ESC
European Society
of Cardiology

Europace (2023) **00**, 1–8
https://doi.org/10.1093/europace/euac260

CLINICAL RESEARCH

Exploring the incremental utility of circulating biomarkers for robust risk prediction of incident atrial fibrillation in European cohorts using regressions and modern machine learning methods

Betül Toprak [1,2], Stephanie Brandt[1], Jan Brederecke [1], Francesco Gianfagna [3,4], Julie K.K. Vishram-Nielsen [5,6], Francisco M. Ojeda [1], Simona Costanzo [7], Christin S. Börschel[1,2], Stefan Söderberg [8], Ioannis Katsoularis[8], Stephan Camen [1,2], Erkki Vartiainen [9], Maria Benedetta Donati [7], Jukka Kontto [9], Martin Bobak [10], Ellisiv B. Mathiesen [11], Allan Linneberg [5,12], Wolfgang Koenig [13,14,15], Maja-Lisa Løchen [16], Augusto Di Castelnuovo [4], Stefan Blankenberg[1,2], Giovanni de Gaetano [7], Kari Kuulasmaa [9], Veikko Salomaa [9], Licia Iacoviello [3,7], Teemu Niiranen [9,17], Tanja Zeller[1,2,18], and Renate B. Schnabel [1,2]* on behalf of the BiomarCaRE and AFFECT-EU Consortia

[1]Department of Cardiology, University Heart and Vascular Center Hamburg, University Medical Center Hamburg-Eppendorf, Martinistraße 52, 20246 Hamburg, Germany; [2]German Centre for Cardiovascular Research (DZHK), Partner Site, Hamburg/Kiel/Luebeck, Potsdamer Straße 58, 10785 Berlin, Germany; [3]Department of Medicine and Surgery, Research Center in Epidemiology and Preventive Medicine (EPIMED), University of Insubria, Via Rossi 9, 21100 Varese, Italy; [4]Mediterranea Cardiocentro, Via Orazio 2, 80122 Napoli, Italy; [5]Center for Clinical Research and Prevention, Bispebjerg and Frederiksberg Hospital, The Capital Region of Denmark, Nordre Fasanvej 57, 2000 Frederiksberg, Denmark; [6]Department of Cardiology, Rigshospitalet, University Hospital of Copenhagen, Blegdamsvej 9, 2100 Copenhagen, Denmark; [7]Department of Epidemiology and Prevention, IRCCS Neuromed, Via dell´ Elettronica, 86077 Pozzilli, Italy; [8]Department of Public Health and Clinical Medicine, and Heart Centre, Umeå University, SE-901 87 Umeå, Sweden; [9]Department of Public Health and Welfare, Finnish Institute for Health and Welfare, POB 30, Mannerheimintie 166, 00271 Helsinki, Finland; [10]Department of Epidemiology and Public Health, University College London, 1-19 Torrington Place, London, WC1E 7HB, UK; [11]Brain and Circulation Research Group, Department of Clinical Medicine, UiT The Arctic University of Norway, Hansine Hansens veg 18, 9019 Tromsø, Norway; [12]Department of Clinical Medicine, Faculty of Health and Medical Sciences, University of Copenhagen, Blegdamsvej 3B, 2200 Copenhagen, Denmark; [13]German Heart Centre Munich, Technical University of Munich, Lazarettstraße 36, 80636 Munich, Germany; [14]German Centre for Cardiovascular Research (DZHK), Partner Site Munich Heart Alliance, Biedersteinerstraße 29, 80802 Munich, Germany; [15]Institute of Epidemiology and Medical Biometry, University of Ulm, Helmholtzstraße 22, 89081 Ulm, Germany; [16]Department of Community Medicine, UiT The Arctic University of Norway, Hansine Hansens veg 18, 9019 Tromsø, Norway; [17]Department of Medicine, Turku University Hospital and University of Turku, Kiinamyllynkatu 4-8, 20521 Turku, Finland; and [18]University Center of Cardiovascular Science, University Heart and Vascular Center Hamburg, Martinistraße 52, 20246 Hamburg, Germany

Received 16 September 2022; accepted after revision 23 November 2022

Aims To identify robust circulating predictors for incident atrial fibrillation (AF) using classical regressions and machine learning (ML) techniques within a broad spectrum of candidate variables.

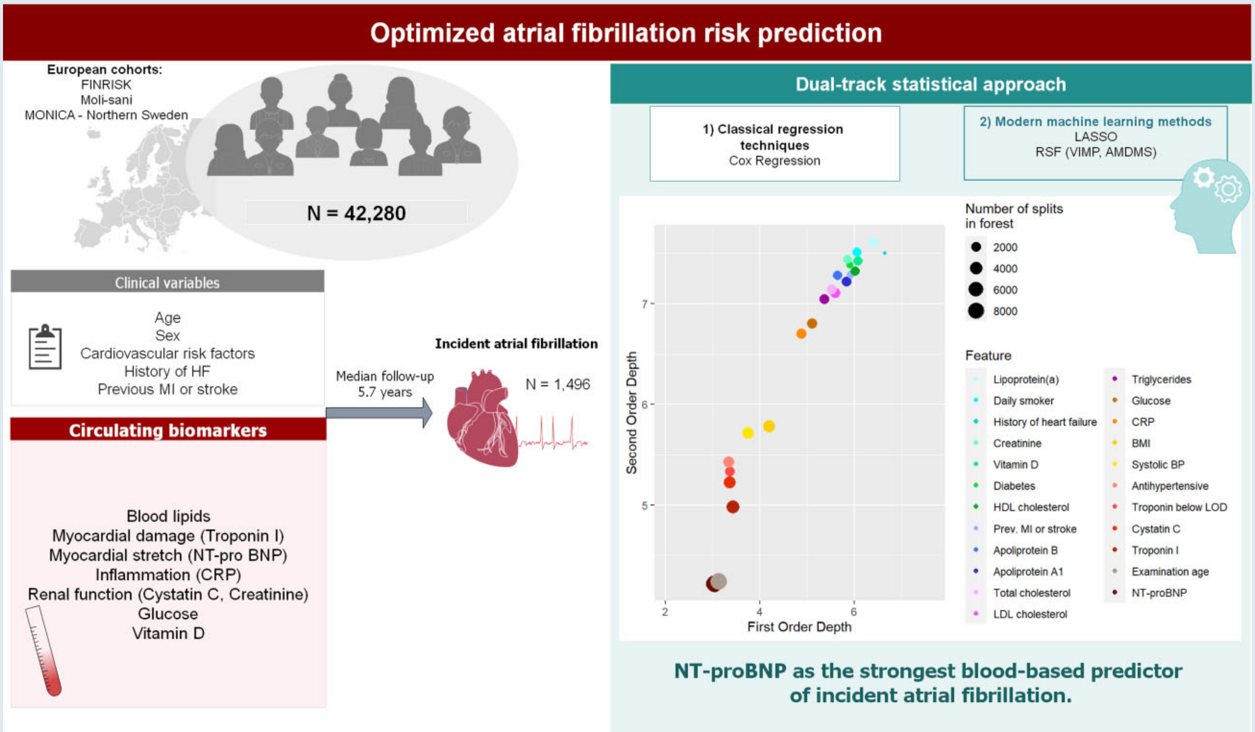
Methods and results In pooled European community cohorts ($n = 42\,280$ individuals), 14 routinely available biomarkers mirroring distinct pathophysiological pathways including lipids, inflammation, renal, and myocardium-specific markers (N-terminal pro B-type natriuretic peptide [NT-proBNP], high-sensitivity troponin I [hsTnI]) were examined in relation to incident AF using Cox regressions and distinct ML methods. Of 42 280 individuals (21 843 women [51.7%]; median [interquartile range, IQR] age, 52.2 [42.7, 62.0] years), 1496 (3.5%) developed AF during a median follow-up time of 5.7 years. In multivariable-adjusted Cox-regression analysis, NT-proBNP was the strongest circulating predictor of incident AF [hazard ratio (HR) per standard deviation (SD), 1.93 (95% CI,

