

1 **Development and validation of a neural network-based** 2 **survival model for mortality in ischemic heart disease**

3 Peter C. Holm^a, Amalie D. Haue^{a,b}, David Westergaard^{a, c, d}, Timo Röder^a, Karina Banasik^{a,c},
4 Vinicius Tragante^e, Alex H. Christensen^{b,f,g}, Laurent Thomas^{h,i}, Therese H. Nøstⁱ, Anne-Heidi
5 Skogholtⁱ, Kasper K. Iversen^{f,g}, Frants Pedersen^{b,f}, Dan E. Høfsten^{b,f}, Ole B. Pedersen^{f,j}, Sisse
6 Rye Ostrowski^{f,k}, Henrik Ullum^{f,k,l}, Mette N. Svendsen^m, Iben M. Gjødsbøl^m, Thorarinn
7 Gudnasonⁿ, Daníel F. Guðbjartsson^e, Anna Helgadóttir^e, Kristian Hveemⁱ, Lars V. Køber^{b,f},
8 Hilma Holm^e, Kari Stefansson^{e,o}, Søren Brunak^{a,p, q}, and Henning Bundgaard^{b,f}

9 ^a Novo Nordisk Foundation Center for Protein Research, University of Copenhagen,
10 Blegdamsvej 3B, DK-2200 Copenhagen, Denmark

11 ^b Department of Cardiology, The Heart Center, Copenhagen University Hospital, Rigshospitalet,
12 Blegdamsvej 9, DK-2100 Copenhagen, Denmark

13 ^c Department Obstetrics and Gynecology, Copenhagen University Hospital, Kettegård Alle 30,
14 DK-2650 Hvidovre, Denmark

15 ^d Methods and Analysis, Statistics Denmark, Sejrøgade 11, DK-2100 Copenhagen, Denmark

16 ^e deCODE genetics, Sturlugata 8, 102 Reykjavik, Iceland

17 ^f Department of Clinical Medicine, Faculty of Health and Medical Sciences, University of
18 Copenhagen, Blegdamsvej 3B, DK-2200 Copenhagen, Denmark

19 ^g Department of Cardiology, Copenhagen University Hospital, Herlev-Gentofte Hospital,
20 Borgmester Ib Juuls Vej 1, DK-2730 Herlev, Denmark

21 ^h Department of Clinical and Molecular Medicine, Norwegian University of Science and
22 Technology, 7491 Trondheim, Norway

23 ⁱ K.G. Jebsen Center for Genetic Epidemiology, Department of Public Health, Norwegian
24 University of Science and Technology, 7491 Trondheim, Norway

25 ^j Department of Clinical Immunology, Zealand University Hospital, DK-4600 Køge, Denmark

26 ^k Department of Clinical Immunology, Copenhagen University Hospital, Rigshospitalet,
27 Blegdamsvej 9, DK-2100 Copenhagen, Denmark.

1

1 ^lStatens Serum Institut, Artillerivej 5, DK-2300 Copenhagen, Denmark

2 ^m Department of Public Health, University of Copenhagen, Øster Farimagsgade 5, DK-1353
3 Copenhagen, Denmark

4 ⁿLaeknasetrid Cardiology Clinic, Thonglabakka 1, 109 Reykjavík, Iceland

5 ^o Faculty of Medicine, University of Iceland, Vatnsmyrarvegur 16, Reykjavik 101, Iceland

6 ^p Copenhagen University Hospital, Rigshospitalet, Blegdamsvej 9, DK-2100 Copenhagen,
7 Denmark

8 ✉ Correspondence: Søren Brunak (soren.brunak@cpr.ku.dk)

9

1 **Abstract**

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

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

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

16 **Findings:** On the test set, the tdAUC was 0.88 (95% CI 0.86-0.90, case count, cc=196) at six
17 months, 0.88(0.86-0.90, cc=261) at one year, 0.84(0.82-0.86, cc=395) at three years, and
18 0.82(0.80-0.84, cc=763) at five years. On the same data, GRACE2.0 had a lower performance:
19 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.
20 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 patient¹⁻³. 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) guidelines¹. 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)⁴⁻⁷.

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
13 imputation of missing data or expert feature engineering^{8,9}, leveraging the multitude of
14 heterogeneous healthcare data stored in modern EHRs and national registries in the development
15 of clinical decision support tools.

16 ML-based approaches have shown promising results for risk-estimation in cardiology with better
17 performance than traditional models¹⁰⁻¹³. 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
20 alone¹¹. Similarly, Mohammad and colleagues developed and validated a neural network to
21 predict 1-year mortality and re-admission for heart failure after incident myocardial infarction
22 with greater discrimination than the GRACE2.0 score¹². However, the majority of the secondary
23 risk prediction models based on ML have not used time-to-event analysis. One notable exception

1 is the model presented by Steele et al, which however have not been externally validated¹⁰. By
2 not using survival analysis, previous ML-based analyses omit data points with incomplete
3 follow-up (censoring) and thereby effectively prevents a model from distinguishing between
4 “died after a week” and “died after 10 months” which we believe is of obvious clinical interest.

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

1 **Methods**

2 **Data foundation**

3 The algorithm PMHnet was developed using a cohort constructed from the Danish National
4 Patient Registry (NPR) and the Eastern Danish Heart Registry (EDHR)¹⁴. The EDHR contains
5 structured information on all coronary artery angiographies (CAGs) performed in the Capital
6 Region of Denmark and Region Zealand. The cohort was linked to a population-wide EHR
7 database that covers Eastern Denmark from 1st of January 2006 to the 7th of July 2016 (BTH),
8 and genotype data from the Copenhagen Hospital Biobank Cardiovascular Diseases study (CHB-
9 CVD)¹⁵⁻¹⁷. The BTH dataset fully covered Eastern Denmark (2.6 million patients). Outcomes
10 were obtained from the Central Person Registry and the Danish Register for Causes of Death^{18,19}.
11 Data sources were linked using encrypted Danish personal identification numbers¹⁹.

12 **Selection criteria and model development**

13 First, we identified all adult Danish citizens (>18 years of age) in NPR with an ICD-10 code for
14 IHD (I20-I25) who had undergone their first CAG between Jan 1, 2006, and Jun 1, 2016,
15 demonstrating one-, two-, or three-vessel disease (1-3VD) or diffuse atheromatosis (DIF).
16 Vascular disease is here defined as stenosis above 50%²⁰. For patients fulfilling these criteria
17 (n=39,746), we used the date of the CAG as the index date and included five years of follow-up.
18 Patients were followed until either death or censoring, whichever came first. Using the hold-out
19 method, the derivation data was randomly divided into a training set (n=34,746) and a test set
20 (n=5,000) used for model development and independent assessment of performance,
21 respectively²¹ (Figure 1, Table 1).

