


# Development and validation of a neural network-based survival model for mortality in ischemic heart disease

Peter C. Holm<sup>a</sup>, Amalie D. Haue<sup>a,b</sup>, David Westergaard<sup>a, c, d</sup>, Timo Röder<sup>a</sup>, Karina Banasik<sup>a,c</sup>, Vinicius Tragante<sup>e</sup>, Alex H. Christensen<sup>b,f,g</sup>, Laurent Thomas<sup>h,i</sup>, Therese H. Nøst<sup>i</sup>, Anne-Heidi Skogholt<sup>i</sup>, Kasper K. Iversen<sup>f,g</sup>, Frants Pedersen<sup>b,f</sup>, Dan E. Høfsten<sup>b,f</sup>, Ole B. Pedersen<sup>f,j</sup>, Sisse Rye Ostrowski<sup>f,k</sup>, Henrik Ullum<sup>f,k,l</sup>, Mette N. Svendsen<sup>m</sup>, Iben M. Gjødsbøl<sup>m</sup>, Thorarinn Gudnason<sup>n</sup>, Daníel F. Guðbjartsson<sup>e</sup>, Anna Helgadóttir<sup>e</sup>, Kristian Hveem<sup>i</sup>, Lars V. Kjøber<sup>b,f</sup>, Hilma Holm<sup>e</sup>, Kari Stefansson<sup>e,o</sup>, Søren Brunak<sup>a,p, </sup>, and Henning Bundgaard<sup>b,f</sup>

<sup>a</sup> Novo Nordisk Foundation Center for Protein Research, University of Copenhagen, Blegdamsvej 3B, DK-2200 Copenhagen, Denmark

<sup>b</sup> Department of Cardiology, The Heart Center, Copenhagen University Hospital, Rigshospitalet, Blegdamsvej 9, DK-2100 Copenhagen, Denmark

<sup>c</sup> Department Obstetrics and Gynecology, Copenhagen University Hospital, Kettegård Alle 30, DK-2650 Hvidovre, Denmark

<sup>d</sup> Methods and Analysis, Statistics Denmark, Sejrøgade 11, DK-2100 Copenhagen, Denmark

<sup>e</sup> deCODE genetics, Sturlugata 8, 102 Reykjavik, Iceland

<sup>f</sup> Department of Clinical Medicine, Faculty of Health and Medical Sciences, University of Copenhagen, Blegdamsvej 3B, DK-2200 Copenhagen, Denmark

<sup>g</sup> Department of Cardiology, Copenhagen University Hospital, Herlev-Gentofte Hospital, Borgmester Ib Juuls Vej 1, DK-2730 Herlev, Denmark

<sup>h</sup> Department of Clinical and Molecular Medicine, Norwegian University of Science and Technology, 7491 Trondheim, Norway

<sup>i</sup> K.G. Jebsen Center for Genetic Epidemiology, Department of Public Health, Norwegian University of Science and Technology, 7491 Trondheim, Norway

<sup>j</sup> Department of Clinical Immunology, Zealand University Hospital, DK-4600 Køge, Denmark

<sup>k</sup> Department of Clinical Immunology, Copenhagen University Hospital, Rigshospitalet, Blegdamsvej 9, DK-2100 Copenhagen, Denmark.

- 1 <sup>l</sup> Statens Serum Institut, Artillerivej 5, DK-2300 Copenhagen, Denmark
- 2 <sup>m</sup> Department of Public Health, University of Copenhagen, Øster Farimagsgade 5, DK-1353
- 3 Copenhagen, Denmark
- 4 <sup>n</sup> Laeknasetrid Cardiology Clinic, Thonglabakka 1, 109 Reykjavík, Iceland
- 5 <sup>o</sup> Faculty of Medicine, University of Iceland, Vatnsmyrarvegur 16, Reykjavik 101, Iceland
- 6 <sup>p</sup> Copenhagen University Hospital, Rigshospitalet, Blegdamsvej 9, DK-2100 Copenhagen,
- 7 Denmark
- 8 ✉ Correspondence: Søren Brunak (soren.brunak@cpr.ku.dk)

9

## Abstract

**Background:** Current risk prediction models for ischemic heart disease (IHD) use a limited set of established risk factors and are based on classical statistical techniques. Using machine-learning techniques and including a broader panel of features from electronic health records (EHRs) may improve prognostication.

**Objectives:** Developing and externally validating a neural network-based time-to-event model (PMHnet) for prediction of all-cause mortality in IHD.

**Methods:** We included 39,746 patients (training: 34,746, test: 5,000) with IHD from the Eastern Danish Heart Registry, who underwent coronary angiography (CAG) between 2006-2016. Clinical and genetic features were extracted from national registries, EHRs, and biobanks. The feature-selection process identified 584 features, including prior diagnosis and procedure codes, laboratory test results, and clinical measurements. Model performance was evaluated using time-dependent AUC (tdAUC) and the Brier score. PMHnet was benchmarked against GRACE Risk Score 2.0 (GRACE2.0), and externally validated using data from Iceland (n=8,287). Feature importance and model explainability were assessed using SHAP analysis.

**Findings:** On the test set, the tdAUC was 0.88 (95% CI 0.86-0.90, case count, cc=196) at six months, 0.88(0.86-0.90, cc=261) at one year, 0.84(0.82-0.86, cc=395) at three years, and 0.82(0.80-0.84, cc=763) at five years. On the same data, GRACE2.0 had a lower performance: 0.77 (0.73-0.80) at six months, 0.77(0.74-0.80) at one year, and 0.73(0.70-0.75) at three years. PMHnet showed similar performance in the Icelandic data.

1    **Conclusion:** PMHnet significantly improved survival prediction in patients with IHD compared  
2    to GRACE2.0. Our findings support the use of deep phenotypic data as precision medicine tools  
3    in modern healthcare systems.

4    **Keywords:** ischemic heart disease, prediction models, survival analysis, artificial intelligence,  
5    neural networks, GRACE

## 1   **Introduction**

2   In patients with ischemic heart disease (IHD), improved clinical application of the wide array of  
3   prognostic risk factors and disease markers may inform treatment options for the individual  
4   patient<sup>1-3</sup>. For example, the updated version of Global Registry of Acute Coronary Events  
5   (GRACE) score (GRACE2.0) received a class IIa recommendation for assessing risk and  
6   management of patients with non-ST-elevation myocardial infarction (nSTEMI) in the 2020  
7   European Society of Cardiology (ESC) guidelines<sup>1</sup>. However, GRACE2.0 and other traditional  
8   risk scoring schemes in IHD such as Framingham and Thrombolysis in Myocardial Infarction  
9   (TIMI) use a limited set of input features (<10) and likely underutilize most data available in  
10   modern electronic health records (EHRs)<sup>4-7</sup>.

11   Integrating a richer set of input features could be overcome with machine learning (ML) models  
12   such as neural networks. These can capture non-linear interactions without the need for imputation  
13   of missing data or expert feature engineering<sup>8,9</sup>, leveraging the multitude of heterogeneous  
14   healthcare data stored in modern EHRs and national registries in the development of clinical  
15   decision support tools.

16   ML-based approaches have shown promising results for risk-estimation in cardiology with better  
17   performance than traditional models<sup>10-13</sup>. In patients with stable IHD, Motwani et al. showed that  
18   an ML algorithm combining clinical variables with imaging variables from coronary CT  
19   angiography predicted 5-year all-cause mortality better than models using clinical metrics alone<sup>11</sup>.  
20   Similarly, Mohammad and colleagues developed and validated a neural network to predict 1-year  
21   mortality and re-admission for heart failure after incident myocardial infarction with greater  
22   discrimination than the GRACE2.0 score<sup>12</sup>. However, the majority of the secondary risk prediction  
23   models based on ML have not used time-to-event analysis. One notable exception is the model

presented by Steele et al, which however have not been externally validated<sup>10</sup>. By not using survival analysis, previous ML-based analyses omit data points with incomplete follow-up (censoring) and thereby effectively prevents a model from distinguishing between “died after a week” and “died after 10 months” which we believe is of obvious clinical interest.

To overcome these limitations, we describe the development and validation of a neural network-based survival model, PMHnet, for predicting all-cause mortality in patients with IHD using 584 different features extracted from population-wide healthcare registries and complete EHRs. We identified several influential features which previously have been omitted from risk prediction of patients with IHD.

## Methods

### Data foundation

The algorithm PMHnet was developed using a cohort constructed from the Danish National Patient Registry (NPR) and the Eastern Danish Heart Registry (EDHR)<sup>14</sup>. The EDHR contains structured information on all coronary artery angiographies (CAGs) performed in the Capital Region of Denmark and Region Zealand. The cohort was linked to a population-wide EHR database that covers Eastern Denmark from 1<sup>st</sup> of January 2006 to the 7<sup>th</sup> of July 2016 (BTH), and genotype data from the Copenhagen Hospital Biobank Cardiovascular Diseases study (CHB-CVD)<sup>15–17</sup>. The BTH dataset fully covered Eastern Denmark (2.6 million patients). Outcomes were obtained from the Central Person Registry and the Danish Register for Causes of Death<sup>18,19</sup>. Data sources were linked using encrypted Danish personal identification numbers<sup>19</sup>.

### Selection criteria and model development

First, we identified all adult Danish citizens (>18 years of age) in NPR with an ICD-10 code for IHD (I20-I25) who had undergone their first CAG between Jan 1, 2006, and Jun 1, 2016, demonstrating one-, two-, or three-vessel disease (1–3VD) or diffuse atheromatosis (DIF). Vascular disease is here defined as stenosis above 50%<sup>20</sup>. For patients fulfilling these criteria (n=39,746), we used the date of the CAG as the index date and included five years of follow-up. Patients were followed until either death or censoring, whichever came first. Using the hold-out method, the derivation data was randomly divided into a training set (n=34,746) and a test set (n=5,000) used for model development and independent assessment of performance, respectively<sup>21</sup> (Figure 1, Table 1).

For each of the 39,746 patients, we reduced the available features prior to index event (i.e. first CAG between Jan 1, 2006, and Jun 1, 2016) to a smaller set based on prevalence such that e.g., a diagnosis code could be found in at least 5% of the training set. The final set of 584 features was separated into five different categories: *ClinicalOne* (8 features), *ClinicalTwo* (15 features), *Diagnoses* (322 features), *Procedures* (154 features), and *Biochemical* (85 features) (Table 2). *ClinicalOne* included the same eight input features as used by GRACE2.0 and *ClinicalTwo* had 14 additional clinical features (Table 2) that were selected based on availability. Features were defined using data recorded prior to the index date, except for creatinine, cardiac biomarkers for ischemic heart disease, and blood pressure where high missingness led us to allow measurements obtained after the CAG in cases of missingness (7-day threshold for cardiac biomarkers, and 21-day threshold for the others). *Diagnoses* included ICD-10 codes registered in NPR and similarly, *Procedures* consisted of procedure codes (surgery and examinations such as X-rays) registered in NPR. *Biochemical* contained results of in-hospital blood tests. Additional details on feature extraction, missingness, pre-processing, encoding, and categories can be found in supplementary methods. The amounts of missingness across the training and test set have been tabulated in Table S1. Feature categories and encodings are available in the appendices. Missing values were left missing and encoded as such in PMHnet; meaning that for categorial variables all values were zero in case of missingness and for continuous variables missing values were assigned the mean variable (details are available in the Supplementary Material).

## **External validation data from Iceland**

For the external validation cohort, we identified Icelandic adults who had undergone CAG at the only interventional cardiology center in Iceland, Landspítali– The National University Hospital in Reykjavík<sup>22</sup>. We obtained data collected prospectively between January 1, 2007, and December



31, 2017. Information on ICD-10 diagnoses and procedure codes were aggregated from the Landspítali, from registers kept by the Directorate of Health: The Register of Primary Health Contacts, the Register of Contacts with Medical Specialists in Private Practice and the Causes of Death Register, as well as at recruitment for deCODE studies. Biochemical assay measurements were obtained from the three largest clinical laboratories in Iceland, with measurements performed at: (i) Landspítali; (ii) The Laboratory in Mjódd, Reykjavík, Iceland; and (iii) Akureyri Hospital, the regional hospital in North Iceland.

## **Polygenic risk scores**

Polygenic risk scores (PRSs) for patients with genotypes available through the CHB-CVD<sup>17,23</sup> (37.4% of the cohort) were calculated using the LDpred2 framework, implemented in the R package bigsnpr (v1.5.2) with R version 3.5.052<sup>24</sup>. PRSs were calculated based on GWAS summary statistics data from 19 traits relevant for cardiometabolic health, obtained from 17 GWAS meta-analyses. List of meta-analyses and details on the PRS calculations is included in the Supplementary Material.

## **Machine learning model architecture and development**

To model time-to-event data and allow for censoring, we used the generic discrete-time survival model for neural networks described by Gensheimer and Narasimhan<sup>25</sup>. In this model, follow-up time is divided into a fixed number of intervals and the model estimates a conditional hazard for each interval, i.e., the probability of dying in that time interval given that the patient is still alive at the end of the preceding interval. PMHnet uses 30 intervals separated in time such that event times in the training data are evenly distributed across all intervals. To obtain predictions between breakpoints in the discretization grid, we assumed that the probability density function was

constant in each time interval, and we thus interpolated using a piecewise linear function<sup>26</sup>. The implementation applied the PyTorch machine-learning framework using the authors' Keras version as a reference<sup>25</sup>.

We used a feed-forward neural network and tested various hyperparameters. The output layer was a fully connected sigmoid activated layer that outputs conditional hazards for each of the 30 different time points. We added dropout to each of the hidden layers to regularize the network and prevent over-fitting. The number of layers, neurons, learning-rate, and dropout rate for each layer were fine-tuned through hyperparameter optimization using the Optuna optimization framework, with a five-fold cross validation<sup>27</sup>. The hyperparameter search space and the best trial for the complete model is included in table S3. The neural networks were trained using stochastic gradient descent, with a constant learning rate, to minimize the negative log-likelihood.