* Corresponding author at: University Heart and Vascular Center Hamburg, University Medical Center Hamburg-Eppendorf, Martinistraße 52, 20246 Hamburg, Germany. Tel: +49 40 7410 53979; fax: +49 40 7410 53622. *E-mail address:* r.schnabel@uke.de

1.82–2.04); $P < 0.001$]. Further, hsTnI [HR per SD, 1.18 (95% CI, 1.13–1.22); $P < 0.001$], cystatin C [HR per SD, 1.16 (95% CI, 1.10–1.23); $P < 0.001$], and C-reactive protein [HR per SD, 1.08 (95% CI, 1.02–1.14); $P = 0.012$] correlated positively with incident AF. Applying various ML techniques, a high inter-method consistency of selected candidate variables was observed. NT-proBNP was identified as the blood-based marker with the highest predictive value for incident AF. Relevant clinical predictors were age, the use of antihypertensive medication, and body mass index.

Conclusion Using different variable selection procedures including ML methods, NT-proBNP consistently remained the strongest blood-based predictor of incident AF and ranked before classical cardiovascular risk factors. The clinical benefit of these findings for identifying at-risk individuals for targeted AF screening needs to be elucidated and tested prospectively.

Graphical Abstract



AMDMS, Averaged minimal depth of a maximal subtree; BMI, body mass index; BP, blood pressure; CRP, C-reactive protein; HDL, high-density lipoprotein; HF, heart failure; LASSO, Least absolute shrinkage and selection operator; LDL, low-density lipoprotein; LOD, limit of detection; MI, myocardial infarction; NT-proBNP, N-terminal pro B-type natriuretic peptide; RSF, Random survival forest; VIMP, variable importance.

Keywords Atrial fibrillation • Biomarkers • Risk Prediction • Machine learning • Epidemiology • Community

What's new?

- Elaborated statistical approach combining classical regressions and distinct modern machine learning (ML) methods to identify robust blood-based predictors of incident atrial fibrillation (AF).
- Using different variable selection methods, N-terminal pro B-type natriuretic peptide (NT-proBNP) remained the strongest blood-based predictor of incident AF across a broad spectrum of routinely available candidate biomarkers in a large population-based European cohort including 42 280 individuals.
- Biomarker-enriched, multivariable risk prediction models may offer great potential to further improve risk stratification for targeted AF screening.

Introduction

Atrial fibrillation (AF) is highly prevalent in ageing populations and associated with substantial cardiovascular morbidity and mortality,[1] which renders this disease a major health issue in Europe and worldwide. Risk prediction has become a cornerstone of epidemiological research to identify individuals at risk of developing AF as well as subsequent complications, including stroke and heart failure (HF), and to guide population-wide screening and prevention. Clinical risk indicators such as obesity, physical inactivity, hypertension, alcohol use, and prevalent cardiovascular disease have been reported to be strongly related to new-onset AF. However, they merely account for 50% of the population-attributable risk of AF.[2]

Circulating biomarkers serve as objective, quantitative measures of manifest or even subclinical pathophysiological conditions related to AF. Owing to the multifactorial nature of AF, prior community-based studies examined candidate biomarkers that reflect the blood-based fingerprint of major pathways in the pathogenesis of AF, i.e. neurohumoral stress, cardiomyocyte damage, cardiac remodeling, inflammation, oxidative stress, and renal function. N-terminal pro B-type natriuretic peptide (NT-proBNP) reflecting myocardial stretch and C-reactive protein (CRP) as an indicator of inflammatory activity consistently remained the strongest predictors of incident AF in multivariable-adjusted models.[3]

Overall, most of these analyses include the measurement of only a single or a small selection of blood biomarkers, or incorporate emerging biomarkers with limited actual clinical application, and are mostly based on traditional regression models as part of a hypothesis-driven approach. In contrast, machine learning (ML) methods that are data-driven and do not necessarily need a pre-specified model structure, but instead build an optimal model from the information available in the data, have emerged as a valuable tool in cardiovascular risk prediction.[4]

We therefore investigated the incremental predictive value of 14 established and routinely available blood biomarkers in relation to incident AF in a head-to-head comparison in three large community-based cohorts across Europe by applying a dual-track analytical approach including (i) conventional regression-based and (ii) modern ML-based models for optimized and robust risk prediction of incident AF.

