and the first six PCs of population structures as covariates.

**CP and MCCP.** CP estimates the confidence of predicting the class y to a new object x, given a training set of z(x₁, y₁), z(x₂, y₂), ..., z(xᵢ, yᵢ), where xᵢ is generally a vector and yᵢ is a class label indicating the class to which the xᵢ belongs. A measurable function (equation [2]) quantifies how unusual (non-conformal) the xᵢ is in comparison with the training set,

\[ NCM = -y \cdot \delta(x) \]  

where NCM is nonconformity measure, y is the all-possible nonzero classes, e.g., (1, −1) and δ(x) is the decision value obtained from the decision function of the fitted model, e.g., LR in the present study. As in the definition of the inductive conformal prediction of CP, the training set is further split into a proper training set ((x₁,...,xᵢ) and a calibration set set ((xᵢ+1,...,xᵢ+n), where i is less than n. NCM is calculated for both calibration and test sets based on the model trained on the proper training set. We assume that both training and calibration sets are independent and identically distributed. Using the only calibration set alone, the probability value of an individual i to be the class y is calculated as follows

\[ pᵢ = \frac{1}{N_{cal} + 1} \sum_{i=1}^{N_{cal}} \delta(x) \]  

Note that all samples in the calibration set are included to compute the probability values for all possible classes. This may be problematic when the data are imbalanced. To address this issue in the classification model, the MCCP was introduced. As shown in Fig. 3, MCCP restricts Eq. 3 with the calibration set within sample class (e.g., case or control alone in binary classifications). By choosing an expected error α ∈ [0, 1], for every test sample, a
predicting region outputs the following:

$$\Gamma^* = \{ y \in Y : p^* > \alpha \}$$  (4)

where Y is the set of possible classes, \(p_i\) is the probability value when an individual is in a class. In the binary classification problem, we set \(p_i\) as the probability value when an individual is in one class, e.g., case, and \(p_o\) as another class, e.g., control. Unlike other classification frameworks, where a prediction is always provided and is a unique class, the prediction region \(\Gamma^*\) is a set and it can be empty or contain one or more classes. Confidence, credibility, and prediction of MCCP are defined as follows: confidence: \(\text{sup}[1 - \alpha : \Gamma^* \subseteq \{1\}]\), that is the greatest 1 - \(\alpha\) for which \(\Gamma^*\) is a unique class, e.g., case or control. In the problem of binary classifications, it is also equal to 1 - \(\min(p_o, p_i)\).

Credibility: \(\text{sup}[\alpha: \Gamma^* = \{1\}]\), i.e., the smallest \(\alpha\) for which \(\Gamma^*\) is empty. It is also equivalent to \(\min(p_o, p_i)\) in the binary classifications.

Prediction: \(\Gamma^*\) when 1 - \(\alpha\) is equal to the confidence, i.e., \(\alpha = \min(p_o, p_i)\) in binary classifications.

As an example of interpretation, given an output \(p^*_0\) of 0.01 and \(p^*_1\) of 0.8 for a test individual, the individual will be predicted to be a case with a credibility of 0.8 and confidence of 0.99.

Calibration assessment. A reliability curve (observed error versus expected error) was used to assess calibrations by MCCP and an LR model. First, we divided data into five folds. MCCP and LR were built on four folds and used to make predictions on the remaining fold. This procedure was repeated five times to make sure all samples were covered. The observed error was measured as the proportion of incorrect predictions against true cases or controls status at an expected error rate using the estimated MCCP probability values and probabilities from the LR, respectively. At a given error rate (so-called expected error rate) \(\alpha\), the observed error for LR and MCCP was computed by Eqs. (5) and (6), respectively.

$$\text{Observed}_{\text{err}}(\alpha) = \sum_{y \in \{1\}} (1 - \alpha) \times y$$  (5)

$$\text{Observed}_{\text{err}}(\text{MCCP}) = \sum_{y \in \{1\}} p^*_y \times y$$  (6)

Evaluation metrics. Performances of MCCP are measured by validity and coverage (also termed efficiency). A valid prediction means that the frequency of errors (i.e., the fraction of true values outside the prediction region) is no more than \(\alpha\) at a chosen error rate \(\alpha\). The validity can be calculated for all class objects as well as for objects of one specific class. Coverage is defined as the percentage of unique class \(\Gamma^*\), which is also the proportion of samples predicted as case or control, as shown in this study. Varying error rate \(\alpha\) from 0 to 1, observed error, coverage, AUC, PPV, and NPV are computed using LR adjusted for age, genetic sex, batches of genotyping arrays, and the first six PCs of population structures. Confidence intervals of AUC, PPV, and NPV were calculated using the pROC package within R.

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

All GWAS summary statistics used in this study are publicly available in the following repositories: Coronary Artery Disease Genomics, International Inflammation from 1000 Genomes Project can be accessed at ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/.

References

1. Buxton, A. et al. The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. Nucleic Acids Res. 47, D1005–D1012 (2019).

2. Easton, D. F. et al. Genome-wide association study identifies novel breast cancer susceptibility loci. Nature 447, 1087–1093 (2007).

3. Al Olama, A. A. et al. A meta-analysis of 87,040 individuals identifies 23 new susceptibility loci for prostate cancer. Nat. Genet. 46, 1103–1109 (2014).

4. Nelson, C. P. et al. Association analyses based on false discovery rate implicate new loci for coronary artery disease. Nat. Genet. 49, 1385–1391 (2017).

5. Schizophrenia Working Group of the Psychiatric Genomics Consortium. Biological insights from 108 schizophrenia-associated genetic loci. Nature 511, 421–427 (2014).

6. Sunjiangming/PRS_MCCP and are archived in Zenodo (https://doi.org/10.5281/zenodo.4661464#). Additional software PLINK v1.0b5.2, KING 2.2.4, GCTA v1.25.2 and R packages (glmnet,2.0-16, doParallel,1.0.14, foreach,1.4.4, caret,6.0-82, impute,1.56.0, rms,5.1-2, pROC1.16.2, Zeigl,5.1.7) were used in this study.

Received: 30 April 2020; Accepted: 16 July 2021; Published online: 06 September 2021
A list of members appears in the Supplementary Information.