## **Model evaluation and validation**

Using the hold-out test set and the external validation data from Iceland, performance of PMHnet was evaluated through assessment of both model discrimination and calibration. We used time-dependent area under the receiver operating characteristic curve (tdAUC) as the main measure of discrimination, but also calculated the Brier score that can be used to assess both discrimination and calibration<sup>28-30</sup>. Calibration was also analyzed graphically by comparing the predicted risks with the estimated actual risks<sup>28</sup>. The Score function from the riskRegression R package was used to compute performance measures and compare models. For comparisons between two competing models, the Score function gives p-values that correspond to Wald tests on the standard errors obtained using an estimate of the influence functions following Blanche et al.<sup>30</sup>.

To benchmark PMHnet we calculated the GRACE2.0 for all patients in the hold-out test set, using the GRACE2.0 webtool. We extracted the javascript source code for the GRACE2.0 webtool using the developer tools in Google Chrome. The javascript code was then manually converted to an R package for automatized computation of GRACE2.0 on the entire cohort. The eight variables used in GRACE2.0 were available for 51.4% of the cohort. Since GRACE2.0 does not allow for missing features, we imputed missing variables using the missForest R package<sup>31</sup>. Imputed values were only used for calculating the GRACE2.0 score. In addition to the conventional GRACE2.0 score, we re-fitted the GRACE2.0 score to our training data using the PMHnet architecture. This corresponds to model\_2 in Figure 4 that only uses the features from *ClinicalOne* as its input.

For independent validation of PMHnet we used, as described above, Icelandic EHR data from 8,287 patients. Of the 584 features identified in the Danish derivation cohort, we found matching data for 404 features in the Icelandic data. A down-scaled model was re-trained on the Danish training set to make comparisons.

## **Explainability and effect of missing features**

To investigate the impact of different features on model predictions and to provide model explanations, we calculated Shapley additive explanation (SHAP) values for all features and patients in the training set using the SHAP python package<sup>32</sup>. In the model explanations, a negative SHAP value for a given feature means that the feature pulls the prediction towards mortality and vice versa for positive SHAP values relative to the median prediction. The magnitude of the SHAP value is percentage points.

To assess how resilient the model was in the event of missing data, missingness was introduced in the test data by replacing all values of a given feature with the median value of that feature in the

training data. The predictions were then compared with the predictions of PMHnet. Resiliency was then quantified using change in tdAUC (discrimination) and Brier scores (calibration), where the predictions of PMHnet were compared to that of model with artificially introduced values (a total of 584 comparisons).

As the network was not trained on the Icelandic data before external validation, no SHAP analysis was performed on the predictions on the Icelandic data.

## **Statistical analysis**

Categorical features are reported as counts (%) and continuous features as mean [95% CI], 95% CIs are obtained from standard deviations or through bootstrapping. Time-dependent AUCs and Brier scores were calculated using the `riskRegression` R-package<sup>28</sup>. Likewise, model comparisons were obtained from the same package. All statistical analyses and visualizations were performed using R version 4.1.

## **Data access and ethics approvals**

The study was approved by The National Ethics Committee (1708829, ‘Genetics of CVD’—a genome-wide association study on repository samples from CHB), The Danish Data Protection Agency (ref: 514-0255/18-3000, 514-0254/18-3000, SUND-2016-50), The Danish Health Data Authority (ref: FSEID-00003724 and FSEID-00003092), and The Danish Patient Safety Authority (3-3013-1731/1/). Danish personal identifiers were pseudonymised prior to any analysis.

The study was approved by the Data Protection Authority of Iceland and the National Bioethics Committee of Iceland (VSN-15-114). Icelandic participants that donated biological samples

provided informed consent. Personal identities of the participants were encrypted with a third-party system provided by the Data Protection Authority of Iceland.

Study design, methods, and results were reported in agreement with the TRIPOD statement<sup>33,34</sup> and following the STROBE recommendations<sup>35</sup>.

## **Funding**

Novo Nordisk Foundation (grant agreements: NNF14CC0001 and NNF17OC0027594) – Hellerup, Denmark; NordForsk (*PM Heart*; grant agreement: 90580) – Oslo, Norge; and the Innovation Foundation (*BigTempHealth*; grant agreement: 5153-00002B) – Aarhus, Denmark.

## Results

The derivation cohort of 39,746 Danish patients with IHD were randomly subdivided into a training (N=34,746) and a test set (N=5,000) (Table 1). At inclusion the patients mean age (95%-CIs) was 66.0 years [65.7; 66.4] (67.3% males) in the training set and 66.2 years [66.0;66.3] (68.2% males) in the test set. The distribution of the degree of coronary artery disease was similar in the two groups (distributions of patients presenting with one-, two-, or three-vessel disease or diffuse atherosclerosis, respectively).

The Kaplan-Meier estimate of five-year survival (all-cause) was 81.8% [81.4; 82.2] for the training set and 82.5 [81.3; 83.6] for the test set (Figure S1, Table S2). The restricted mean follow-up time was 1,635 days ( $\pm 2.58$ ) for the training set and 1,635 days ( $\pm 6.49$ ) for the test set.

## PMHnet model predictions

In the internal validation using the hold-out test set, the complete PMHnet model had tdAUCs of 0.88 [0.86; 0.90] at six months (case count, cc=196); 0.88 [0.86; 0.90] at one year (cc=261), 0.84 [0.82; 0.86] at three years (cc=395) (Figure 2), and 0.82 [0.80; 0.84] at five years (cc=763). In comparison, the corresponding values for the conventional GRACE2.0 score on the same dataset were 0.77 [0.74; 0.80] at six months, 0.77 [0.74; 0.80] at one year, and 0.73 [0.71; 0.75] at three years. For the re-fitted GRACE2.0 score, tdAUCs were 0.79 [0.76; 0.83] at six months, 0.78 [0.75; 0.81] at one year, and 0.76 [0.74; 0.78] at three years. Since GRACE2.0 features had to be imputed for 48.6% of the population, we also evaluated the performance on the subset without missingness, which were largely the same (Figure 3, dashed line). GRACE2.0 is not designed for providing predictions after three years and was therefore not evaluated beyond that time-point. The

1 difference in tdAUCs between PMHnet and either of the two GRACE2.0 models was significant  
2 at each of the three prediction horizons (Table 3).

3 As an additional visual test of discrimination, we constructed five different risk strata using the 5-  
4 year predicted survival (defined as 90%, 75%, 50%, and 25%) and examined the observed survival  
5 (Figure 4), which showed good separation between the five strata. PMHnet was found to be well-  
6 calibrated as seen from Figure 2B and the calculated Brier scores of 3.2% [2.8; 3.6] at six months,  
7 4.1% [3.7; 4.5] at one year, 7.6% [7.1; 8.1] at three years, and 11.3% [10.5; 12.0] at five years.

8 In comparison, GRACE2.0 had Brier scores of 3.7% [3.3; 4.1] at six months, 4.9% [4.4; 5.4] at  
9 one year, 9.9% [9.3; 10.5] at three years. Re-fitting GRACE2.0 with the PMHnet architecture  
10 considerably improved the calibration as evident from the calibration curve which was comparable  
11 to that of PMHnet (Figure 2B). The differences in three-year predicted risk between PMHnet and  
12 the two GRACE2.0 scores for all patients in the test set are shown in Figure S2.

### 13 **External validation of PMHnet**

14 For external validation the cohort of 8,287 patients from Iceland was used (Table 1, *validation*  
15 *set*). Due to data availability a modified, and down-scaled version of PMHnet was used (404  
16 features) on the Icelandic data. The tdAUCs were 0.87 [0.84;0.90] at six months, 0.84 [0.81;0.87]  
17 at one year, and 0.81 [0.79;0.83] at three years. In comparison, the performance of the down-scaled  
18 version on the Danish (internal, hold-out) test set was 0.87 [0.85;0.90] at six months, 0.87  
19 [0.85;0.89] at one year, and 0.82 [0.80;0.85] at three years. The predictive performances were  
20 highly concordant in the Icelandic data, and the model maintained good calibration for the six  
21 months and one year predictions, but was found to be miscalibrated in the three year predictions  
22 (Figure S3).

Influence of the different input feature categories on the PMHnet predictionTo assess the importance of the different input feature categories systematically, we trained five intermediate survival models using the PMHnet architecture (Figure 3). For example, *model-1* used diagnosis codes as input features only, *model-2* used the clinical variables known from GRACE only. The other intermediate versions were trained using different combinations of the feature categories. Figure 3 shows the model discrimination at the three different prediction horizons for the six different models fitted using the PMHnet architecture. The performance of the intermediate model based on diagnosis codes only (*model-1*) was similar to the performance of the re-fitted GRACE2.0 model (*model-2*). Interestingly, combining diagnosis codes with the GRACE2.0 clinical features (*Diagnoses + ClinicalOne*) in *model-3*, there was an overall increase in AUC at three years, which would suggest a synergistic effect. With the addition of the *ClinicalTwo*-data, the  $\Delta$ tdAUC between *model-3* and *model-4* was  $3.4\text{e-}2$  [ $1.8\text{e-}2$ ;  $5.0\text{e-}2$ ] at six months,  $3.2\text{e-}2$  [ $1.8\text{e-}2$ ;  $4.7\text{e-}2$ ] at one year, and  $1.3\text{e-}2$  [ $0.3\text{e-}2$ ;  $2.3\text{e-}2$ ] at three years. Similarly, adding *Biochemical* to the input features in *model-5* associated with significant  $\Delta$ tdAUCs of  $2.1\text{e-}2$  [ $0.2\text{e-}2$ ;  $3.6\text{e-}2$ ] at six months,  $1.4\text{e-}2$  [ $0.2\text{e-}2$ ;  $2.7\text{e-}2$ ] at one year, and a non-significant  $\Delta$ tdAUC of  $0.9\text{e-}2$  [ $-0.1\text{e-}2$ ;  $1.9\text{e-}2$ ] at three years. The gain in discrimination from *model-5* to the complete *model-6* (PMHnet) by adding *Procedures* to the input features was only significant at the one-year prediction horizon.

### Testing inclusion of polygenic risk scores in PMHnet

For 37.4% of the cohort, we had genotype information available and used that to calculate 19 different polygenic risk scores (PRS) that all related to different cardiometabolic traits. Limiting both the training set and the test set to only individuals with genotype data (37.4%), we tested adding the PRS scores to *model-1* (*Diagnoses*), *model-2* (*Diagnoses + ClinicalOne*), and *model-6*



(*All Features*) and evaluated the model performance (Figure S4). Addition of the PRS did not significantly improve either model discrimination at any time-point (Figure S4A).

### **Explainability analysis using SHAP**

We performed SHAP-analyses of the five-year model predictions to quantify the impact of the different features on PMHnet predictions. SHAP is a technique rooted in cooperative game theory that provides an estimate of feature impact on the model output. It quantifies the contribution of each feature to the prediction outcome, allowing for a better understanding of feature importance and model behavior. By considering the interactions and dependencies between features, SHAP analysis provides insights into the specific factors influencing the model's decision-making process and aids in identifying key drivers and relationships within the model<sup>36</sup>. Across the five different feature categories, biochemical test results and diagnosis codes were most impactful, while the category *ClinicalOne* was least impactful (Figure 5A). The most impactful diagnosis code was chronic obstructive pulmonary disease (ICD10: J44) and thus ranked higher than classical risk factors for IHD, such as type 2 diabetes. At the feature level, age, the number of affected vessels (1–3VD, or DIF), and smoking were on-average the most predictive features. The top-25 features in terms of average model impact (average magnitude of SHAP-value), included ten features from *Biochemical*, nine from *ClinicalTwo*, three from *ClinicalOne*, two from *Diagnoses*, and one from *Procedures* (Figure 5A). To further examine the impact on model prediction for the two most impactful features, number of affected *vessels* and *age*, we constructed SHAP dependence plots (Figure 5B)<sup>37</sup>. For *age* we observed non surprisingly that higher age pulled the prediction towards non-survival, and that age was estimated to add/remove anywhere between -25 and 20 percent points to the predicted 5-year survival. Similarly for *vessels*, more affected vessels impacted the predicted survival negatively. For both features, we noted that identical feature values not always

1 impacted the survival to the same extent. This vertical dispersion in both plots represent interaction  
2 effects with the other included features<sup>37</sup>. Although our SHAP analyses reveal that many features  
3 have lower impact relative to the most impactful features, the aggregated sum of the many low-  
4 impact features (560 lowest) outweighs many of the well-known risk-factors (25 highest). For this  
5 reason, we did not attempt to limit the number of included features. Extending the analysis of  
6 feature-level model impact to the rest of the top-25 features, we constructed a summary plot of the  
7 SHAP-values that shows the distribution of model impacts across feature values (Figure 6).  
8 Interestingly, we see that for several features the knowledge that the feature is missing has a  
9 pronounced impact on the model predictions. As an example, we see that if an electrocardiogram  
10 has not been obtained (or recorded in the database) then the model predictions are pulled towards  
11 non-survival.

## 12 **Patient-level feature importance**

13 Finally, we also generated individual explanations for three example patients in the test set and  
14 show the SHAP values for the nine most impactful features (Figure 7). The patients were randomly  
15 selected from the subsets of patients with a prediction in the intervals (0.25; 0.5], (0.5, 0.75], and  
16 (0.75; 1]. For *patient 1*, with the worst 5-year prognosis, *age* was not among the nine most  
17 impactful features. Instead, the explainability algorithm highlights diagnosis codes and  
18 biochemical values. For *patient 2* with a 5-year predicted risk of 73%, the most impactful feature  
19 was *age* (78 years), which pulled the prediction towards mortality. The impact of *age* was however  
20 largely cancelled by *rest*, which constitutes the aggregated sum of all the features not among the  
21 nine most predictive. For *patient 3*, the feature *age* was again highlighted as the most impactful,  
22 but in this case, it impacted the prognosis positively. Apart for a history of cigarette smoking, none  
23 of the highlighted features had negative impact on the prognosis.