Methods

Study population

The present study includes three community-based cohorts (FINRISK 1997, Moli-sani, and Northern Sweden MONICA) from the Monica Risk, Genetics, Archiving and Monograph (MORGAM) (https://www.thl.fi/morgam/)/Biomarker for Cardiovascular Risk Assessment across Europe (BiomarCaRE) (http://www.biomarcare.eu) projects with available information on AF status at baseline and follow-up. From the original cohorts comprising a total of $n = 43\,219$ individuals with performed examinations between 1986 and 2010, we excluded $n = 492$ individuals with self-reported and/or physician-diagnosed history of AF/atrial flutter and/or prior ICD-8-9- or -10 coding for AF/atrial flutter (FINRISK 1997, $n = 82$; Moli-sani, $n = 313$; Northern Sweden MONICA, $n = 97$), as well as $n =$ 447 individuals with missing follow-up information on incident AF (all from Moli-sani), leaving a final study population of $n = 42\,280$ participants free of AF/atrial flutter at baseline for the analyses. Full details on the enrollment and follow-up procedures of each included cohort are available in the *Supplementary material online.*

The present study complies with the Declaration of Helsinki. Local ethics committees have approved all participating studies. Written informed consent was provided by participants. The authors had full access to the data and take responsibility for its integrity.

Risk factors and follow-up

For each cohort, the following variables were available from baseline visits: age, sex, body mass index (BMI), systolic and diastolic blood pressure, antihypertensive medication, daily smoking, diabetes mellitus, history of HF, myocardial infarction (MI), and stroke. Information on cholesterol-lowering medication were available in all cohorts except for the first among the Northern Sweden MONICA cohorts, which began to recruit participants in 1986.

The diagnosis of AF during follow-up was based on questionnaire information, national hospital discharge registry data including data on ambulatory visits to hospitals, and available information on concomitant incident AF as extracted from causes of death registry data. The last follow-up was performed between 2009 and 2010. Details on the follow-up period per cohort are provided in the Supplementary material online, *Table S3*.

Biomarker quantification

In total, 14 different biomarkers were quantified from stored blood samples: myocardium-specific markers [NT-proBNP, high-sensitivity troponin I (hsTnI)], lipids [total cholesterol, low-density lipoprotein (LDL), high-density lipoprotein (HDL), triglycerides, lipoprotein(a) [Lp(a)], apolipoprotein A1 (Apo A1) and B (Apo B)], renal parameters [creatinine, cystatin C], CRP, glucose, and vitamin D. Biomarker measurements were available in a large proportion of the study population; the numbers of missing values per biomarker and cohort are shown in the Supplementary material online, *Table S2*. In $n = 40\,645$ individuals, NT-proBNP was measured on the ELECSYS 2010 platform using an electrochemiluminescence immunoassay (Roche Diagnostics), with an analytical range given as 5–35 000 ng/L, and intra- and inter-assay coefficients of variation of 1.38% and 2.58%, respectively. Measurement details of all other biomarkers are provided in the *Supplementary material online.*

Statistical analysis

Continuous variables are presented as median (25th, 75th percentile) and binary variables as absolute and relative frequencies. Missing values were handled by multiple imputation using the method of chained equations.[5] If exact biomarker values were missing because they were below (above) the detection level, they were set to the minimum (maximum) value of the respective assay for all analyses. A total of 10 imputed data sets was produced and results were pooled following Rubin's rules.[6]

To achieve a more normal distribution, values of some of the blood-based biomarkers were log-transformed [NT-proBNP, hsTnI, triglycerides, Lp(a), creatinine, CRP, glucose, vitamin D], only for cystatin C power-transformation with a lambda of 0.375 was applied. For hsTnI, a bivariate approach was used.[7] Values below the limit of detection (LOD) of 1.9 ng/L were set to the minimum detectable value (1.9 ng/L), and a respective indicator variable was included in all hsTnI-related models. The continuous hsTnI values, either unchanged or those set to 1.9 ng/L, were also log-transformed. Before investigating their association with incident AF, we used Spearman correlations to examine whether the circulating biomarkers correlate with each other. As Spearman correlation only uses ranks, the indicator variable for hsTnI was excluded and values that were originally below the LOD of 1.9 ng/L were set to 1.8 ng/L.

Cox regressions were performed to examine the association of each biomarker and the occurrence of new-onset AF during follow-up. A Cox model was computed for each biomarker and adjusted for the following clinical covariates: BMI, systolic blood pressure, antihypertensive medication, daily smoking, diabetes mellitus, history of HF, previous MI, or stroke. Sex and cohort served as stratification variables. Age was used as the time scale. The association of AF with each of the investigated biomarkers was quantified by hazard ratio (HR) per one standard deviation (SD) increase with the respective 95% confidence interval (95% CI). Multivariable-adjusted Cox-regression analyses are also provided separately for each of the cohorts. As a sensitivity analysis, a random marker coefficient was added to account for possible heterogeneity in the different cohorts with regard to the association of the biomarkers with time-to-AF. As supplementary and proof-of-concept analyses, we explored the potential additive utility of investigated biomarkers by C-index with 10-fold cross-validation and net reclassification improvement (NRI) using the Kaplan–Meier method as suggested by Pencina *et al.*[8] A two-sided *P*-value of <0.05 was considered statistically significant. All statistical analyses were performed using R statistical software version 4.0.5 (http://www.R-project.org).

Machine learning

ML methods address analytical challenges including non-linearity, heterogeneous interactions, and the handling of a large amount of candidate predictors, which hamper the real-world value of traditional regression-based models.[4] We used different ML techniques, those that amend traditional regression models through their variable selection utility (least absolute shrinkage and selection operator, LASSO), as well as tree-based methods (random survival forest, RSF), in order to detect robust predictors of incident AF as identified by different variable selection procedures. These ML methods were applied on one of the imputed datasets. All investigated blood-based biomarkers and clinical risk factors were included in the models. In LASSO, age was used as a time scale and candidate predictors were ranked by their respective HR per SD. In contrast, age was included as a candidate predictor in RSF models given the nature of the implementation used for the RSF method. Where applicable, five-fold cross-validation was performed in the model-building process.

