Prediction of first cardiovascular disease event in 2.9 million individuals using Danish administrative healthcare data

a nationwide, registry-based derivation and validation study

Christensen, Daniel Mølager; Phelps, Matthew; Gerds, Thomas; Malmborg, Morten; Schjerning, Anne-Marie; Strange, Jarl Emanuel; El-Chouli, Mohamad; Larsen, Lars Bruun; Fosbøl, Emil; Køber, Lars; Torp-Pedersen, Christian; Mehta, Suneela; Jackson, Rod; Gislason, Gunnar

Published in:
European heart journal open

DOI (link to publication from Publisher):
10.1093/ehjopen/oeab015

Publication date:
2021

Document Version
Publisher's PDF, also known as Version of record

ESC European Society of Cardiology

European Heart Journal Open (2021) **00**, 1–10
doi:10.1093/ehjopen/oeab015

ORIGINAL ARTICLE
Epidemiology & prevention

Prediction of first cardiovascular disease event in 2.9 million individuals using Danish administrative healthcare data: a nationwide, registry-based derivation and validation study

Daniel Mølager Christensen [1,*], Matthew Phelps [1], Thomas Gerds[1,2], Morten Malmborg [1], Anne-Marie Schjerning[1,3], Jarl Emanuel Strange[4], Mohamad El-Chouli[1], Lars Bruun Larsen[5,6], Emil Fosbøl[7], Lars Køber [7], Christian Torp-Pedersen[8,9], Suneela Mehta[10,11], Rod Jackson[10], and Gunnar Gislason[1,4]

[1]The Danish Heart Foundation, Vognmagergade 7, 3rd Floor, Copenhagen 1120, Denmark [2]Department of Biostatistics, University of Copenhagen, Øster Farimagsgade 5, Copenhagen 1014, Denmark [3]Department of Cardiology, Zealand University Hospital, Sygehusvej 10, Roskilde 4000, Denmark [4]Department of Cardiology, Copenhagen University Hospital Herlev and Gentofte, Kildegårdsvej 28, Hellerup 2900, Denmark [5]Research Unit of General Practice, University of Southern Denmark, J. B. Winsløws Vej 9A, Odense 5000, Denmark [6]Steno Diabetes Center Sjælland, Region of Zealand, Birkevænget 3, 3rd floor, Holbæk 4300, Denmark [7]Department of Cardiology, Rigshospitalet, Blegdamsvej 9, Copenhagen 2100, Denmark [8]Department of Clinical Research, Nordsjaellands Hospital, Dyrehavevej 29, Hillerød 3400, Denmark [9]Department of Cardiology, Aalborg University Hospital, Hobrovej 18-22, Aalborg 9100, Denmark [10]Section of Epidemiology and Biostatistics, University of Auckland, Park Ave 22-30, Grafton, Auckland, New Zealand; and [11]Waitematā and Auckland District Health Boards, Shea Tce 15, Level 2, Takapuna, Auckland City 0622, New Zealand

Received 15 July 2021; editorial decision 30 July 2021; accepted 31 July 2021; online publish-ahead-of-print 2 August 2021
Handling editor: Karolina Szummer

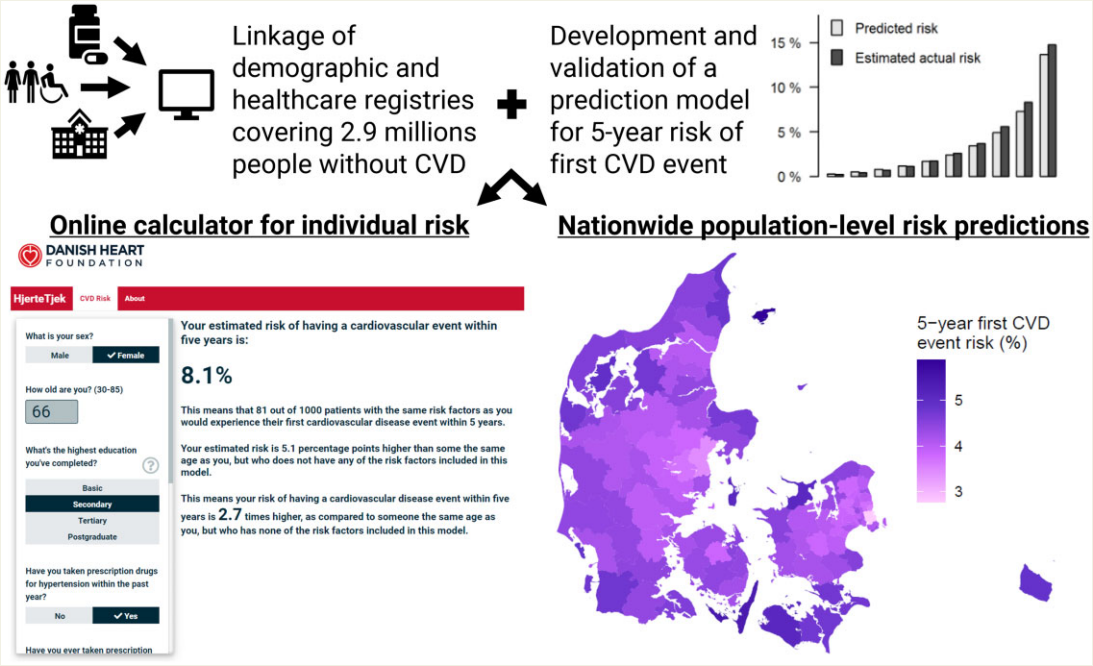
Aims The aim of this study was to derive and validate a risk prediction model with nationwide coverage to predict the individual and population-level risk of cardiovascular disease (CVD).

Methods and results All 2.98 million Danish residents aged 30–85 years free of CVD were included on 1 January 2014 and followed through 31 December 2018 using nationwide administrative healthcare registries. Model predictors and outcome were pre-specified. Predictors were age, sex, education, use of antithrombotic, blood pressure-lowering, glucose-lowering, or lipid-lowering drugs, and a smoking proxy of smoking-cessation drug use or chronic obstructive pulmonary disease. Outcome was 5-year risk of first CVD event, a combination of ischaemic heart disease, heart failure, peripheral artery disease, stroke, or cardiovascular death. Predictions were computed using cause-specific Cox regression models. The final model fitted in the full data was internally-externally validated in each Danish Region. The model was well-calibrated in all regions. Area under the receiver operating characteristic curve (AUC) and Brier scores ranged from 76.3% to 79.6% and 3.3 to 4.4. The model was superior to an age-sex benchmark model with differences in AUC and Brier scores ranging from 1.2% to 1.5% and -0.02 to -0.03. Average predicted risks in each Danish municipality ranged from 2.8% to 5.9%. Predicted risks for a 66-year old ranged from 2.6% to 25.3%. Personalized predicted risks across ages 30–85 were presented in an online calculator (https://hjerteforeningen.shinyapps.io/cvd-risk-manuscript/).