## Discussion

In this study, we developed a feature-rich neural network-based survival algorithm, PMHnet, for prediction of all-cause mortality in patients with IHD using data from 34,746 Danish patients. With the aim of providing predictions that can be used to guide treatment and care, the model was developed to operate with an index date immediately after the diagnosis-confirming coronary angiography and with a prediction horizon of five-years. The model was tested using data from 5,000 Danish patients and externally validated using data from 8,288 Icelandic patients. We found that PMHnet had excellent discrimination with tdAUCs ranging from 0.88 at six months to 0.82 at five years. Similar results were found on the external Icelandic data, which confirms that the model and its deep feature foundation generalized well to novel patients and a different healthcare setting. Evaluated on the Danish data, we found the model to be well-calibrated with predicted probabilities accurately reflecting the observed proportions, also in different risk strata.

To aid the clinical interpretation of model predictions, we used SHAP-values to highlight the most impactful features and to explain how the different features affect the prediction for the individual patient. Model explainability is important for evaluating the model output and is paramount for the clinical adaption of any ML-model<sup>38</sup>, including focus on features that are clinically actionable, i.e. modifiable versus non-modifiable factors.

Compared to the GRACE2.0 score, which is widely considered the gold-standard risk-stratification tool in current clinical use for predicting mortality after acute coronary syndrome (ACS)<sup>39</sup>, PMHnet had superior discrimination and calibration. However, it is important to note that there are differences between the intended patient populations for the two models. The GRACE2.0 score has been developed using a derivation cohort of patients with STEMI, n-STEMI, and unstable angina with time of initial admission as time-zero<sup>5</sup>. In contrast, our study used time at coronary

angiography as its baseline and was applied to all patients with IHD and coronary artery pathology ranging from diffuse atheromatosis to three-vessel disease. The validity of GRACE for a cohort with such characteristics has not been established. To provide a direct comparison of the two algorithms, we used the GRACE2.0 features and re-fitted the GRACE2.0 using our training data. The re-fitted GRACE2.0 score had the same model discrimination and better model calibration than the original version (Fig. 2), but still had inferior prediction compared to PMHnet. Moreover, PMHnet includes assessment of personalized risk predictions (Figure 7), meaning that the most impactful prognostic features might vary from patient to patient. It remains important to stress, that the model displays correlations only, but we argue that individualized risk predictions are of increasing importance in healthcare systems of increasing complexity with more and more patients surviving many years with a burden of multiple chronic diseases. To support safe and effective treatment in such healthcare systems, knowledge of modifiable as well as unmodifiable risk factors for disease progression is paramount to decide the better treatment option on a case to case basis. The above observations are in good agreement with the current literature on ML-based secondary risk-stratification models, which have found machine learning models to offer better performance than simpler, existing scores in current clinical use<sup>10–13,40,41</sup>. A transition towards feature-rich models that utilize more of the available data can therefore be an advantage. The 584 features used in our final model are neither bespoke nor specifically collected for this decision-support application, and instead represent clinical information gathered during routine work-up, management and treatment of patients. Using models that rely on several hundred features means that the current practice of manually entering data into a webtool becomes very impractical<sup>42</sup>. Instead, novel risk-prediction models need to be fully integrated in the EHR systems such that data can be automatically pulled and integrated. In the present study, we demonstrate that PMHnet can

1 be scaled and used in a different healthcare system and provide evidence that the results are  
2 generalizable. However, clinical implementation of PMHnet is beyond the scope of the present  
3 study.

4 Among previously published machine-learning models for secondary prediction in IHD, our study  
5 has several strengths. Firstly, whereas almost all the existing literature uses binary classification,  
6 the use of survival or time-to-event models is less explored. One notable exception is the model  
7 reported by Steele et al.<sup>10</sup> which employed random survival forests and elastic net Cox regression  
8 to predict mortality in patients (n=80,000) with a history of coronary artery disease. One of the  
9 defining characteristics of survival models is the ability to handle censored data<sup>43</sup>, which in other  
10 types of models would have been left out. In addition, survival models can distinguish between  
11 “died after a week” and “died after 10 months” which e.g., would be identical in a 1-year binary  
12 classification model. To the best of our knowledge, we are the first to use neural network-based  
13 survival models in this context. Secondly, we externally validated our model using data from a  
14 different country. Most models in the literature were not suboptimal externally validated<sup>10,11,40,41</sup>,  
15 and among those that have been, only two used data from a different country<sup>12,13</sup>. Demonstrating  
16 that a given model can accurately predict beyond borders is crucial for generalizability. Thirdly,  
17 our model can predict all-cause mortality up to five years after the index date. Probably owing to  
18 the fact that survival models have not been used, most models in the domain exclusively operate  
19 with a prediction horizon of one or two years<sup>12,13,40,41</sup>. Longer prediction horizons might be  
20 necessary for long-term disease management.

21 Although the features we have used represent data collected from a typical clinical workflow and  
22 therefore should be generally applicable, inter-regional and inter-national differences in clinical  
23 practice may affect what data is available and when. This is for instance exemplified by differences

1 in diagnostic work-up, or timely access to coronary angiography, but may also relate to differences  
2 in access to previous medical data. Such aspects could affect the generalizability of our included  
3 features, but with internationally accepted treatment guidelines the differences should be minimal.  
4 Possible solutions are to reduce slightly the complexity of models to better match the intersection  
5 of features available across countries/regions and/or re-fitting and possibly retraining the model  
6 each time it is deployed in a new setting. In our external validation we used data from Iceland,  
7 which in an international perspective is very similar to Denmark when comparing healthcare  
8 systems<sup>44,45</sup>. For that reason, we downscaled the model slightly to account for data availability, but  
9 re-training on Icelandic data was not deemed necessary. Using the Icelandic data, we found the  
10 model on average to be well-calibrated, but from visual inspection of the calibration curve found  
11 evidence of miscalibration as the model was found to overestimate the risk for some patients which  
12 would suggest that further adjustment of the model is needed were it to be deployed in Iceland.  
13 However, since miscalibration does not affect the accurate risk-stratification of patients, we did  
14 not pursue that issue further in this study. For future studies, we note that techniques such as Platt-  
15 scaling or isotonic regression could be used to remedy calibration-issues<sup>46</sup>. Overall, we argue that  
16 it is an inherent strength of the study that no strict feature selection criteria were applied. First,  
17 neural networks are by design capable of handling correlated data and second, an increasing data-  
18 rich healthcare system demands a better usage of the data being generated. By means of the  
19 explainability analysis, the study succeeds in showcasing that dependencies are present. We  
20 acknowledge that an inherent limitation of the study is that it only unmasks correlations, and that  
21 the explainability step indeed is influenced by the dependencies.

22 The dynamic nature of clinical environments can lead to deterioration of model performance<sup>47,48</sup>.  
23 Advances in treatment and diagnosis mean that the baseline risk of patients with ischemic heart

disease could change over time, which in turn would lead to a drift of model calibration. As an example, the logistic EuroSCORE<sup>49</sup>, a pan-European risk-stratification model for cardiac surgery published in 2003 was since its inception gradually found to overestimate mortality<sup>50</sup>. The EuroSCORE has since been replaced by EuroSCORE II which for the time being has remedied these issues<sup>51</sup>. This type of systematic decline of model performance has important implications for our model as well and necessitates that the model performance is continuously monitored to ensure acceptable up-to-date performance. The need for real-time monitoring of model performance is another strong argument for risk-stratification tools to be tightly integrated within EHR systems.

As a secondary analysis in this study, we tested adding a panel of polygenic risk scores to the input features in PMHnet and assessed how they might impact model performance. The 19 different genetic risk scores were included based on being related to cardiometabolic health and covered traits such as *blood pressure*, and *total cholesterol*, but also included *heart failure* and *acute myocardial infarction*. Where for example the *acute myocardial infarction* risk score is developed for primary risk prediction, i.e., disease development, our model is concerned with secondary risk prediction, i.e., modelling risk for those who already have the disease. It is known that CAD PRS associate with events in both primary and secondary event populations, however, these PRSs are PRSs of prevalent diseases, not mortality. Whether or not a primary risk score is useful in a secondary risk context is clear upfront, but as parts of the underlying disease process are known to be shared between the two, we hypothesized that the score could be used in our context. Focusing the analysis only on the subset of patients for which we could obtain genotypes (n=13,449, 37.4%), we found no significant difference in performance after adding the PRS scores to either *model-1 (Diagnoses)* and *model-2 (Diagnoses + ClinicalOne)*. As the 585 features include the prior disease

1 history recorded over more than 25 years, our interpretation is that a “realized” life-course disease  
2 trajectory is more informative than the germline risk that can be calculated from the genotype.  
3 Moreover, the disease trajectory also holds information on life-course exposures and may therefore  
4 also include exposure information that quite implicitly is related to genetic data only. The version  
5 of PMHnet trained on the disease history only, with and without genetics, indicates that there is  
6 little gain from the PRS tested in this case. As we did not have genetic data for the full cohort, we  
7 cannot exclude that our interpretation is affected by this aspect.

8 Prospective studies are needed to ascertain how feature-rich risk-stratification methods can be used  
9 to alter, guide, and hopefully improve treatment. The ability to accurately predict high-risk is  
10 useful for identifying patients that may benefit from more extensive treatment and more frequent  
11 visits at the hospital. In contrast, accurate identification of low-risk patients may potentially be  
12 used to limit work-up, extent of pharmacological treatment and follow-up content and intensity  
13 and thereby prevent potential harmful overtreatment. Striking the correct balance between over-  
14 and undertreatment can contribute to the advancement of precision medicine, and here a well-  
15 calibrated and highly discriminative risk-prediction model can serve as an important tool.

16 In routine cardiology, a multitude of diagnostic, prognostic, and treatment-related scores are  
17 applied. However, all the presently applied scores are based on a very limited number of features.  
18 The present findings indicate a significantly added value of applying far more features and of  
19 introducing machine-learning. Thus, precision treatments in cardiology may benefit from using  
20 more features and machine-learning to replace the present scores.



## **Data availability statement**

Due to national and EU regulations, the datasets used for model development and validation cannot be made publicly available. Research groups with access to secure and dedicated computing environments can request access to the source data registries via application to the Danish Health Data Authority.

## **Conflicts of interest**

Søren Brunak reports ownerships in Intomics, Hoba Therapeutics, Novo Nordisk, Lundbeck, and ALK; and managing board memberships in Proscion and Intomics. Henning Bundgaard reports ownership in Novo Nordisk and has received lecture fees from Amgen, BMS, MSD and Sanofi. The following co-authors are employed by deCODE genetics/Amgen, Inc: Vinicius Tragante, Daníel F. Guðbjartsson, Anna Helgadóttir, Hilma Holm, and Kari Stefansson.

## **Acknowledgements**

We acknowledge Mette Hartlev, Franziska Walder, Mette Gørtz, and Katharina Ó Cathaoir for helpful comments and discussions in the writing of this manuscript.

## References

1. Collet, J.-P. *et al.* 2020 ESC Guidelines for the management of acute coronary syndromes in patients presenting without persistent ST-segment elevation: The Task Force for the management of acute coronary syndromes in patients presenting without persistent ST-segment elevation of the European Society of Cardiology (ESC). *Eur. Heart J.* **42**, 1289–1367 (2021).
2. Knuuti, J. *et al.* 2019 ESC Guidelines for the diagnosis and management of chronic coronary syndromes: The Task Force for the diagnosis and management of chronic coronary syndromes of the European Society of Cardiology (ESC). *Eur. Heart J.* **41**, 407–477 (2020).
3. Steg, Ph. G. *et al.* ESC Guidelines for the management of acute myocardial infarction in patients presenting with ST-segment elevation. *Eur. Heart J.* **33**, 2569–2619 (2012).
4. Wilson, P. W. F. *et al.* Prediction of Coronary Heart Disease Using Risk Factor Categories. *Circulation* **97**, 1837–1847 (1998).
5. Fox, K. A. A. *et al.* Should patients with acute coronary disease be stratified for management according to their risk? Derivation, external validation and outcomes using the updated GRACE risk score. *BMJ Open* **4**, e004425 (2014).
6. Hung, J. *et al.* Performance of the GRACE 2.0 score in patients with type 1 and type 2 myocardial infarction. *Eur. Heart J.* **42**, 2552–2561 (2020).
7. Antman, E. M. *et al.* The TIMI Risk Score for Unstable Angina/Non-ST Elevation MI. *JAMA* **284**, 835 (2000).
8. Rajkomar, A., Dean, J. & Kohane, I. Machine Learning in Medicine. *N. Engl. J. Med.* **380**, 1347–1358 (2019).

- 1 9. Topol, E. J. High-performance medicine: the convergence of human and artificial intelligence.  
2 *Nat. Med.* **25**, 44–56 (2019).
- 3 10. Steele, A. J., Denaxas, S. C., Shah, A. D., Hemingway, H. & Luscombe, N. M. Machine  
4 learning models in electronic health records can outperform conventional survival models for  
5 predicting patient mortality in coronary artery disease. *PLOS ONE* **13**, e0202344 (2018).
- 6 11. Motwani, M. *et al.* Machine learning for prediction of all-cause mortality in patients with  
7 suspected coronary artery disease: a 5-year multicentre prospective registry analysis. *Eur.*  
8 *Heart J.* ehw188 (2016) doi:10.1093/eurheartj/ehw188.
- 9 12. Mohammad, M. A. *et al.* Development and validation of an artificial neural network algorithm  
10 to predict mortality and admission to hospital for heart failure after myocardial infarction: a  
11 nationwide population-based study. *Lancet Digit. Health* **4**, e37–e45 (2022).
- 12 13. D’Ascenzo, F. *et al.* Machine learning-based prediction of adverse events following an acute  
13 coronary syndrome (PRAISE): a modelling study of pooled datasets. *The Lancet* **397**, 199–  
14 207 (2021).
- 15 14. Özcan, C. *et al.* The Danish Heart Registry. *Clin. Epidemiol.* **8**, 503–508 (2016).
- 16 15. Schmidt, M. *et al.* The Danish National Patient Registry: a review of content, data quality, and  
17 research potential. *Clin. Epidemiol.* 449 (2015) doi:10.2147/clep.s91125.
- 18 16. Nielsen, A. B. *et al.* Survival prediction in intensive-care units based on aggregation of long-  
19 term disease history and acute physiology: a retrospective study of the Danish National Patient  
20 Registry and electronic patient records. *Lancet Digit. Health* **1**, e78–e89 (2019).
- 21 17. Sørensen, E. *et al.* Data Resource Profile: The Copenhagen Hospital Biobank (CHB). *Int. J.*  
22 *Epidemiol.* **50**, 719–720e (2020).