For RSF, the used split-criterion was maximally selected rank statistics as suggested by Wright et al.[9] Parameters were tuned using a grid search and out-of-bag errors of the models, which were applied to find the optimal RSF model. Tuned parameters were the number of predictors that were randomly sampled at each split and the minimum number of data points per node. The sample fraction was 0.632, representing the amount of data that is used in each tree. A total of 1000 trees were grown in each model. Variables were then ranked by their variable importance (VIMP) in the final RSF model.

As feature strength is difficult to assess with the traditional method of VIMP in RSF, we additionally performed another variable selection method for tree-based models, the minimal depth of a maximal subtree averaged over forest (AMDMS)—analysis to further examine the predictive ability of a variable in the random forest model.[10] This metric was visualized by plots displaying the first- and second-order depth of all clinical variables and biomarkers.

Results

Baseline characteristics

The characteristics and biomarker levels of 42 280 individuals free of AF at baseline are displayed in *Table 1*, and of each individual cohort in the Supplementary material online, *Table S1*. The median age of study participants was 52.2 years (age range 24–97 years at baseline), 48.3% were men. At baseline, 20.3% of the total cohort were daily smokers, 5.7% had a known diabetes, 1.1% a history of HF, and 4% a previous MI or stroke. The median values for NT-proBNP and hsTnI were 44.1 and 2.3 ng/L, respectively.

During the follow-up period comprising a median of 5.7 years, $n = 1496$ cases (3.5%) of incident AF were documented. Follow-up information by cohort are provided in the Supplementary material online, *Table S3*.

Correlations of circulating biomarkers

Spearman correlations of all biomarkers are provided in the Supplementary material online, *Figure S1*. Most blood lipids correlated positively with each other. NT-proBNP correlated moderately only with cystatin C ($\rho = 0.27$).

Association of different biomarkers and incident AF using classical statistics

After adjustment for clinical covariates, NT-proBNP showed the strongest association (HR per SD 1.93, 95% CI 1.82–2.04, $P < 0.001$) with the occurrence of new-onset AF. HsTnI (HR per SD 1.18, 95% CI 1.13–1.22, $P < 0.001$), cystatin C (HR per SD 1.16, 95% CI 1.10–1.23, $P < 0.001$), and CRP (HR per SD 1.08, 95% CI 1.02–1.14, $P = 0.012$) were also significantly related to incident AF. In addition, multivariable-adjusted Cox-regression analyses revealed lower levels of total cholesterol (HR per SD 0.90, 95% CI 0.85–0.95, $P < 0.001$) and LDL cholesterol (HR per SD 0.93, 95% CI 0.88–0.98, $P = 0.007$), as well as of Apo B (HR per SD 0.92, 95% CI 0.88–0.98, $P = 0.005$), and triglycerides (HR per SD 0.92, 95% CI 0.87–0.98, $P = 0.009$) to be associated with incident AF (*Figure 1*). There was some heterogeneity in associations across the cohorts (see Supplementary material online, *Table S4*). The pooled associations did not change markedly after accounting for heterogeneity between the cohorts (see Supplementary material online, *Figure S2*).

Moreover, NT-proBNP yielded the highest discriminative ability and reclassification improvement for incident AF when compared to a clinical reference model (see Supplementary material online, *Table S5*).

Table 1 Baseline characteristics of the study population

Variables	Total cohort ($n = 42\,280$)
Men, no. (%)	20 437 (48.3)
Age at baseline (years)	52.2 (42.7, 62.0)
BMI (kg/m^2)	26.9 (24.1, 30.2)
Systolic blood pressure (mmHg)	134.5 (122.0, 150.0)
Diastolic blood pressure (mmHg)	81.0 (75.0, 88.0)
Diabetes mellitus, no. (%)	2399 (5.7)
Daily smoking, no. (%)	8582 (20.3)
Antihypertensive medication, no. (%)	9258 (21.9)
Cholesterol-lowering medication, no. (%)	2705 (6.7)
History of HF, no. (%)	452 (1.1)
Previous MI or stroke, no. (%)	1678 (4.0)
Circulating biomarkers	
NT-proBNP (ng/L)	44.1 (21.1, 84.7)
Troponin I (ng/L)	2.3 (1.9, 3.6)
Troponin I below LOD, no. (%)	16 046 (38.0)
Total cholesterol (mmol/L)	5.5 (4.8, 6.3)
LDL cholesterol (mmol/L)	3.4 (2.8, 4.1)
HDL cholesterol (mmol/L)	1.4 (1.2, 1.7)
Triglycerides (mmol/L)	1.2 (0.9, 1.7)
Lipoprotein(a) (mg/dL)	9.0 (3.9, 20.4)
Apolipoprotein A1 (g/L)	1.5 (1.3, 1.7)
Apolipoprotein B (g/L)	1.0 (0.8, 1.2)
Creatinine (mg/dL)	0.8 (0.7, 0.9)
Cystatin C (mg/L)	0.9 (0.8, 1.0)
CRP (mg/L)	1.3 (0.6, 2.9)
Glucose (mmol/L)	5.0 (4.6, 5.5)
Vitamin D (ng/mL)	16.6 (11.8, 22.7)

