Cardiovascular risk prediction in men and women aged under 50 years using routine care data

Os, H.J.A. van; Kanning, J.P.; Bonten, T.N.; Rakers, M.M.; Putter, H.; Numans, M.E.; ... ; Wermer, M.J.H.

Journal of the American Heart Association

ORIGINAL RESEARCH

Cardiovascular Risk Prediction in Men and Women Aged Under 50 Years Using Routine Care Data

Hendrikus J. A. van Os, MD; Jos P. Kanning, MSc; Tobias N. Bonten, MD, PhD; Margot M. Rakers, MD; Hein Putter, PhD; Mattijs E. Numans, MD, PhD; Ynte M. Ruigrok, MD, PhD; Rolf H. H. Groenwold, MD, PhD; Marieke J. H. Wermer, MD, PhD

BACKGROUND: Prediction models for risk of cardiovascular events generally do not include young adults, and cardiovascular risk factors differ between women and men. Therefore, this study aimed to develop prediction models for first-ever cardiovascular event risk in men and women aged 30 to 49 years.

METHODS AND RESULTS: We included patients aged 30 to 49 years without cardiovascular disease from a Dutch routine care database. Outcome was defined as first-ever cardiovascular event. Our reference models were sex-specific Cox proportional hazards models based on traditional cardiovascular predictors, which we compared with models using 2 predictor subsets with the 20 or 50 most important predictors based on the Cox elastic net model regularization coefficients. We assessed the C-index and calibration curve slopes at 10 years of follow-up. We stratified our analyses based on 30- to 39-year and 40- to 49-year age groups at baseline. We included 542 141 patients (mean age 39.7, 51% women). During follow-up, 10 767 cardiovascular events occurred. Discrimination of reference models including traditional cardiovascular predictors was moderate (women: C-index, 0.648 [95% CI, 0.645–0.652]; men: C-index, 0.661 [95%CI, 0.658–0.664]). In women and men, the Cox proportional hazard models including 50 most important predictors resulted in an increase in C-index (0.030 and 0.012, respectively), and a net correct reclassification of 3.7% of the events in women and 1.2% in men compared with the reference model.

CONCLUSIONS: Sex-specific electronic health record-derived prediction models for first-ever cardiovascular events in the general population aged <50 years have moderate discriminatory performance. Data-driven predictor selection leads to identification of nontraditional cardiovascular predictors, which modestly increase performance of models.

Key Words: cardiovascular risk ■ prediction ■ sex differences ■ young adults

Cardiovascular events are a leading cause of disability and death worldwide.[1] In the last half century cardiovascular event-related mortality decreased continually. However, opportunities in primary prevention of cardiovascular events are still being missed.[2] Currently in Europe, decisions on preventive interventions in adults without prior cardiovascular disease (CVD) aged 40 to 69 years are based on the absolute 10-year risk of cardiovascular events, resulting from the Systematic COronary Risk Evaluation 2 (SCORE2) prediction model.[3] Early identification of individuals at high risk of cardiovascular events is beneficial, because atherosclerosis is a chronic process that starts early in life.[4] Therefore, early treatment of risk factors is beneficial, and accurate risk estimates applicable to younger people are required.[5]

Evidence on sex differences between cardiovascular risk factors is mounting, which pleads for including sex-specific risk factors such as preeclampsia and combined oral contraceptive pill use in prediction models.[6] Derivation of sex-specific models for the

Correspondence to: Hendrikus J. A. van Os, MD, Leiden University Medical Center, Department of Neurology, Albinusdreef 2, 2333 ZA Leiden, The Netherlands. Email: h.j.a.van_os@lumc.nl

Supplemental Material is available at https://www.ahajournals.org/doi/suppl/10.1161/JAHA.122.027011

For Sources of Funding and Disclosures, see page 9.

JAHA is available at: www.ahajournals.org/journal/jaha

CLINICAL PERSPECTIVE

What Is New?

- Sex-specific electronic health record-derived prediction models for first-ever cardiovascular events in the general population aged <50 years have moderate discriminatory performance and are well-calibrated.
- Data-driven predictor selection leads to identification of nontraditional cardiovascular predictors, which modestly increase discriminatory performance of models and correct reclassification of events, mostly in women.

What Are the Clinical Implications?

- Sex-specific electronic health record-derived prediction models could be used to identify subgroups of patients <50 years that are at increased risk of first-ever cardiovascular events. These patients could then be invited to the primary care practice center for further cardiovascular risk assessment including measurement of, for example, systolic blood pressure and total and high-density lipoprotein cholesterol.
- For patients aged 30 to 39 years, our results call for further research into defining meaningful thresholds of 10-year risk of first-ever cardiovascular events, as they are not yet specified in current guidelines.

Nonstandard Abbreviations and Acronyms

ATC	Anatomical Therapeutic Chemical (classification System)
ICPC	International Classification of Primary Care

prediction of cardiovascular risk in young individuals requires a large sample size. Pooling electronic health record (EHR) data results in large prospective cohorts, offering a great opportunity for the derivation of prediction models.[7] The QRISK3 prediction model for the risk of cardiovascular events is an example of leveraging information from the EHR, and has been successfully externally validated in the general population in the United Kingdom.[8] QRISK3 is a traditional regression model using predictors which are selected based on prior knowledge. However, because EHR-derived cohorts are constituted by both a large sample size and a high number of potentially relevant predictors, complex data-driven modeling techniques may outperform traditional regression models in predicting the risk of cardiovascular event.[9–11]

This study aimed to develop sex-specific prediction models for first-ever cardiovascular event risk in patients aged 30 to 49 years in a primary care setting, using data from a large Dutch EHR-derived population-based cohort. We assessed whether the data-driven selection of predictors and the use of complex prediction models offer an increase in predictive performance, compared with a Cox regression model using only traditional cardiovascular predictors.