- 1 18. Helweg-Larsen, K. The Danish Register of Causes of Death. *Scand. J. Public Health* **39**, 26–  
2 29 (2011).
- 3 19. Schmidt, M., Pedersen, L. & Sørensen, H. T. The Danish Civil Registration System as a tool  
4 in epidemiology. *Eur. J. Epidemiol.* **29**, 541–549 (2014).
- 5 20. Harris, P. J. *et al.* The prognostic significance of 50% coronary stenosis in medically treated  
6 patients with coronary artery disease. *Circulation* **62**, 240–248 (1980).
- 7 21. Arlot, S. & Celisse, A. A survey of cross-validation procedures for model selection. *Stat. Surv.*  
8 **4**, (2010).
- 9 22. Björnsson, E. *et al.* Association of Genetically Predicted Lipid Levels With the Extent of  
10 Coronary Atherosclerosis in Icelandic Adults. *JAMA Cardiol.* **5**, 13–20 (2020).
- 11 23. Laursen, I. H. *et al.* Cohort profile: Copenhagen Hospital Biobank - Cardiovascular Disease  
12 Cohort (CHB-CVDC): Construction of a large-scale genetic cohort to facilitate a better  
13 understanding of heart diseases. *BMJ Open* **11**, e049709 (2021).
- 14 24. Privé, F., Arbel, J. & Vilhjálmsson, B. J. LDpred2: better, faster, stronger. *Bioinforma. Oxf.*  
15 *Engl.* **36**, 5424–5431 (2020).
- 16 25. Gensheimer, M. F. & Narasimhan, B. A scalable discrete-time survival model for neural  
17 networks. *PeerJ* **7**, e6257 (2019).
- 18 26. Kvamme, H. & Borgan, Ø. Continuous and discrete-time survival prediction with neural  
19 networks. *Lifetime Data Anal.* **27**, 710–736 (2021).
- 20 27. Akiba, T., Sano, S., Yanase, T., Ohta, T. & Koyama, M. *Optuna: A Next-generation*  
21 *Hyperparameter Optimization Framework*. <https://arxiv.org/abs/1907.10902> (2019).
- 22 28. Gerds, T. A. & Kattan, M. W. *Medical risk prediction models: with ties to machine learning*.  
23 (CRC Press, 2021).

29. Schumacher, M., Graf, E. & Gerds, T. How to Assess Prognostic Models for Survival Data: A Case Study in Oncology. *Methods Inf. Med.* **42**, 564–571 (2003).
30. Blanche, P. *et al.* Quantifying and comparing dynamic predictive accuracy of joint models for longitudinal marker and time-to-event in presence of censoring and competing risks. *Biometrics* **71**, 102–113 (2015).
31. Stekhoven, D. J. & Bühlmann, P. MissForest--non-parametric missing value imputation for mixed-type data. *Bioinformatics* **28**, 112–118 (2011).
32. Lundberg, S. M. *et al.* From local explanations to global understanding with explainable AI for trees. *Nat. Mach. Intell.* **2**, 56–67 (2020).
33. Collins, G. S., Reitsma, J. B., Altman, D. G. & Moons, K. G. M. Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD). *Circulation* **131**, 211–219 (2015).
34. Collins, G. S. & Moons, K. G. M. Reporting of artificial intelligence prediction models. *The Lancet* **393**, 1577–1579 (2019).
35. von Elm, E. *et al.* The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. *J. Clin. Epidemiol.* **61**, 344–349 (2008).
36. Lundberg, S. M. & Lee, S.-I. A Unified Approach to Interpreting Model Predictions. in *Advances in Neural Information Processing Systems 30* (eds. Guyon, I. *et al.*) 4765–4774 (Curran Associates, Inc., 2017).
37. Lundberg, S. M., Erion, G. G. & Lee, S.-I. Consistent Individualized Feature Attribution for Tree Ensembles. Preprint at <https://doi.org/10.48550/arXiv.1802.03888> (2019).
38. Kundu, S. AI in medicine must be explainable. *Nat. Med.* **27**, 1328–1328 (2021).

39. D'Ascenzo, F. *et al.* TIMI, GRACE and alternative risk scores in Acute Coronary Syndromes: A meta-analysis of 40 derivation studies on 216,552 patients and of 42 validation studies on 31,625 patients. *Contemp. Clin. Trials* **33**, 507–514 (2012).
40. Kwon, J. *et al.* Deep-learning-based risk stratification for mortality of patients with acute myocardial infarction. *PLOS ONE* **14**, e0224502 (2019).
41. Wallert, J., Tomasoni, M., Madison, G. & Held, C. Predicting two-year survival versus non-survival after first myocardial infarction using machine learning and Swedish national register data. *BMC Med. Inform. Decis. Mak.* **17**, 99 (2017).
42. Sharma, V. *et al.* Adoption of clinical risk prediction tools is limited by a lack of integration with electronic health records. *BMJ Health Care Inform.* **28**, e100253 (2021).
43. George, B., Seals, S. & Aban, I. Survival analysis and regression models. *J. Nucl. Cardiol. Off. Publ. Am. Soc. Nucl. Cardiol.* **21**, 686–694 (2014).
44. Einhorn, E. S. Nordic Health Care Systems: Recent Reforms and Current Policy Challenges. *Scand. Stud.* **84**, 106–108 (2012).
45. Kristiansen, I. S. & Pedersen, K. M. [Health care systems in the Nordic countries--more similarities than differences?]. *Tidsskr. Den Nor. Laegeforening Tidsskr. Prakt. Med. Ny Raekke* **120**, 2023–2029 (2000).
46. Niculescu-Mizil, A. & Caruana, R. Predicting good probabilities with supervised learning. in *Proceedings of the 22nd international conference on Machine learning* 625–632 (Association for Computing Machinery, 2005). doi:10.1145/1102351.1102430.
47. Davis, S. E., Lasko, T. A., Chen, G. & Matheny, M. E. Calibration Drift Among Regression and Machine Learning Models for Hospital Mortality. *AMIA. Annu. Symp. Proc.* **2017**, 625–634 (2018).

- 1 48. Jenkins, D. A., Sperrin, M., Martin, G. P. & Peek, N. Dynamic models to predict health  
2 outcomes: current status and methodological challenges. *Diagn. Progn. Res.* **2**, 23 (2018).
- 3 49. Roques, F., Michel, P., Goldstone, A. R. & Nashef, S. a. M. The logistic EuroSCORE. *Eur.*  
4 *Heart J.* **24**, 882–883 (2003).
- 5 50. Hickey, G. L. *et al.* Dynamic trends in cardiac surgery: why the logistic EuroSCORE is no  
6 longer suitable for contemporary cardiac surgery and implications for future risk models. *Eur.*  
7 *J. Cardiothorac. Surg.* **43**, 1146–1152 (2013).
- 8 51. Nashef, S. A. M. *et al.* EuroSCORE II. *Eur. J. Cardiothorac. Surg.* **41**, 734–745 (2012).

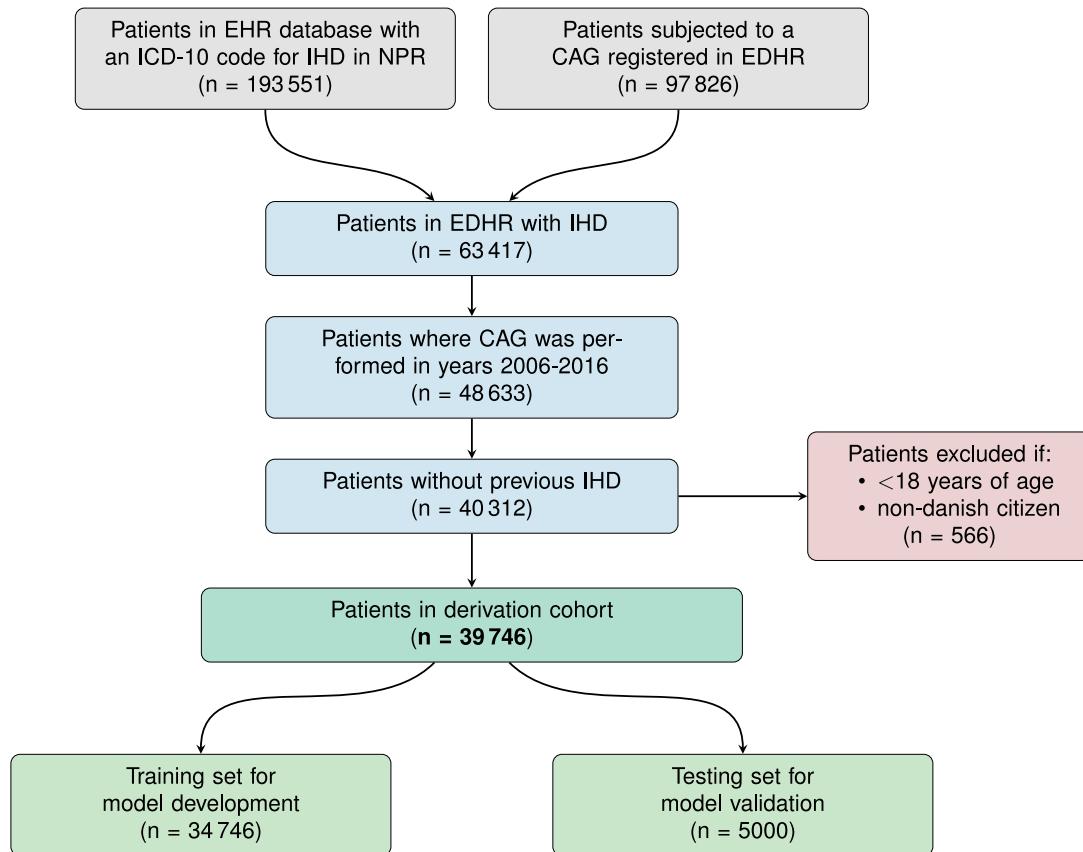


Figure 1: *Flowchart of inclusion of patient with ischemic heart disease.* Flowchart showing how the derivation cohort was identified based on data from NPR and EDHR. CAG: Coronary arteriography. EHRs: Electronic health records. EDHR: Eastern Danish Heart Registry. ICD-10: International classification of diseases, 10th revision. IHD: Ischemic heart disease. NPR: Danish National patient registry.



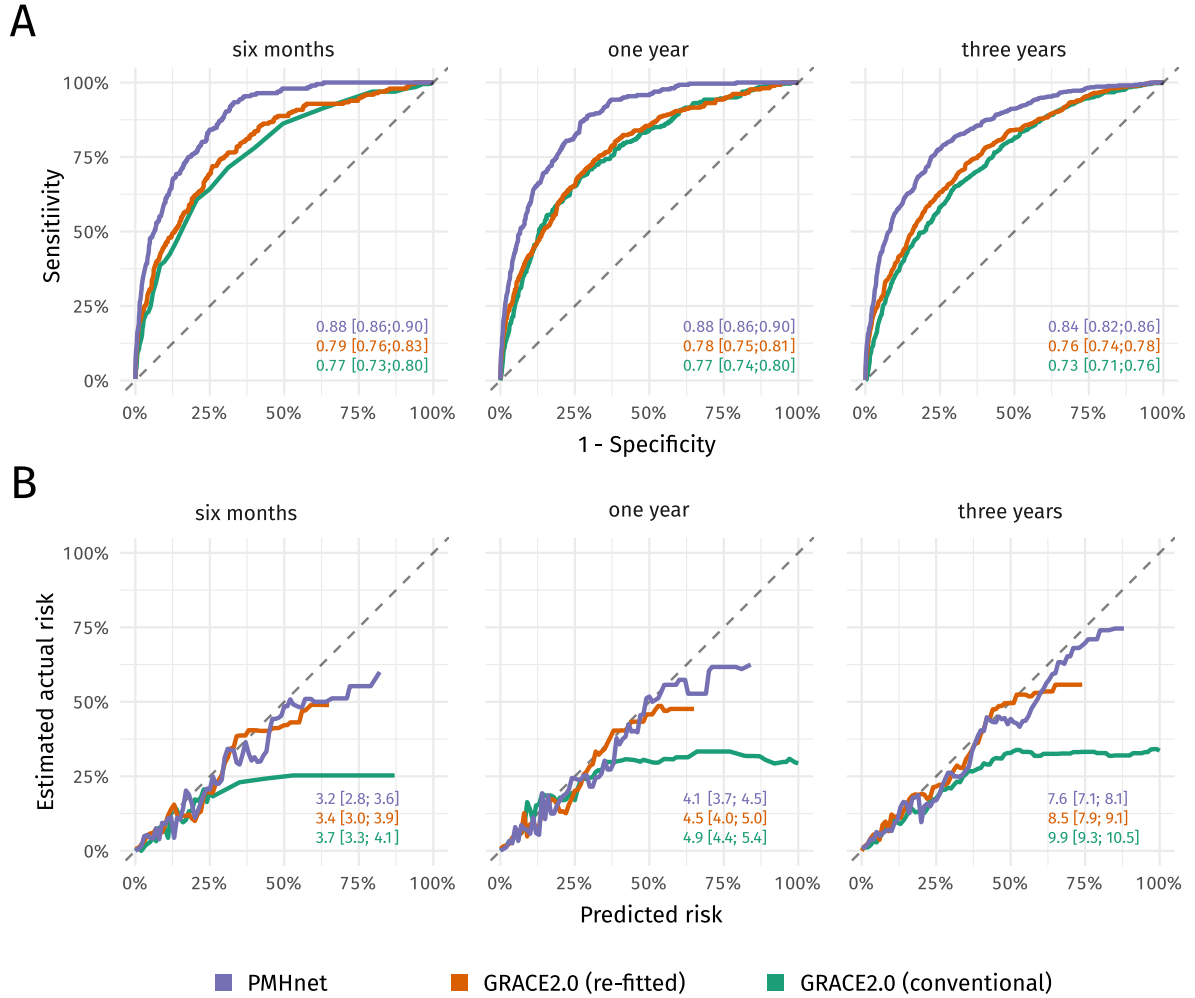


Figure 2: *Model performance of PMHnet and the GRACE2.0 score.* A) Time-dependent receiver operating characteristics (ROC) curves at three different prediction horizons for PMHnet, GRACE2.0 (re-fitted), and GRACE2.0 (conventional). Labels show the time-dependent area under the ROC curves (AUC). GRACE2.0 (re-fitted) is a model that uses the GRACE2.0 input features but uses the PMHnet architecture and is trained using our training data. B) Calibration curves showing the relation between predicted risk and the estimated actual risk. Labels show the Brier score for each of the three models. Lower scores are associated with better calibration and discrimination of predictions.