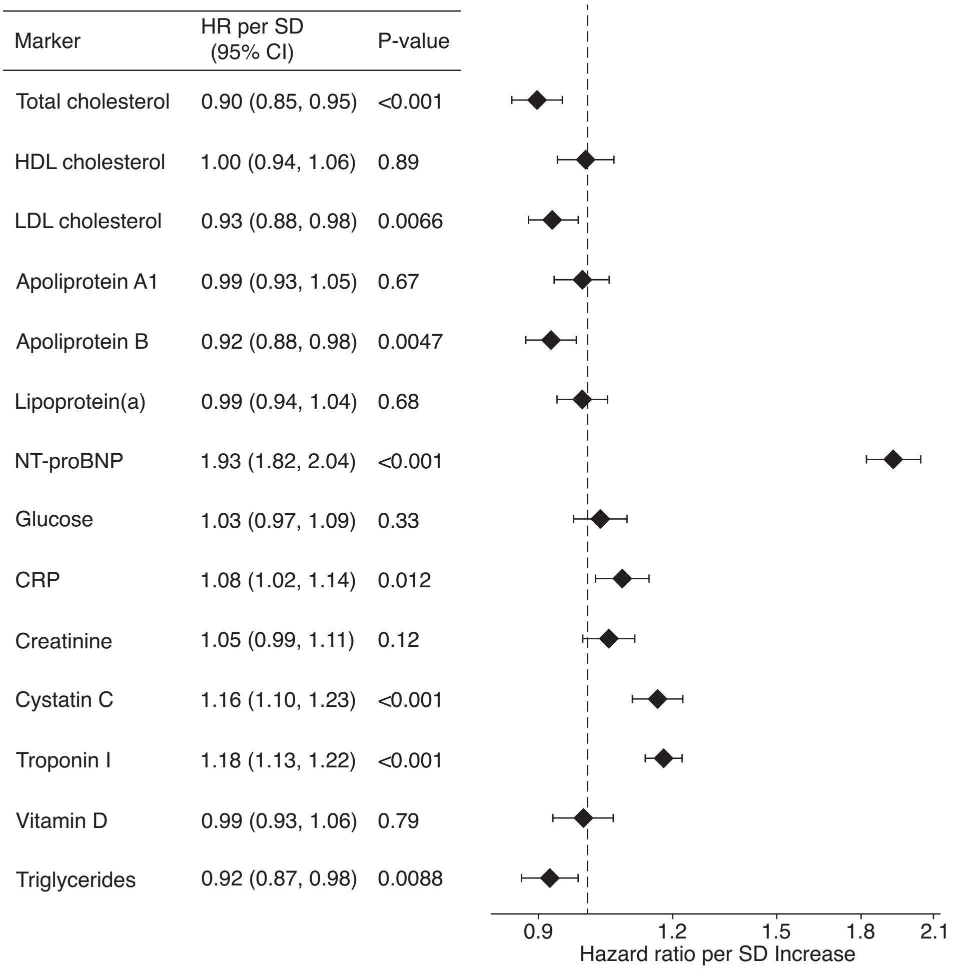
Baseline characteristics are presented as absolute and relative frequencies for categorical variables and as median values (25th, 75th percentile) for continuous variables. Troponin I was measured by a high-sensitivity assay (LOD 1.9 ng/L). BMI, body mass index; CRP, C-reactive protein; HF, heart failure; HDL, high-density lipoprotein; LDL, low-density lipoprotein; LOD, limit of detection; MI, myocardial infarction; NT-proBNP, N-terminal pro B-type natriuretic peptide.

Predictive ability of different biomarkers and clinical covariates for incident AF using machine learning

A head-to-head ranking of the 10 most important clinical and laboratory variables according to each of the ML methods applied is presented in *Table 2*. Ranking lists overlapped between LASSO and the two RSF models (VIMP/AMDMS; *Table 2*).

In the LASSO model, NT-proBNP remained the strongest predictor of incident AF (HR per SD 1.80), followed by BMI (HR per SD 1.24) and antihypertensive medication (HR per SD 1.10; Supplementary material online, *Figure S3A*). Similarly, age (VIMP 0.08) and NT-proBNP (VIMP 0.05) ranked highest among all candidate predictors when RSF was applied (see Supplementary material online, *Figure S3B*).

In AMDMS analysis, NT-proBNP yielded a stronger feature strength than age (*Figure 2*). In both tree-based models, age, antihypertensive medication and BMI emerged as the most important clinical variables. HsTnI was the second most predictive blood-based marker for incident AF in all three ML models (*Table 2*).



Marker	HR per SD (95% CI)	P-value
Total cholesterol	0.90 (0.85, 0.95)	<0.001
HDL cholesterol	1.00 (0.94, 1.06)	0.89
LDL cholesterol	0.93 (0.88, 0.98)	0.0066
Apoliprotein A1	0.99 (0.93, 1.05)	0.67
Apoliprotein B	0.92 (0.88, 0.98)	0.0047
Lipoprotein(a)	0.99 (0.94, 1.04)	0.68
NT-proBNP	1.93 (1.82, 2.04)	<0.001
Glucose	1.03 (0.97, 1.09)	0.33
CRP	1.08 (1.02, 1.14)	0.012
Creatinine	1.05 (0.99, 1.11)	0.12
Cystatin C	1.16 (1.10, 1.23)	<0.001
Troponin I	1.18 (1.13, 1.22)	<0.001
Vitamin D	0.99 (0.93, 1.06)	0.79
Triglycerides	0.92 (0.87, 0.98)	0.0088

Figure 1 Multivariable-adjusted Cox regressions. Association of 14 circulating biomarkers and incident atrial fibrillation. Provided are hazard ratios (HR) per standard deviation (SD) increase and 95% confidence intervals (CI) per model. All models are adjusted for body mass index, systolic blood pressure, antihypertensive medication, daily smoking, diabetes, history of heart failure, previous myocardial infarction or stroke, cohort, age, and sex. Troponin I was measured by a high-sensitivity assay [limit of detection (LOD) 1.9 ng/L]. For the models including Troponin I, the prediction was performed using both, the continuous value and indicator variable. Only the continuous value for Troponin I is presented here. CRP, C-reactive protein; HDL, high-density lipoprotein; LDL, low-density lipoprotein; NT-proBNP, N-terminal pro B-type natriuretic peptide.

Discussion

We investigated the predictive value of 14 different candidate biomarkers for incident AF by different variable selection procedures in three prospective European community-based cohorts. The cardiac biomarkers NT-proBNP and hsTnI, the renal function marker cystatin C, and CRP as an indicator of inflammation were significantly related to incident AF in multivariable-adjusted Cox-regression models. Among these biomarkers, regression-based models revealed NT-proBNP as the strongest blood-based predictor of incident AF. Applying several ML techniques, we observed a high inter-method consistency of selected variables with NT-proBNP being identified as the blood-based marker with the highest predictive value for incident AF. Among all candidate predictors including clinical variables in ML-based analyses, NT-proBNP and age were the two most important variables for the prediction of AF risk.

