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

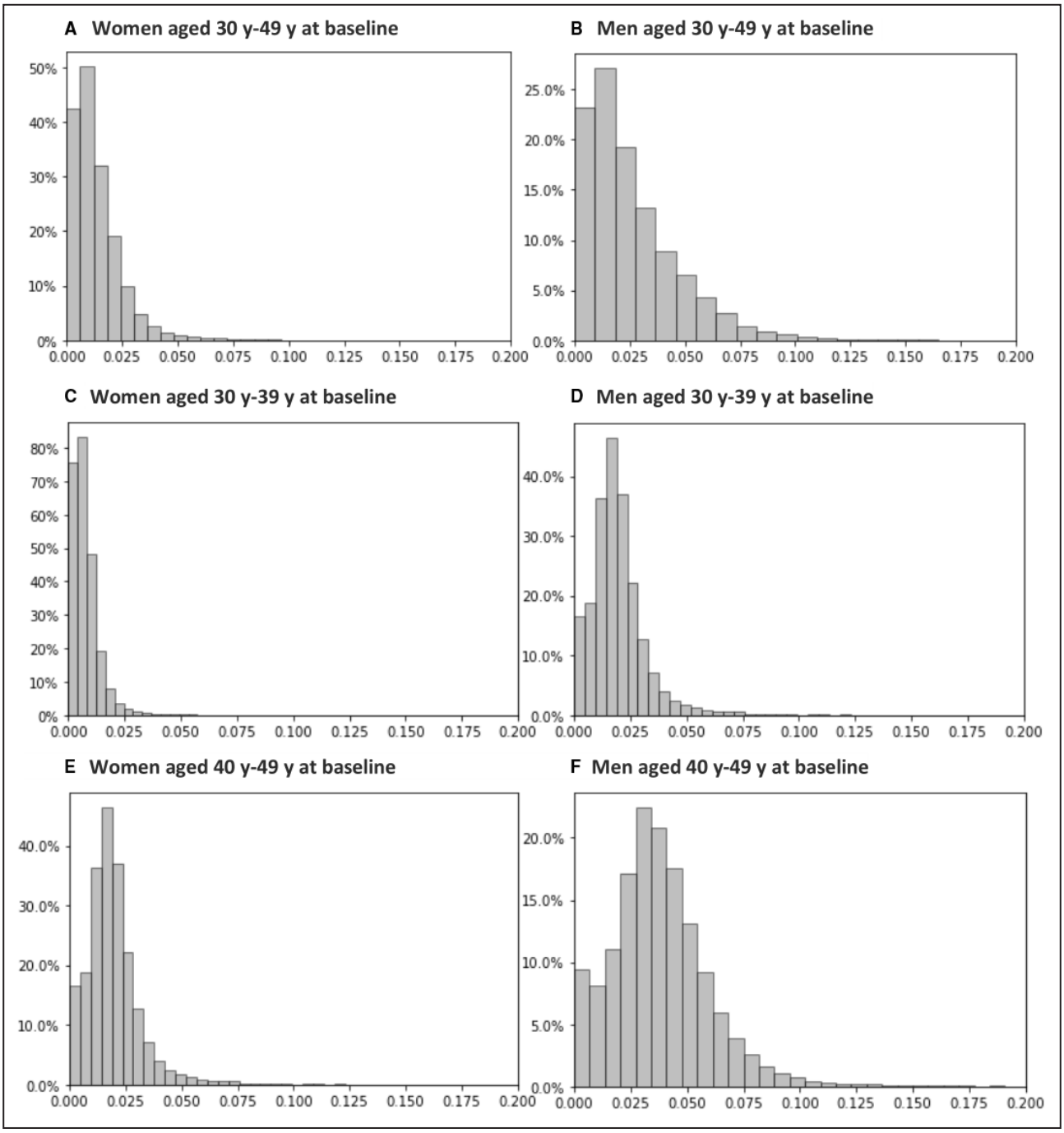

Figure. Absolute 10-year risk predictions of first-ever cardiovascular events including the 50 most important predictors, for women and men stratified by age groups.
A, Women aged 30 to 49 years at baseline. **B**, Men aged 30 to 49 years at baseline. **C**, Women aged 30 to 39 years at baseline. **D**, 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

Supplemental Material

Table S1. Clustering of cardiovascular events, exclusion criteria and predictors

	Predictor	ICPC	ICD-9	ICD-10	ATC
Cardiovascular event definition	stroke, not specified	K90	NULL	I64	NULL
	ischemic stroke	NULL	434	I63	NULL
		NULL	436	NULL	NULL
	hemorrhagic stroke	NULL	431	I61	NULL
		NULL	4329	I629	NULL
	acute myocardial infarction	K75	410	I21	NULL
		NULL	NULL	I22	NULL
History of cardiovascular disease (exclusion)	TIA (including amaurosis fugax)	K89	435	G45	NULL
		NULL	36234	NULL	NULL
	subarachnoid hemorrhage	NULL	430	I60	NULL
	ICA aneurysm unruptured	NULL	44281	I720	NULL
	Other acute ischemic heart diseases	NULL	41189	I24	NULL
		K76	4149	I252	NULL
		NULL	4148	I255	NULL
		NULL	NULL	I256	NULL
		NULL	NULL	I258	NULL
	history of personal cardiovascular disease	NULL	V1254	Z8673	NULL
	Presence of aortocoronary bypass graft	NULL	NULL	Z951	NULL
	angina pectoris	K74	413	I20	NULL
	peripheral arterial disease	K92	44020	I7021	NULL
		NULL	44021	I7022	NULL
		NULL	44022	I7023	NULL
		NULL	44023	I7024	NULL
		NULL	44024	I7025	NULL
		NULL	44029	I7026	NULL
		NULL	NULL	I7029	NULL
	Retinal vascular occlusions	NULL	NULL	H34	NULL
	event specific medication	NULL	NULL	NULL	B01AC04
		NULL	NULL	NULL	B01AC07
		NULL	NULL	NULL	B01AC24
Predictor group	**Predictor**	**ICPC**	**ICD-9**	**ICD-10**	**ATC**
Cardiovascular system	Cardiac complaints	K01	NULL	NULL	NULL
		K02	NULL	NULL	NULL
	Family history of stroke	NULL	V171	Z823	NULL
	Heart palpitations	K04	NULL	R002	NULL
		K05	NULL	NULL	NULL
	Ankle edema	K07	71907	R600	NULL
		NULL	NULL	R609	NULL
	Cardiac arrhythmia medication	NULL	NULL	NULL	C01B
	Cardiac disease medication	NULL	NULL	NULL	C01A