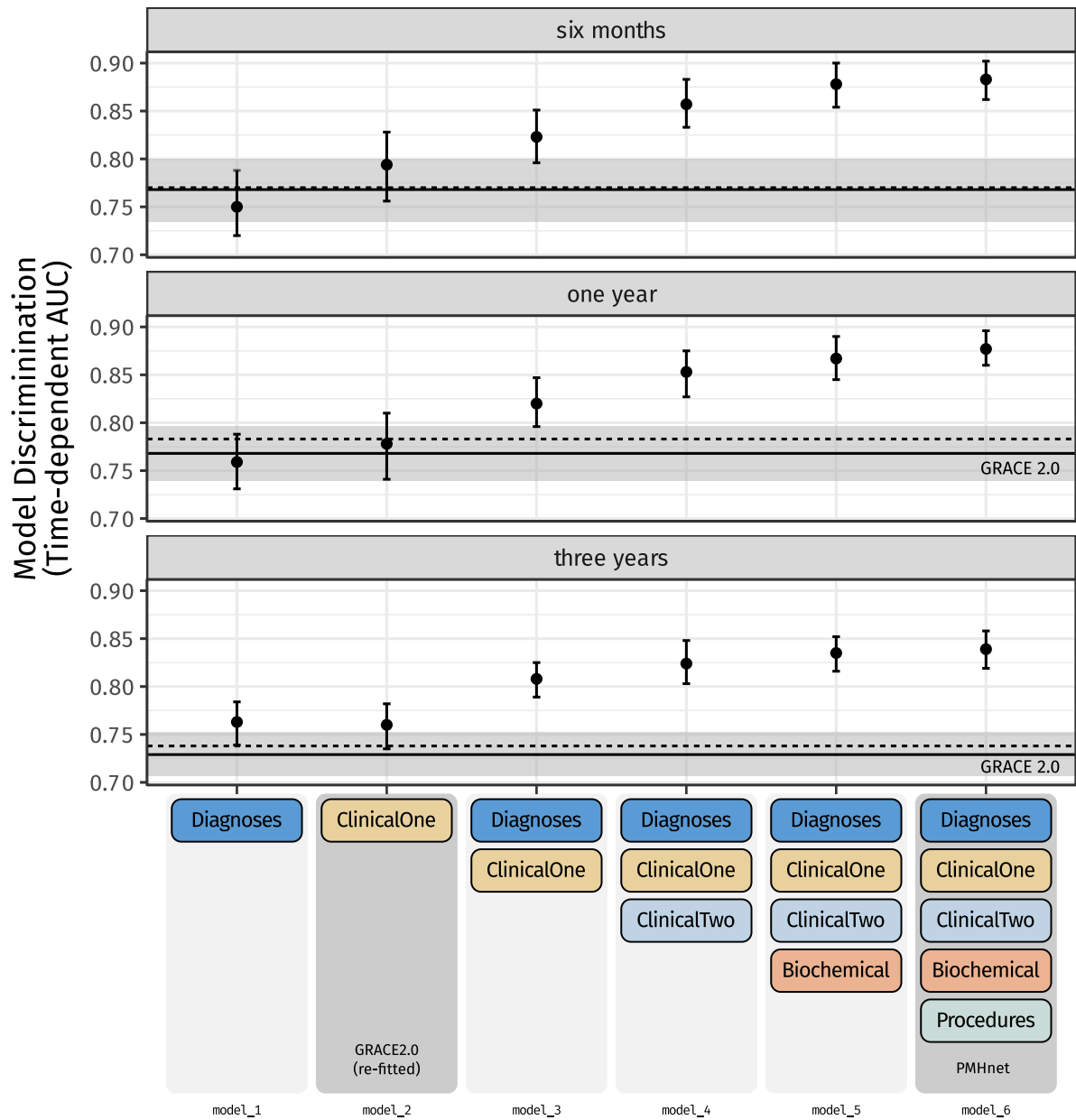


Figure 3: *Model discrimination with increasing number of feature categories.* Time-dependent area under the curve (tdAUC) for various intermediate PMHnet models (and the final one (model 6)) at six months, one year, and three years after the index coronary angiography. Discrimination was evaluated using the hold-out test set. The colored boxes represent the different feature categories that were used as model input in the different models. Horizontal reference lines show the model discrimination of the GRACE2.0 score on the same data. The solid line is the tdAUC of GRACE2.0 on all patients and the dotted line is the tdAUC on the subset of patients where none of the GRACE2.0 input features were missing.

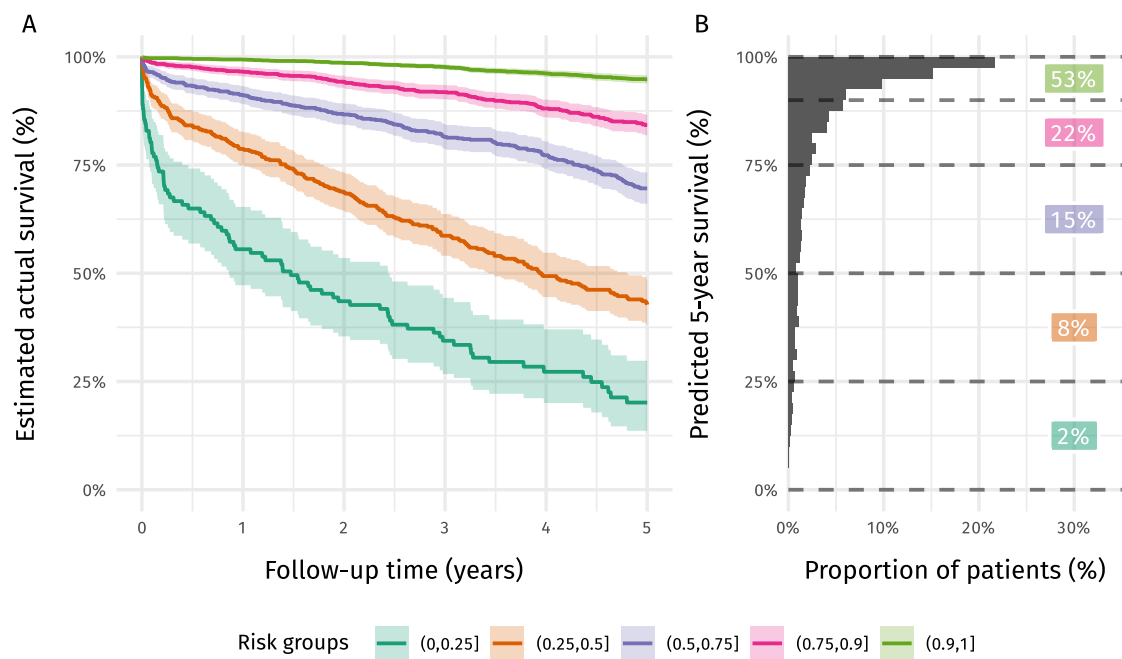


Figure 4: *Observed survival across 5-year predicted risk groups.* A) Estimated actual survival (Kaplan-Meier estimates) for patients in the test set manually stratified into five different risk-groups depending on the predicted survival at five-years by PMHnet. B) Distribution of PMHnet 5-year predicted risk. Vertical lines show the cut-offs that are used to define the risk strata used in A).

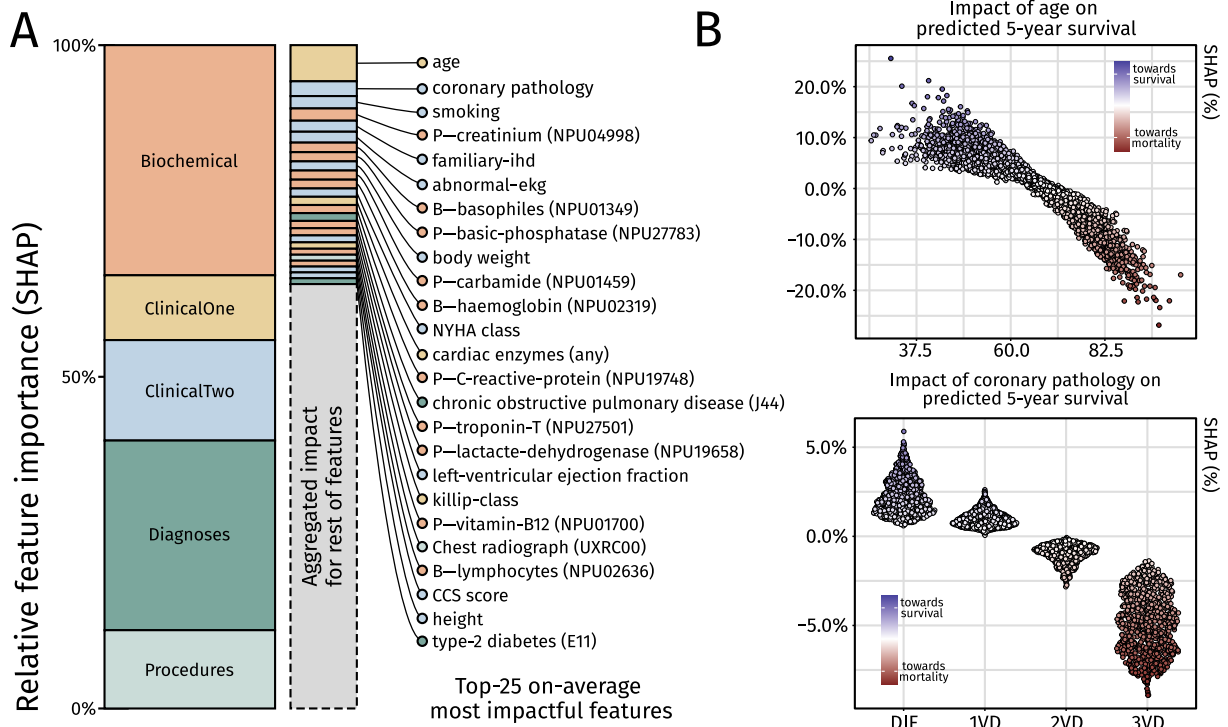


Figure 5: *Overview of feature importance.* Summary of the results from SHAP analysis on the model predictions at five years on patients in the Danish test set. A) Left: Relative feature importance aggregated across the five different feature categories. Biochemical test results and diagnoses were found to affect the model prediction the most. Right: Relative feature importance for all singular features included in the model. Features arranged according to SHAP-values and labels are included for the top 25-most impactful features. Color of features correspond to the feature category in which they belong. SHAP: SHapley Additive exPlanations. B) Relationship between age and SHAP-value with each point showing the SHAP-value for age for a patient in the test-set, and relationship between coronary pathology (e.g. vessel status) and impact on model prediction.