METHODS

Data Source

The research cohort in this study was derived from the STIZON (Stichting Informatievoorziening voor Zorg en Onderzoek) database. STIZON directly receives data from EHRs of a large number of primary care providers throughout the Netherlands.[12] We only selected patients from general practice centers which were localized in catchment areas of hospitals participating in the STIZON network. This enabled us to link hospital *International Classification of Diseases, Ninth and Tenth Revisions* (*ICD-9*) and (*ICD-10*) diagnoses to primary care data. The STIZON data set contains Anatomical Therapeutic Chemical classification system (ATC) medication prescriptions from primary care pharmacies during follow-up time, and International Classification of Primary Care (ICPC) diagnosis codes for clinical entities in principle starting from birth.[13,14] *ICD-9* and *ICD-10* codes were available for all in-hospital diagnoses that occurred during follow-up. Inclusion criteria were an age of 30 to 49 at baseline, and subscription to a STIZON general practice center between January 1, 2007 and December 31, 2020 for at least 1 year, which was required because we defined the 1-year as a run-in period. This run-in period was used for averaging the predictor values of laboratory or vital parameter assessments, if multiple of such measurements were present within this period. Exclusion criteria were CVD, and use of statins or cardiovascular event-specific thrombocyte aggregation inhibitors at baseline. Follow-up time started at the end of the 1 year run-in period (January 1, 2008) or on the first general practice center subscription date after January 1, 2008. Patients were censored at the earliest date of the diagnosis of a first-ever fatal or nonfatal cardiovascular event, noncardiovascular death, deregistration with any practice connected to the STIZON network, or the last upload of computerized data to the STIZON database (December 31, 2020). The ethics review board has provided a statement that this study was not subject to ethics review according to the Medical Research Involving Human Subjects Act wet medisch onderzoek. Because of the sensitive nature of the data collected for this study, data will need to be requested from a third party (STIZON).

Outcome Definition

First-ever cardiovascular events were defined using *ICD-9*, *ICD-10*, or ICPC codes for fatal and nonfatal acute myocardial infarction and stroke (including ischemic, hemorrhagic, and unspecified stroke; Table S1).

Predictors

All predictors which were used for analyses can be found in Table S1. Predictors included demographics, symptoms, and diagnoses other than fatal and nonfatal cardiovascular events, and were based on ICPC, *ICD-9*, and *ICD-10* codes, prescribed medication coded according to the ATC classification, laboratory test results performed in primary care, consultation dates, and frequency.[13,14] In addition, the 4-digit postal code area data were transformed into a socioeconomic status score based on income, education, and occupation of the inhabitants.[15] ICPC, *ICD-9*, and *ICD-10* codes and condition-specific ATC-codes were clustered based on clinical knowledge by 2 domain experts (H.vO. and M.R.) if multiple codes constituted the same clinical entity. An example is the grouping of different types of malignancy diagnoses into an overall malignancy predictor. For computational purposes, we only selected predictors that occurred in at least 0.1% of the total study population across the entire follow-up time, after clustering. All continuous predictors were standardized before analysis. Baseline information was assessed at the end of the 1-year run-in period.

Missing Value Handling

With respect to missing predictor values, we made a distinction between binary predictors—such as registration of a certain diagnosis or prescription of medication—and continuous predictors such as measurements of laboratory parameters or blood pressure. For all binary predictors, we assumed that the absence of an EHR registration meant the absence of the clinical entity itself, and therefore no imputation was performed. However, for continuous predictors such as vital parameters or laboratory assessments, imputation of missing values was required for inclusion in the prediction models. Because in routine health care data the majority of such assessments are only performed in a small subset of the population, the extent of missingness may be large and the underlying mechanism of missingness is likely missing not at random. Because in our data set for all continuous laboratory or vital parameter assessments missingness exceeded 25%, we chose not to impute the missing values to limit the risk of biased predictor value imputations. We used only binary indicators in the analyses, which indicated whether the assessment had been performed or not.

Predictor Selection

We used 2 methods for the selection of predictors which were used to develop prediction models. First, for the reference models we chose the traditional cardiovascular risk factors age, sex, smoking (ever), and either an *ICD-9*, *ICD-10*, or ICPC diagnosis code or condition-specific ATC medication prescription code for hyperlipidemia, hypertension, and diabetes, based on prior evidence.[16] Since we excluded patients who received statin treatment at baseline, hyperlipidemia was based on diagnosis codes only. Second, we used data-driven predictor selection based on a Cox elastic net model (α of 0.00058 for women, α of 0.00072 for men; L1 to L2 regularization penalty ratio: 0.5) to select the most important 20 and 50 predictors based on the absolute regularized coefficients of a sex-specific Cox elastic net model.