* Corresponding author. Tel: +4570250000, Email: dmchristensen@hjerteforeningen.dk

Conclusion A CVD risk prediction model based solely on nationwide administrative registry data provided accurate prediction of personal and population-level 5-year first CVD event risk in the Danish population. This may inform clinical and

Graphical Abstract



Keywords Risk prediction • Risk stratification • Cardiovascular disease • Primary prevention • Registries • Nationwide

Introduction

Cardiovascular disease (CVD) causes immense morbidity, mortality, and economic burden worldwide.[1,2] Predicted CVD risk has been used to inform treatment decisions for preventing CVD for several decades.[3] Current European Society of Cardiology and American College of Cardiology/American Heart Association guidelines recommend opportunistic routine assessment of CVD risk among individuals without prior CVD starting from age 40, using the Systemic Coronary Risk Evaluation (SCORE) and Pooled Cohort Equations (PCE) models, respectively.[3,4] These equations are derived from a range of cohorts, many recruited decades ago, and provide inferior risk prediction in modern era populations where the burden of CVD has reduced due to advances in treatment and prevention.[5] Ideally, risk prediction models should be tailored to the target populations and reflect a contemporary distribution of CVD risk.[5,6]

Large-scale administrative healthcare databases hold great potential for predictive risk modelling by providing data tailored to the target populations. An advantage of such prediction models is that they can be used beyond prediction of individual risk. They can provide countrywide, regionwide, municipality-wide, general practice-level, or even patient-list level risk predictions. This approach may guide allocation of resources and preventive efforts to high-risk areas and subpopulations. Furthermore, administrative data-based risk prediction models can be updated and recalibrated dynamically (i.e. at future time-points) and integrated into electronic health records to aid decision-making at the individual point of care. A further advantage of an administrative data-based risk prediction approach is the absence of clinical and laboratory variables.[7] This eases use in community and primary care settings where clinical and laboratory risk factor levels may not be known[8,9] and expands the concept of guideline-based opportunistic routine CVD risk assessment beyond the setting of a doctor's consultation. As far as we are aware, only one other country (New Zealand) has developed countrywide CVD risk prediction equations based solely on administrative data and these equations were well calibrated with good risk discrimination in national, regional, and ethnic populations.[10]

In the current study, we aimed to use Danish nationwide, routinely collected administrative data from more than 2.98 million individuals

without prior CVD to develop and validate a risk prediction model for the prediction of personal and population-level 5-year risk of first CVD event.

Methods

Study setting and data sources

Denmark is a country in Northern Europe of 42 933 km^2 divided into five geographical administrative regions encompassing 98 municipalities with a total population of 5 627 235 as of 1 January 2014 (Supplementary material online, *Figure S1*). All Danish permanent residents have equal access to fully tax-funded healthcare and education. Utilization of these services is registered for administrative purposes through a unique personal Civil Registration Number. This number allows de-identified linkage of nationwide administrative registries at an individual level for research purposes. For the current study, we linked registries encompassing demographics,[11] hospital contacts,[12] redeemed prescriptions,[13] educational attainment,[14] and causes of death.[15] Data on redeemed prescriptions and hospital contacts were available since 1995.[12,13] Diagnostic codes used for the present study were classified according to International Classification of Diseases, Tenth revision (ICD-10) and drug prescriptions were classified according to the Anatomical Therapeutic Chemical (ATC) Classification System.

Study population

All Danish permanent residents aged 30–85 years alive on 1 January 2014 were included in the study. Exclusion criteria were: (i) previous ambulatory or in-hospital contact with a registered diagnosis of CVD (see Supplementary material online, *Table S1* for further definitions) and (ii) missing data on educational attainment.

Outcome

The outcome of interest was first CVD event, defined as first occurrence of ischaemic heart disease (IHD), ischaemic stroke, haemorrhagic stroke, heart failure (HF), peripheral artery disease (PAD), or cardiovascular death. IHD, ischaemic stroke, haemorrhagic stroke, HF, and PAD were identified from hospital admissions using discharge diagnostic codes. Cardiovascular deaths were identified from death certificates listing a cardiovascular diagnosis among the causes of death. The cardiovascular diagnoses used in the outcome definition have been validated with high positive predictive values: IHD 88–97%, HF 80%, PAD 91–100%, and stroke 80–86%.[16,17] See Supplementary material online, *Table S1* for further descriptions and diagnostic codes.

Predictors

Predictors were pre-specified based on likelihood of being risk factors for CVD. They were selected based on inclusion in previous work by Mehta *et al.*[10] in New Zealand and amended to suit availability of data in Danish administrative registries.

The following variables were identified at baseline for the entire study population: age, sex, education, chronic obstructive pulmonary disease (COPD), and dispensing of smoking-cessation drugs, blood pressure-lowering drugs, glucose-lowering drugs, lipid-lowering drugs, and antithrombotic drugs. Levels of education were defined according to the highest level of education attained and classified as basic (e.g. primary school), secondary (e.g. high school or vocational training), tertiary (e.g. Bachelor's degree), or postgraduate (e.g. Master's degree). Smoking-cessation drug use and history of COPD were combined to one predictor as a proxy for smoking status. Dispensing of glucose-lowering drugs served as a proxy for the presence of either type 1 or type 2 diabetes, which was previously validated with a PPV of 95%.[18] Metformin prescriptions redeemed by females <40 years of age were not included in the glucose-lowering drug definition, as the indication was presumed to be polycystic ovarian syndrome.[19] Further descriptions along with ICD-10 and ATC codes were shown in Supplementary material online, *Table S1*.

Statistical analysis

Patient characteristics were presented as medians with interquartile ranges and frequencies with percentages. Study individuals were followed from study start (1 January 2014) until outcome of interest, non-cardiovascular death (competing risk), emigration, or 31 December 2018, whichever came first. Cause-specific Cox regression was used to predict the 5-year risk of first CVD event with non-cardiovascular death as a competing risk.[20] The model included age (continuous, modelled by restricted cubic splines to allow for non-linear effects), sex (male/female), education (postgraduate, tertiary, secondary, or basic), smoking proxy (yes/no), glucose-lowering drug use (yes/no), blood pressure-lowering drug use (yes/no), lipid-lowering drug use (yes/no), antithrombotic drug use (yes/no), and interactions between age and glucose-lowering drug use, age and blood pressure-lowering drug use, age and lipid-lowering drug use, blood pressure-lowering drug use and glucose-lowering drug use, and antithrombotic drug use and glucose-lowering drug use. Interactions were pre-specified based on clinical plausibility and recently developed CVD risk prediction equations from New Zealand.[5,10] We show results of complete case analyses where subjects with missing information on education are excluded. We presented personalized predicted 5-year risks of first CVD event across ages 30–85 in low-risk and intermediate-risk scenarios with and without each predictor. We presented personalized predicted 5-year risks of first CVD event for a 66-year-old individual to represent a common intermediate- to high-risk person in our population. Finally, we presented average predicted 5-year risk of first CVD event for all individuals without previous CVD aged 30–85 and 66 years, separately, in each Danish municipality.

The final model was developed in the full study population. We performed internal–external validation of the final model by assessing performance in each of the five administrative Regions (Supplementary material online, *Figure S1A*).[21,22] We also assessed model performance in separate age groups. To further validate our model, we split our data into a training set and a testing set based on geography. The testing set encompassed all study participants living in the Capital Region of Denmark and the four remaining administrative Regions served as the training set (Supplementary material online, *Figure S1B*). In addition, we performed internal validation by splitting the data randomly (training set 63%, testing set 37%). We fitted the model in the training sets and assessed predictive performance in the testing sets. To assess discrimination, areas under the receiver operating characteristics curve (AUC) were calculated.[23] To assess overall model performance, Brier scores were calculated.[24,25] To assess model calibration, we plotted deciles of the predicted 5-year risks of first CVD event against the estimated actual risks.[26] We compared model performance to a benchmark model containing only age and sex and reported differences in AUCs and Brier scores. Data management and statistical analyses were performed using R version 3.6.1.