	NULL	NULL	NULL	C01C
	NULL	NULL	NULL	C01D
	NULL	NULL	NULL	C01E
Cardiovascular risk management	K49	NULL	NULL	NULL
Hypertension	K86	401	I10	NULL
	K87	402	I11	NULL
Elevated blood pressure	K85	NULL	NULL	NULL
Diuretics	NULL	NULL	NULL	C03
Vasodilatators	NULL	NULL	NULL	C04
Vasoprotective medication	NULL	NULL	NULL	C05
Betablockers	NULL	NULL	NULL	C07
Calcium channel blockers	NULL	NULL	NULL	C08
RAAS inhibitors	NULL	NULL	NULL	C09
Antihypertensive medication: atypical	NULL	NULL	NULL	C02
Atrial fibrillation	K78	4273	I48	NULL
Oral anticoagulant drugs	NULL	NULL	NULL	B01AA
	NULL	NULL	NULL	B01AB
	NULL	NULL	NULL	B01AE
	NULL	NULL	NULL	B01AF
Embolism	K93	4151	I26	NULL
	NULL	4449	I82	NULL
Obstetric embolism	NULL	67320	O88	NULL
Heart failure	K77	428	I50	NULL
	NULL	NULL	I110	NULL
Cor pulmonale	K82	416	I27	NULL
Heart murmur	K81	NULL	NULL	NULL
Cardiac atherosclerosis	NULL	4144	I2584	NULL
	NULL	4143	NULL	NULL
Other heart disease	K84	NULL		NULL
	NULL	420	I30	NULL
	NULL	421	I31	NULL
	NULL	422	I32	NULL
	NULL	423	I33	NULL
	NULL	424	I40	NULL
	NULL	425	I41	NULL
	NULL	426	I42	NULL
	NULL	427	I43	NULL
	NULL	429	I51	NULL
	NULL	NULL	I52	NULL
	NULL	NULL	I34	NULL
	NULL	NULL	I35	NULL
	NULL	NULL	I36	NULL
	NULL	NULL	I37	NULL
	NULL	NULL	I38	NULL
	NULL	NULL	I39	NULL
Rheumatic heart disease	NULL	390	I00	NULL

		NULL	398	NULL	NULL
		NULL	39890	I01	NULL
		NULL	NULL	I02	NULL
		NULL	NULL	I03	NULL
		NULL	NULL	I04	NULL
		NULL	NULL	I05	NULL
		NULL	NULL	I06	NULL
		NULL	NULL	I07	NULL
		NULL	NULL	I08	NULL
		NULL	NULL	I09	NULL
	Orthostatic hypotension	K88	4580	I951	NULL
	Thrombophlebitis	K94	NULL	I80	NULL
	Varices	K95	NULL	I86	NULL
		NULL	NULL	I83	NULL
		NULL	NULL	I85	NULL
	Hemorroids	K96	NULL	K64	NULL
	Diabetes mellitus	T90	250	E10	NULL
		NULL	NULL	E11	NULL
	Antidiabetic medication	NULL	NULL	NULL	A10A
		NULL	NULL	NULL	A10B
	Prediabetes	A91.05	7902	R73	NULL
	Hyperlipidemia	T93	272	E78	NULL
	Thrombocyte aggregation inhibitors	NULL	NULL	NULL	B01AC06
		NULL	NULL	NULL	B01AC56
		NULL	NULL	NULL	B01AC08
	Statin use	NULL	NULL	NULL	C10A
		NULL	NULL	NULL	C10B
	Metabolic syndrome	NULL	2779	E88	NULL
Airway	COPD	R95	491	J44	NULL
	Asthma	R96	493	J45	NULL
	Cough suppressants	NULL	NULL	NULL	R05D
	Antihistaminic medication for systemic	NULL	NULL	NULL	R06A
	Nasal medication	NULL	NULL	NULL	R01
	Medication for obstructive airway disease	NULL	NULL	NULL	R03A
		NULL	NULL	NULL	R03B
		NULL	NULL	NULL	R03C
		NULL	NULL	NULL	R03D
	Bronchitis	R78	466	J20	NULL
	Allergic rhinitis	R97	NULL	J30	NULL
	Upper airway symptoms	R21	NULL	NULL	NULL
		R07	NULL	NULL	NULL
		R22	NULL	NULL	NULL
		R77	NULL	NULL	NULL
	Pneumonia	R81	480	J12	NULL
		NULL	481	J13	NULL
		NULL	482	J14	NULL