Model Development

The 3 different selections of predictors (traditional cardiovascular risk factors for the reference model, and the 20 and 50 most important predictors based on a Cox elastic net model) were used to develop Cox proportional hazard (PH) models, Cox elastic net models, and random survival forests. Models were developed for women and men separately. Cox elastic net models and random survival forests are more flexible than Cox PH models, because they include hyperparameters. Hyperparameters of Cox elastic net and random survival forests were optimized using predefined hyperparameter grids (Table S2). To account for overfitting and internally validate our findings, we used a nested validation approach. First, the data were randomly split into a derivation and validation set of, respectively, 80% and 20% of the population. Hyperparameter optimization was then performed on the derivation set, using 10-fold cross validation. Overall model performance was assessed using the hold-out validation set. We repeated this process 50 times using bootstrap resampling to assess variability in outcomes and to report empirical 95% CIs. We did consider noncardiovascular death as a competing event, since our population was young and noncardiovascular mortality was expected to be low. Model performance was defined by both model discrimination (concordance index or C-index) and calibration (calibration curve slope at 10 years of follow-up). We expressed change in C-index between reference and other prediction models as difference relative to the full scale of the C-index, which is from 0.5 to 1. Further, we assessed net reclassification using the categorical net reclassification index. We chose a 2.5% 10-year absolute risk of first-ever cardiovascular events as threshold for high cardiovascular risk. This is in line with the European Society of Cardiology guidelines for prevention of CVD in individuals aged

<50 years and implies that risk factor treatment should be considered. Our predefined absolute risk threshold of 2.5% is therefore of clinical importance.[17] In addition, we stratified our analyses based on 2 age groups (30–39 and 40–49 years at baseline). The 30- to 39-year age group is of particular interest, because the SCORE2 model starts at an age of 40 years. For all performance metrics we calculated empirical 95% CI by fitting a new model in each of the 50 bootstrap samples and basing the CI on the SD of the distribution of the performance metrics. Python version 3.10 was used for preprocessing and analysis of data. Our study adhered to the Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis statement for reporting.[18]

RESULTS

We included 542 141 patients aged 30 to 49 years without prior CVD or statin use at baseline in this study, of whom 51% were women. During 5 461 316 person-years of follow-up, a total of 10 767 first-ever cardiovascular events occurred. This resulted in an incidence rate of 19.7 (95% CI, 19.3–20.1) per 10 000 person-years in the total population, 13.6 (95% CI, 13.2–14.0) in women and 26.2 (95% CI, 25.5–26.8) in men. Table 1 shows the baseline characteristics of men and women in the total study population. The average age was 39.7 years (SD±5.7). Systolic blood pressure was assessed in 6.6%, and total serum cholesterol in 2.4% of the total population. We, therefore, discarded continuous measurements and only included indicators of whether tests were performed.

Subsequently, after the data-driven selection of predictors using Cox elastic net models, the 20 most important predictors are shown in Table 2. The 50 most important predictors can be found in Table S3. Substantial differences in predictor importance were observed between women and men. For example, for women, 2 female-specific risk factors (combined oral contraceptive use and intrauterine contraceptive use) are ranked in the top 20. The top 20 most important predictors for women and men, stratified based on the 30- to 39-year and 40- to 49-year age groups, are shown in Table S4.

Discrimination of Cox PH reference models including traditional cardiovascular predictors for both women and men was moderate (women: C-index, 0.648 [95% CI, 0.645–0.652]; men: C-index, 0.661 [95% CI, 0.658–0.664]), and calibration was good (calibration curve slope in women: 0.999 [95% CI, 0.998–1.001]; and in men: 1.001 [95% CI, 0.998–1.004]; Table 3). In women, the Cox PH model, including 50 most important predictors, resulted in an increase in C-index of 0.030 compared with the reference model (20% difference with the reference model relative to the full scale of the C-index). In men, Cox PH model, including 50 most important predictors, also resulted in the relatively largest increase in C-index, although to a lesser extent compared with women (0.012 increase in C-index; 7% difference with the reference model relative to the full scale of the C-index). The more flexible modeling approaches (Cox elastic net and random survival forests) did not perform better than the Cox PH models across any of the different predictor subsets (Table S5).

For women and men, the categorical net reclassification index was assessed for the Cox PH model with

Table 1. Baseline Characteristics for Women and Men

Baseline characteristics	Women (n=276 113)		Men (n=266 028)	
	Cases (n=3800)	Controls (n=272 313)	Cases (n=6915)	Controls (n=259 113)
Demographic features				
Age, y, mean(SD)	42.4 (5.0)	39.5 (5.7)	42.9 (4.8)	39.6 (5.6)
Socioeconomic status score, mean (SD)	0.23 (0.75)	0.31 (0.71)	0.25 (0.74)	0.30 (0.72)
Follow-up time, y, median (IQR)	6.6 (3.8–9.4)	11.0 (8.3–13.0)	6.9 (4.0–9.6)	11.0 (8.0–13.0)
Cardiovascular risk factors, n (%)				
Smoking, current	154 (4.1)	4897 (1.8)	264 (3.8)	5087 (2.0)
Hyperlipidemia	32 (0.8)	761 (0.3)	69 (1.0)	1261 (0.5)
Hypertension	157 (4.1)	3896 (1.4)	168 (2.4)	3339 (1.3)
Diabetes	43 (1.1)	1163 (0.4)	67 (1.0)	1295 (0.5)
Measurements, n (%)*				
Systolic blood pressure	485 (12.8)	20 823 (7.6)	526 (7.6)	13 907 (5.4)
Serum glucose	133 (3.5)	8245 (3.0)	171 (2.5)	4463 (1.7)
Total serum cholesterol	318 (8.4)	13 585 (5.0)	468 (6.8)	12 150 (4.7)

Cases=patients who experienced a first-ever cardiovascular event during follow-up; controls=all other patients. IQR indicates interquartile range.
*Any laboratory or vital parameter measurement during the 1-year run-in period.