Ethics

Studies based on pseudonymized registry data do not require ethical approval in Denmark. The data responsible institution (Capital Region of Denmark) approved the current study (approval number P-2019-537).

Table 1 Characteristics and follow-up of the study population

	Female (n = 1 558 026)	Male (n = 1 425 081)
Age (years), median [IQR]	52 [41, 64]	50 [41, 62]
Education		
Postgraduate	134 124 (8.6)	152 807 (10.7)
Tertiary	418 575 (26.9)	270 467 (19.0)
Secondary	612 468 (39.3)	673 176 (47.2)
Basic	392 859 (25.2)	328 631 (23.1)
Smoking proxy	114 185 (7.3)	93 779 (6.6)
Glucose-lowering drugs	69 998 (4.5)	79 388 (5.6)
Blood pressure-lowering drugs	352 783 (22.6)	264 452 (18.6)
Lipid-lowering drugs	178 623 (11.5)	151 037 (10.6)
Antithrombotic drugs	80 540 (5.2)	75 749 (5.3)
Follow-up		
Time (years), median [IQR]	5 [5, 5]	5 [5, 5]
CVD event	49 545 (3.2)	70 195 (4.9)
Non-fatal event		
Heart failure	5857 (0.4)	7565 (0.5)
Acute coronary syndrome	15 319 (1.0)	28 711 (2.0)
Ischaemic stroke	13 034 (0.8)	16 090 (1.1)
Transient ischaemic attack	5248 (0.3)	5612 (0.4)
Haemorrhagic stroke	3659 (0.2)	3309 (0.2)
Peripheral artery disease	1766 (0.1)	3271 (0.2)
Death		
CV	9564 (0.6)	11 251 (0.8)
Non-CV	60 965 (3.9)	63 383 (4.4)
Time to CVD event (years), median [IQR]	2.5 [1.3, 3.8]	2.5 [1.3, 3.8]

CV, cardiovascular; CVD, cardiovascular disease; IQR, interquartile range.

Results

On 1 January 2014, the Danish population comprised 3 598 511 persons aged 30–85 years. After applying the exclusion criteria of history of CVD (408 529, 11.4%) and missing information on education (206 875, 5.7%), 2 983 107 individuals were included in the final study population. Median ages were 52 and 50 years for females and males, respectively (*Table 1*, see Supplementary material online, *Figure S2* for the age distribution of the study population). Approximately a quarter of the population had basic education as their highest level of education with a higher proportion among the older age groups (Supplementary material online, *Figure S3*). Percentages of other predictors ranged from 4.5% (females, glucose-lowering drug use) to 22.6% (females, blood pressure-lowering drug use). Presence of all predictors increased with age, except for male sex (decreased with age) and lipid-lowering drug use (peaked among individuals in their seventies) (Supplementary material online, *Figure S3*). In the overall study population, 119 740 (4.0%) developed a first CVD event during a median follow-up of 2.5 years. The median age at baseline for those who had a CVD event was 66 years (Supplementary material online, *Table S2*). IHD was the most frequently recorded component of the combined outcome of first CVD event, followed by ischaemic stroke (Supplementary material online, *Figure S4*).

Low-risk and intermediate-risk scenarios

To illustrate the impact of each predictor on risk of first CVD event, we showed predicted risks across ages 30–85 years in low-risk scenarios, adding each predictor, and intermediate-risk scenarios, adding or removing each predictor (*Figure 1*). A low-risk scenario (*Figure 1A*), i.e. a combination of predictors resulting in low expected risk, was defined as a female with postgraduate education without any other predictors. An intermediate-risk scenario (*Figure 1B*), i.e. a combination of predictors resulting in an intermediate expected risk, was defined as a male with secondary education, smoking proxy, and lipid-lowering drug use. The number of individuals in the study population fulfilling the low-risk scenario and the intermediate-risk scenario criteria are shown in Supplementary material online, *Figure S5*. Glucose-lowering drug use, blood pressure-lowering drug use, smoking proxy, male sex, antithrombotic drug use, and lower education predicted a higher 5-year risk of first CVD event. Lipid-lowering drug use predicted a higher risk among younger individuals and a lower risk among those above 52 years of age. For comparison with other populations and CVD prediction models, adjusted hazard ratios for first CVD event, with age as a two-level categorical variable, were presented (Supplementary material online, *Figure S6*).