1 For each of the 39,746 patients, we reduced the available features prior to index event (i.e. first
2 CAG between Jan 1, 2006, and Jun 1, 2016) to a smaller set based on prevalence such that e.g., a
3 diagnosis code could be found in at least 5% of the training set. The final set of 584 features was
4 separated into five different categories: *ClinicalOne* (8 features), *ClinicalTwo* (15 features),
5 *Diagnoses* (322 features), *Procedures* (154 features), and *Biochemical* (85 features) (Table 2).
6 *ClinicalOne* included the same eight input features as used by GRACE2.0 and *ClinicalTwo* had
7 14 additional clinical features (Table 2) that were selected based on availability. Features were
8 defined using data recorded prior to the index date, except for creatinine, cardiac biomarkers for
9 ischemic heart disease, and blood pressure where high missingness led us to allow measurements
10 obtained after the CAG in cases of missingness (7-day threshold for cardiac biomarkers, and 31-
11 day threshold for the others). *Diagnoses* included ICD-10 codes registered in NPR and similarly,
12 *Procedures* consisted of procedure codes (surgery and examinations such as X-rays) registered in
13 NPR. *Biochemical* contained results of in-hospital blood tests. Additional details on feature
14 extraction, missingness, and pre-processing can be found in supplementary methods. The
15 amounts of missingness across the training and test set have been tabulated in Table S1. Missing
16 values were left missing and encoded as such in PMHnet.

17 **External validation data from Iceland**

18 For the external validation cohort, we identified Icelandic adults who had undergone CAG at the
19 only interventional cardiology center in Iceland, Landspítali— The National University Hospital
20 in Reykjavík²². We obtained data collected prospectively between January 1, 2007, and
21 December 31, 2017. Information on ICD-10 diagnoses and procedure codes were aggregated
22 from the Landspítali, from registers kept by the Directorate of Health: The Register of Primary
23 Health Contacts, the Register of Contacts with Medical Specialists in Private Practice and the

1 Causes of Death Register, as well as at recruitment for deCODE studies. Biochemical assay
2 measurements were obtained from the three largest clinical laboratories in Iceland, with
3 measurements performed at: (i) Landspítali; (ii) The Laboratory in Mjódd, Reykjavík, Iceland;
4 and (iii) Akureyri Hospital, the regional hospital in North Iceland.

5 **Polygenic risk scores**

6 Polygenic risk scores (PRSs) for patients with genotypes available through the CHB-CVD^{17,23}
7 (31.4% of the cohort) were calculated using the LDpred2 framework, implemented in the R
8 package bigsnpr (v1.5.2) with R version 3.5.052²⁴. PRSs were calculated based on GWAS
9 summary statistics data from 19 traits relevant for cardiometabolic health, obtained from 17
10 GWAS meta-analyses. List of meta-analyses and details on the PRS calculations is included in
11 the Supplementary Material.

12 **Machine learning model architecture and development**

13 To model time-to-event data and allow for censoring, we used the generic discrete-time survival
14 model for neural networks described by Gensheimer and Narasimhan²⁵. In this model, follow-up
15 time is divided into a fixed number of intervals and the model estimates a conditional hazard for
16 each interval, i.e., the probability of dying in that time interval given that the patient is still alive
17 at the end of the preceding interval. PMHnet uses 30 intervals separated in time such that event
18 times in the training data are evenly distributed across all intervals. To obtain predictions
19 between breakpoints in the discretization grid, we assumed that the probability density function
20 was constant in each time interval, and we thus interpolated using a piecewise linear function²⁶.
21 The implementation applied the PyTorch machine-learning framework using the authors' Keras
22 version as a reference²⁵.

1 We used a feed-forward neural network and tested various hyperparameters. The output layer
2 was a fully connected sigmoid activated layer that outputs conditional hazards for each of the 30
3 different time points. We added dropout to each of the hidden layers to regularize the network
4 and prevent over-fitting. The number of layers, neurons, learning-rate, and dropout rate for each
5 layer were fine-tuned through hyperparameter optimization using the Optuna optimization
6 framework, with a five-fold cross validation²⁷. The hyperparameter search space and the best
7 trial for the complete model is included in table S3. The neural networks were trained using
8 stochastic gradient descent, with a constant learning rate, to minimize the negative log-
9 likelihood.

10 **Model evaluation and validation**

11 Using the hold-out test set and the external validation data from Iceland, performance of PMHnet
12 was evaluated through assessment of both model discrimination and calibration. We used time-
13 dependent area under the receiver operating characteristic curve (tdAUC) as the main measure of
14 discrimination, but also calculated the Brier score that can be used to assess both discrimination
15 and calibration²⁸⁻³⁰. Calibration was also analyzed graphically by comparing the predicted risks
16 with the estimated actual risks²⁸. The Score function from the riskRegression R package was
17 used to compute performance measures and compare models. For comparisons between two
18 competing models, the Score function gives p-values that correspond to Wald tests on the
19 standard errors obtained using an estimate of the influence functions following Blanche et al.³⁰.

20 To benchmark PMHnet we calculated the GRACE2.0 for all patients in the hold-out test set,
21 using the GRACE2.0 webtool. We extracted the javascript source code for the GRACE2.0
22 [webtool](#) using the developer tools in Google Chrome. The javascript code was then manually

1 converted to an R package for automatized computation of GRACE2.0 on the entire cohort. The
2 eight variables used in GRACE2.0 were available for 51.4% of the cohort. Since GRACE2.0
3 does not allow for missing features, we imputed missing variables using the *missForest* R
4 package³¹. Imputed values were only used for calculating the GRACE2.0 score. In addition to
5 the conventional GRACE2.0 score, we re-fitted the GRACE2.0 score to our training data using
6 the PMHnet architecture. This corresponds to *model_2* in Figure 4 that only uses the features
7 from *ClinicalOne* as its input.

8 For independent validation of PMHnet we used, as described above, Icelandic EHR data from
9 8,287 patients. Of the 584 features identified in the Danish derivation cohort, we found matching
10 data for 404 features in the Icelandic data. A down-scaled model was re-trained on the Danish
11 training set to make comparisons.

12 **Explainability and effect of missing features**

13 To investigate the impact of different features on model predictions and to provide model
14 explanations, we calculated Shapley additive explanation (SHAP) values for all features and
15 patients in the training set using the SHAP python package³².