Table 2. Top 20 Most Important Predictors for Women and Men Separately

Predictor	Coef.*
Women (n=276 113)	
Age, y	0.416
Socioeconomic status score	0.115
Combined oral contraceptive use	0.070
Antirheumatic medication	0.060
Gastroesophageal reflux medication	0.053
Smoking: current	0.052
Acetylsalicylic acid use	0.052
Comorbidity count	0.049
RAAS inhibitors	0.045
Beta-blockers	0.043
Calcium channel blockers	0.040
Blood pressure measured last year	0.032
Dermatological complaints	0.031
Intrauterine contraceptive use	0.030
Hyperlipidemia	0.029
Antibiotic use	0.028
Depression	0.027
HIV/AIDS	0.024
Female sex organ complaints and symptoms	0.023
Diabetes	0.023
Men (n=266 028)	
Age, y	0.533
Socioeconomic status score	0.101
Smoking: current	0.069
Antirheumatic medication	0.067
Diabetes	0.039
Practice nurse contact for somatic complaints	0.035
RAAS inhibitors	0.033
Psoriasis	0.031
Gastroesophageal reflux medication	0.027
Comorbidity count	0.026
Hyperlipidemia	0.019
Epilepsia	0.019
Calcium channel blockers	0.018
Oral anticoagulant drugs	0.016
Esophageal disorders	0.014
Allergic rhinitis	0.014
Antibiotic use	0.014
Alcohol use	0.014
Kidney failure	0.014
Male sex organ complaints	0.014

*Absolute, regularized coefficient of Cox elastic net models (women: alpha=0.00058; men: alpha=0.00062).

†Comorbidity count: simple count of chronic conditions per patient, listed in Table S2. RAAS indicates renin-angiotensin-aldosterone system.

50 most important predictors versus the reference Cox PH model. For women, net correct reclassification was 3.7% for events (95% CI, 3.2%–4.2%), and 0.0% for nonevents (95% CI, −0.1% – 0.1%); and for men, net correct reclassification for events was 1.2% (95% CI, 0.8% – 1.6%), and −0.8% (95% CI, −1.1% to −0.4%) for nonevents. Absolute risks for the Cox PH model with 50 most important predictors are shown for women and men (Figure).

After stratification of the 30- to 39-year and 40- to 49-year age groups at baseline, discriminatory performance was attenuated in the 30- to 39-year age group, and further decreased in the 40- to 49-year age group, for all Cox PH models in both women and men (Table 3).

DISCUSSION

We found that in an EHR-derived population-based cohort of primary care patients aged between 30 to 49 years, sex-specific prediction models for first-ever cardiovascular events had moderate discriminatory performance and were well calibrated. Compared with the reference Cox PH models, the Cox PH models based on the 50 most important predictors had better discriminatory performance in both women and men and were well calibrated. In women the improvement in discrimination was more substantial as compared with men, and the net correct reclassification of events was 3.7%. The more complex modeling methods Cox elastic net and random survival forests did not result in improvements in discrimination or calibration compared with the reference model, regardless of the predictor subset that was chosen. After stratification of the age groups at baseline, we found that discriminatory performance was attenuated in the 30- to 39-year age group, and further decreased in the 40- to 49-year age group. This was as expected, because we restricted the range of age, which is the most important predictor for cardiovascular events.

Several previous studies reported on the prediction of cardiovascular events using large EHR-derived data sets and complex data-driven models. One study which used data from the Clinical Practice Research Datalink database (n=378 256 patients between 30 and 84 years at baseline) found that a neural network substantially outperformed a reference logistic regression model (C-index: 0.764 versus 0.728), and correctly reclassified 7.6% of events. However, no survival models were used which limits the possibilities for valid clinical implementation. Another study included 423 604 UK Biobank participants and deployed an automated machine learning pipeline named AutoPrognosis. Compared with a Cox PH reference model which included only traditional cardiovascular predictors, a machine learning ensemble method including all 473 predictors resulted in a C-index of 0.774 versus 0.734 of the reference models, and a net correct reclassification

Table 3. Discrimination and Calibration of Sex-Specific Prediction Models for Different Predictor Subsets, Stratified by Age Group