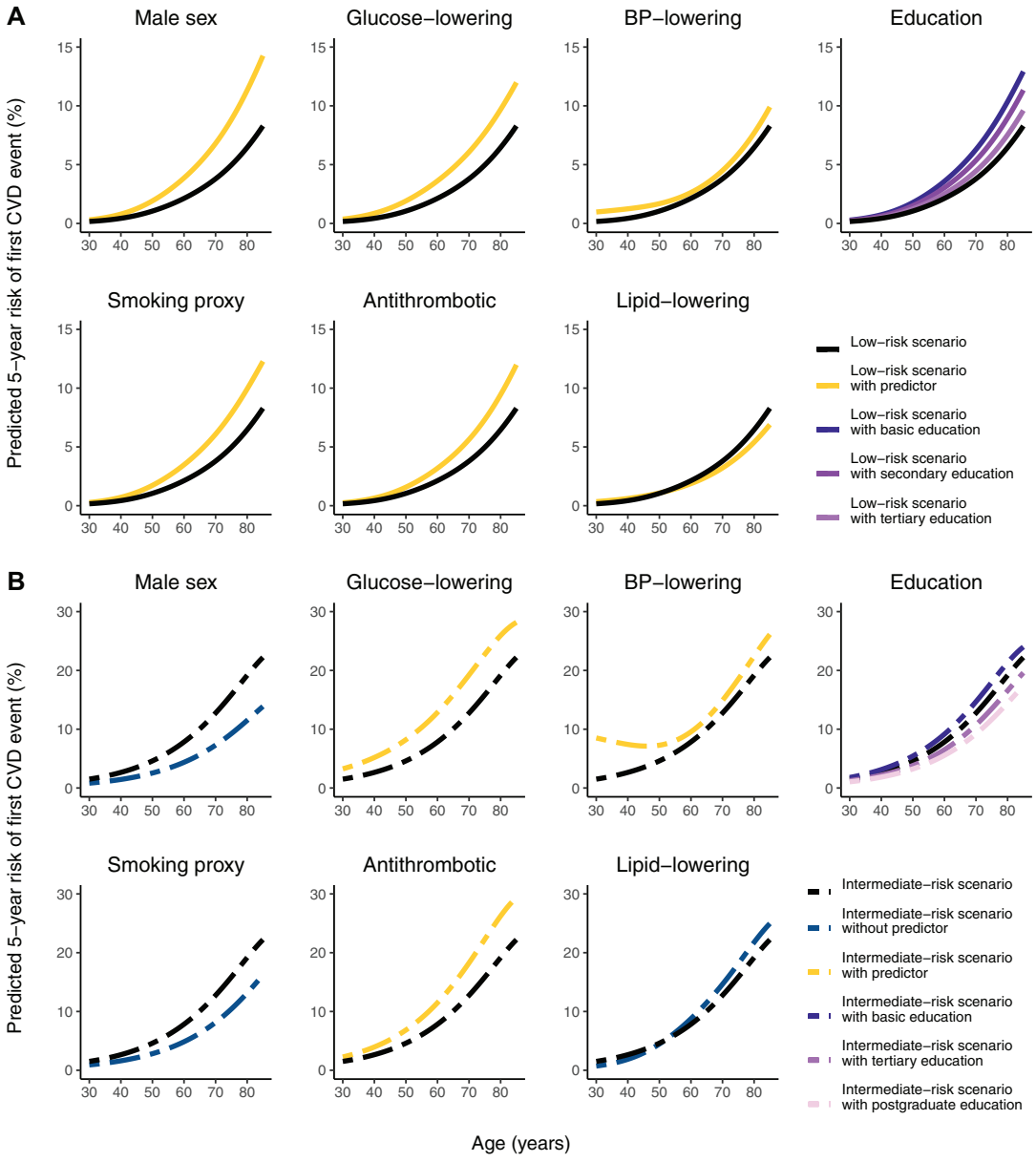


Figure 1 Predicted 5-year risk of first CVD event across ages 30–85 in low-risk and intermediate-risk scenarios with and without each predictor. The low-risk scenario (*A*) was defined as female sex, postgraduate education, and absence of all other predictors. The intermediate-risk scenario (*B*) was defined as male sex, basic education, smoking proxy = yes, and lipid-lowering drugs = yes. BP, blood pressure; CVD, cardiovascular disease.

Personalized risk predictions

To illustrate the personalized predicted 5-year risk of first CVD event for a given individual, we showed predicted risks for a 66-year old with all possible combinations of predictors (*Figure 2*). The highest predicted risk for a 66-year old was 25.3% for a male with basic education, glucose-lowering drug use, blood pressure-lowering drug use, smoking proxy, and antithrombotic drug use. The lowest predicted risk for a 66-year old was 2.6% for a female with postgraduate education and lipid-lowering drug use. Predicted 5-year risks of first CVD event in the overall population ranged from 0.17% to 38.4%. The most common predictor combinations, excluding the low-risk scenario, were male sex, smoking proxy, and secondary education in the lowest age groups, and female sex, blood pressure-lowering drug use, and basic education in the highest age groups (Supplementary material online, *Figure S7*). To estimate the personalized predicted 5-year risk of first CVD event at any age, for any combination of predictors, we provided an online risk calculator (https://hjerteforeningen.shinyapps.io/cvd-risk-manuscript/).

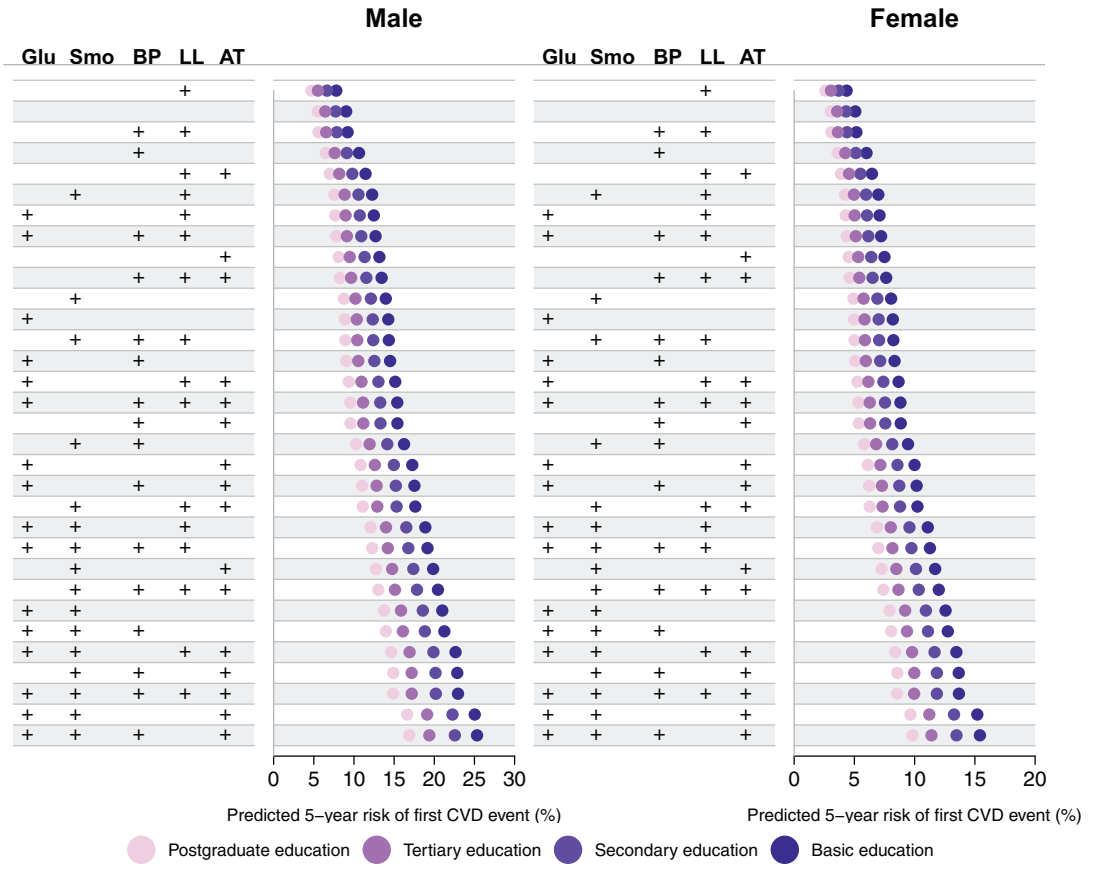


Figure 2 Predicted 5-year risk of first CVD event for all combinations of predictors for a 66-year-old individual. BP, blood pressure-lowering drugs; CVD, cardiovascular disease; Glu, glucose-lowering drugs; Smo, smoking proxy; AT, antithrombotic drugs; LL, lipid-lowering drugs; Edu, education; Sec, secondary; Bas, basic; Ter, tertiary.

Average risk in municipalities

We showed average predicted risks of first CVD event in each of the 98 Danish municipalities overall and for 66-year olds (*Figure 3*). The average predicted municipality risk ranged from 2.8% (Copenhagen municipality) to 5.9% (Laesoe municipality) (*Figure 3A*). The average predicted risk of first CVD event among 66-year olds ranged from 5.9% (Rudersdal municipality) to 7.5% (Vesthimmerland municipality) (*Figure 3B*).

Validation

The model was well-calibrated and had good discrimination in all five regions with AUCs ranging from 76.3% to 79.6% (*Figure 4*). Brier scores ranged from 3.3 to 4.4. Visual inspections of the calibration plots showed a small underestimation of risk from decile five in the Capital Region of Denmark and Region Zealand, whereas risk was slightly overestimated in all deciles in the remaining three regions (*Figure 4*). The model performed better than a benchmark model containing only age and sex in all five regions. Differences in AUC ranged from 1.2% to 1.5% and differences in Brier scores ranged from -0.02 to -0.03. Characteristics of the populations in each Region are shown in Supplementary material online, *Table S3*. Geographical and random split validation showed similarly good model performance (Supplementary material online, *Table S4* and Supplementary material online, *Figure S8*). Model validation in subgroups by age bands showed good calibration in most age bands. Calibration was suboptimal in the oldest age band (80–85 years) with overestimation of risk in the lower deciles and underestimation in the two highest deciles (Supplementary material online, *Figure S9*).