16 To assess how resilient the model was in the event of missing data, missingness was introduced
17 in the test data by replacing all values of a given feature with the median value of that feature in
18 the training data. The predictions were then compared with the predictions of PMHnet.
19 Resiliency was then quantified using change in tAUC (discrimination) and Brier scores
20 (calibration), where the predictions of PMHnet were compared to that of model with artificially
21 introduced values (a total of 584 comparisons).

1 **Statistical analysis**

2 Categorical features are reported as counts (%) and continuous features as mean [95% CI], 95%
3 CIs are obtained from standard deviations or through bootstrapping. Time-dependent AUCs and
4 Brier scores were calculated using the `riskRegression` R-package²⁸. Likewise, model
5 comparisons were obtained from the same package. All statistical analyses and visualizations
6 were performed using R version 4.1.

7 **Data access and ethics approvals**

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

14 The study was approved by the Data Protection Authority of Iceland and the National Bioethics
15 Committee of Iceland (VSN-15-114). Icelandic participants that donated biological samples
16 provided informed consent. Personal identities of the participants were encrypted with a third-
17 party system provided by the Data Protection Authority of Iceland.

18 Study design, methods, and results were reported in agreement with the TRIPOD statement^{33,34}
19 and following the STROBE recommendations³⁵.

1 **Funding**

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

1 **Results**

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

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

11 **PMHnet model predictions**

12 In the internal validation using the hold-out test set, the complete PMHnet model had tAUCs of
13 0.88 [0.86; 0.90] at six months, 0.88 [0.86; 0.90] at one year, 0.84 [0.82; 0.86] at three years
14 (Figure 2), and 0.82 [0.80; 0.84] at five years. In comparison, the corresponding values for the
15 conventional GRACE2.0 score on the same dataset were 0.77 [0.74; 0.80] at six months, 0.77
16 [0.74; 0.80] at one year, and 0.73 [0.71; 0.75] at three years. For the re-fitted GRACE2.0 score,
17 tAUCs 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]
18 at three years. Since GRACE2.0 features had to be imputed for 48.6% of the population, we also
19 evaluated the performance on the subset without missingness, which were largely the same
20 (Figure 3, dashed line). GRACE2.0 is not designed for providing predictions after three years
21 and was therefore not evaluated beyond that time-point. The difference in tAUCs between

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

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

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

14 **External validation of PMHnet**

15 For external validation the cohort of 8,287 patients from Iceland was used (Table 1, *validation*
16 *set*). Due to data availability a modified, and down-scaled version of PMHnet was used (404
17 features) on the Icelandic data. The tdAUCs were 0.86 [0.84;0.90] at six months, 0.84
18 [0.81;0.87] at one year, and 0.81 [0.79;0.83] at three years. In comparison, the performance of
19 the down-scaled version on the Danish (internal, hold-out) test set was 0.87 [0.85;0.90] at six
20 months, 0.87 [0.85;0.89] at one year, and 0.82 [0.80;0.85] at three years. The predictive
21 performances were highly concordant, while model calibration was slightly worse in the
22 Icelandic data (Figure S3).

1 Influence of the different input feature categories on the PMHnet prediction

2 To assess the importance of the different input feature categories systematically, we trained five
3 intermediate survival models using the PMHnet architecture (Figure 3). For example, *model-1*
4 used diagnosis codes as input features only, *model-2* used the clinical variables known from
5 GRACE only. The other intermediate versions were trained using different combinations of the
6 feature categories. Figure 3 shows the model discrimination at the three different prediction
7 horizons for the six different models fitted using the PMHnet architecture. The performance of
8 the intermediate model based on diagnosis codes only (*model-1*) was similar to the performance
9 of the re-fitted GRACE2.0 model (*model-2*). Interestingly, combining diagnosis codes with the
10 GRACE2.0 clinical features (*Diagnoses + ClinicalOne*) in *model-3*, there was an overall
11 increase in AUC at three years, which would suggest a synergistic effect. With the addition of
12 the *ClinicalTwo*-data, the Δ AUC between *model-3* and *model-4* was 3.4 [1.8; 5.0] at six months,
13 3.2 [1.8; 4.7] at one year, and 1.3 [0.3; 2.3] at three years. Similarly, adding *Biochemical* to the
14 input features in *model-5* associated with significant Δ AUCs of 2.1 [0.2; 3.6] at six months, 1.4
15 [0.2; 2.7] at one year, and a non-significant Δ AUC of 0.9 [-0.1; 1.9] at three years. The gain in
16 discrimination from *model-5* to the complete *model-6* (PMHnet) by adding *Procedures* to the
17 input features was only significant at the one-year prediction horizon.

18 **Testing inclusion of polygenic risk scores in PMHnet**

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

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

3 **Explainability analysis using SHAP**

4 We performed SHAP-analyses of the five-year model predictions to quantify the impact of the
5 different features on PMHnet predictions. SHAP is a technique rooted in cooperative game
6 theory that provides an estimate of feature impact on the model output. It quantifies the
7 contribution of each feature to the prediction outcome, allowing for a better understanding of
8 feature importance and model behavior. By considering the interactions and dependencies
9 between features, SHAP analysis provides insights into the specific factors influencing the
10 model's decision-making process and aids in identifying key drivers and relationships within the
11 model³⁶. Across the five different feature categories, biochemical test results and diagnosis codes
12 were most impactful, while the category *ClinicalOne* was least impactful (Figure 5A). The most
13 impactful diagnosis code was chronic obstructive pulmonary disease (ICD10: J44) and thus
14 ranked higher than classical risk factors for IHD, such as type 2 diabetes. At the feature level,
15 age, the number of affected vessels (1–3VD, or DIF), and smoking were on-average the most
16 predictive features. The top-25 features in terms of average model impact (average magnitude of
17 SHAP-value), included ten features from *Biochemical*, nine from *ClinicalTwo*, three from
18 *ClinicalOne*, two from *Diagnoses*, and one from *Procedures* (Figure 5A). To further examine the
19 impact on model prediction for the two most impactful features, number of affected *vessels* and
20 *age*, we constructed SHAP dependence plots (Figure 5B)³⁷. For *age* we observed non
21 surprisingly that higher age pulled the prediction towards non-survival, and that age was
22 estimated to add/remove anywhere between -25 and 20 percent points to the predicted 5-year
23 survival. Similarly for *vessels*, more affected vessels impacted the predicted survival negatively.