Age range, y	Predictors	Women (n=276 113)				Men (n=266 028)			
		Performance metrics (95% CI)				Performance metrics (95% CI)			
		C-index	Δ C-stat*	Δ C-stat†	Calibration curve slope at 10 y	C-index	Δ C-stat*	Δ C-stat†	Calibration curve slope at 10 y
30–49	Baseline	0.648 (0.645–0.652)	Ref.	Ref.	0.999 (0.998–1.001)	0.661 (0.658–0.664)	Ref.	Ref.	1.001 (0.998–1.004)
	20	0.674 (0.671–0.677)	0.026	18%	1.000 (0.998–1.003)	0.673 (0.670–0.676)	0.012	7%	1.000 (0.998–1.002)
	50	0.678 (0.675–0.681)	0.03	20%	1.000 (0.997–1.002)	0.673 (0.671–0.675)	0.012	7%	1.001 (0.998–1.004)
30–39	Baseline	0.605 (0.601–0.609)	Ref.	Ref.	1.000 (0.998–1.003)	0.608 (0.604–0.612)	Ref.	Ref.	1.000 (0.998–1.003)
	20	0.651 (0.646–0.654)	0.049	47%	1.000 (0.997–1.003)	0.629 (0.625–0.633)	0.021	19%	1.001 (0.998–1.004)
	50	0.658 (0.654–0.663)	0.053	50%	0.999 (0.998–1.002)	0.629 (0.626–0.633)	0.021	19%	0.999 (0.996–1.002)
40–49	Baseline	0.572 (0.568–0.576)	Ref.	Ref.	0.999 (0.998–1.002)	0.578 (0.574–0.583)	Ref.	Ref.	1.001 (0.998–1.004)
	20	0.619 (0.615–0.623)	0.047	65%	1.000 (0.997–1.003)	0.600 (0.596–0.605)	0.022	28%	1.000 (0.997–1.003)
	50	0.624 (0.619–0.628)	0.052	72%	1.000 (0.997–1.002)	0.601 (0.597–0.605)	0.023	29%	1.001 (0.998–1.004)

Baseline traditional cardiovascular predictors: age, hypertension, antihypertensive medication, diabetes, hyperlipidemia, with Cox proportional hazard model using baseline predictors as reference model.
*Difference in C-statistic compared with the reference model.
†Difference in C-statistic compared with the reference model relative to full scale.

of events of 12.5%. An important difference with our study is that the UK Biobank contained relatively complete information on continuous predictors such as systolic blood pressure and total cholesterol.

In general, improvement in model performance may be attributable to (1) information gain resulting from including more predictors, or (2) modeling gain which is the ability of models to capture nonlinear associations or interactions among predictors.[19] In our study, the gain of complex (random survival forests) versus simple (Cox PH) models appeared to be limited. Random survival forests performed slightly more poorly compared with Cox regression models, potentially because random forests methods are prone to overfitting.[20] We do seem to find information gain by including predictors which are ranked as most important according to Cox elastic net models. This indicates that data-driven predictor selection results in the identification of valuable nontraditional cardiovascular predictors which increase predictive performance, such as socioeconomic status score and hormonal contraceptive use in women specifically. Because Cox PH and Cox elastic net models have a similar performance, Cox PH models would be preferred for clinical use since they can be interpreted more easily.[21]

Limitations and Strengths

Our study has several limitations. First, EHRs are designed to record data that are routinely collected during the clinical workflow to streamlining patient care, and not for the purpose of research.[22] Despite standardization using universal ICPC, *ICD* and ATC coding, previous research shows substantial underreporting in clinical diagnosis codes and large variability in interpractice data quality.[23] Underreporting leads to misclassification in predictors and outcome. Misclassification is not a problem in prediction research if the measurement error is similar in development compared with the deployment setting. Misclassification of the outcome may, however, lead to a biased estimation of absolute risk.[24] Fatal cardiovascular events could only be identified if they occurred in-hospital using *ICD-9* or *ICD-10* codes. It is possible that our study incidence of these events has been underestimated. Cardiovascular mortality comprises a quarter of all total CVD events. Prior research shows that the discriminating ability of prediction models did not differ between the fatal and non-fatal cardiovascular events.[25] Further, to optimally exclude patients with a history of cardiovascular events at baseline, we excluded patients with prescriptions of thrombocyte aggregation inhibitors which were specific for cardiovascular events (clopidogrel, dipyridamole, ticagrelor) at baseline. We did not include acetylsalicylic acid in this definition because of its prescription as analgesic