The application of artificial intelligence (AI) for cardiovascular risk prediction has gained increasing attention during recent years. In AF research, there has been a significant increase of AI use with a special focus on AF detection, e.g. by electrocardiographic screening, wearable devices. Less efforts were made for predicting incident AF, although the early identification of at-risk individuals may potentially prevent future AF-related sequelae.[11] In the era of precision medicine, ML-based algorithms are able to handle multifaceted datasets with a large number of potential predictor variables rendering them advantageous over traditional regression methods.[4] Circulating biomarkers, in particular, bear the potential of unraveling the complex pathophysiology of AF development and thus, could enhance risk prediction by mirroring the individual's susceptibility for AF. However, ML-based analyses of proteomics for risk prediction of new-onset AF are scarce. In our study, we explored a wide range of established and novel biomarkers for the prediction of incident AF using multiple classical and ML-based approaches in a large population-based cohort of over 42 000 individuals across Europe. Natriuretic peptides including B-type natriuretic peptide, its precursor fragment NT-proBNP, and atrial natriuretic peptide, have been consistently reported to be related to prevalent as well as

Table 2 Ranking of the 10 most important clinical variables and biomarkers by (A) regularized Cox regression (Least absolute shrinkage and selection operator; LASSO), (B) Random Survival Forest (RSF): (1) Variable Importance (VIMP) and (2) Averaged minimal depth of a maximal subtree (AMDMS)

Variable	**Regularized Cox regression**	**RSF**	
	LASSO[a]	**VIMP**	**AMDMS**
1	**NT-proBNP**	**Age at baseline**	**NT-proBNP**
2	**BMI**	**NT-proBNP**	**Age at baseline**
3	**Antihypertensive medication**	Troponin I below LOD	**Troponin I**
4	History of HF	**Troponin I**	Cystatin C
5	**Troponin I**	**Antihypertensive medication**	Troponin I below LOD
6	Daily smoker	Cystatin C	**Antihypertensive medication**
7	HDL cholesterol	Systolic blood pressure	Systolic blood pressure
8	Apolipoprotein A1	**BMI**	**BMI**
9	Apolipoprotein B	CRP	CRP
10	Lipoprotein(a)	Glucose	Glucose

Provided is the variable ranking according to three selection models. Items in the first set of the 10 most important variables, which overlap between the models, are marked in bold.
[a]In the LASSO model, age was used as a time scale and variables were ranked by their respective hazard ratios (HR) per standard deviation (SD). Troponin I was measured by a high-sensitivity assay (LOD 1.9 ng/L). BMI, body mass index; CRP, C-reactive protein; HDL, high-density lipoprotein; HF, heart failure; LOD, limit of detection; NT-proBNP, N-terminal pro B-type natriuretic peptide.

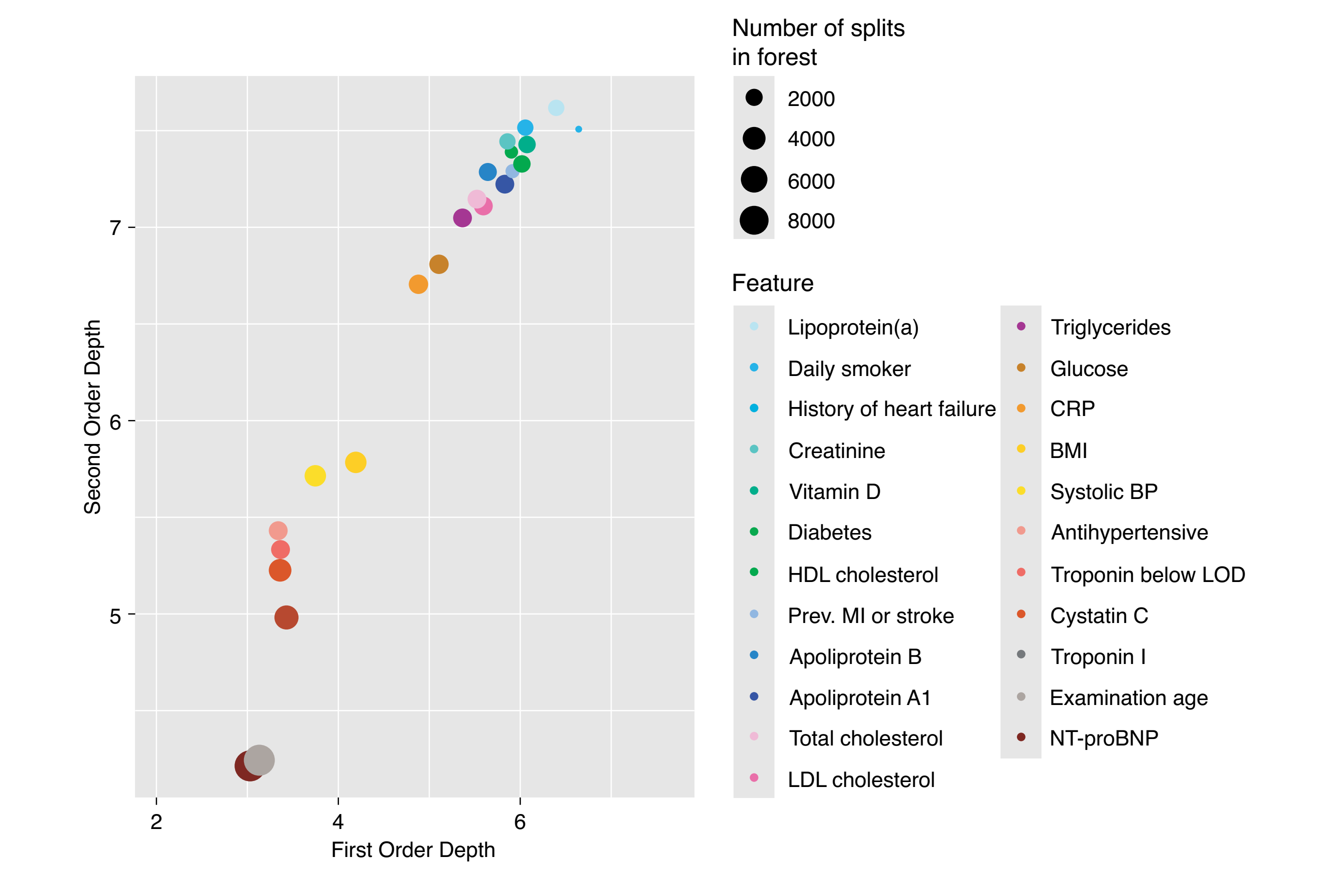


Figure 2 Predictive ability of clinical variables and biomarkers for incident atrial fibrillation using minimal depth of a maximal subtree averaged over forest (AMDMS) analysis. Provided are the first- and second-order depths and number of splits in forest per variable. Troponin I was measured by a high-sensitivity assay (LOD 1.9 ng/L). BMI, body mass index; BP, blood pressure; CRP, C-reactive protein; HDL, high-density lipoprotein; LDL, low-density lipoprotein; LOD, limit of detection; MI, myocardial infarction; NT-proBNP, N-terminal pro B-type natriuretic peptide.