		NULL	483	J15	NULL
		NULL	484	J16	NULL
		NULL	485	J17	NULL
		NULL	486	J18	NULL
	Upper respiratory disease	R74	NULL	J31	NULL
		R75	NULL	J32	NULL
		R76	NULL	J33	NULL
		NULL	NULL	J34	NULL
		NULL	NULL	J35	NULL
		NULL	NULL	J36	NULL
		NULL	NULL	J37	NULL
		NULL	NULL	J38	NULL
		NULL	NULL	J39	NULL
	Influenza	R80	487	J09	NULL
		NULL	NULL	J10	NULL
		NULL	NULL	J11	NULL
	Hyperventilation	R98	NULL	NULL	NULL
	Streptangina	R72	NULL	NULL	NULL
	Sarcoidosis	B99.02	NULL	D86	NULL
	Pertussis	R71	NULL	A37	NULL
	Lung diseases due to external agents	NULL	NULL	J60	NULL
		NULL	NULL	J61	NULL
		NULL	NULL	J62	NULL
		NULL	NULL	J63	NULL
		NULL	NULL	J64	NULL
		NULL	NULL	J65	NULL
		NULL	NULL	J66	NULL
		NULL	NULL	J67	NULL
		NULL	NULL	J68	NULL
		NULL	NULL	J69	NULL
		NULL	NULL	J70	NULL
	Tuberculosis	A70	NULL	A15	NULL
		NULL	NULL	A17	NULL
		NULL	NULL	A18	NULL
		NULL	NULL	A19	NULL
	Emphysema	NULL	NULL	J43	NULL
	Bronchiectasy	NULL	NULL	J47	NULL
	Interstitial lung disease	NULL	516	J84	NULL
Dermatology	Herpes zoster	S70	NULL	B02	NULL
	Dermatitis	S87	NULL	L20	NULL
		S88	NULL	L21	NULL
		NULL	NULL	L23	NULL
		NULL	NULL	L24	NULL
		NULL	NULL	L25	NULL
		NULL	NULL	L26	NULL
		NULL	NULL	L28	NULL

		NULL	NULL	L29	NULL
	Psoriasis	S91	6961	L40	NULL
	Non-pressure chronic ulcer	S97	NULL	L97	NULL
	Dermatological complaints	S03	680	NULL	NULL
		S04	681	NULL	NULL
		S05	682	NULL	NULL
		S18	683	NULL	NULL
		S22	684	NULL	NULL
		S29	685	NULL	NULL
		S74	686	NULL	NULL
		S75	NULL	NULL	NULL
		S79	NULL	NULL	NULL
		S81	NULL	NULL	NULL
		S82	NULL	NULL	NULL
		S84	NULL	NULL	NULL
		S93	NULL	NULL	NULL
		S96	NULL	NULL	NULL
		S98	NULL	NULL	NULL
		S99	NULL	NULL	NULL
		F02	NULL	NULL	NULL
		F03	NULL	NULL	NULL
		F04	NULL	NULL	NULL
		F73	NULL	NULL	NULL
		F99	NULL	NULL	NULL
		D05	NULL	NULL	NULL
ENT	Vertigo	H82	NULL	H80	NULL
		N17	NULL	H81	NULL
		NULL	NULL	H82	NULL
	Hearing problems	H02	NULL	NULL	NULL
		H03	NULL	NULL	NULL
		H77	NULL	NULL	NULL
		H84	NULL	NULL	NULL
		H83	NULL	NULL	NULL
		H85	NULL	NULL	NULL
	Deafness	H86	NULL	NULL	NULL
	Otitis	H71	NULL	H65	NULL
		H72	NULL	H66	NULL
		H74	NULL	H76	NULL
		H70	NULL	NULL	NULL
Gastroenterology	Hepatitis	D72	NULL	B15	NULL
		NULL	NULL	B16	NULL
		NULL	NULL	B17	NULL
		NULL	NULL	B18	NULL
		NULL	NULL	B19	NULL
	Esophagal disease	D84	NULL	K20	NULL
		NULL	NULL	K21	NULL

		NULL	NULL	K22	NULL
		NULL	NULL	K23	NULL
	Appendicitis	D88	NULL	K35	NULL
		NULL	NULL	K36	NULL
		NULL	NULL	K37	NULL
	Diverticulitis	K57	NULL	D92	NULL
	Irritable bowel syndrome	D93	NULL	K58	NULL
	Inflammatory bowel disease	D94	NULL	K50	NULL
		NULL	NULL	K51	NULL
		NULL	NULL	K52	NULL
	Cholecystitis	D98	NULL	K80	NULL
		NULL	NULL	K81	NULL
		NULL	NULL	K82	NULL
		NULL	NULL	K83	NULL
		NULL	NULL	K84	NULL
	Pancreatitis	D99.04	NULL	K85	NULL
	Coeliac disease	D99.06	NULL	K900	NULL
	Gastric ulcus	D85	NULL	K25	NULL
		D86	NULL	K26	NULL
		NULL	NULL	K27	NULL
		NULL	NULL	K28	NULL
	Gastroenterological complaints	D01	NULL	NULL	NULL
		D02	NULL	NULL	NULL
		D03	NULL	NULL	NULL
		D06	NULL	NULL	NULL
		D09	NULL	NULL	NULL
		D10	NULL	NULL	NULL
		D11	NULL	NULL	NULL
		D12	NULL	NULL	NULL
		D16	NULL	NULL	NULL
		D21	NULL	NULL	NULL
		D87	NULL	NULL	NULL
		D78	NULL	NULL	NULL
		D29	NULL	NULL	NULL
		D70	NULL	NULL	NULL
		D18	NULL	NULL	NULL
	Constipation medication	NULL	NULL	NULL	A06A
	Propulsives	NULL	NULL	NULL	A03F
	Drugs for acid related disorders	NULL	NULL	NULL	A02A
	Gastroesophageal reflux medication	NULL	NULL	NULL	A02B
Immunology	Rheumatoid arthritis	L88	714	M05	NULL
	HIV/AIDS	B90	42	B20	J05AR
		NULL	V08	Z21	NULL
	Selective immunosuppressants	NULL	NULL	NULL	L04AA
	Antiinflammatory drugs	NULL	NULL	NULL	M01C
		NULL	NULL	NULL	M01A