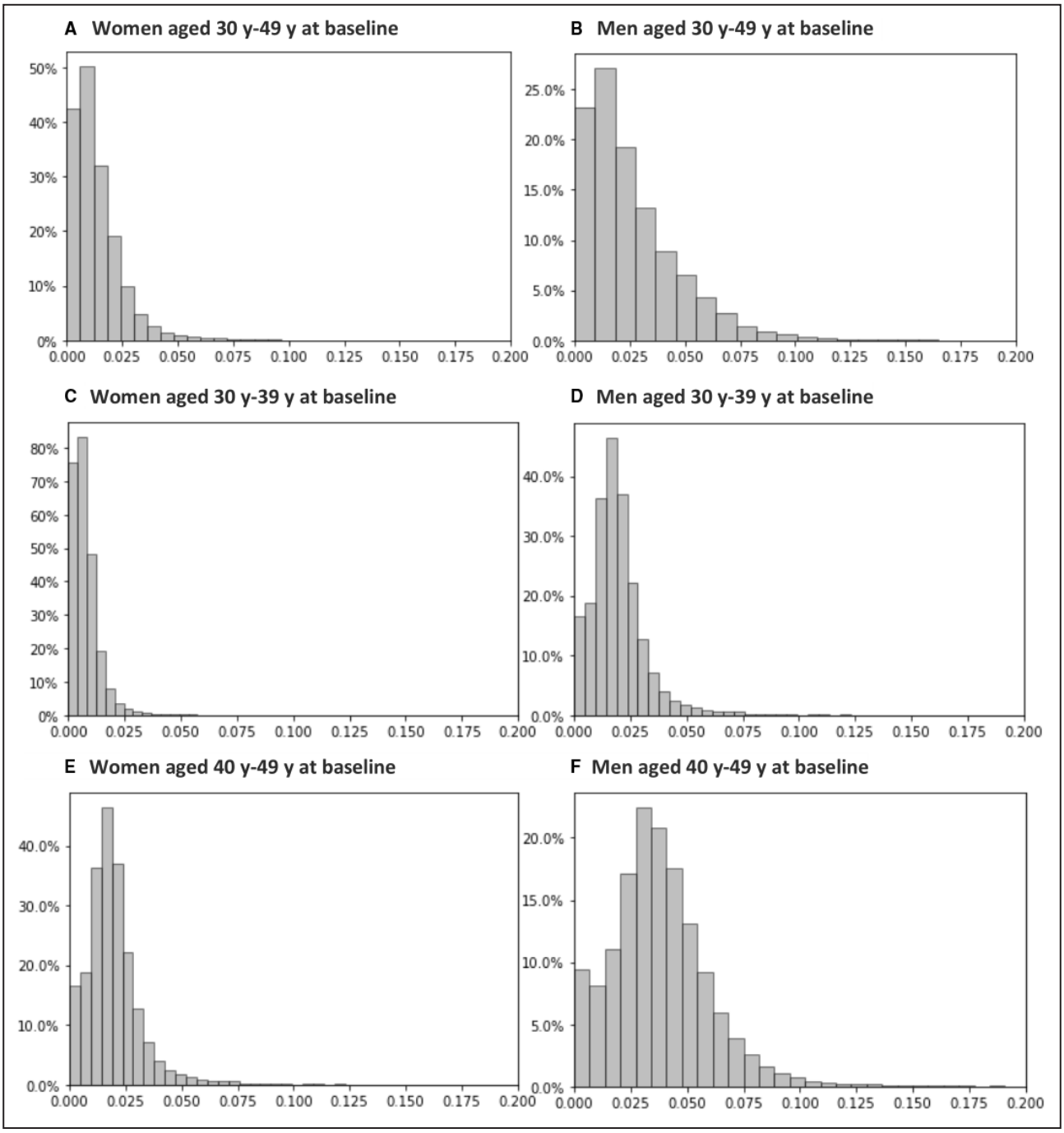

Figure. Absolute 10-year risk predictions of first-ever cardiovascular events including the 50 most important predictors, for women and men stratified by age groups.
A, Women aged 30 to 49 years at baseline. **B**, Men aged 30 to 49 years at baseline. **C**, Women aged 30 to 39 years at baseline. **D**, Men aged 30 to 39 years at baseline. **E**, Women aged 40 to 49 years at baseline. **F**, Men aged 40 to 49 years at baseline. On the x-axis the predicted probabilities from prediction models including the 50 most important predictors are shown, and on the y-axis the fraction (%) of the total population in each bin. All histograms have a bin size of 100.

in the study period, hence specificity for cardiovascular events was low.[26] In addition, we did not develop lifetime risk models in this cohort of young patients, because of the risk of misclassification in predictors and outcome may aggravate cohort effects. Second, we did not take noncardiovascular death into account as a competing risk because we assessed a young patient cohort at a maximum of 49 years at baseline. In this population, the cumulative incidence of noncardiovascular death was small (0.6%) compared with

the entire population, limiting the competing risk effect on the estimation of stroke risk. It should however be noted that registration of mortality in our EHR data is of suboptimal quality. Third, the reference Cox PH model did not include continuous laboratory or vital parameter measurements such as systolic blood pressure and total serum cholesterol, which limits the head-to-head comparison with commonly used models such as SCORE2.[3] However, such a comparison was not the purpose of this study. In addition, because we use data-driven selection of predictors, we identified predictor representations other than continuous measurements of blood pressure and cholesterol that did not require imputation. This is an advantage because of the often high extent of missingness of measurement data in the EHR. Fourth, our study population excluded patients receiving statin at baseline, which limits its use in patients already receiving statin treatment. However, our prediction models are specifically suited to support preventive interventions such as initiation of statin treatment, similar to the QRISK3 study in the United Kingdom, which is also based on EHR data.[8] We did not choose to exclude patients who received antihypertensive but not statin treatment at baseline, since in these patients the clinical decision on the initiation of statin treatment is also relevant and our models could be used for this decision. Fifth, although the continuous net reclassification index is a more sensitive measure to assess model reclassification, we chose the categorical net reclassification index because the 10-year risk threshold of 2.5% represents a clinically relevant threshold.

Strengths of this study include the large sample size of a cohort of patients aged <50 years at baseline, which is to our best knowledge among the largest to date. This offered a unique possibility to study data driven methods for the prediction of cardiovascular events in young patients. Furthermore, all predictors used in our models are directly available in the EHR, which facilitates implementation of the models directly in clinical practice. In addition, the linking of primary care and hospital diagnosis codes in the STIZON cohort enables validation of the cardiovascular outcome. Further, the data-driven predictor selection procedure results in that our models leverage predictive information from predictors other than continuous measurements of traditional cardiovascular predictors. Therefore, it is not necessary to impute these continuous measurements, which were missing in the vast majority of patients in our population.