1 For both features, we noted that identical feature values not always impacted the survival to the
2 same extent. This vertical dispersion in both plots represent interaction effects with the other
3 included features³⁷. Although our SHAP analyses reveal that many features have lower impact
4 relative to the most impactful features, the aggregated sum of the many low-impact features (560
5 lowest) outweighs many of the well-known risk-factors (25 highest). For this reason, we did not
6 attempt to limit the number of included features. Extending the analysis of feature-level model
7 impact to the rest of the top-25 features, we constructed a summary plot of the SHAP-values that
8 shows the distribution of model impacts across feature values (Figure 6).

9 **Patient-level feature importance**

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

1 **Discussion**

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

14 To aid the clinical interpretation of model predictions, we used SHAP-values to highlight the
15 most impactful features and to explain how the different features affect the prediction for the
16 individual patient. Model explainability is important for evaluating the model output and is
17 paramount for the clinical adaption of any ML-model³⁸, including focus on features that are
18 clinically actionable, i.e. modifiable versus non-modifiable factors.

19 Compared to the GRACE2.0 score, which is widely considered the gold-standard risk-
20 stratification tool in current clinical use for predicting mortality after acute coronary syndrome
21 (ACS)³⁹, PMHnet had superior discrimination and calibration. However, it is important to note
22 that there are differences between the intended patient populations for the two models. The
23 GRACE2.0 score has been developed using a derivation cohort of patients with STEMI, n-

1 STEMI, and unstable angina with time of initial admission as time-zero⁵. In contrast, our study
2 used time at coronary angiography as its baseline and was applied to all patients with IHD and
3 coronary artery pathology ranging from diffuse atheromatosis to three-vessel disease. The
4 validity of GRACE for a cohort with such characteristics has not been established. To provide a
5 direct comparison of the two algorithms, we used the GRACE2.0 features and re-fitted the
6 GRACE2.0 using our training data. The re-fitted GRACE2.0 score had the same model
7 discrimination and better model calibration than the original version (Fig. 2), but still had inferior
8 prediction compared to PMHnet.

9 The above observations are in good agreement with the current literature on ML-based
10 secondary risk-stratification models, which have found machine learning models to offer better
11 performance than simpler, existing scores in current clinical use^{10-13,40,41}. A transition towards
12 feature-rich models that utilize more of the available data can therefore be an advantage. The 584
13 features used in our final model are neither bespoke nor specifically collected for this decision-
14 support application, and instead represent clinical information gathered during routine work-up,
15 management and treatment of patients. Using models that rely on several hundred features means
16 that the current practice of manually entering data into a webtool becomes very impractical⁴².
17 Instead, novel risk-prediction models need to be fully integrated in the EHR systems such that
18 data can be automatically pulled and integrated.

19 Among previously published machine-learning models for secondary prediction in IHD, our
20 study has several strengths. Firstly, whereas almost all the existing literature uses binary
21 classification, the use of survival or time-to-event models is less explored. One notable exception
22 is the model reported by Steele et al.¹⁰ which employed random survival forests and elastic net
23 Cox regression to predict mortality in patients (n=80,000) with a history of coronary artery

1 disease. One of the defining characteristics of survival models is the ability to handle censored
2 data⁴³, which in other types of models would have been left out. In addition, survival models can
3 distinguish between “died after a week” and “died after 10 months” which e.g., would be
4 identical in a 1-year binary classification model. To the best of our knowledge, we are the first to
5 use neural network-based survival models in this context. Secondly, we externally validated our
6 model using data from a different country. Most models in the literature were not suboptimal
7 externally validated^{10,11,40,41}, and among those that have been, only two used data from a different
8 country^{12,13}. Demonstrating that a given model can accurately predict beyond borders is crucial
9 for generalizability. Thirdly, our model can predict all-cause mortality up to five years after the
10 index date. Probably owing to the fact that survival models have not been used, most models in
11 the domain exclusively operate with a prediction horizon of one or two years^{12,13,40,41}. Longer
12 prediction horizons might be necessary for long-term disease management.

13 Although the features we have used represent data collected from a typical clinical workflow and
14 therefore should be generally applicable, inter-regional and inter-national differences in clinical
15 practice may affect what data is available and when. This is for instance exemplified by
16 differences in diagnostic work-up, or timely access to coronary angiography, but may also relate
17 to differences in access to previous medical data. Such aspects could affect the generalizability
18 of our included features, but with internationally accepted treatment guidelines the differences
19 should be minimal. Possible solutions are to reduce slightly the complexity of models to better
20 match the intersection of features available across countries/regions and/or re-fitting and possibly
21 retraining the model each time it is deployed in a new setting. In our external validation we used
22 data from Iceland, which in an international perspective is very similar to Denmark when
23 comparing healthcare systems^{44,45}. For that reason, we downscaled the model slightly to account

1 for data availability, but re-training on Icelandic data was not deemed necessary. That being said,
2 using the Icelandic data, we found evidence of miscalibration as the model was found to
3 overestimate the risk for some patients and this suggests that further adjustment of the model is
4 needed were it to be deployed in Iceland. However, since miscalibration does not affect the
5 accurate risk-stratification of patients, we did not pursue that issue further in this study. For
6 future studies, we note that techniques such as Platt-scaling or isotonic regression could be used
7 to remedy calibration-issues⁴⁶.

8 The dynamic nature of clinical environments can lead to deterioration of model performance^{47,48}.
9 Advances in treatment and diagnosis mean that the baseline risk of patients with ischemic heart
10 disease could change over time, which in turn would lead to a drift of model calibration. As an
11 example, the logistic EuroSCORE⁴⁹, a pan-European risk-stratification model for cardiac surgery
12 published in 2003 was since its inception gradually found to overestimate mortality⁵⁰. The
13 EuroSCORE has since been replaced by EuroSCORE II which for the time being has remedied
14 these issues⁵¹. This type of systematic decline of model performance has important implications
15 for our model as well and necessitates that the model performance is continuously monitored to
16 ensure acceptable up-to-date performance. The need for real-time monitoring of model
17 performance is another strong argument for risk-stratification tools to be tightly integrated within
18 EHR systems.