	Systemic corticosteroid drugs	NULL	NULL	NULL	H02A
		NULL	NULL	NULL	H02B
Lifestyle	Folate deficiency anemia	B81	NULL	D52	NULL
	Iron deficiency anemia	B80	NULL	D50	NULL
	Anemia: other causes	B82	NULL	D64	NULL
	Antianemic preparations	NULL	NULL	NULL	B03A
		NULL	NULL	NULL	B03B
		NULL	NULL	NULL	B03X
	Alcohol intake	P15	NULL	F10	NULL
		NULL	NULL	Y90	NULL
	Smoking	P17	3051	Z716	N07BA
		NULL	V1582	Z720	NULL
		NULL	NULL	F17	NULL
	Drug abuse	P19	NULL	F19	NULL
		NULL	NULL	F55	NULL
		NULL	NULL	F11	NULL
		NULL	NULL	F12	NULL
		NULL	NULL	F13	NULL
		NULL	NULL	F14	NULL
		NULL	NULL	F15	NULL
		NULL	NULL	F16	NULL
		NULL	NULL	F18	NULL
	Weight increase	T07	NULL	NULL	NULL
	Weight loss	T08	NULL	NULL	NULL
	Overweight	T82	NULL	E66	NULL
		T83	NULL	NULL	NULL
	Vitamin deficiency	T91	NULL	NULL	NULL
	Personal history of self harm	NULL	NULL	Z915	NULL
Male-specific	Complaints of the male genital system	Y04	NULL	NULL	NULL
		Y81	NULL	NULL	NULL
		Y13	NULL	NULL	NULL
	Complaints prostate	Y06	NULL	NULL	NULL
		Y85	NULL	NULL	NULL
	Infertility	N46	6069	N469	NULL
	Erectile dysfunction	P08.01	NULL	N52	G04BE
		Y07	NULL	NULL	NULL
Musculoskeletal	Bone fracture	L72	NULL	M484	NULL
		L73	NULL	M495	NULL
		L74	NULL	NULL	NULL
		L75	NULL	M843	NULL
		L76	NULL	M844	NULL
		NULL	NULL	M907	NULL
		NULL	NULL	M966	NULL
		NULL	NULL	S02	NULL
		NULL	NULL	S12	NULL
		NULL	NULL	S22	NULL

	NULL	NULL	S32	NULL
	NULL	NULL	S42	NULL
	NULL	NULL	S52	NULL
	NULL	NULL	S62	NULL
	NULL	NULL	S72	NULL
	NULL	NULL	S82	NULL
	NULL	NULL	S92	NULL
	NULL	NULL	T02	NULL
	NULL	NULL	T08	NULL
	NULL	NULL	T10	NULL
	NULL	NULL	T12	NULL
	NULL	NULL	T142	NULL
Osteoporosis	L95	NULL	M80	M05BA
	NULL	NULL	M81	NULL
Arthritis	L84	NULL	M15	NULL
	L89	NULL	M16	NULL
	L90	NULL	M17	NULL
	L91	NULL	M18	NULL
	NULL	NULL	M19	NULL
Gout	T92	NULL	M1A	NULL
	NULL	NULL	M10	NULL
Movement related complaints	L01	NULL	NULL	NULL
	L03	NULL	NULL	NULL
	L99	NULL	NULL	NULL
	L92	NULL	NULL	NULL
	L15	NULL	NULL	NULL
	L08	NULL	NULL	NULL
	L17	NULL	NULL	NULL
	L81	NULL	NULL	NULL
	L78	NULL	NULL	NULL
	L19	NULL	NULL	NULL
	L98	NULL	NULL	NULL
	L80	NULL	NULL	NULL
	L79	NULL	NULL	NULL
	L10	NULL	NULL	NULL
	L83	NULL	NULL	NULL
	L93	NULL	NULL	NULL
	L05	NULL	NULL	NULL
	L86	NULL	NULL	NULL
	L14	NULL	NULL	NULL
	L12	NULL	NULL	NULL
	L04	NULL	NULL	NULL
	L02	NULL	NULL	NULL
	L77	NULL	NULL	NULL
	L13	NULL	NULL	NULL
	L11	NULL	NULL	NULL

		L09	NULL	NULL	NULL
		L16	NULL	NULL	NULL
		L29	NULL	NULL	NULL
		L07	NULL	NULL	NULL
		L87	NULL	NULL	NULL
		L85	NULL	NULL	NULL
Neurology	Headache	N01	7840	G44	NULL
		N02	NULL	G51	NULL
		N90	NULL	NULL	NULL
		N92	NULL	NULL	NULL
	Migraine	N89	346	G43	N02C
	Migraine without aura	NULL	3461	G430	NULL
		NULL	3462	G434	NULL
		NULL	3463	G435	NULL
		NULL	3464	G436	NULL
		NULL	3466	G437	NULL
		NULL	3467	G438	NULL
		NULL	3468	G439	NULL
		NULL	3469	G43A	NULL
		NULL	NULL	G43B	NULL
		NULL	NULL	G43C	NULL
	Migraine with aura	NULL	3460	G431	NULL
		NULL	3465	NULL	NULL
	Epilepsia	N88	NULL	G40	N03A
	Sensibility disorder	N06	NULL	G25	NULL
	Restless legs syndrome	N04	NULL	NULL	NULL
	Head injury and concussion	N79	NULL	NULL	NULL
		N80	NULL	NULL	NULL
	Polyneuropathy	N94	NULL	G61	NULL
		N05	NULL	G62	NULL
		NULL	NULL	G63	NULL
		NULL	NULL	G64	NULL
	Parkinson's disease	N87	NULL	G20	N04
	Diseases of myoneural junction and muscle	N99	NULL	G70	NULL
		NULL	NULL	G71	NULL
		NULL	NULL	G72	NULL
		NULL	NULL	G73	NULL
	Multiple sclerosis	N86	NULL	G35	NULL
		NULL	NULL	G36	NULL
		NULL	NULL	G37	NULL
	Opioid use	NULL	NULL	NULL	N01A
	Logal analgesics	NULL	NULL	NULL	N01B
	Analgesics: other	NULL	NULL	NULL	N02B
Ophthalmology	Cataract	F92	NULL	H25	NULL
		NULL	NULL	H26	NULL
		NULL	NULL	H28	NULL