incident AF.[3] They are largely cardiac-specific and good indicators of myocardial stress, even in case of subclinical pathology. However, natriuretic peptides are not specific for atrial myopathy, which is known to be the pathophysiological substrate and hallmark of AF. Instead, they are also elevated in other cardiac conditions and may signal cardiac affection and thus, a certain susceptibility to AF. Our results prove the robustness of NT-proBNP as a predictor of incident AF at multiple levels—in AMDMS analysis even ranking before age. In the setting of AF screening, the STROKESTOP II already provided evidence for NT-proBNP-stratified systematic screening of a population-based cohort including 75-/76-year-old individuals in the Stockholm region with an increased AF detection yield by almost 20-fold using a single-time point electrocardiogram.[12] Future research will show, whether the application of NT-proBNP-guided and refined AF screening will translate into a reduction of clinical endpoints and AF-related complications.

In our analyses, hsTnI was shown to be another biomarker related to incident AF. Troponin is a cardiac-specific structural protein and an indicator of myocardial injury. Albeit being predictive for cardiovascular disease and death in the general population, only modest additive value was observed for the prediction of new-onset AF in prior analyses.[13] In our cohort, we were able to confirm the previously observed, rather moderate association of hsTnI with AF.

We also investigated blood lipids as candidate markers of incident AF as they are modifiable and may be efficiently targeted by lipid-lowering medication. Our data demonstrated total cholesterol driven by the LDL fraction, Apo B, and triglycerides to be inversely related with incident AF. The inverse association of blood lipids is counterintuitive as dyslipidemia is a well-established risk factor of most other cardiovascular diseases. However, our observations were comparable to previous investigations of the relation between blood lipids and incident AF. The underlying pathophysiology has largely remained unclear to date. A potential mechanism has been attributed to a membrane-stabilizing effect of cholesterol on atrial excitability by the modification of ion channel density and function.[14,15]

Currently, a systematic screening of AF at population level has not been implemented broadly, inter alia, due to cost-effectiveness issues.[16] A personalized approach by estimating an individual's AF risk based on promising multivariable models with consecutive assignment of individuals to different risk strata with targeted screening regimes may constitute a viable way. Our findings with remarkable inter-method consistency with regard to the ranking of clinical and circulating candidate variables thus may lay the ground for future efforts to integrate clinical and omics data for efficiently selecting at-risk individuals for primary AF screening.

Limitations and strengths

Our study is limited by the method of AF ascertainment because intermittent, often oligo- or asymptomatic AF episodes may have been missed. Thus, misclassification of true AF cases as non-AF may have attenuated the observed associations. Furthermore, our sample comprised individuals of European descent. Biomarker concentrations, e.g. natriuretic peptides, have been shown to be higher in African-Americans. However, the predictive ability has been comparable in different ethnicities.[17] Further, serial measurements of biomarkers were not available, although concentrations may be variable over time.

Strengths of the present study include the large number of individuals with baseline information and adjudicated AF data, as well as the measurement of 14 different and well-established circulating markers for a comprehensive head-to-head comparison. The application of a dual approach including both, classical regression-based and several modern ML analyses, enabled for the identification of robust markers for incident AF and provided an internal validation of our results.

Emerging biomarkers were not considered as candidate predictors because they are not routinely and universally available and might have hampered the translation of our results into clinical settings. Despite the excitement about their ability to predict incident AF, most of these markers neither have been externally nor prospectively validated, sufficiently tested in clinical practice, nor compared to existing biomarkers such as the natriuretic peptides. Given their broad availability and routine measurement, biomarkers such as NT-proBNP alone or in combination with clinical covariates permit real-world applicability for AF risk prediction and are more likely to become the cornerstone of targeted screening in the near future.[3] In addition, our results may provide a benchmark for future studies assessing upcoming biomarker candidates for AF from targeted approaches and omics technologies.

Conclusions

Using regression-based models and modern ML techniques to amend classical statistics, NT-proBNP was consistently shown to be the strongest blood-based marker for incident AF across a broad range of routinely available biomarkers in European population-based cohorts. Major clinical predictors included age, antihypertensive medication and BMI. Multivariable, AI-based risk prediction models integrating such robust predictors may further optimize and personalize risk stratification and AF screening in the future.

Supplementary material

Supplementary material is available at *Europace* online.

Acknowledgements

We thank the participants and staff of the cohorts for their continuing dedication and efforts.

Funding

R.B.S. has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme under the grant agreement No 648131, from the European Union's Horizon 2020 research and innovation programme under the grant agreement no. 847770 (AFFECT-EU) and German Center for Cardiovascular Research (DZHK e.V.) (81Z1710103 and 81Z0710114); German Ministry of Research and Education (BMBF 01ZX1408A) and ERACoSysMed3 (031L0239). The MORGAM Project has received funding from EU projects MORGAM (Biomed, BMH4-CT98-3183), Genom EUtwin (FP5, QLG2-CT-2002-01254), ENGAGE (FP7, HEALTH -F4-2007-201413), CHANCES (FP7, HEALTH-F3-2010-242244), BiomarCaRE (FP7, HEALTH-F2-2011-278913), euCanSHare (Horizon 2020, No. 825903) and AFFECT-EU (Horizon 2020, No. 847770); and Medical Research Council, London (G0601463, No. 80983: Biomarkers in the MORGAM Populations). This has supported central coordination, workshops and part of the activities of the MORGAM Data Centre, the MORGAM Laboratories and the MORGAM Participating Centres. The FINRISK surveys were mainly funded by budgetary funds of the National Institute for Health and Welfare, Finland (THL). Additional funding has been obtained from numerous non-profit foundations. V.S. has been supported by the Finnish Foundation for Cardiovascular Research. T.N. has been funded by the Finnish Foundation for Cardiovascular Research, the Emil Aaltonen Foundation and the Academy of Finland [grant number 321351]. K.K. received funding from the European Union. The Northern Sweden MONICA study was mainly funded by the Norr- and Västerbotten county councils. S.S. has been supported by the Västerbotten county council (ALF) and by the Swedish Heart and Lung Foundation. G.d.G and S.Co. were supported by a grant from the Ministero della Salute, Italy. Bando Ricerca Finalizzata 2018 [RF-2018-12367074].