19 As a secondary analysis in this study, we tested adding a panel of polygenic risk scores to the
20 input features in PMHnet and assessed how they might impact model performance. The 19
21 different genetic risk scores were included based on being related to cardiometabolic health and
22 covered traits such as *blood pressure*, and *total cholesterol*, but also included *heart failure* and
23 *acute myocardial infarction*. Where for example the *acute myocardial infarction* risk score is

1 developed for primary risk prediction, i.e., disease development, our model is concerned with
2 secondary risk prediction, i.e., modelling risk for those who already have the disease. It is known
3 that CAD PRS associate with events in both primary and secondary event populations, however,
4 these PRSs are PRSs of prevalent diseases, not mortality. Whether or not a primary risk score is
5 useful in a secondary risk context is clear upfront, but as parts of the underlying disease process
6 are known to be shared between the two, we hypothesized that the score could be used in our
7 context. Focusing the analysis only on the subset of patients for which we could obtain
8 genotypes (n=13,449, 37.4%), we found no significant difference in performance after adding the
9 PRS scores to either *model-1 (Diagnoses)* and *model-2 (Diagnoses + ClinicalOne)*. As the 585
10 features include the prior disease history recorded over more than 25 years, our interpretation is
11 that a “realized” life-course disease trajectory is more informative than the germline risk that can
12 be calculated from the genotype. Moreover, the disease trajectory also holds information on life-
13 course exposures and may therefore also include exposure information that quite implicitly is
14 related to genetic data only. The version of PMHnet trained on the disease history only, with and
15 without genetics, indicates that there is little gain from the PRS tested in this case. As we did not
16 have genetic data for the full cohort, we cannot exclude that our interpretation is affected by this
17 aspect.

18 Prospective studies are needed to ascertain how feature-rich risk-stratification methods can be
19 used to alter, guide, and hopefully improve treatment. The ability to accurately predict high-risk
20 is useful for identifying patients that may benefit from more extensive treatment and more
21 frequent visits at the hospital. In contrast, accurate identification of low-risk patients may
22 potentially be used to limit work-up, extent of pharmacological treatment and follow-up content
23 and intensity and thereby prevent potential harmful overtreatment. Striking the correct balance

1 between over- and undertreatment can contribute to the advancement of precision medicine, and
2 here a well-calibrated and highly discriminative risk-prediction model can serve as an important
3 tool.

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

1 **Data availability statement**

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

6 **Conflicts of interest**

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

12 **Acknowledgements**

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

1 **References**

- 2 1. Collet, J.-P. *et al.* 2020 ESC Guidelines for the management of acute coronary syndromes in
3 patients presenting without persistent ST-segment elevation: The Task Force for the
4 management of acute coronary syndromes in patients presenting without persistent ST-
5 segment elevation of the European Society of Cardiology (ESC). *Eur. Heart J.* **42**, 1289–
6 1367 (2021).
- 7 2. Knuuti, J. *et al.* 2019 ESC Guidelines for the diagnosis and management of chronic coronary
8 syndromes: The Task Force for the diagnosis and management of chronic coronary
9 syndromes of the European Society of Cardiology (ESC). *Eur. Heart J.* **41**, 407–477 (2020).
- 10 3. Steg, Ph. G. *et al.* ESC Guidelines for the management of acute myocardial infarction in
11 patients presenting with ST-segment elevation. *Eur. Heart J.* **33**, 2569–2619 (2012).
- 12 4. Wilson, P. W. F. *et al.* Prediction of Coronary Heart Disease Using Risk Factor Categories.
13 *Circulation* **97**, 1837–1847 (1998).
- 14 5. Fox, K. A. A. *et al.* Should patients with acute coronary disease be stratified for management
15 according to their risk? Derivation, external validation and outcomes using the updated
16 GRACE risk score. *BMJ Open* **4**, e004425 (2014).
- 17 6. Hung, J. *et al.* Performance of the GRACE 2.0 score in patients with type 1 and type 2
18 myocardial infarction. *Eur. Heart J.* **42**, 2552–2561 (2020).
- 19 7. Antman, E. M. *et al.* The TIMI Risk Score for Unstable Angina/Non–ST Elevation MI.
20 *JAMA* **284**, 835 (2000).
- 21 8. Rajkomar, A., Dean, J. & Kohane, I. Machine Learning in Medicine. *N. Engl. J. Med.* **380**,
22 1347–1358 (2019).

- 1 9. Topol, E. J. High-performance medicine: the convergence of human and artificial
2 intelligence. *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
10 algorithm to predict mortality and admission to hospital for heart failure after myocardial
11 infarction: a 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,
17 and research potential. *Clin. Epidemiol.* 449 (2015) doi:10.2147/cep.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
20 Patient 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.*
8 *Surv.* **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. & Vilhjálmsón, 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).

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

- 1 39. D'Ascenzo, F. *et al.* TIMI, GRACE and alternative risk scores in Acute Coronary
2 Syndromes: A meta-analysis of 40 derivation studies on 216,552 patients and of 42
3 validation studies on 31,625 patients. *Contemp. Clin. Trials* **33**, 507–514 (2012).
- 4 40. Kwon, J. *et al.* Deep-learning-based risk stratification for mortality of patients with acute
5 myocardial infarction. *PLOS ONE* **14**, e0224502 (2019).
- 6 41. Wallert, J., Tomasoni, M., Madison, G. & Held, C. Predicting two-year survival versus non-
7 survival after first myocardial infarction using machine learning and Swedish national
8 register data. *BMC Med. Inform. Decis. Mak.* **17**, 99 (2017).
- 9 42. Sharma, V. *et al.* Adoption of clinical risk prediction tools is limited by a lack of integration
10 with electronic health records. *BMJ Health Care Inform.* **28**, e100253 (2021).
- 11 43. George, B., Seals, S. & Aban, I. Survival analysis and regression models. *J. Nucl. Cardiol.*
12 *Off. Publ. Am. Soc. Nucl. Cardiol.* **21**, 686–694 (2014).
- 13 44. Einhorn, E. S. Nordic Health Care Systems: Recent Reforms and Current Policy Challenges.
14 *Scand. Stud.* **84**, 106–108 (2012).
- 15 45. Kristiansen, I. S. & Pedersen, K. M. [Health care systems in the Nordic countries--more
16 similarities than differences?]. *Tidsskr. Den Nor. Laegeforening Tidsskr. Prakt. Med. Ny*
17 *Raekke* **120**, 2023–2029 (2000).
- 18 46. Niculescu-Mizil, A. & Caruana, R. Predicting good probabilities with supervised learning. in
19 *Proceedings of the 22nd international conference on Machine learning* 625–632
20 (Association for Computing Machinery, 2005). doi:10.1145/1102351.1102430.
- 21 47. Davis, S. E., Lasko, T. A., Chen, G. & Matheny, M. E. Calibration Drift Among Regression
22 and Machine Learning Models for Hospital Mortality. *AMIA. Annu. Symp. Proc.* **2017**, 625–
23 634 (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.
7 *Eur. 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).
- 9
- 10

1 **Figure legends**

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

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

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

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

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

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

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

1 **Table legends**

2 **Table 1: Cohort characteristics for training, test, and external validation set.** For continuous
3 features (age, height, etc.) the mean is given along with 95% bootstrap confidence intervals (CI)
4 and, if applicable, the amount of missingness in parentheses – mean [low, high] (missingness).
5 The mean and CI are calculated using only the non-missing features. For categorical features the
6 raw counts and relative frequencies are both specified. The comorbidities are defined from the
7 ICD-10 codes that had been assigned to a given patient prior to the index date. Medication is
8 defined from prescriptions prior to index date. Lipid-lowering drugs: C10, anti-hypertensive
9 drugs: C02, C03, C07, C08 and C09, type-2 diabetes drugs: A10B, insulin: A10A. See
10 supplementary methods for further details.