	Conjunctivitis	F70	NULL	H10	NULL
	Visual disorders	F83	NULL	NULL	NULL
		F84	NULL	NULL	NULL
		F94	NULL	NULL	NULL
	Glaucoma	F93	NULL	H40	NULL
		NULL	NULL	H42	NULL
	Ophthalmologic medication	NULL	NULL	NULL	S01
		NULL	NULL	NULL	S02
Psychiatry	Sleaplessness	P06	NULL	NULL	NULL
	Memory disorders	P20	NULL	NULL	NULL
	Dementia	P70	NULL	F01	N06D
		NULL	NULL	F02	NULL
		NULL	NULL	F03	NULL
	ADD/ADHD	NULL	NULL	NULL	N06B
	Anxiety or stress related disorders	P74	300	F40	N05B
		P75	NULL	F41	N05C
		P79	NULL	F42	NULL
		P82	NULL	F43	NULL
		NULL	NULL	F44	NULL
		NULL	NULL	F45	NULL
		NULL	NULL	F48	NULL
	Psychotic disorders	P72	298	F20	N05A
		P73	NULL	F21	NULL
		P98	NULL	F22	NULL
		NULL	NULL	F23	NULL
		NULL	NULL	F24	NULL
		NULL	NULL	F25	NULL
		NULL	NULL	F28	NULL
		NULL	NULL	F29	NULL
	Psychiatric complaints: other	P99	NULL	NULL	NULL
		P02	NULL	NULL	NULL
		P78	NULL	NULL	NULL
		P03	NULL	NULL	NULL
		P29	NULL	NULL	NULL
		P01	NULL	NULL	NULL
		P04	NULL	NULL	NULL
	Bipolar or manic disorders	P73.02	296	F30	NULL
		NULL	NULL	F31	NULL
	Depressive disorder	P76	311	F32	N06A
		NULL	NULL	F33	NULL
	Mood disorder: unspecified	NULL	NULL	F34	NULL
	Intellectual disabilities	NULL	NULL	F70	NULL
		NULL	NULL	F71	NULL
		NULL	NULL	F72	NULL
		NULL	NULL	F73	NULL
		NULL	NULL	F78	NULL

		NULL	NULL	F79	NULL
	Delirium	P71	2930	F05	NULL
	Personality disorders	P80	301	F60	NULL
		NULL	NULL	F63	NULL
		NULL	NULL	F64	NULL
		NULL	NULL	F65	NULL
		NULL	NULL	F66	NULL
		NULL	NULL	F68	NULL
		NULL	NULL	F69	NULL
Socio-economic	Financial problems	Z08	NULL	NULL	NULL
		Z10	NULL	NULL	NULL
		Z01	NULL	NULL	NULL
	Other socially related problems	Z29	NULL	Z73	NULL
		Z05	NULL	Z55	NULL
		Z12	NULL	Z56	NULL
		Z14	NULL	Z57	NULL
		Z16	NULL	Z59	NULL
		Z18	NULL	Z60	NULL
		Z19	NULL	Z62	NULL
		Z20	NULL	Z63	NULL
		Z21	NULL	Z64	NULL
		Z22	NULL	Z65	NULL
		Z25	NULL	NULL	NULL
Family history	Oncological diagnosis	A29.02	NULL	Z80	NULL
		A29.03	NULL	NULL	NULL
		A29.04	NULL	NULL	NULL
	Cardiovascular disease	A29.01	NULL	Z824	NULL
		NULL	NULL	Z823	NULL
	Diabetes mellitus	A29.05	NULL	NULL	NULL
	Hyperlipidemia	A29.06	NULL	Z834	NULL
	Psychiatric disease	NULL	NULL	Z81	NULL
Urology	Miction related problems	U02	NULL	NULL	NULL
		U01	NULL	NULL	NULL
		U05	NULL	NULL	NULL
	Hematuria	U06	NULL	NULL	NULL
	Cystitis & pyelonephritis	U71	NULL	N30	NULL
		U70	NULL	N10	NULL
	Kidney failure	U99	585	N17	NULL
		NULL	NULL	N18	NULL
		NULL	NULL	N19	NULL
	Nephrosis	U88	NULL	N00	NULL
		NULL	NULL	N01	NULL
		NULL	NULL	N02	NULL
		NULL	NULL	N03	NULL
		NULL	NULL	N04	NULL
		NULL	NULL	N05	NULL

		NULL	NULL	N06	NULL
		NULL	NULL	N07	NULL
		NULL	NULL	N08	NULL
	Urolithiasis	U95	NULL	N20	NULL
		NULL	NULL	N21	NULL
		NULL	NULL	N22	NULL
		NULL	NULL	N23	NULL
	Urine incontinence	U04	NULL	N394	NULL
Women-specific factors	Precocious puberty	T99.05	2591	E301	NULL
	Ovarian dysfunction	T99.06	256	E28	NULL
	Gestational diabetes mellitus	W84.02	64800	O249	NULL
		NULL	6488	O244	NULL
	Diabetes during pregnancy	NULL	6480	O240	NULL
		NULL	NULL	O241	NULL
		NULL	NULL	O243	NULL
		NULL	NULL	O248	NULL
	Poor fetal growth	W84.04	6565	O365	NULL
	Pregnancy complicating risk factors	W77	NULL	NULL	NULL
	Hemorrhage during pregnancy	NULL	640	O46	NULL
	Complications during birth	W92	NULL	NULL	NULL
		NULL	641	O61	NULL
		NULL	642	O62	NULL
		NULL	643	O63	NULL
		NULL	644	O64	NULL
		NULL	645	O65	NULL
		NULL	646	O66	NULL
		NULL	647	O68	NULL
		NULL	648	O69	NULL
		NULL	649	O70	NULL
		NULL	768	O72	NULL
		NULL	NULL	O73	NULL
		NULL	NULL	O74	NULL
		NULL	NULL	O75	NULL
		NULL	NULL	O76	NULL
		NULL	NULL	O77	NULL
	Encounter for supervision of normal pregnancy	W78	V22	Z34	NULL
	Infertility	W15	628	N97	NULL
	Preeclampsia or eclampsia	W81.02	6424	O14	NULL
		W81.03	6425	O15	NULL
		NULL	6426	O11	NULL
		NULL	6427	NULL	NULL
	Gestational hypertension	NULL	6423	O13	NULL
	Hypertensive disorders during pregnancy	NULL	6420	O16	NULL
		NULL	6421	O10	NULL