Discussion

We developed and validated a novel risk prediction model for estimation of the 5-year risk of first CVD event in 2.98 million Danish residents using administrative data from nationwide population-based registries. Predictors included in the model were age, sex, education, glucose-lowering drug use, blood pressure-lowering drug use, antithrombotic drug use, lipid-lowering drug use, and a smoking proxy. The model was well-calibrated in geographical regions and age bands. We provided examples of the utility of our model for prediction of personalized and population-level risk. We created a web

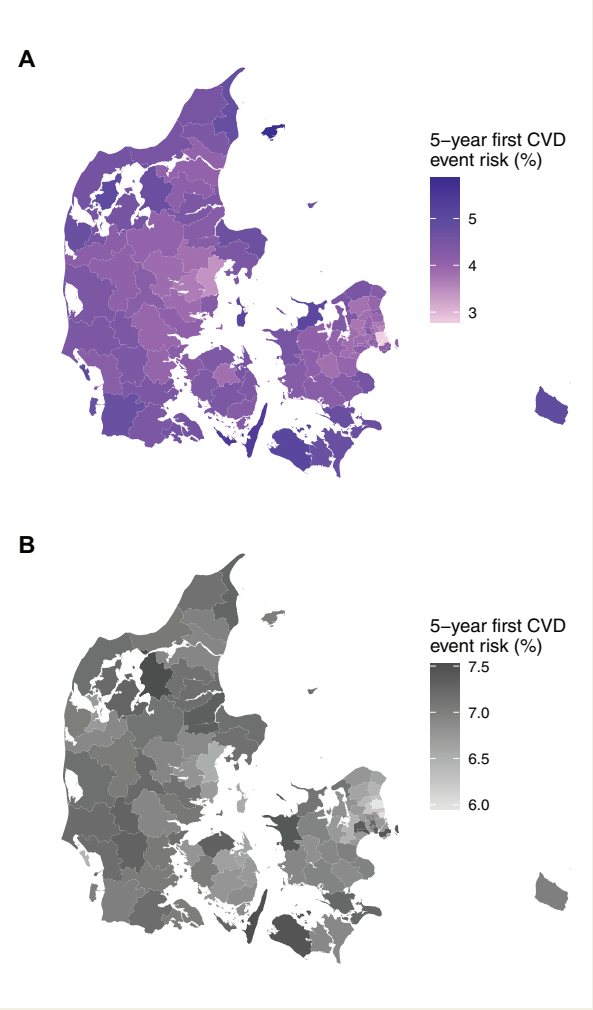


Figure 3 Average predicted 5-year risk of first cardiovascular disease event in the 98 Danish municipalities. (*A*) The average risk for all 30–85-year olds in each municipality. (*B*) The average risk for 66-year olds in each municipality.

calculator based on our risk prediction model intended for use in community settings that can be used easily by laypersons owing to the absence of laboratory and clinical variables.[8,9]

All Danish residents are assigned a Civil Registration Number at birth or immigration, by which demographic data and healthcare service usage data is recorded. Hence, the 2.98 million participants in this study encompassed virtually the entire Danish population aged 30–85. We had complete data on all predictors, except for education, and all outcome variables as well as minimal censoring. This provided an ideal data set for derivation and validation of a CVD risk prediction model for the Danish population. That was the main strength of our study, as it eliminated issues with generalizability and external validity, although generalizability to other countries may be limited. However, other countries or regions, with databases similar to those described in the present study, can develop CVD risk prediction models tailored specifically to their populations by utilizing our approach. Contemporary or updated risk prediction models are needed, because prediction models based on older cohorts are likely to overestimate risk in modern day populations, as treatment advances and changes in risk factors in recent decades have led to a lower incidence and mortality of CVD.[27–29]

A drawback of using administrative data to fit risk prediction models is the lack of laboratory variables and more specific clinical data for personalized risk prediction. The original forms of the most widely known risk prediction models in preventive cardiology, e.g. SCORE,[30] Framingham Risk Score,[31] and PCE,[32] typically incorporate lipid levels and blood pressure. A benefit of that approach is that these, as well as other modifiable CVD risk factors, are directly accounted for in the models. Thus, they present a clear target for risk factor modification, and changes in risk can be communicated directly by the clinician to the patient, e.g. by showing predicted risks at lower lipid levels or blood pressures. However, laboratory and clinical variables do not necessarily result in more accurate CVD risk prediction per se, as demonstrated in previous studies directly comparing models with and without laboratory measures,[7,31,33] since non-modifiable factors such as sex, age, and sociodemographic factors may capture up to 80% of the prognostic performance in cardiovascular risk models.[34] The inclusion of variables such as lipid levels would preclude the use of our model for population-level risk predictions, as such variables are not routinely collected at the population level. Furthermore, the disadvantages of only including universally available variables in our administrative data approach are outweighed by the absence of several weaknesses associated with conventional approaches to risk prediction modelling, namely, nonrepresentative study samples, few events in predictor combination strata, and many variables with missing data. In addition, our web calculator for personalized risk prediction is mainly intended for use in the community setting where laboratory and clinical risk factor levels may not be known. End-users of our online risk calculator are alerted of their cardiovascular risk and prompted to pursue individual targeting of their modifiable risk factors.

We followed the Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis and The REporting of studies Conducted using Observational Routinely collected health Data (RECORD) Statement recommendations for developing our novel risk prediction model.[35,36] As such, we prespecified predictors for inclusion in our model that were known risk factors for CVD and were commonly used in previous risk scores, albeit modified to suitability in our administrative databases. We chose a parsimonious number of predictors for our model to reduce complexity and avoid overfitting.

Sex, age, glucose-lowering drug use, blood pressure-lowering drug use, and a smoking proxy were included to serve as proxies for well-documented predictors of CVD.[37] Socioeconomic position is increasingly recognized as an important predictor of poor health outcomes, and recently developed CVD risk models such as QRISK3, PREDICT, and the models developed by Mehta *et al.* incorporated socioeconomic measures.[5,6,10] We chose education as a simple measure of socioeconomic position. As discussed, improvements in preventive treatment have reduced the risk of CVD in recent decades, and Danish patients are increasingly prescribed antithrombotic and lipid-lowering drugs for