11 (#): Diffuse atheromatosis could not be defined with complete certainty for 28.4% of the
12 Icelandic data, and coronary pathology was therefore set as NA in such cases.

1 *COPD: Chronic obstructive pulmonary disease, ICD: Implantable cardioverter-defibrillator,*
2 *LVEF: Left-ventricular ejection fraction. PM: Permanent pacemaker.*

3
4 *Table 2: **Input features used for model development.** The different features were organized in*
5 *five different categories each representing different domains. The clinical characteristics were*
6 *divided into two subgroups where `ClinicalOne` contains the features used in the `GRACE2.0`*
7 *score.*

8
9 *Table 3: **Difference in discrimination between `PMHnet` and the `GRACE2.0` score.** For ΔAUC ,*
10 *we obtain 95% CIs (in brackets) and p-values from the `Score` function in the R package*
11 *`riskRegression`³⁸.*

Patients in EHR database with an ICD-10 code for IHD in NPR
(n = 193 551)

Patients subjected to a CAG registered in EDHR
(n = 97 826)

Patients in EDHR with IHD
(n = 63 417)

Patients where CAG was performed in years 2006-2016
(n = 48 633)

Patients without previous IHD
(n = 40 312)

Patients excluded if:

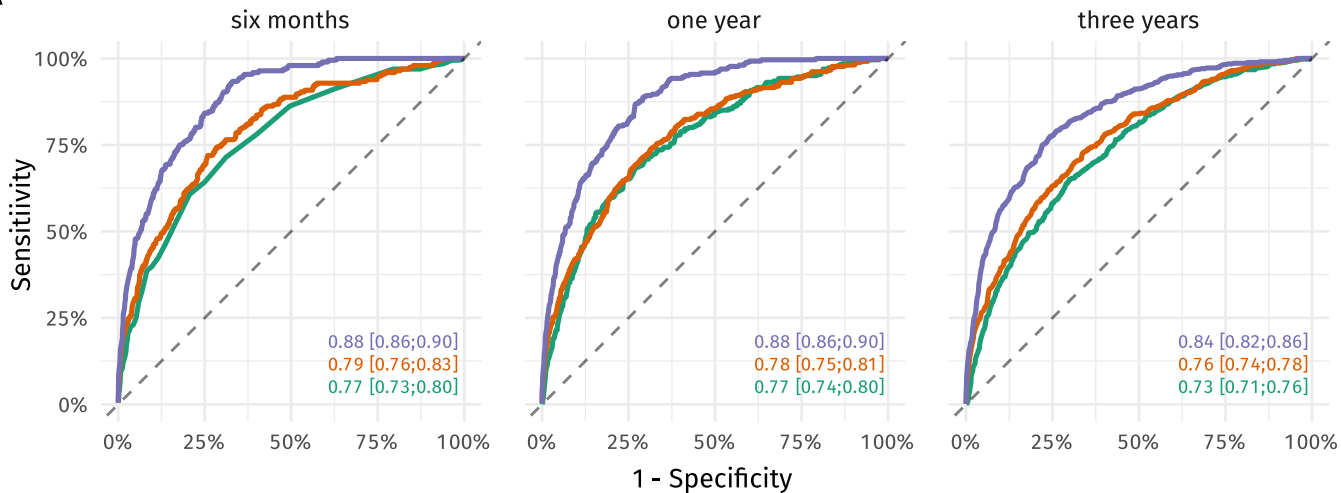
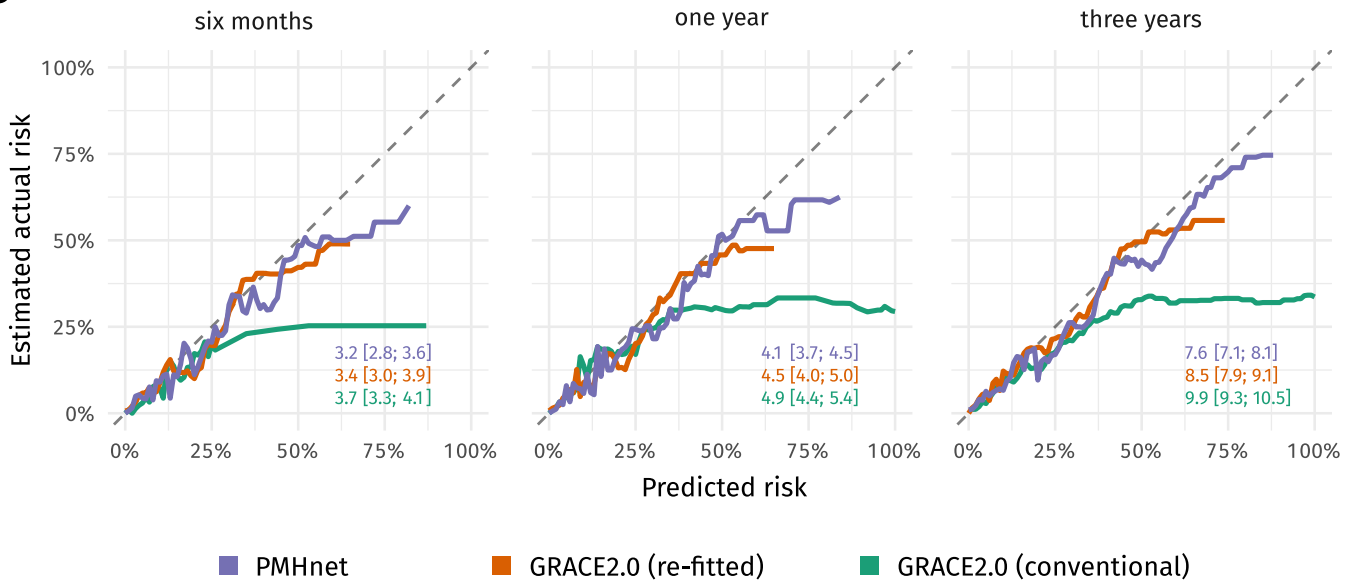
- <18 years of age
- non-danish citizen