	NULL	6422	NULL	NULL
	NULL	6429	NULL	NULL
Risk factors during pregnancy	W84.03	NULL	P05	NULL
	W84.04	NULL	P07	NULL
	W84.05	NULL	P08	NULL
	W84.06	NULL	P09	NULL
	W84.07	NULL	NULL	NULL
	W84.08	NULL	NULL	NULL
Birth	W90	V270	Z370	NULL
	NULL	NULL	Z372	NULL
	NULL	V272	Z373	NULL
	NULL	V273	Z374	NULL
	NULL	V274	Z375	NULL
	NULL	V275	Z376	NULL
	NULL	V276	Z377	NULL
	NULL	V277	Z379	NULL
	NULL	650	O80	NULL
Trauma during birth	NULL	NULL	P10	NULL
	NULL	NULL	P11	NULL
	NULL	NULL	P12	NULL
	NULL	NULL	P13	NULL
	NULL	NULL	P14	NULL
	NULL	NULL	P15	NULL
Preterm birth	NULL	6442	O60	NULL
Stillbirth	W91	7799	P95	NULL
	W93	V271	Z371	NULL
Abortion	W82	632	O01	NULL
	W83	634	O02	NULL
	NULL	633	O04	NULL
	NULL	635	NULL	NULL
	NULL	637	O07	NULL
	NULL	639	O08	NULL
Complaints female reproductive organs	X04	NULL	NULL	NULL
	X11	NULL	NULL	NULL
	X13	NULL	NULL	NULL
	X14	NULL	NULL	NULL
	X15	NULL	NULL	NULL
	X16	NULL	NULL	NULL
	X17	NULL	NULL	NULL
	X18	NULL	NULL	NULL
	X84	NULL	NULL	NULL
	X87	NULL	NULL	NULL
Complaints breasts	X18	NULL	NULL	NULL
	X19	NULL	NULL	NULL
	X20	NULL	NULL	NULL
	X88	NULL	NULL	NULL

		99.02	NULL	NULL	NULL
		99.03	NULL	NULL	NULL
		99.05	NULL	NULL	NULL
		X02	6253	N943	NULL
		X03	NULL	N944	NULL
		X09	NULL	N945	NULL
		X10	NULL	N946	NULL
	Irregular menstruation	X05	NULL	NULL	NULL
		X06	NULL	NULL	NULL
		X07	626	N91	NULL
		X08	NULL	N92	NULL
	Hysterectomy	NULL	NULL	Z90710	NULL
		NULL	NULL	NULL	NULL
		NULL	NULL	NULL	NULL
		NULL	NULL	NULL	NULL
		NULL	NULL	NULL	NULL
		NULL	NULL	NULL	NULL
		NULL	NULL	NULL	NULL
		NULL	NULL	NULL	NULL
		NULL	NULL	NULL	NULL
		NULL	NULL	NULL	NULL
	Hormonal replacement therapy	NULL	NULL	NULL	G03CA
		NULL	NULL	NULL	G03FA
		NULL	NULL	NULL	G03FB
	Intrauterine device	NULL	NULL	NULL	G02B
	Contraceptive use	W11	V259	Z920	G02B
		NULL	NULL	Z30	G03A
		NULL	NULL	NULL	G03HB
		NULL	NULL	NULL	G03AA
		NULL	NULL	NULL	G03AB
Oncology	Ongological diagnosis	A79	140	C00	NULL
		B72	141	C01	NULL
		B73	142	C02	NULL
		B74	143	C03	NULL
		D74	144	C04	NULL
		D75	145	C05	NULL
		D76	146	C06	NULL
		D77	147	C07	NULL
		D71	148	C08	NULL
		N74	149	C09	NULL
		R84	150	C10	NULL
		R85	151	C11	NULL
		S77	152	C12	NULL
		T71	153	C13	NULL
		U75	154	C14	NULL
		U76	155	C15	NULL

U77	156	C16	NULL
W72	157	C17	NULL
X75	158	C18	NULL
X76	159	C19	NULL
X77	160	C20	NULL
Y77	161	C21	NULL
Y78	162	C22	NULL
NULL	163	C23	NULL
NULL	164	C24	NULL
NULL	165	C25	NULL
NULL	166	C26	NULL
NULL	167	C30	NULL
NULL	168	C31	NULL
NULL	169	C32	NULL
NULL	170	C33	NULL
NULL	171	C34	NULL
NULL	172	C35	NULL
NULL	173	C36	NULL
NULL	174	C37	NULL
NULL	175	C38	NULL
NULL	176	C39	NULL
NULL	177	C40	NULL
NULL	178	C41	NULL
NULL	179	C43	NULL
NULL	180	C44	NULL
NULL	181	C45	NULL
NULL	182	C46	NULL
NULL	183	C47	NULL
NULL	184	C48	NULL
NULL	185	C49	NULL
NULL	186	C50	NULL
NULL	187	C51	NULL
NULL	188	C52	NULL
NULL	189	C53	NULL
NULL	190	C54	NULL
NULL	191	C55	NULL
NULL	192	C56	NULL
NULL	193	C57	NULL
NULL	194	C58	NULL
NULL	195	C60	NULL
NULL	196	C61	NULL
NULL	197	C62	NULL
NULL	198	C63	NULL
NULL	199	C64	NULL
NULL	200	C65	NULL
NULL	201	C66	NULL