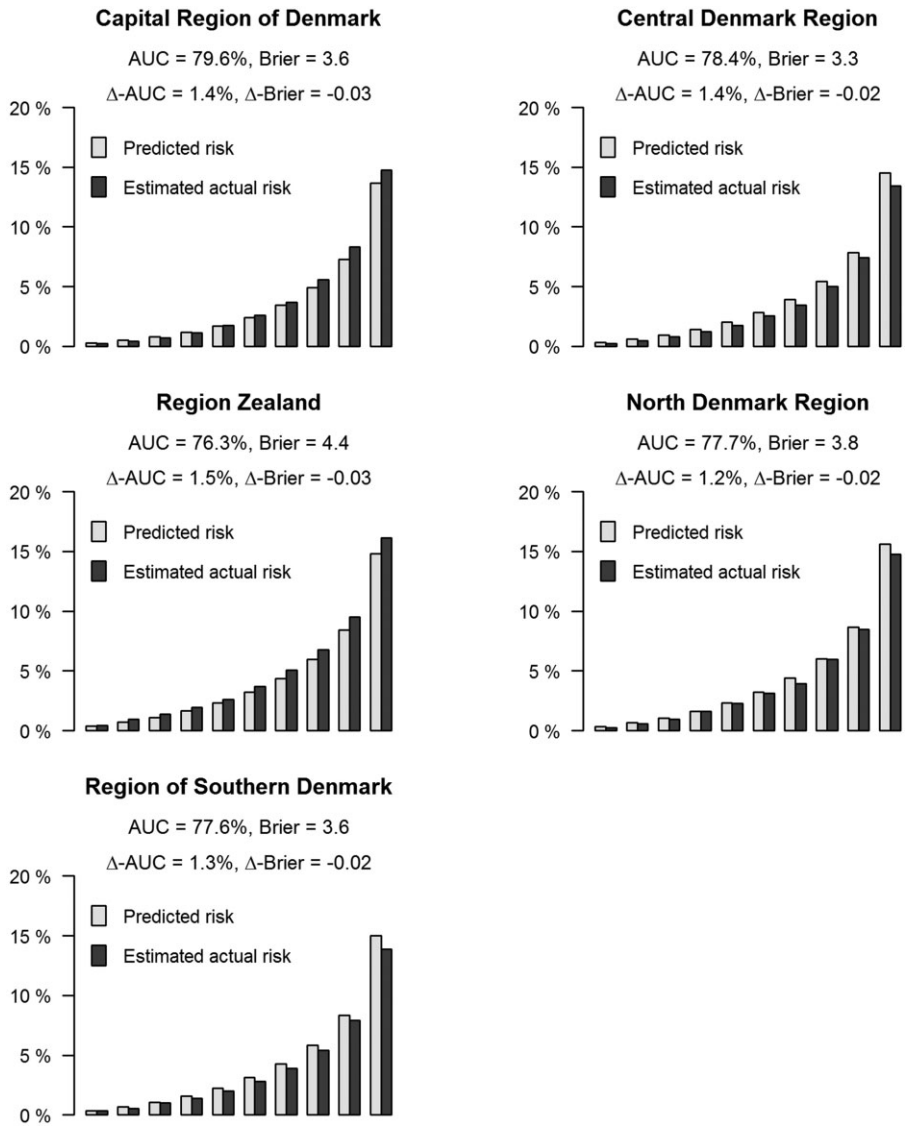


Figure 4 Calibration plots and discrimination metrics across the five regions. The predicted risk is plotted in deciles against the estimated actual risk. Differences in AUC and Brier scores compared to an age-sex benchmark model is presented as delta-AUC and delta-Brier. AUC, area under the receiver operating characteristic curve.

primary prevention of CVD.[38] To account for this, we included lipid-lowering and antithrombotic drug use at baseline in our model. Notably, in our data, we lacked information on body mass index and lifestyle factors such as physical activity-level, which are known and important modifiable risk factors for CVD. In spite of this limitation, our model showed good predictive performance, since sociodemographic factors (i.e. age and education) are surrogates for exposure to CVD risk factors, such as obesity, throughout the lifespan.[39,40] Our approach to variable selection can be replicated to identify candidate predictors in other databases with different data structure and population characteristics.

The main outcome differed from SCORE, which is the currently recommended CVD risk prediction tool in Denmark. We chose 5 years rather than 10 years as our prediction horizon, which enabled us to develop our prediction model in a contemporary (2014) cohort with 5 years follow-up and included non-fatal CVD events in our composite endpoint. Fatal CVD does not adequately capture CVD events, as mortality following CVD has decreased during the last decades, especially in high-income European countries.[1] We included HF and haemorrhagic stroke as CVD events, as they lead to high morbidity and largely share the same risk factors as atherosclerotic CVD.[41,42] The direction of the associations between predictors and the outcome were as expected.

Our smoking proxy consisting of either smoking-cessation drug use or COPD has not been validated. The sensitivity for identifying individuals who smoked was likely low, as the prevalence of our smoking proxy was only 7%, whereas the daily smoking prevalence in Denmark has been reported to be 17%.[43] However, we presume that the PPV of our smoking proxy was high as smoking-cessation drugs are not approved for any other indication and previous Danish population-based studies found that 78% of COPD patients were smokers.[44] The hazard ratios for our smoking proxy were similar to those reported in previous multivariable CVD risk prediction models.[5,6] Lipid-lowering drug use predicted a reduced risk of CVD in older age groups and an increased risk in younger age groups. This may be because lipid-lowering treatment was prescribed to younger patients with a high-risk indication, whereas the reduced risk in the older age groups reflected the expected treatment benefit. Antithrombotic drug use predicted an overall increased risk, as they were only indicated for individuals with an elevated CVD risk.[3] Similarly, blood pressure-lowering drug use predicted a higher risk of first CVD event overall, but the risk was attenuated in middle-aged and older age groups, which likely reflects a more severe indication for treatment in the younger age groups. Newer glucose-lowering drugs with benefits on cardiovascular outcomes were not widely used at the time of our cohort inclusion in 2014.[45] Thus, our model should be adjusted in the forthcoming years to reflect the change in risk distribution as these drugs become more common in the treatment of type 2 diabetes. The ability to easily adjust the model prospectively (e.g. as done in the UK with QRISK) highlights a strength of the present approach.[6]

We wanted to have a study population that was as inclusive as possible for population-level CVD risk prediction and, therefore, chose to include a broad age range (30–85 years). Nonetheless, previous studies found that CVD risk prediction models developed in mainly young and middle-aged persons predicted individual risk in older age groups poorly, partly due to the competing risk of non-cardiovascular death.[46,47] We handled the competing risk in our statistical modelling. Yet, our model had suboptimal calibration in the 80–85-year age band, which warrants a more cautious interpretation of the risk predictions presented for this age group.

The municipality-level risk predictions that we provided were an example of how our model can be applied beyond prediction of individual risk. The models developed by Mehta *et al.*[48] have also recently been used to identify quality improvement opportunities in the utilization of cardiovascular preventive pharmacotherapy across a country (New Zealand) and in sub-populations. Identifying high-risk areas and subpopulations, may guide public health-level interventions such as allocation of resources and targeted preventive efforts.

Conclusion

A CVD risk prediction model based solely on nationwide administrative registry data provided accurate prediction of 5-year first CVD event risk in the entire Danish population. We supplied examples of both personal and population-level use of the model. The model can be used to facilitate community-based, clinical, and public health-level primary prevention. An online risk calculator based on our risk prediction model is freely available (https://hjerteforeningen.shinyapps.io/cvd-risk-manuscript/).

Lead author biography

Daniel Mølager Christensen is a medical doctor with interests in cardiology, pidemiology, and medical risk prediction. He is currently a PhD-fellow at the Danish Heart Foundation conducting research using the Danish national healthcare registries.

Supplementary material

Supplementary material is available at *European Heart Journal Open* online.

Funding

The present study was fully supported by a research grant from the Danish Heart Foundation (grant identification number 20-R146-A9798).