(n = 566)

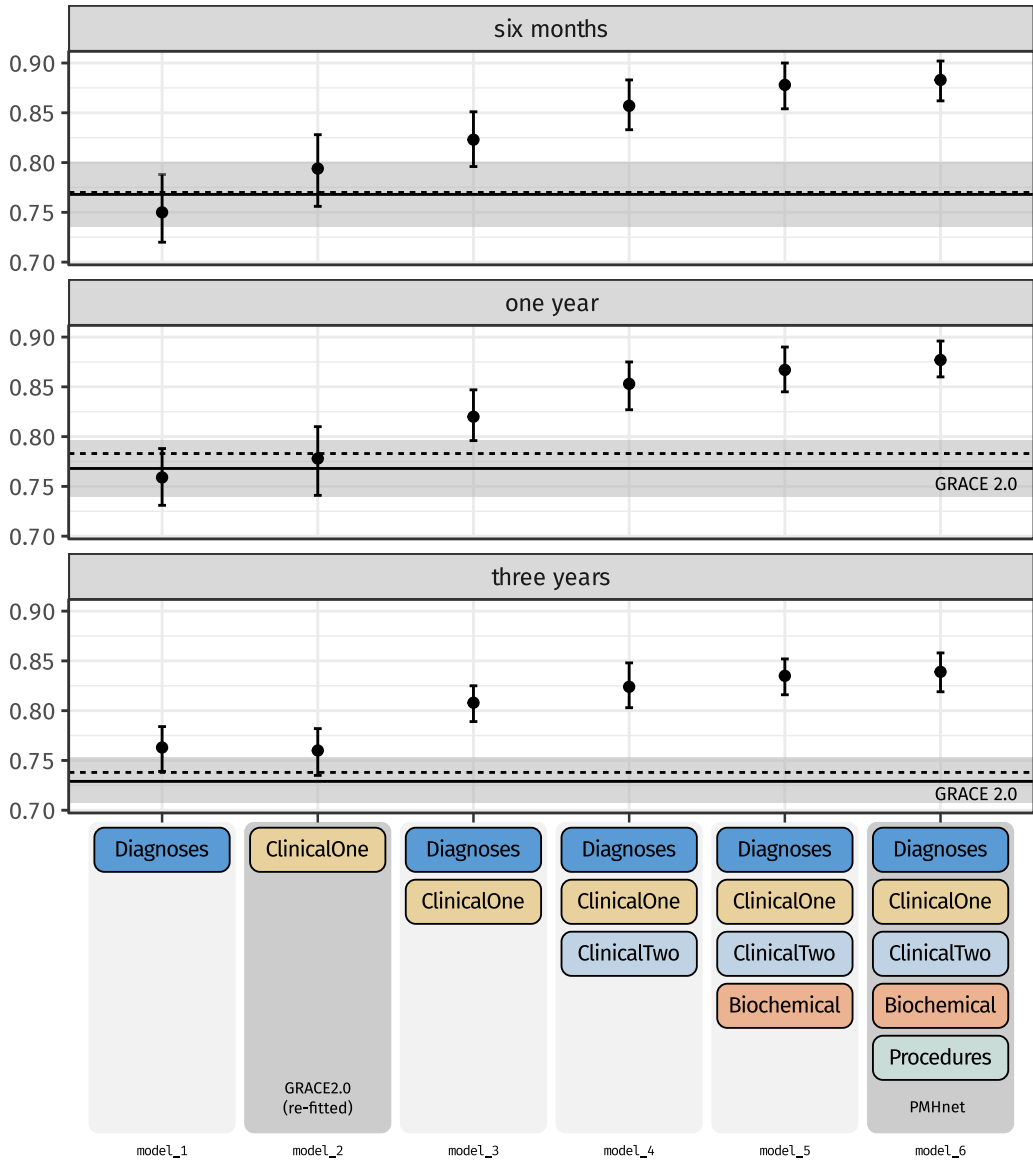
Patients in derivation cohort
(n = 39 746)

Training set for model development
(n = 34 746)

Testing set for model validation
(n = 5 000)

A**B**

Model Discrimination
(Time-dependent AUC)



GRACE 2.0

GRACE 2.0

Diagnoses

ClinicalOne

Diagnoses

ClinicalOne

Diagnoses

ClinicalOne

ClinicalTwo

Diagnoses

ClinicalOne

ClinicalTwo

Biochemical

Diagnoses

ClinicalOne

ClinicalTwo

Biochemical

Procedures

PMHnet

model_1

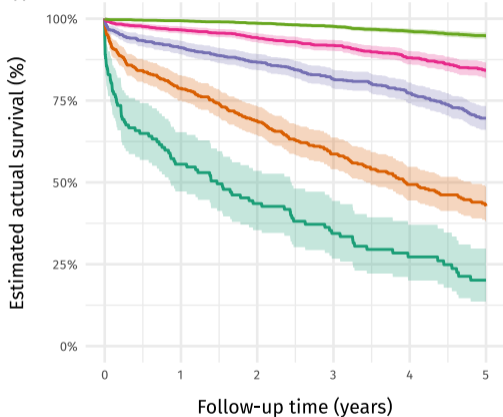
model_2

model_3

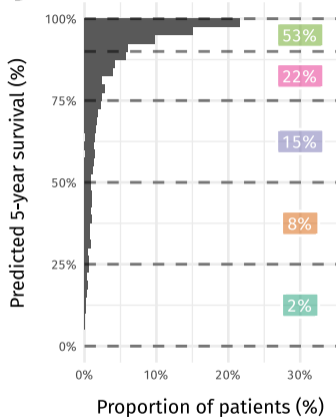
model_4

model_5

model_6

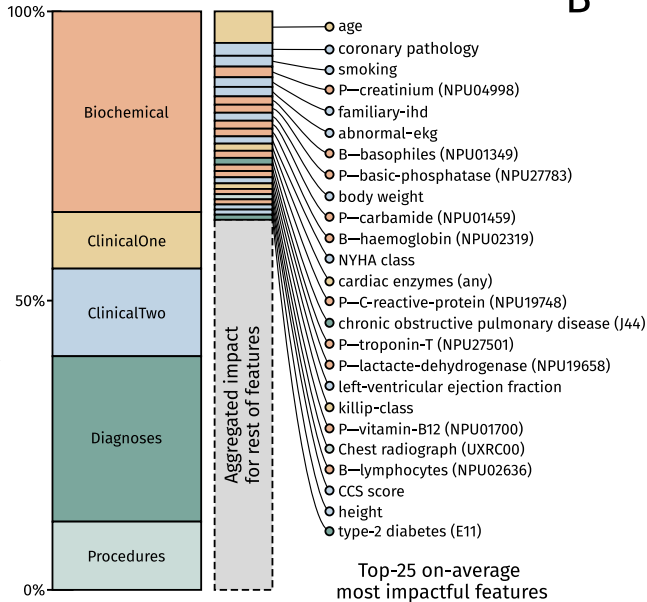
A

Risk groups (0,0.25] (0.25,0.5] (0.5,0.75] (0.75,0.9] (0.9,1]