		NULL	202	C67	NULL
		NULL	203	C68	NULL
		NULL	204	C69	NULL
		NULL	205	C70	NULL
		NULL	206	C71	NULL
		NULL	207	C72	NULL
		NULL	208	C73	NULL
		NULL	209	C74	NULL
		NULL	230	C75	NULL
		NULL	231	C76	NULL
		NULL	232	C77	NULL
		NULL	233	C78	NULL
		NULL	234	C79	NULL
		NULL	235	C80	NULL
		NULL	236	C7A	NULL
		NULL	237	C7B	NULL
		NULL	238	C81	NULL
		NULL	239	C82	NULL
		NULL	NULL	C83	NULL
		NULL	NULL	C84	NULL
		NULL	NULL	C85	NULL
		NULL	NULL	C86	NULL
		NULL	NULL	C87	NULL
		NULL	NULL	C88	NULL
		NULL	NULL	C89	NULL
		NULL	NULL	C90	NULL
		NULL	NULL	C91	NULL
		NULL	NULL	C92	NULL
		NULL	NULL	C93	NULL
		NULL	NULL	C94	NULL
		NULL	NULL	C95	NULL
		NULL	NULL	C96	NULL
		NULL	NULL	Z85	NULL
	Chemotherapy	NULL	NULL	NULL	L01
Other	Minor complaints in primary care	S20	NULL	NULL	NULL
		S95	NULL	NULL	NULL
		S09	NULL	NULL	NULL
		S11	NULL	NULL	NULL
		F16	NULL	NULL	NULL
		F72	NULL	NULL	NULL
		S15	NULL	NULL	NULL
		F76	NULL	NULL	NULL
		F71	NULL	NULL	NULL
		H13	NULL	NULL	NULL
		S10	NULL	NULL	NULL
		S94	NULL	NULL	NULL

		H04	NULL	NULL	NULL
	Hypothyroidism	T86	NULL	E01	H03A
		NULL	NULL	E02	H03C
		NULL	NULL	E03	NULL
	Hyperthyroidism	T85	NULL	E05	H03B
	Fever	A03	NULL	NULL	NULL
	Tiredness	A04	NULL	NULL	NULL
		A05	NULL	NULL	NULL
	Syncope	A06	NULL	NULL	NULL
	Allergy	A12	NULL	NULL	NULL
	Wish for euthanesia	A20	NULL	NULL	NULL
	Mononucleic infection	A75	NULL	NULL	NULL
	Trauma, unspecified	A80	NULL	NULL	NULL
	Adverse drug reaction	A85	NULL	NULL	NULL
	Transplanted organ and tissue status	A87.02	NULL	Z94	NULL
	Stoma	A87.01	NULL	NULL	NULL
	Death	A96	NULL	NULL	NULL
Medication use: other	Antibiotic use	NULL	NULL	NULL	J01A
		NULL	NULL	NULL	J01C
		NULL	NULL	NULL	J01D
		NULL	NULL	NULL	J01E
		NULL	NULL	NULL	J01F
		NULL	NULL	NULL	J01G
		NULL	NULL	NULL	J01M
		NULL	NULL	NULL	J01X
	Vaccine	R44	NULL	NULL	NULL
	Mineral supplements	NULL	NULL	NULL	A12A
		NULL	NULL	NULL	A12B
		NULL	NULL	NULL	A12C
	Antiprozoal drugs	NULL	NULL	NULL	P01
		NULL	NULL	NULL	J07BB
	Vitamin supplements	NULL	NULL	NULL	A11A
		NULL	NULL	NULL	A11C
		NULL	NULL	NULL	A11D
		NULL	NULL	NULL	A11E
		NULL	NULL	NULL	A11G
		NULL	NULL	NULL	A11H
		NULL	NULL	NULL	A11J
	Topical antibiotics	NULL	NULL	NULL	D06
	Topical steroid drugs	NULL	NULL	NULL	D07A
		NULL	NULL	NULL	D07B
		NULL	NULL	NULL	D07C
		NULL	NULL	NULL	D07X
	Antifungal medication	NULL	NULL	NULL	D01A
	Antipruritic medication	NULL	NULL	NULL	D04A

Antiseptic medication	NULL	NULL	NULL	D08

Table S2. Comorbidity count.

Comorbidities
Hypertension
Diabetes mellitus
Hyperlipidemia
Atrial fibrillation
Cardiac arrythmia other than atrial fribrillation
Cardiac valve disorders
Obstructive airway disease
Heart failure
Asthma
Tuberculosis
Vertigo
Deafness
Inflammatory bowel disease
Interstitial lung disease
Rheumatoid arthritis
HIV/AIDS
Sarcoidosis
Gout
Hepatitis
Osteoporosis
Arthritis
Migraine
Epilepsia
Parkinson's disease
Kidney failure
Other chronic neurological disordes**
Multiple sclerosis
Visus complaints
Dementia
Anxiety disorder
Manic disorder
Psychotic disorder
Depression
Neoplasm
Organ transplant

*The comorbidity count is a simple count of the occurrence of the above chronic conditions
**Listed inTable S1.

Table S3. Hyperparameter ranges for optimization of Cox ElasticNet and random survival forest models.

Hyperparameters	Range	
	Minimum	Maximum
Cox ElasticNet		
Alpha	0.00001	1
L1/L2 regularization ratio	0.5	1
Random Survival Forests		
Number of trees	50	200
Maximum tree depth	3	7
Minimum number of samples required to split	2	5
Minimum number of samples required to be at a leaf node	2	5

Table S4. Top 20 most important predictors for women and men separately, for age groups 30–39 and 40–49 years.