Clinical Implications

Our EHR-derived models will not replace traditional models such as SCORE2 but could be used in a 2-step population health approach. First, at any given time point our models can automatically identify patient subgroups at increased risk for first-ever cardiovascular events above the absolute 10-year risk cut-off as specified by the European Society of Cardiology prevention guideline. Second, these patient subgroups could be invited to the primary care practice center for further cardiovascular risk assessment including measurement of systolic blood pressure and total and high-density lipoprotein cholesterol, after which traditional models such as SCORE2 could be used to estimate individualized risk. A previous modeling study found that such stepped strategy may result in more cost-effective cardiovascular risk management than the current opportunistic screening.[27] The European Society of Cardiology guidelines state 2.5% 10-year risk of cardiovascular events as the threshold between moderate and high risk for women and men aged <50 years, high risk being an indication for preventive pharmacotherapeutics. Although for patients <50 years in our cohort absolute 10-year risks are generally low, our data-driven models can be used to automatically identify patients whose absolute risk reaches the 2.5% risk cut-off. In women, we found that the Cox PH model with 50 most important predictors resulted in a net correct reclassification of events (3.7%) around this risk cut-off compared with the reference model. Although this percentage is low, application on a large scale could lead to sufficient clinical impact to justify the use of a relatively more complex model. After stratification based on the 30- to 39-year and 40-to-49-year age groups, we found that men and women between the age of 30 to 39 years at baseline had substantially lower absolute risks of cardiovascular events compared with those aged between 40 and 49 years. However, since the European Society of Cardiology guidelines use the SCORE2 model which does not include patients under 40 years, the absolute risk threshold of 2.5% likely is too high for individuals between the age of 30 to 39 years. Therefore, to define meaningful thresholds that can guide preventive therapy, we call for further research into the age group of 30 to 39 years. The focus may in this context not be pharmacotherapeutic, but rather on lifestyle interventions for prevention of CVD. In addition, for the 30- to 39-year age group lifetime risk estimation may further help in risk communication and interpretation. However, we should first invest in the creation of higher quality longitudinal data sources to derive valid lifetime risk prediction models. In addition, data-driven predictor selection has led to the identification of important nontraditional cardiovascular predictors such as socioeconomic status score and NSAID use. After stratifying for age subgroups, we found differences in the ranking of the 20 predictors that were most important in our prediction models. For example, in both women and men aged 30 to 39 years at baseline, the relative

importance of NSAID use further increased compared with the 40- to 49-year age group.

CONCLUSIONS

Sex-specific EHR-derived prediction models for first-ever cardiovascular events in the general population aged <50 years have moderate discriminatory performance and are well calibrated. Data-driven predictor selection leads to identification of nontraditional cardiovascular predictors, which modestly increase discriminatory performance of models and correct reclassification of events, mostly in women.

ARTICLE INFORMATION

Received June 1, 2022; accepted January 17, 2023.

Affiliations

Department of Neurology (H.J.A.v.O., M.J.H.W.), National eHealth Living Lab (H.J.A.v.O., T.N.B., M.M.R.), Department of Public Health & Primary Care (H.J.A.v.O., T.N.B., M.E.N.), Department of Biomedical Data Sciences (H.P., R.H.H.G.), and Department of Clinical Epidemiology (R.H.H.G.), Leiden University Medical Center, Leiden, The Netherlands Department of Neurology and Neurosurgery, UMC Utrecht Brain Center, Utrecht, The Netherlands (J.P.K., Y.M.R.)

Sources of Funding

Dr Van Os was funded by a personal Dekker Junior Clinical Scientist Grant (2018T082) and the Innovation Grant (2018T016) from the Dutch Heart Foundation, and a ZonMw Gender & Prevention Grant (555003014). Dr Wermer was supported by a personal VIDI grant from ZonMW, a Dekker Clinical Established Investigator Grant from the Netherlands Heart Foundation, and a Fellowship grant from the Netherlands Brain Foundation. Dr Ruigrok has received funding from the European Research Council under the European Union's Horizon 2020 research and innovation program (grant agreement No. 852173).

Disclosures

None.