Conflicts of interest: E.F. reported an independent research grant unrelated to the current research from The Novo Nordisk Foundation. LK reported speaker's honorarium from Novo, Novartis, Astra Zeneca, and Boehringer. C.T.P. reported research grants from Bayer and Novo Nordisk not related to this study. All other authors had no conflicts of interest to disclose.

Data availability statement

No additional data are available as access to Danish registry data is granted on an individual basis by the relevant authorities.

References

1. Timmis A, Townsend N, Gale CP, Torbica A, Lettino M, Petersen SE, Mossialos EA, Maggioni AP, Kazakiewicz D, May HT, De Smedt D, Flather M, Zuhlke L, Beltrame JF, Huculeci R, Tavazzi L, Hindricks G, Bax J, Casadei B, Achenbach S, Wright L, Vardas P; European Society of Cardiology. European Society of Cardiology: cardiovascular disease statistics 2019. *Eur Heart J* 2020;**41**:12–85.
2. Virani SS, Alonso A, Benjamin EJ, Bittencourt MS, Callaway CW, Carson AP, Chamberlain AM, Chang AR, Cheng S, Delling FN, Djousse L, Elkind MSV, Ferguson JF, Fornage M, Khan SS, Kissela BM, Knutson KL, Kwan TW, Lackland DT, Lewis TT, Lichtman JH, Longenecker CT, Loop MS, Lutsey PL, Martin SS, Matsushita K, Moran AE, Mussolino ME, Perak AM, Rosamond WD, Roth GA, Sampson UKA, Satou GM, Schroeder EB, Shah SH, Shay CM, Spartano NL, Stokes A, Tirschwell DL, VanWagner LB, Tsao CW; American Heart Association Council on Epidemiology and Prevention Statistics Committee and Stroke Statistics Subcommittee. Heart disease and stroke statistics-2020 update: a report from the American Heart Association. *Circulation* 2020;**141**:e139–e596.
3. Piepoli MF, Hoes AW, Agewall S, Albus C, Brotons C, Catapano AL, Cooney MT, Corra U, Cosyns B, Deaton C, Graham I, Hall MS, Hobbs FDR, Lochen ML, Lollgen H, Marques-Vidal P, Perk J, Prescott E, Redon J, Richter DJ, Sattar N, Smulders Y, Tiberi M, van der Worp HB, van Dis I, Verschuren WMM, Binno S; Group ESCSD. 2016 European Guidelines on cardiovascular disease prevention in clinical practice: The Sixth Joint Task Force of the European Society of Cardiology and Other Societies on Cardiovascular Disease Prevention in Clinical Practice (constituted by representatives of 10 societies and by invited experts). Developed with the special contribution of the European Association for