Women				Men			
30–39 years (n = 135902)		40–49 years (n = 140211)		30–39 years (n = 128660)		40–49 years (n = 137368)	
Predictor	**Coef.***	**Predictor**	**Coef.***	**Predictor**	**Coef.***	**Predictor**	**Coef.***
Age	0.215	Age	0.114	Age	0.322	Age	0.211
NSAID use	0.093	Socioeconomic status score	0.079	Socioeconomic status score	0.091	Socioeconomic status score	0.068
Socioeconomic status score	0.085	Combined oral contraceptive use	0.070	NSAID use	0.085	Smoking: current	0.061
Comorbidity count***	0.080	Smoking: current	0.059	Smoking: current	0.082	NSAID use	0.057
Betablocker use	0.055	RAAS inhibitors	0.058	Comorbidity count***	0.064	Practice nurse contact	0.047
Acetylsalicylic acid use	0.053	Acetylsalicylic acid use	0.053	RAAS inhibitors	0.047	Psoriasis	0.035
Calcium channel blockers	0.052	Gastroesophageal reflux medication	0.048	Practice nurse contact	0.037	RAAS inhibitors	0.033
Gastroesophageal reflux medication	0.051	Antibiotic use	0.040	Vitamin deficiency	0.036	Calcium channel blockers	0.024
Systemic corticosteroid use	0.040	Calcium channel blockers	0.039	Acetylsalicylic acid use	0.032	Comorbidity count***	0.023
Medication for obstructive airway disease	0.037	Depression	0.038	Gastroesophageal reflux medication	0.031	Gastroesophageal reflux medication	0.020
RAAS inhibitors	0.026	Betablocker use	0.037	Propulsive medication	0.030	Epilepsia	0.019
Vitamin status measured**	0.025	Blood pressure measured	0.036	Hyperlipidemia	0.027	Insulin use	0.019
Blood pressure measured**	0.024	Comorbidity count***	0.035	Upper airway symptoms	0.025	Antiseptic use	0.019
Smoking: current	0.023	Dermatological complaints	0.033	Diabetes mellitus	0.025	Antidiabetic medication (non-insulin)	0.017
Osteoporosis	0.022	Hyperlipidemia	0.032	Antihistaminic use	0.021	History of pneumonia	0.016
Glucose measured last year	0.017	NSAID use	0.028	Headache	0.021	Esophageal disorders	0.013
Diabetes mellitus	0.014	Influenza vaccine administered	0.027	Creatinine measured**	0.021	Syncope	0.013
Palpitations	0.011	Insulin use	0.023	Neurological sensory disease	0.020	Oral anticoagulant drugs	0.013
Hyperventilation	0.011	Incontinence	0.021	Parkinson's disease	0.019	Vitamin prescription	0.013
Infertility	0.009	Hyperventilation	0.020	Personality disorder	0.019	Creatinine measurement**	0.012

*Absolute, regularized coefficient of Cox elastic net models

**Presence of at least one measurement during the run-in period

***Comorbidity count: simple count of chronic conditions per patient, enlisted in Table S2.

Table S5. Discrimination and calibration of sex-specific prediction models for different predictor subsets.

		Women (n = 276,113) Performance metrics (95% CI)				Men (n = 266,028) Performance metrics (95% CI)			
Models	**Predictors**	**C-index**	**Δ C-stat.***	**Δ C-stat.****	**Calibration curve slope at 10 years**	**C-index**	**Δ C-stat***	**Δ C-stat.****	**Calibration curve slope at 10 years**
Cox - Proportional Hazard	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)
	10	0.671 (0.668 - 0.674)	0.023	16%	1.000 (0.996 - 1.004)	0.671 (0.669 - 0.675)	0.010	6%	1.000 (0.999 - 1.002)
	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)	0.030	20%	1.000 (0.997 - 1.002)	0.673 (0.671 - 0.675)	0.012	7%	1.001 (0.998 - 1.004)
Cox ElasticNet	Baseline	0.649 (0.646 - 0.652)	0.001	1%	1.002 (0.997 - 1.005)	0.660 (0.648 - 0.663)	-0.001	-1%	1.002 (0.999 - 1.003)
	10	0.668 (0.664 - 0.671)	0.020	14%	1.001 (0.997 - 1.002)	0.671 (0.668 - 0.674)	0.010	6%	1.001 (0.998 - 1.004)
	20	0.671 (0.668 - 0.674)	0.023	16%	1.000 (0.997 - 1.002)	0.672 (0.669 - 0.675)	0.011	7%	1.001 (0.998 - 1.004)
	50	0.675 (0.672 - 0.678)	0.027	18%	0.999 (0.998 - 1.000)	0.672 (0.669 - 0.676)	0.011	7%	0.999 (0.998 - 1.000)
Random survival forest	Baseline	0.645 (0.648 - 0.656)	-0.003	-2%	0.999 (0.998 - 1.001)	0.659 (0.656 - 0.662)	-0.002	-1%	0.998 (0.995 - 1.000)
	10	0.663 (0.660 - 0.667)	0.015	10%	0.998 (0.995 - 1.000)	0.669 (0.666 - 0.672)	0.008	5%	0.997 (0.994 - 0.999)
	20	0.667 (0.664 - 0.671)	0.019	13%	0.997 (0.994 - 0.999)	0.671 (0.669 - 0.674)	0.010	6%	0.996 (0.993 - 0.998)
	50	0.670 (0.667 - 0.673)	0.022	15%	0.993 (0.990 - 0.996)	0.672 (0.670 - 0.675)	0.011	7%	0.992 (0.989 - 0.993)

Baseline traditional cardiovascular predictors: age, hypertension, antihypertensive medication, diabetes mellitus, hyperlipidemia, with Cox PH 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