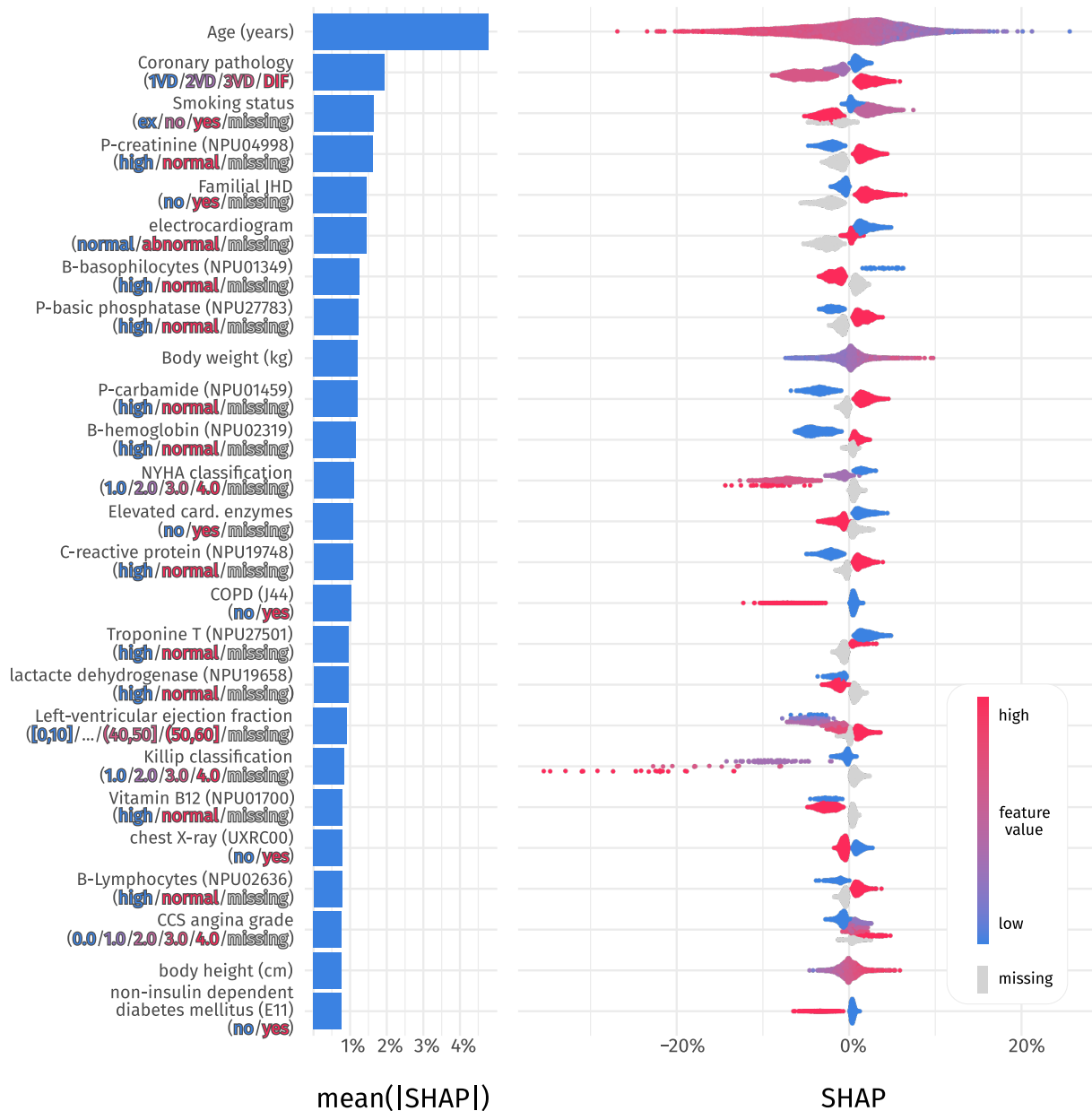


Figure 6: *SHAP* summary plot for top 25 most impactful features. Left: average magnitude of model impact. Y-axis labels specifies the feature name and inside parentheses is shown the unit or the factor levels, for continuous and categorical features, respectively. Right: Distribution of feature impacts across the test set. For continuous features, the colors correspond to the feature value ranging from blue (smallest) to red (largest). For categorical features, the factor levels are colored from blue to red. Grey indicates missingness.

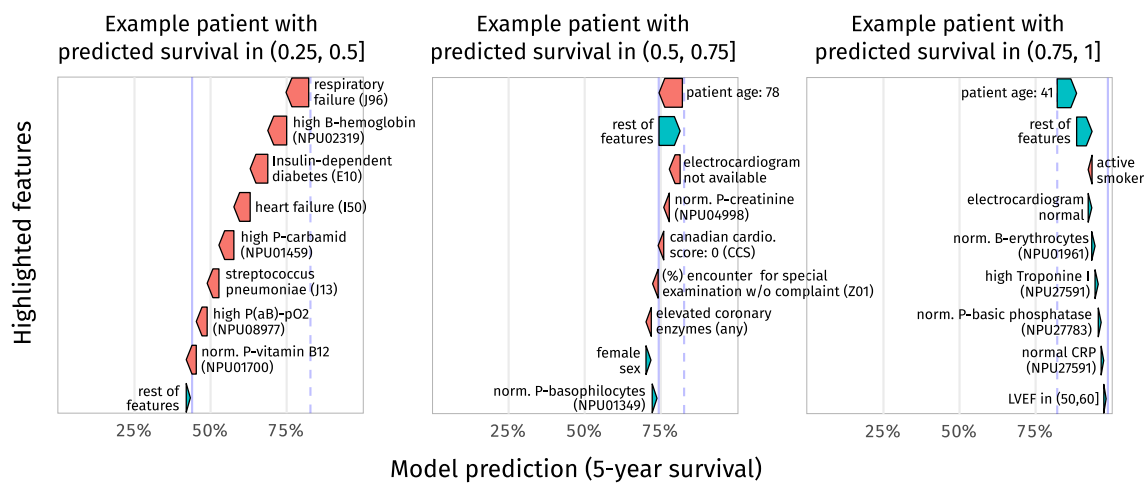


Figure 7: *Patient-level model explanations with SHAP.* Model predictions for three different representative patients with a predicted 5-year survival of 42%, 73%, and 98% with SHAP-explanations showing the estimated impact on model prediction. The patient data have been slightly adjusted to make them non-identifiable. Blue represents features that contribute positively to the prediction. Red represents features that contribute negatively to the prediction. SHAP: SHapley Additive exPlanations. The dashed line is the median prediction from the training set patient and the solid line is the prediction for each of the example patients.

	Training (n=34,746)	Test (n=5,000)	Validation (n=8,288)
Male Sex	23413/67.3%	3410/68.2	5876/70.9%
Age (year)[95%CI]	66.0 [65.7;66.4]	66.2 [66.0;66.3]	66.0 [65.8; 66.2]
Height (cm) [95%CI] (NA)	172.8 [172.7;172.9] (6.5%)	173 [172.7;173.2] (6.3%)	174.3 [174.1; 174.5]
Weight (kg) [95%CI] (NA)	81.3 [81.1;81.5] (4.3%)	81.7 [81.1;82.1] (4.1%)	87.2 [86.8; 87.5]
Comorbidities (ICD10)			
Diabetes	4635/13.3%	694/13.9	1035/12.5%
Hypertension	788/2.27%	105/2.1%	216/2.6%
COPD	2586/7.44%	389/7.78%	715/8.6%
Dyslipidemia	4291/12.3%	641/12.8%	381/4.6%
Medication (ATC)			
lipid-lowering drugs	18188/52.3%	2679/53.6%	6592/79.5%
anti-hypertensive drugs	25684/73.9%	3679/73.6%	7494/90.4%
type-2 diabetes drugs	5575/16.0%	808/16.2%	1048/12.6%
insulin	2088/6.01%	320/6.4%	158/1.9%
P-creatinine, mmol/l [95%CI] (NA) (*)	84.8 [84.5;85.2] (10.1%)	84.2 [83.4;85.1] (10.8%)	91.7 [90.6; 92.9] (8.3%)
P-creatinine NPU04998 (disc.)			-
Within ref. range	15,490/44.6%	2,188/43.8%	
Above ref. range	4,437/12.8%	659/13.2%	
Data missing	14,819/42.6%	2,153/43.1%	
B-basophiles NPU01349 (disc.)			-
Within ref. range	12,982/37.4%	1,816/36.3%	
Above ref. range	192/0.56%	28/0.56%	
Data missing	21,572/62.1% <sup>6</sup>	3,156/63.1%	
P-basic phosphatase NPU27783 (disc.):			-
Within ref. range	12,597 / 36.3%	1,737/34.7%	
Above ref. range	1,783/5.13%	280/5.6%	
Data missing	20,366/58.6%	2,983/59.7%	
Elevated cardiac biomarkers (†)	15531/44.7% (38.7%)	2158/43.2% (40%)	1599/19.3% (39.6%)
Abnormal EKG:			-
Yes	14,679/42.2%	2,043/40.9%	
No	10,64/30.6%	1,554/31.1%	
Data missing	9,426/27.1%	1,403/28.1%	
Familial IHD:			-
Known family history	10,558/30.4%	1,495/29.9%	
No known family history	18,067/52.0%	2,655/53.1%	
Data missing	6,121/17.6%	850/17%	
Heart rate, bpm [95%CI] (NA)	74.7 [74.5;74.9] (20.2%)	74.2 [73.7;74.8] (21.1%)	69.6 [69.3; 70.0] (23.7%)
Blood pressure			
Systolic, mmHg [95%CI] (NA)	139.0 [138.7;139.2] (17.6%)	138.9 [138.2;139.7] (17.9%)	148.1 [147.6; 148.6] (14.0%)
Diastolic, mmHg [95%CI] (NA)	77.9 [77.7;78.1] (33.5%)	77.7 [77.2;78.2] (34%)	85.5 [85.2; 85.8] (14.0%)
Cardiac arrest at admission?	513/1.5%	69/1.4%	130/1.6%
ICD or PM	557/1.6%	88/1.8%	188/2.3%
Killip class			
Killip: 1	14297/41.1%	2085/41.7%	3469/41.9%
Killip: 2	455/1.3%	62/1.2%	1028/12.4%
Killip: 3	138/0.4%	13/0.3%	191/2.3%
Killip: 4	170/0.5%	22/0.4%	135/1.6%
Data missing	19686/56.7%	2818/56.4%	3465/41.8%
Left-ventricular ejection fraction (%)			-
LVEF: [0,10]	100/0.3%	16/0.3%	
LVEF: (10,20]	675/1.9%	78/1.6%	43/0.5%
LVEF: (20,30]	1305/3.8%	171/3.4%	166/2.0%
LVEF: (30,40]	1760/5.1%	258/5.2%	308/3.7%
LVEF: (40,50]	3915/11.3%	530/10.6%	758/9.1%
LVEF: (50,60]	11068/31.9%	1576/31.5%	2175/26.3%
Data missing	15923/45.8%	2371/47.4%	4838/58.4%
Smoking status			
Active smoker	10237/29.5%	1497/29.9%	1551/18.8%
Former smoker	11961/34.4%	1758/35.2%	4331/52.2%
Never smoked	9347/26.9%	1296/25.9%	2316/28.0%
Data missing	3201/9.2%	449/9.0%	101/1.2%
Coronary pathology			
Diffuse atheromatosis	9288/26.7%	1331/26.6%	652/7.9%
1 vessel disease	13310/38.3%	1936/38.7%	2740/33.0%
2 vessel disease	6469/18.6%	945/18.9%	1424/17.2%
3 vessel disease	5679/16.3%	788/15.8%	1133/13.6%
Data missing	-	-	28.37% (#)

Table 1: Cohort characteristics for training, test, and external validation set. The comorbidities are defined from the ICD-10 codes that had been assigned to a given patient prior to the index date. (Continued next page)

Table 1: (Continued) Medication is defined from prescriptions prior to index date. Lipid-lowering drugs: C10, anti-hypertensive drugs: C02, C03, C07, C08 and C09, type-2 diabetes drugs: A10B, insulin: A10A. 95%CI: 95% confidence intervals. ATC: Anatomical Therapeutic Chemical Code. ICD-10: International classification of Disease, 10th Revision. NA: Not applicable. COPD: Chronic obstructive pulmonary disease, ICD: Implantable cardioverter-defibrillator, LVEF: Left-ventricular ejection fraction. PM: Permanent pacemaker. (\*): See supplementary methods for details on differences between continuous and discrete classification of creatinine. (#): Diffuse atheromatosis could not be defined with complete certainty for 28.4% of the Icelandic data, and coronary pathology was therefore set as NA in such cases.



Category	Features
<i>ClinicalOne</i> (CRACE2.0)	Age, pulse, systolic blood pressure, cardiac arrest at presentation (yes/no), abnormal cardiac enzymes (yes/no), Killip-class, creatinine, ST-segment deviation (yes/no)
<i>ClinicalTwo</i>	Abnormal ECG (yes/no), CCS class, diastolic blood pressure, coronary artery dominance (R/L/B), familial IHD (yes/no), height, weight, ICD-device or PM (yes/no), ischemia test, LVEF, NYHA class, sex, smoking status, coronary pathology,
<i>Diagnoses</i>	322 different level-3 ICD-10 diagnosis codes.
<i>Procedures</i>	154 different NOMESCO procedure codes corresponding to various radiological examinations and surgical procedures
<i>Biochemical</i>	85 different lab tests with results categorized as <i>below</i> , <i>within</i> , or <i>above</i> the reference range

Table 2: *Input features used for model development.* The different features were organized in five different categories each representing different domains. The clinical characteristics were divided into two subgroups where ClinicalOne contains the features used in the GRACE2.0 score.

	6 months		1 year		3 years	
Comparison	$\Delta$ AUC (%)	p-value	$\Delta$ AUC (%)	p-value	$\Delta$ AUC (%)	p-value
PMHnet vs. GRACE2.0 (conventional)	11.5 [9.0; 14.0]	1e-9	10.9 [8.6; 13.2]	6e-21	10.3 [8.3; 12.4]	1e-22
PMHnet vs. GRACE2.0 (re-fitted)	8.9 [6.5; 11.2]	2e-13	8.9 [6.5; 11.2]	2e-18	7.6 [5.8; 9.4]	1e-6

Table 3: *Difference in discrimination between PMHnet and the GRACE2.0 score.* For  $\delta$ AUC, we obtain 95% CIs (in brackets) and p-values from the Score function in the R package riskRegression<sup>38</sup>.

# Supplementary material

## Supplementary methods

### Feature inclusion and pre-processing of the Danish training and test set

For each patient we extracted the following data: all diagnosis codes (ICD-10) and procedure codes (SKS/NOMESCO) assigned to at least 1% of the cohort between 1st of January 1994 and the time of the coronary angiography (from NPR); all results from laboratory tests taken on at least 5% of the cohort between 5 years prior-to and up until the time of the coronary angiography (from BTH); 23 other clinical features such as sex, age, smoking status, coronary pathology, etc. (from EDHR + BTH). Moreover, for 37.4% of the cohort a panel of 19 different PRSs was included (see below). In case of repeated tests/measurements/assignments the one closest to the time of the coronary angiography was used. No data originating after the index coronary angiography were used. Laboratory test results and sex- and age adjusted (if applicable) reference ranges were originally stored in the regional laboratory information management systems “Labka” and “BCC” and in this study obtained through BTH<sup>43</sup>. Tests were either annotated in accordance with the Nomenclature, Properties and Units ontology (NPU) or various local coding systems<sup>44</sup>. Using the sex- and age adjusted reference ranges, results of the biochemical tests were discretized to the categories *below*, *within*, *above*, and if a given test had not been taken a category called *missing*. Discretization of biochemical tests using the adjusted reference ranges was a pragmatic alternative to normalizing the many different biochemical tests in a manner that adequately takes both intra- and interdepartmental variance into account,

which both might be related to for instance differences in patient population and/or equipment and machinery.