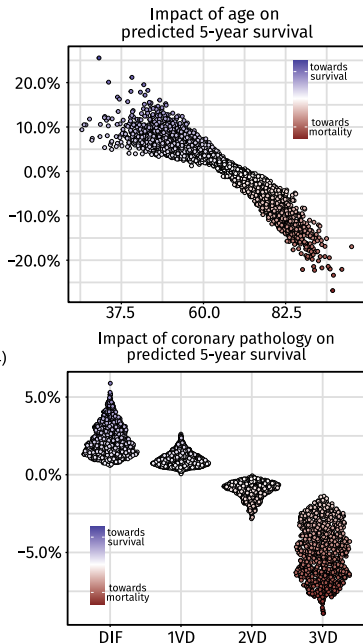
B

A

Relative feature importance (SHAP)



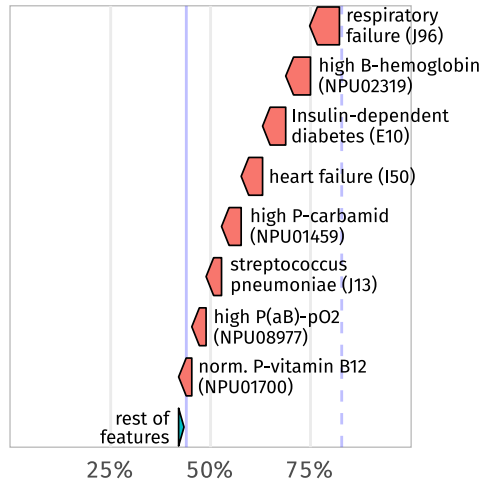
B



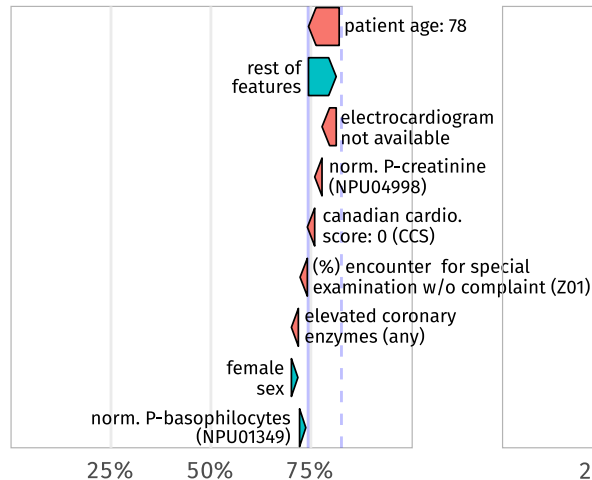


Highlighted features

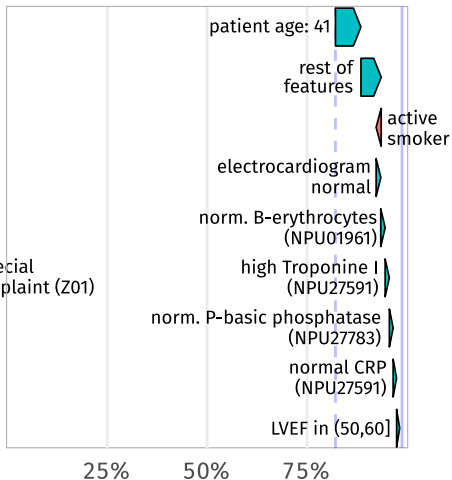
Example patient with predicted survival in (0.25, 0.5]



Example patient with predicted survival in (0.5, 0.75]



Example patient with predicted survival in (0.75, 1]



Model prediction (5-year survival)

	Training (n=34,746)	Test (n=5,000)	Validation (n=8,288)
<i>General characteristics</i>			
Male Sex	23413/67.3%	3410/68.2	5876/70.9%
Age (year)	66.0 [65.7;66.4]	66.2 [66.0;66.3]	66.0 [65.8; 66.2]
Height (cm)	172.8 [172.7;172.9] (6.5%)	173 [172.7;173.2] (6.3%)	174.3 [174.1; 174.5]
Weight (kg)	81.3 [81.1;81.5] (4.3%)	81.7 [81.1;82.1] (4.1%)	87.2 [86.8; 87.5]
<i>Comorbidities (ICD10)</i>			
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%
<i>Medication (ATC)</i>			
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%
<i>P-creatinine (mmol/l)</i>	84.8 [84.5;85.2] (10.1%)	84.2 [83.4;85.1] (10.8%)	91.7 [90.6; 92.9] (8.3%)
<i>Elevated cardiac biomarkers (CKMB, TnI, or TnT)</i>			
	15531/44.7% (38.7%)	2158/43.2% (40%)	1599/19.3% (39.6%)
<i>Heart rate, bpm</i>	74.7 [74.5;74.9] (20.2%)	74.2 [73.7;74.8] (21.1%)	69.6 [69.3; 70.0] (23.7%)
<i>Blood pressure</i>			
Systolic, mmHg	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	77.9 [77.7;78.1] (33.5%)	77.7 [77.2;78.2] (34%)	85.5 [85.2; 85.8] (14.0%)
<i>Cardiac arrest at admission?</i>	513/1.5%	69/1.4%	130/1.6%
<i>ICD or PM</i>	557/1.6%	88/1.8%	188/2.3%
<i>Killip class</i>			
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%
Killip: NA	19686/56.7%	2818/56.4%	3465/41.8%
<i>Left-ventricular ejection fraction (%)</i>			
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%
LVEF: NA	15923/45.8%	2371/47.4%	4838/58.4%
<i>Smoking status</i>			
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%
Unknown	3201/9.2%	449/9.0%	101/1.2%
<i>Coronary pathology</i>			
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%
NA	-	-	28.37% (#)

Category	Features
<i>ClinicalOne</i> (GRACE2.0)	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

	6 months		1 year		3 years	
Comparison	Δ AUC (%)	p-value	Δ AUC (%)	p-value	Δ 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