Cardiovascular Prevention & Rehabilitation (EACPR). *Eur Heart J* 2016;**37**:2315–2381.
4. Arnett DK, Blumenthal RS, Albert MA, Buroker AB, Goldberger ZD, Hahn EJ, Himmelfarb CD, Khera A, Lloyd-Jones D, McEvoy JW, Michos ED, Miedema MD, Munoz D, Smith SC, Jr., Virani SS, Williams KA, Sr., Yeboah J, Ziaeian B. 2019 ACC/AHA Guideline on the primary prevention of cardiovascular disease: a report of the American College of Cardiology/American Heart Association Task Force on Clinical Practice Guidelines. *J Am Coll Cardiol* 2019;**74**:e177–e232.
5. Pylypchuk R, Wells S, Kerr A, Poppe K, Riddell T, Harwood M, Exeter D, Mehta S, Grey C, Wu BP, Metcalf P, Warren J, Harrison J, Marshall R, Jackson R. Cardiovascular disease risk prediction equations in 400000 primary care patients in New Zealand: a derivation and validation study. *Lancet* 2018;**391**:1897–1907.
6. Hippisley-Cox J, Coupland C, Brindle P. Development and validation of QRISK3 risk prediction algorithms to estimate future risk of cardiovascular disease: prospective cohort study. *BMJ* 2017;**357**:j2099.
7. Gaziano TA, Young CR, Fitzmaurice G, Atwood S, Gaziano JM. Laboratory-based versus non-laboratory-based method for assessment of cardiovascular disease risk: the NHANES I Follow-up Study cohort. *Lancet* 2008;**371**:923–931.
8. Bonner C, Fajardo MA, Hui S, Stubbs R, Trevena L. Clinical validity, understandability, and actionability of online cardiovascular disease risk calculators: systematic review. *J Med Internet Res* 2018;**20**:e29.
9. Bosomworth NJ. Practical use of the Framingham risk score in primary prevention: Canadian perspective. *Can Fam Physician* 2011;**57**:417–423.
10. Mehta S, Jackson R, Pylypchuk R, Poppe K, Wells S, Kerr AJ. Development and validation of alternative cardiovascular risk prediction equations for population health planning: a routine health data linkage study of 1.7 million New Zealanders. *Int J Epidemiol* 2018;**47**:1571–1584.
11. Pedersen CB. The Danish Civil Registration System. *Scand J Public Health* 2011;**39**:22–25.
12. Schmidt M, Schmidt SA, Sandegaard JL, Ehrenstein V, Pedersen L, Sorensen HT. The Danish National Patient Registry: a review of content, data quality, and research potential. *Clin Epidemiol* 2015;**7**:449–490.
13. Pottegard A, Schmidt SAJ, Wallach-Kildemoes H, Sorensen HT, Hallas J, Schmidt M. Data resource profile: the Danish National Prescription Registry. *Int J Epidemiol* 2017;**46**:798–798f.
14. Jensen VM, Rasmussen AW. Danish Education Registers. *Scand J Public Health* 2011;**39**:91–94.
15. Helweg-Larsen K. The Danish Register of Causes of Death. *Scand J Public Health* 2011;**39**:26–29.
16. Schmidt M, Andersen LV, Friis S, Juel K, Gislason G. Data resource profile: Danish Heart Statistics. *Int J Epidemiol* 2017;**46**:1368–1369g.
17. Krarup LH, Boysen G, Janjua H, Prescott E, Truelsen T. Validity of stroke diagnoses in a National Register of Patients. *Neuroepidemiology* 2007;**28**:150–154.
18. Carstensen B, Kristensen JK, Marcussen MM, Borch-Johnsen K. The National Diabetes Register. *Scand J Public Health* 2011;**39**:58–61.
19. Malmborg M, Schmiegelow MDS, Norgaard CH, Munch A, Gerds T, Schou M, Kistorp C, Torp-Pedersen C, Hlatky MA, Gislason G. Does type 2 diabetes confer higher relative rates of cardiovascular events in women compared with men? *Eur Heart J* 2020;**41**:1346–1353.
20. Benichou J, Gail MH. Estimates of absolute cause-specific risk in cohort studies. *Biometrics* 1990;**46**:813–826.
21. Steyerberg EW, Harrell FE. Jr. Prediction models need appropriate internal, internal-external, and external validation. *J Clin Epidemiol* 2016;**69**:245–247.
22. Steyerberg EW, Vergouwe Y. Towards better clinical prediction models: seven steps for development and an ABCD for validation. *Eur Heart J* 2014;**35**:1925–1931.
23. Gerds TA, Cai T, Schumacher M. The performance of risk prediction models. *Biom J* 2008;**50**:457–479.
24. Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N, Pencina MJ, Kattan MW. Assessing the performance of prediction models: a framework for traditional and novel measures. *Epidemiology* 2010;**21**:128–138.
25. Gerds TA, Schumacher M. Consistent estimation of the expected Brier score in general survival models with right-censored event times. *Biom J* 2006;**48**:1029–1040.
26. Gerds TA, Andersen PK, Kattan MW. Calibration plots for risk prediction models in the presence of competing risks. *Stat Med* 2014;**33**:3191–3203.
27. Schmidt M, Jacobsen JB, Lash TL, Botker HE, Sorensen HT. 25 year trends in first time hospitalisation for acute myocardial infarction, subsequent short and long term mortality, and the prognostic impact of sex and comorbidity: a Danish nationwide cohort study. *BMJ* 2012;**344**:e356.
28. Grey C, Jackson R, Schmidt M, Ezzati M, Asaria P, Exeter DJ, Kerr AJ. One in four major ischaemic heart disease events are fatal and 60% are pre-hospital deaths: a national data-linkage study (ANZACS-QI 8). *Eur Heart J* 2017;**38**:172–180.
29. Cook NR, Ridker PM. Calibration of the pooled cohort equations for atherosclerotic cardiovascular disease: an update. *Ann Intern Med* 2016;**165**:786–794.
30. Conroy RM, Pyörälä K, Fitzgerald AP, Sans S, Menotti A, De Backer G, De Bacquer D, Ducimetière P, Jousilahti P, Keil U, Njølstad I, Oganov RG, Thomsen T, Tunstall-Pedoe H, Tverdal A, Wedel H, Whincup P, Wilhelmsen L, Graham IM; SCORE project group. Estimation of ten-year risk of fatal cardiovascular disease in Europe: the SCORE project. *Eur Heart J* 2003;**24**:987–1003.
31. D'Agostino RB, Vasan RS, Pencina MJ, Wolf PA, Cobain M, Massaro JM, Kannel WB. General cardiovascular risk profile for use in primary care: the Framingham Heart Study. *Circulation* 2008;**117**:743–753.
32. Goff DC, Lloyd-Jones DM, Bennett G, Coady S, D'Agostino RB, Gibbons R, Greenland P, Lackland DT, Levy D, O'Donnell CJ, Robinson JG, Schwartz JS, Shero ST, Smith SC, Sorlie P, Stone NJ, Wilson PWF. 2013 ACC/AHA Guideline on the assessment of cardiovascular risk. *J Am Coll Cardiol* 2014;**63**:2935–2959.
33. McGorrian C, Yusuf S, Islam S, Jung H, Rangarajan S, Avezum A, Prabhakaran D, Almahmeed W, Rumboldt Z, Budaj A, Dans AL, Gerstein HC, Teo K, Anand SS; INTERHEART Investigators. Estimating modifiable coronary heart disease risk in multiple regions of the world: the INTERHEART Modifiable Risk Score. *Eur Heart J* 2011;**32**:581–589.
34. Pencina MJ, Navar AM, Wojdyla D, Sanchez RJ, Khan I, Elassal J, D'Agostino RB, Peterson ED, Sniderman AD. Quantifying importance of major risk factors for coronary heart disease. *Circulation* 2019;**139**:1603–1611.
35. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *BMJ* 2015;**350**:g7594.
36. Benchimol EI, Smeeth L, Guttmann A, Harron K, Moher D, Petersen I, Sørensen HT, von Elm E, Langan SM; RECORD Working Committee. The REporting of studies Conducted using Observational Routinely-collected health Data (RECORD) statement. *PLoS Med* 2015;**12**:e1001885.
37. Yusuf S, Hawken S, Ôunpuu S, Dans T, Avezum A, Lanas F, McQueen M, Budaj A, Pais P, Varigos J, Lisheng L. Effect of potentially modifiable risk factors associated with myocardial infarction in 52 countries (the INTERHEART study): case-control study. *Lancet* 2004;**364**:937–952.
38. Jorgensen ME, Andersson C, Olsen AM, Juel K, Mortensen PE, Jorgensen E, Tilsted HH, von Kappelgaard LM, Torp-Pedersen C, Gislason GH. Danish trends in pharmacotherapy, comorbidities, and demographics in patients referred for coronary angiography: what changed during a decade? *Eur Heart J Cardiovasc Pharmacother* 2015;**1**:157–165.
39. Wald NJ, Simmonds M, Morris JK. Screening for future cardiovascular disease using age alone compared with multiple risk factors and age. *PLoS One* 2011;**6**:e18742.
40. Clark AM, DesMeules M, Luo W, Duncan AS, Wielgosz A. Socioeconomic status and cardiovascular disease: risks and implications for care. *Nat Rev Cardiol* 2009;**6**:712–722.
41. Zhang Y, Tuomilehto J, Jousilahti P, Wang Y, Antikainen R, Hu G. Lifestyle factors on the risks of ischemic and hemorrhagic stroke. *Arch Intern Med* 2011;**171**:1811–1818.
42. Djoussé L, Driver JA, Gaziano JM. Relation between modifiable lifestyle factors and lifetime risk of heart failure. *JAMA* 2009;**302**:394–400.
43. Danish smoking habits year report 2018—the Danish Health Authority (in Danish). https://www.sst.dk/en/English/publications/2021/Danish-smoking-habits-2020 (1 September, 2021)
44. Thomsen M, Nordestgaard BG, Vestbo J, Lange P. Characteristics and outcomes of chronic obstructive pulmonary disease in never smokers in Denmark: a prospective population study. *Lancet Respir Med* 2013;**1**:543–550.
45. Palmer SC, Tendal B, Mustafa RA, Vandvik PO, Li S, Hao Q, Tunnicliffe D, Ruospo M, Natale P, Saglimbene V, Nicolucci A, Johnson DW, Tonelli M, Rossi MC, Badve SV, Cho Y, Nadeau-Fredette A-C, Burke M, Faruque LI, Lloyd A, Ahmad N, Liu Y, Tiv S, Millard T, Gagliardi L, Kolanu N, Barmanray RD, McMorrow R, Raygoza Cortez AK, White H, Chen X, Zhou X, Liu J, Rodríguez AF, González-Colmenero AD, Wang Y, Li L, Sutanto S, Solis RC, Díaz González-Colmenero F, Rodriguez-Gutierrez R, Walsh M, Guyatt G, Strippoli GFM. Sodium-glucose cotransporter protein-2 (SGLT-2) inhibitors and glucagon-like peptide-1 (GLP-1) receptor agonists for type 2 diabetes: systematic review and network meta-analysis of randomised controlled trials. *BMJ* 2021;**372**:m4573.
46. Mehta S, Jackson R, Poppe K, Kerr AJ, Pylypchuk R, Wells S. How do cardiovascular risk prediction equations developed among 30-74 year olds perform in older age groups? A validation study in 125 000 people aged 75-89 years. *J Epidemiol Community Health* 2020;**74**:527–533.
47. Wells S, Pylypchuk R, Mehta S, Kerr A, Selak V, Poppe K, Grey C, Jackson R. Performance of CVD risk equations for older patients assessed in general practice: a cohort study. *New Z Med J* 2020;**133**:32–55.)
48. Mehta S, Zhao J, Poppe K, Kerr AJ, Wells S, Exeter DJ, Selak V, Grey C, Jackson R. Cardiovascular preventive pharmacotherapy stratified by predicted cardiovascular risk: a national data linkage study. *Eur J Prevent Cardiol* 2021;