Consequences of the choices above, including discretization were alluded to in Table 1.

Here it is evident that the missingness of the continuous creatine feature creatinine was less than that of the discretized version because we allow measurements 21-days in the “future” relative to time zero. For the discretized version, we did not include those measurements, and consequently so missingness consequently higher. All continuous features were Z-score normalized and missing values were encoded with a value of zero. All categorical features were one-hot encoded and an additional category for missing values was added if applicable<sup>25</sup>.

During model development, we tested and optimized various feature specific hyperparameters. We tested including diagnosis codes as either level-4, level-3, block, or chapter codes representing different steps in the ICD-10 hierarchy<sup>45</sup>. With the hypothesis that codes and results assigned many years prior to time zero time might carry less information, we introduced three “shelf-life” hyperparameters and filtered out diagnosis codes, procedure codes, and biochemical tests assigned/taken more than  $n$  years prior to the coronary angiography.

## Cohort characteristics

For the cohort characteristics presented in Table 1, comorbidities were defined from the ICD-10 codes assigned before or at the index date: *diabetes* was defined as E10 or E11; *hypertension* was defined as I10, I11, I12, I13, I14, or I15; *COPD* was defined as J44; and *dyslipidemia* was defined as E78.0, E78.1, E78.2, E78.3, E78.4, E78.5, and E78.9. The

medication use was defined from prescriptions given prior to or at the index-date, using the same definition as used in Kiiskinen et al. (1), that is: *lipid-lowering drugs* is ATC class C10; *anti-hypertensive drugs* is C02, C03, C07, C08, and C09; *type-2 diabetes drugs* is A10B; and *insulin* is A10A.

For continuous features (age, height, etc.) the mean is given along with 95% bootstrap confidence intervals (CI) and, if applicable, the amount of missingness in parentheses – mean [low, high] (missingness). The mean and CI are calculated using only the non-missing features. For categorical features the raw counts and relative frequencies are both specified.

## **Downscaled version for external validation**

Due to data availability, the model was downscaled for the external validation on Icelandic data. The features that were left out of the model in the down-scaling was all procedure codes (n=154) since they were encoded using another non-compatible coding scheme; five biochemical test which was not used in the Icelandic system; and seven different clinical features for which data could not be obtained: “arrest”, “enzymes”, “abnormal-ekg”, “abnormal-qrs-st”, “coronary artery dominance”, “ischemia test”, and “NYHA-class”.

## **Polygenic risk scores**

PRSs were calculated based on GWAS summary statistics data from 19 traits relevant for cardiometabolic health, obtained from 17 GWAS meta-analyses: acute myocardial infarction(2), atrial fibrillation(3), coronary artery disease(4,5), heart failure(4), dilated cardiomyopathy(6), hypertrophic cardiomyopathy(6), systolic and diastolic blood

pressure(7), stroke(8), total cholesterol and triglyceride levels(9), non-alcoholic fatty liver disease(10), immune-mediated inflammatory diseases(11), chronic kidney disease(12), body mass index(13), type 2 diabetes with adjustment for BMI(14), frailty index(15).

Autosomal genotypes from 188,462 individuals in the CHB Cardiovascular Disease Cohort were filtered to only include variants present in the HapMap3 set of 1,120,696 reference variants. Any missing genotype information was conservatively imputed to be the affected locus' reference allele. We identified a set of 978,246 genotyped variants present in both genotype and GWAS summary statistics data that we subsequently subjected to Ldpred2's recommended standard deviation quality control. After variant matching and quality control, a mean of 969,607 (S.D. 8,777) variants remained for per-chromosome risk score calculation for each of the 19 traits. We used the Ldpred2-auto algorithm with 30 Gibbs sampling chains, 1,000 burn-in iterations and 500 iterations after burn-in. The initial values for the 30 sampling chains were a) the LDSC regression estimate for heritability  $h^2$  (same for all chains); b) one of 30 initial values for the proportion of causal variants  $p$ , evenly spaced on a logarithmic scale from  $10^{-4}$  to 0.9. Final per-chromosome effect sizes were calculated from each set of 30 sampling chains (per trait and chromosome) through a three-step process, which serves to ensure that the model (spanning 30 chains) converged: 1) computing the standard deviations of each chains' predicted scores, 2) keeping only the chains within three median absolute deviations from the median standard deviation, 3) averaging the effect sizes of the remaining chains. Across the 418 per-chromosome models (19 traits times 22 chromosomes), 26.89 chains were included in the final score on average. The lowest number of included chains was 18. Finally, the resulting per-chromosome risk scores were added together into genome-wide polygenic scores.

## **Neural network architecture**

The neural network architecture is a relatively simple feed-forward neural network with between one and three densely connected hidden layers with rectified linear units as the activation function and with a dropout layer after each hidden layer. The output layer has 30 outputs – one for each time bin – and is being fed through a softmax activation function to produce probabilities. The neural network architecture is sketched out in figure S5.

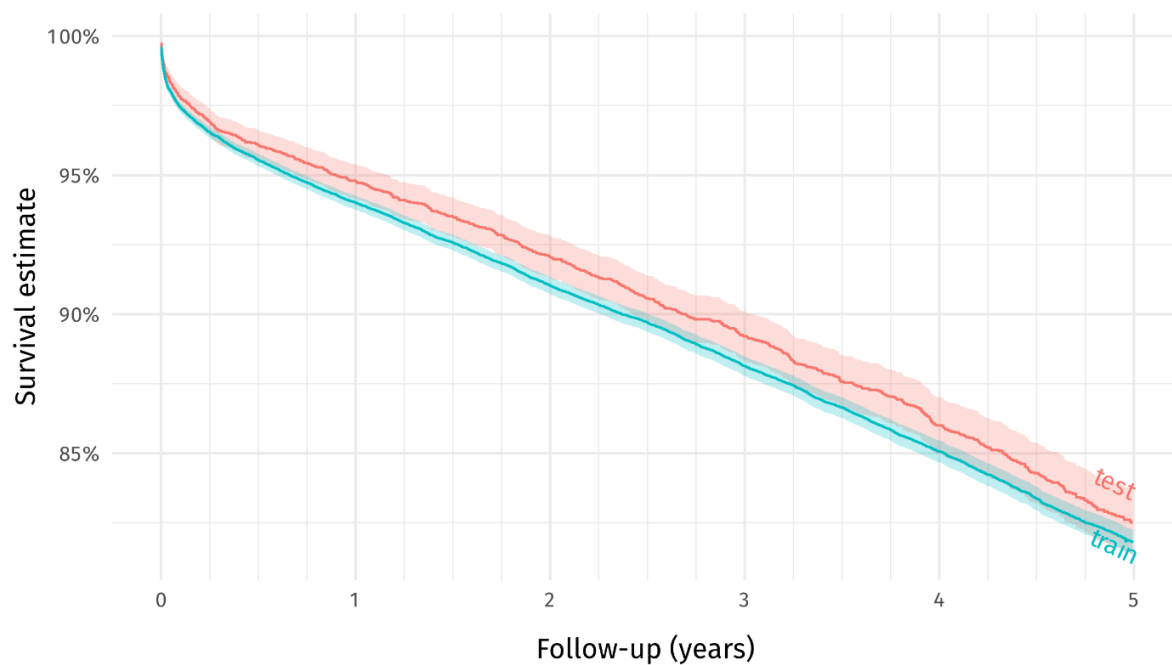
## **Hyperparameter optimization**

For hyperparameter optimization we used the Optuna hyperparameter optimization package for python (16). For all hyperparameter sweeps, we ran 2500 trials. For each trial, hyperparameters were sampled using the TPE (Tree-structured Parzen Estimator) algorithm as implemented in Optuna using 400 startup trials and otherwise all other arguments left on default settings(17). To prune unpromising trials, we used the hyperband pruner, with a minimum resource of 20, and a reduction factor of 3.

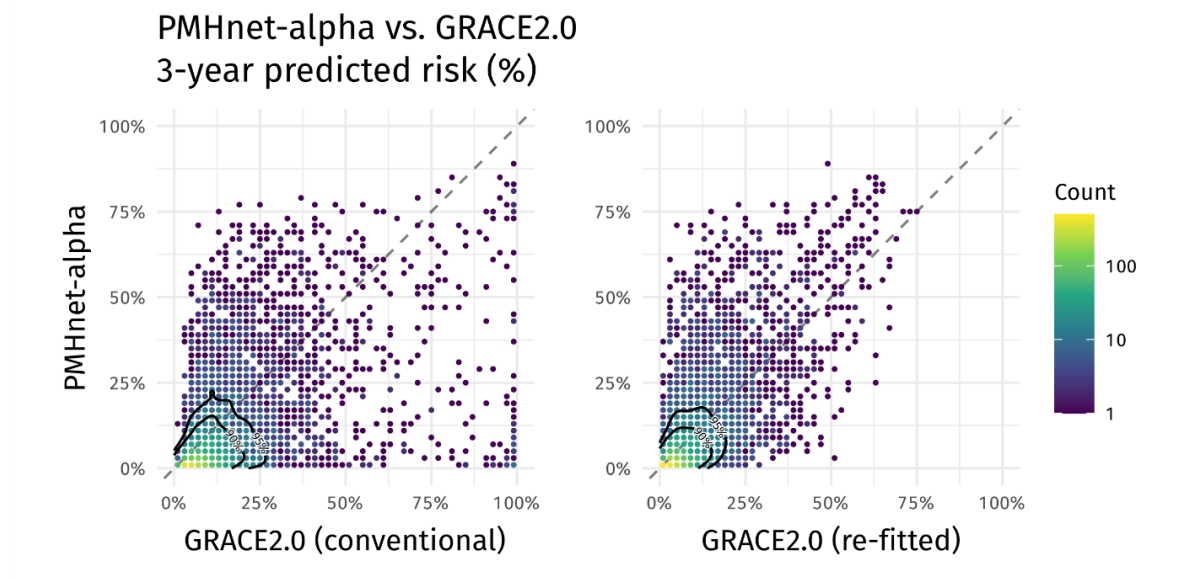
## **Resiliency of PMHnet to missingness**

Finally, the resiliency of the model was assessed as described in Methods. Across the 584 features, age was the only single feature, where the change in tdAUC changed significantly when artificially removed. In the case of model calibration, missingness of six individual features significantly affected the performance. These features were troponins, LVEF, age, smoking status, abnormal ECG, and coronary pathology. Taken together, these results are evidence that PMHnet is resilient to missing data.

## 1    **Supplementary figures**



2  
3    *Figure S1: **Kaplan-Meier estimates for the PMHnet derivation cohort.** Kaplan-Meier*  
4    *estimates for the training (blue) and testing (red) set of 34,749 and 5,000 patients,*  
5    *respectively.*



1

2 **Figure S2: Difference in predicted risk between PMHnet and GRACE2.0.** Binned

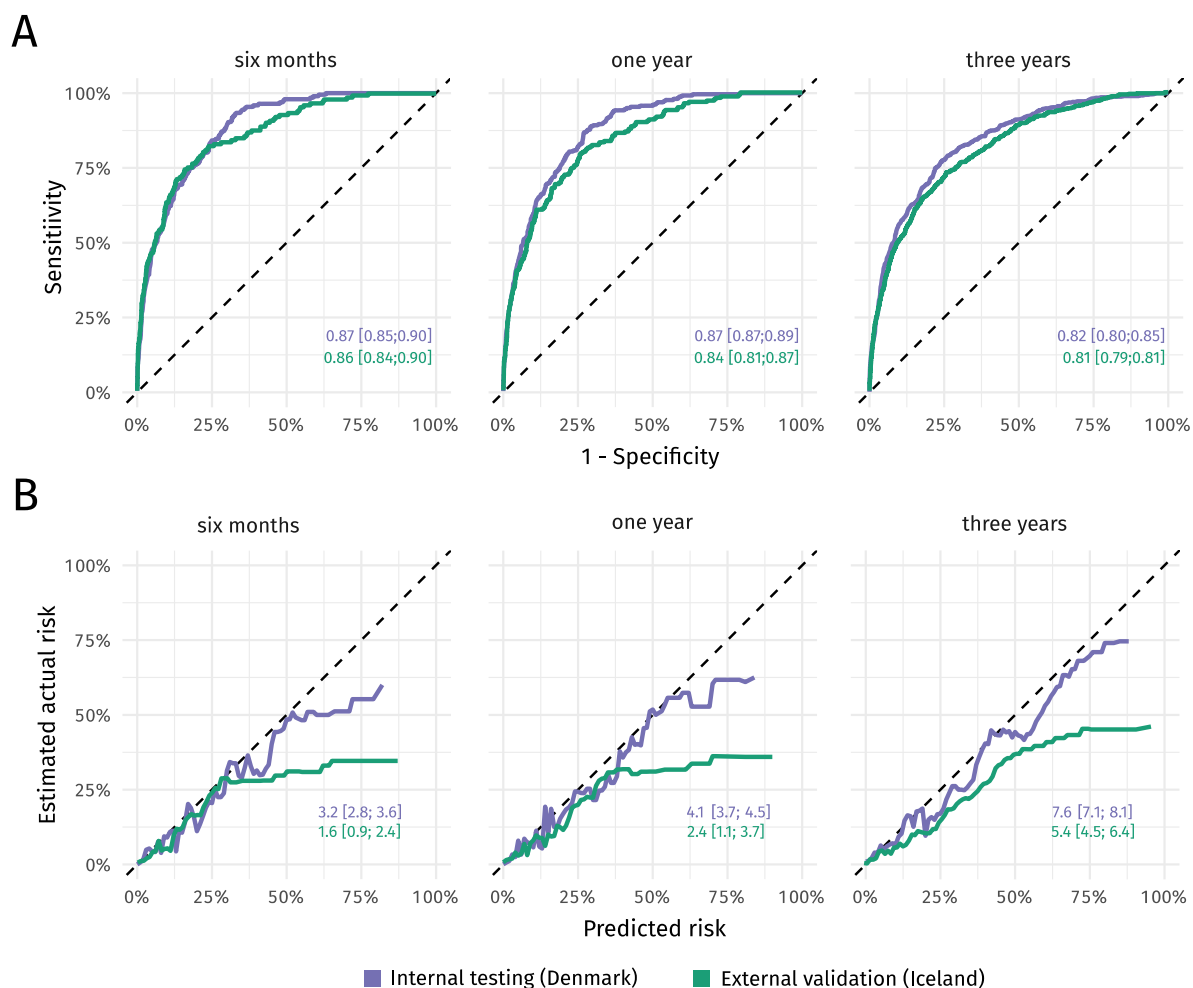
3 scatterplot showing the difference between the 3-year predicted risk by PMHnet (y-axis) and