Supplemental Material

Table S1–S5

REFERENCES

1. *Global status report on noncommunicable diseases 2014*. World Health Organization. 2014. https://apps.who.int/iris/handle/10665/148114 Accesssed February 27, 2023.
2. van der Ende MY, Sijtsma A, Snieder H, van der Harst P. Letter to editor: reply on question of Marques Jr et al. regarding the paper entitled: "The lifelines cohort study: prevalence and treatment of cardiovascular disease and risk factors." *Int J Cardiol*. 2019;294:57. doi: 10.1016/j.ijcard.2019.06.026
3. Score Working Group. Score2 risk prediction algorithms: new models to estimate 10-year risk of cardiovascular disease in Europe. *Eur Heart J*. 2021;42:2439–2454. doi: 10.1093/eurheartj/ehab369
4. Ference BA, Ginsberg HN, Graham I, Ray KK, Packard CJ, Bruckert E, Hegele RA, Krauss RM, Raal FJ, Schunkert H, et al. Low-density lipoproteins cause atherosclerotic cardiovascular disease. 1. Evidence from genetic, epidemiologic, and clinical studies. A consensus statement from the European Atherosclerosis Society Consensus Panel. *Eur Heart J*. 2017;38:2459–2472. doi: 10.1093/eurheartj/ehx144
5. Graham IM, Di Angelantonio E, Visseren F, De Bacquer D, Ference BA, Timmis A, Halle M, Vardas P, Huculeci R, Cooney MT, et al. Systematic coronary risk evaluation (score): JACC focus seminar 4/8. *J Am Coll Cardiol*. 2021;77:3046–3057. doi: 10.1016/j.jacc.2021.04.052
6. Appelman Y, van Rijn BB, Ten Haaf ME, Boersma E, Peters SA. Sex differences in cardiovascular risk factors and disease prevention. *Atherosclerosis*. 2015;241:211–218. doi: 10.1016/j.atherosclerosis.2015.01.027
7. Ohno-Machado L. Sharing data from electronic health records within, across, and beyond healthcare institutions: current trends and perspectives. *J Am Med Inform Assoc*. 2018;25:1113. doi: 10.1093/jamia/ocy116
8. 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. doi: 10.1136/bmj.j2099
9. Steele AJ, Denaxas SC, Shah AD, Hemingway H, Luscombe NM. Machine learning models in electronic health records can outperform conventional survival models for predicting patient mortality in coronary artery disease. *PLoS One*. 2018;13:e0202344. doi: 10.1371/journal.pone.0202344
10. Weng SF, Reps J, Kai J, Garibaldi JM, Qureshi N. Can machine-learning improve cardiovascular risk prediction using routine clinical data? *PLoS One*. 2017;12:e0174944. doi: 10.1371/journal.pone.0174944
11. Alaa A, Schaar M. Autoprognosis: automated clinical prognostic modeling via Bayesian optimization with structured kernel learning. 2018. Proceedings of the 35 th International Conference on Machine. Learning, Stockholm, Sweden, PMLR 80, 2018. URL: https://www.vanderschaar-lab.com/papers/ICML2018_AP.pdf (accessed 27 February 2023)
12. Kuiper JG, Bakker M, Penning-van Beest FJA, Herings RMC. Existing data sources for clinical epidemiology: the pharmo database network. *Clin Epidemiol*. 2020;12:415–422. doi: 10.2147/CLEP.S247575
13. Lamberts H, Wood M, eds. *ICPC, International Classification of Primary Care*. Oxford University Press; 1987.
14. *WHO Collaborating Centre for Drug Statistics Methodology, ATC Classification Index with DDDs, 2023*. 2022. https://www.whocc.no/atc_ddd_index/ Accessed 27 February 2023.
15. Sociaal Cultureel Planbureau. Available at: http://www.scp.nl/Onderzoek/Lopend_onderzoek/A_Z_alle_lopende_onderzoeken/Statusscores. Accessed July 5, 2021.
16. Damen JA, Hooft L, Schuit E, Debray TP, Collins GS, Tzoulaki I, Lassale CM, Siontis GCM, Chiocchia V, Roberts C, et al. Prediction models for cardiovascular disease risk in the general population: systematic review. *BMJ*. 2016;353:i2416. doi: 10.1136/bmj.i2416
17. Visseren FLJ, Mach F, Smulders YM, Carballo D, Koskinas KC, Back M, Benetos A, Biffi A, Boavida JM, Capodanno D, et al. 2021 ESC Guidelines on cardiovascular disease prevention in clinical practice. *Eur Heart J*. 2021;42:3227–3337. doi: 10.1093/eurheartj/ehab484
18. 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. doi: 10.1136/bmj.g7594
19. Alaa AM, Bolton T, Di Angelantonio E, Rudd JHF, van der Schaar M. Cardiovascular disease risk prediction using automated machine learning: a prospective study of 423,604 UK Biobank participants. *PLoS One*. 2019;14:e0213653. doi: 10.1371/journal.pone.0213653
20. Ishwaran HKU, Blackstone EH, Lauer MS. Random survival forests. *Ann Appl Stat*. 2008;2:841–860. doi: 10.1214/08-AOAS169
21. James G, Witten D, Hastie T, Tibshirani R. *An Introduction to Statistical Learning*. 1st ed. Springer; 2013. doi: 10.1007/978-1-4614-7138-7
22. Spasoff RA. *Epidemiologic Methods for Health Policy*. Oxford University Press I; 1999.
23. de Lusignan S, Valentin T, Chan T, Hague N, Wood O, van Vlymen J, Dhoul N. Problems with primary care data quality: osteoporosis as an exemplar. *Inform Prim Care*. 2004;12:147–156. doi: 10.14236/jhi.v12i3.120
24. Pajouheshnia R, van Smeden M, Peelen LM, Groenwold RHH. How variation in predictor measurement affects the discriminative ability and transportability of a prediction model. *J Clin Epidemiol*. 2019;105:136–141. doi: 10.1016/j.jclinepi.2018.09.001
25. van Dis I, Geleijnse JM, Boer JM, Kromhout D, Boshuizen H, Grobbee DE, van der Schouw YT, Monique Verschuren WMM. Effect of including nonfatal events in cardiovascular risk estimation, illustrated with data from The Netherlands. *Eur J Prev Cardiol*. 2014;21:377–383. doi: 10.1177/2047487312443485
26. Stichting Farmaceutische Kengetallen Pijnstilling op recept. 2008. Pharmaceutisch Weekblad, Jaargang 143 Nr 39.https://www.sfk.nl/publicaties/PW/2008/2008-39.html Accessed 27 February 2023
27. Crossan C, Lord J, Ryan R, Nherera L, Marshall T. Cost effectiveness of case-finding strategies for primary prevention of cardiovascular disease: a modelling study. *Br J Gen Pract*. 2017;67:e67–e77. doi: 10.3399/bjgp16X687973