4 GRACE2.0 (conventional) [left-panel, x-axis] or GRACE2.0 (re-fitted) [right-panel, x-axis].

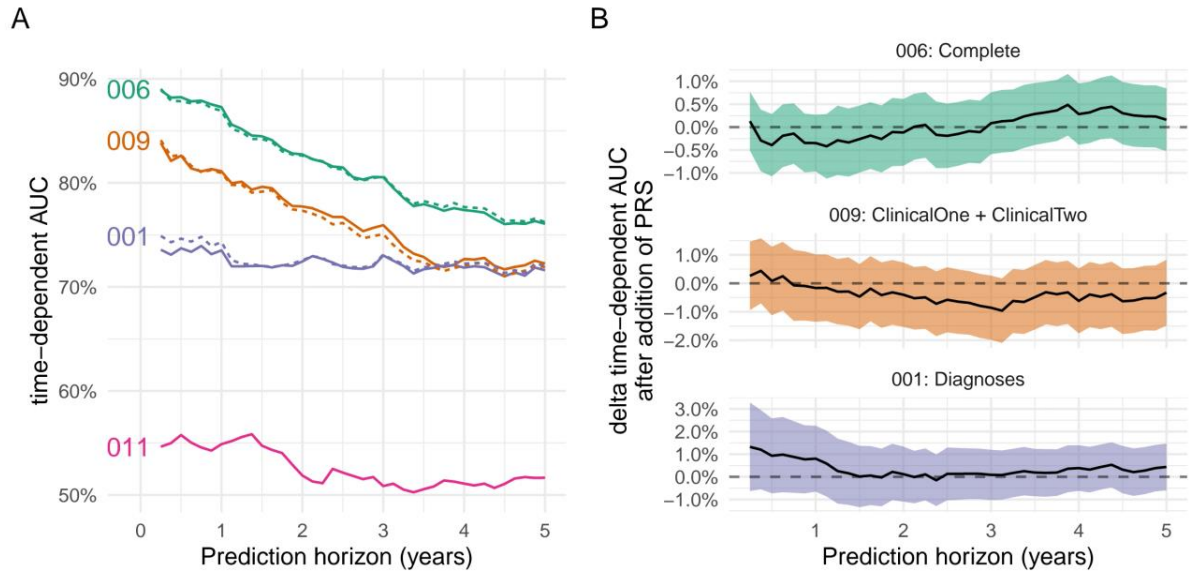
5 Points are colored according to how many patients in the test set fall in that bin. Contour lines

6 indicate 90%- and 95%-point densities.



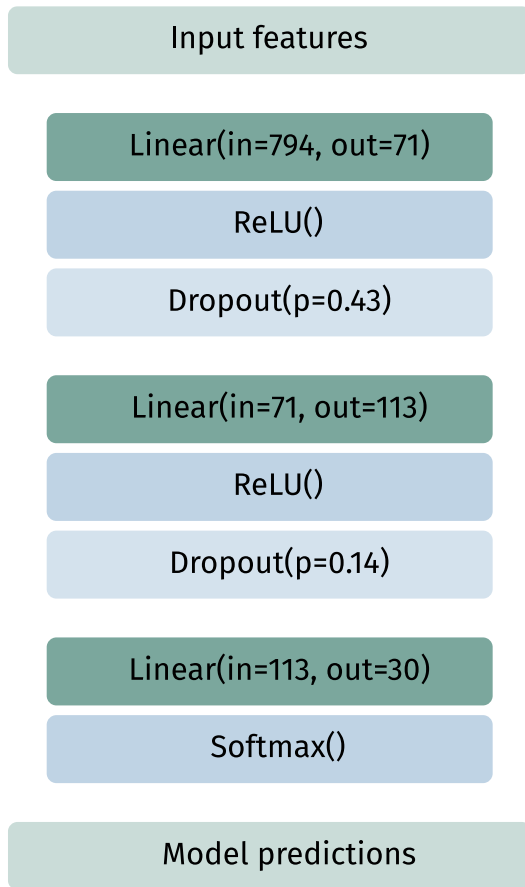


**Figure S3: External validation of PMHnet on Icelandic data.** A) Time-dependent ROC curves at three different prediction horizons for PMHnet evaluated on the Danish test set and the Icelandic external validation set, respectively. Text labels show the corresponding tAUC scores. B) Calibration curves showing the relation between predicted risk and the estimated actual risk. Labels show the Brier score for each of datasets. Lower scores are associated with better calibration and discrimination of predictions. Data for both type of plots and scores were generated using riskRegression and visualized using ggplot2.



**Figure S4: Effect of including polygenic risk scores to clinical features.**

[A] The time-dependent AUC values for different models evaluated on the subset of the hold-out test set where genetic information could be obtained. From top to bottom, the first two lines shows the performance of the complete model ("006") with and without PRS features added. The next sets of two correspond to a model only using clinical features ("009", ClinicalOne + ClinicalTwo) and a model only using diagnosis codes ("001"). Finally, the last line is the performance of a model that is using PRSs as the sole predictors. [B] A formal test of  $\Delta t d A U C$  obtained by adding PRS information to each of the models in panel A). Ribbon shows confidence interval obtained through the "Score" function from the riskRegression R-package.



1

2 **Figure S5: Neural network architecture of PMHnet (full version).**

3 *Using the PyTorch machine learning framework for python, the illustrated neural network*  
4 *architecture was implemented. The architecture hyperparameters, number of layers; number*  
5 *of units; and droprate, were determined from the hyperparameter search performed using the*  
6 *Optuna hyperparameter optimization framework as detailed in the methods section.*

7

## 1 Supplementary tables

2 *Table S1: Missingness across clinical features for the Danish derivation data.*

<b><i>ClinicalOne</i> features</b>	<b>training</b>	<b>testing</b>
Systolic blood pressure	18%	18%
Elevated cardiac enzymes	39%	40%
Cardiac arrest	0%	0%
Age	0%	0%
STEMI	0%	0%
Heart rate	20%	21%
Creatinine	10%	11%
Killip-class	57%	56%
<b><i>ClinicalTwo</i> features</b>	<b>training</b>	<b>testing</b>
Abnormal ECG	0%	0%
Abnormal QRS-ST	0%	0%
Canadian cardiovascular score	8.6%	8.8%
Diastolic blood pressure	33.5%	34%
Dominance	3.7%	3.8%
Familial IHD	0%	0%
Height	6.4%	6.4%
ICD-or-PM	0%	0%
Ischemia test	0%	0%
LVEF	0%	0%
NYHA	75.9%	76.8%
Sex	0%	0%
Smoking	9.2%	9.0%
Vessel status	0%	0%
Weight	4.3%	4.1%

*Table S2: Outcomes and Kaplan-Meier estimates across training, test, and external validation data.*

<i>Outcomes</i>	Training set	Test set	External validation set
RMST(1825), days	1635, SE: 2.58	1651, SE: 6.49	1697, SE: 3.82
KM-estimate, 6 months	95.5% [95.3; 95.8]	96.1% [95.5; 96.6]	98.1% [97.9; 98.4]97.3%
KM-estimate, 1 year	94.0% [93.8; 94.3]	94.8% [94.2; 95.4]	[97.0; 97.7]93.7% [93.2;
KM-estimate, 3 years	88.1% [87.8; 88.5]	89.2% [88.3; 90.1]	94.2]89.6% [88.9; 90.2]
KM-estimate, 5 years	81.8% [81.4; 82.2]	82.5% [81.3; 83.6]	

*Table S3: Hyperparameters search space and final configuration*

<i>Hyperparameter</i>	<i>Search space</i>	<i>Best trial</i>
<i>Biochemical inclusion window</i>		<i>2.5 years</i>
<i>Diagnosis inclusion window</i>		<i>13.5 years</i>
<i>Procedures inclusion window</i>		<i>3.5 years</i>
<i>Number of hidden layers</i>		<i>2</i>
<i>Number of units per hidden layer</i>		<i>Layer 1: 71, Layer 2: 113</i>
<i>Droprate per hidden layer</i>		<i>Layer 1: 42.8% Layer 2: 14.0%</i>
<i>Learning rate</i>		
<i>Momentum</i>		

## References

1. Kiiskinen T, Helkkula P, Krebs K, Karjalainen J, Saarentaus E, Mars N, et al. Genetic predictors of lifelong medication-use patterns in cardiometabolic diseases. *Nat Med.* 2023 Jan;29(1):209–18.
2. Jiang L, Zheng Z, Fang H, Yang J. A generalized linear mixed model association tool for biobank-scale data. *Nat Genet.* 2021 Nov;53(11):1616–21.
3. Nielsen JB, Thorolfsdottir RB, Fritsche LG, Zhou W, Skov MW, Graham SE, et al. Biobank-driven genomic discovery yields new insight into atrial fibrillation biology. *Nat Genet.* 2018 Sep;50(9):1234–9.
4. Nikpay M, Goel A, Won HH, Hall LM, Willenborg C, Kanoni S, et al. A comprehensive 1000 Genomes–based genome-wide association meta-analysis of coronary artery disease. *Nat Genet.* 2015 Oct;47(10):1121–30.
5. van der Harst P, Verweij N. Identification of 64 Novel Genetic Loci Provides an Expanded View on the Genetic Architecture of Coronary Artery Disease. *Circ Res.* 2018 Feb 2;122(3):433–43.
6. Sakaue S, Kanai M, Tanigawa Y, Karjalainen J, Kurki M, Koshihara S, et al. A cross-population atlas of genetic associations for 220 human phenotypes. *Nat Genet.* 2021 Oct;53(10):1415–24.
7. Hoffmann TJ, Ehret GB, Nandakumar P, Ranatunga D, Schaefer C, Kwok PY, et al. Genome-wide association analyses using electronic health records identify new loci influencing blood pressure variation. *Nat Genet.* 2017 Jan;49(1):54–64.
8. Malik R, Chauhan G, Traylor M, Sargurupremraj M, Okada Y, Mishra A, et al. Multi-ancestry genome-wide association study of 520,000 subjects identifies 32 loci associated with stroke and stroke subtypes. *Nat Genet.* 2018 Apr;50(4):524–37.
9. Surakka I, Horikoshi M, Mägi R, Sarin AP, Mahajan A, Lagou V, et al. The impact of low-frequency and rare variants on lipid levels. *Nat Genet.* 2015 Jun;47(6):589–97.
10. Anstee QM, Darlay R, Cockell S, Meroni M, Govaere O, Tiniakos D, et al. Genome-wide association study of non-alcoholic fatty liver and steatohepatitis in a histologically characterised cohort☆. *J Hepatol.* 2020 Sep 1;73(3):505–15.
11. Acosta-Herrera M, Kerick M, González-Serna D, Consortium MG, Consortium SG, Wijmenga C, et al. Genome-wide meta-analysis reveals shared new loci in systemic seropositive rheumatic diseases. *Ann Rheum Dis.* 2019 Mar 1;78(3):311–9.

12. Wuttke M, Li Y, Li M, Sieber KB, Feitosa MF, Gorski M, et al. A catalog of genetic loci associated with kidney function from analyses of a million individuals. *Nat Genet.* 2019 Jun;51(6):957–72.
13. Yengo L, Sidorenko J, Kemper KE, Zheng Z, Wood AR, Weedon MN, et al. Meta-analysis of genome-wide association studies for height and body mass index in ~700000 individuals of European ancestry. *Hum Mol Genet.* 2018 Oct 15;27(20):3641–9.
14. Mahajan A, Taliun D, Thurner M, Robertson NR, Torres JM, Rayner NW, et al. Fine-mapping type 2 diabetes loci to single-variant resolution using high-density imputation and islet-specific epigenome maps. *Nat Genet.* 2018 Nov;50(11):1505–13.
15. Atkins JL, Jylhävä J, Pedersen NL, Magnusson PK, Lu Y, Wang Y, et al. A genome-wide association study of the frailty index highlights brain pathways in ageing. *Aging Cell.* 2021;20(9):e13459.
16. Akiba T, Sano S, Yanase T, Ohta T, Koyama M. Optuna: A Next-generation Hyperparameter Optimization Framework [Internet]. arXiv; 2019 Jul. Report No.: 1907.10902. Available from: <https://arxiv.org/abs/1907.10902>
17. Bergstra J, Bardenet R, Bengio Y, Kégl B. Algorithms for Hyper-Parameter Optimization. In: *Advances in Neural Information Processing Systems* [Internet]. Curran Associates, Inc.; 2011 [cited 2023 Jun 14]. Available from: [https://proceedings.neurips.cc/paper\\_files/paper/2011/hash/86e8f7ab32cfd12577bc2619bc635690-Abstract.html](https://proceedings.neurips.cc/paper_files/paper/2011/hash/86e8f7ab32cfd12577bc2619bc635690-Abstract.html)