Conflict of interest: R.B.S. has received lecture fees and advisory board fees from BMS/Pfizer outside the submitted work. B.T. receives project-related funding by the German Heart Foundation and Ernst und Berta Grimmke-Stiftung unrelated to the submitted work. V.S. reports an honorarium for consulting from Sanofi. He also has research collaboration with Bayer Ltd (all unrelated to the present study). T.N. reports speaker honoraria from Servier outside the submitted work. S.S. received honoraria from Actelion (now Janssen) outside the submitted work. Furthermore, S.S. is a member of the advisory board of Actelion (now Janssen) and of Novartis. I.K. has received research grants by the Heart Foundation in Northern Sweden and Arnerska Research Foundation, both unrelated to the present study. S.B. reports honoraria to his institution from Abbott, Bayer, Siemens, Singulex, and Roche, outside the submitted work. S.B. also reports personal fees from Abbott, Bayer, Siemens, Novartis, Thermo Fisher, Astra Zeneca, AMGEN, Medtronic, and Pfizer. S.Co. has received honoraria for lecturing from The Dutch Beer Institute foundation—The Brewers of Europe, outside the submitted work. S.Co. and A.D.C. are the PI and Co-PI, respectively, of a research grant from ERAB (the European Foundation for Alcohol Research, EA 1767, completed in January 2020), which is unrelated to the present study. G.d.G. is a member of the International Scientific Forum on Alcohol Research. The members of the Forum receive no financial support. The Forum itself receives no support from any other organization or company in the alcoholic beverage industry. G.d.G. was also a consultant to the Web Newsletter of Assobirra, the Italian Association of the Beer and Malt Industries and is a corresponding member of the non-profit Accademia Italiana della Vite e del Vino. G.d.G. reports personal fees for given lectures at the 8th European Beer and Health Symposium (2017), Beer and Health Initiative (The Dutch Beer Institute Foundation—The Brewers of Europe), outside the submitted work. All remaining authors have declared no conflicts of interest.

Data availability

The data underlying this article will be shared on reasonable request to the corresponding author.

References

1. Schnabel RB, Yin X, Gona P, Larson MG, Beiser AS, McManus DD *et al.* 50 Year trends in atrial fibrillation prevalence, incidence, risk factors, and mortality in the Framingham Heart Study: a cohort study. *Lancet* 2015;**386**:154–62.
2. Börschel CS, Ohlrogge AH, Geelhoed B, Niiranen T, Havulinna AS, Palosaari T *et al.* Risk prediction of atrial fibrillation in the community combining biomarkers and genetics. *Europace* 2021;**23**:674–81.
3. Sinner MF, Stepas KA, Moser CB, Krijthe BP, Aspelund T, Sotoodehnia N *et al.* B-type natriuretic peptide and C-reactive protein in the prediction of atrial fibrillation risk: the CHARGE-AF consortium of community-based cohort studies. *Europace* 2014;**16**: 1426–33.
4. Goldstein BA, Navar AM, Carter RE. Moving beyond regression techniques in cardiovascular risk prediction: applying machine learning to address analytic challenges. *Eur Heart J* 2017;**38**:1805–14.
5. Buuren S, Groothuis-Oudshoorn K. Mice: multivariate imputation by chained equations in R. *J Stat Softw* 2011;**45**.
6. Rubin DB. Inference and missing data. *Biometrika* 1976;**63**:581–92.
7. Royston P, Sauerbrei W. *Multivariable Model-Building: A Pragmatic Approach to Regression Anaylsis Based on Fractional Polynomials for Modelling Continuous Variables*. Hoboken, NJ: John Wiley & Sons; 2008;Vol. 777.
8. Pencina MJ, D'Agostino RB Sr., Steyerberg EW. Extensions of net reclassification improvement calculations to measure usefulness of new biomarkers. *Stat Med* 2011;**30**: 11–21.
9. Wright MN, Dankowski T, Ziegler A. Unbiased split variable selection for random survival forests using maximally selected rank statistics. *Stat Med* 2017;**36**:1272–84.
10. Hemant I, Kogalur UB, Gorodeski EZ, Minn AJ, Lauer MS. High-dimensional variable selection for survival data. *J Am Stat Assoc* 2010;**105**:205–17.
11. Schnabel RB, Marinelli EA, Arbelo E, Boriani G, Boveda S, Buckley CM *et al.* Early diagnosis and better rhythm management to improve outcomes in patients with atrial fibrillation: the 8th AFNET/EHRA consensus conference. *Europace* 2022.
12. Kemp Gudmundsdottir K, Fredriksson T, Svennberg E, Al-Khalili F, Friberg L, Frykman V *et al.* Stepwise mass screening for atrial fibrillation using N-terminal B-type natriuretic peptide: the STROKESTOP II study. *Europace* 2020;**22**:24–32.
13. Rienstra M, Yin X, Larson MG, Fontes JD, Magnani JW, McManus DD *et al.* Relation between soluble ST2, growth differentiation factor-15, and high-sensitivity troponin I and incident atrial fibrillation. *Am Heart J* 2014;**167**:109–115.e2.
14. Lopez FL, Agarwal SK, Maclehose RF, Soliman EZ, Sharrett AR, Huxley RR *et al.* Blood lipid levels, lipid-lowering medications, and the incidence of atrial fibrillation: the atherosclerosis risk in communities study. *Circ Arrhythm Electrophysiol* 2012;**5**:155–62.
15. Mora S, Akinkuolie AO, Sandhu RK, Conen D, Albert CM. Paradoxical association of lipoprotein measures with incident atrial fibrillation. *Circ Arrhythm Electrophysiol* 2014; **7**:612–9.
16. Hindricks G, Potpara T, Dagres N, Arbelo E, Bax JJ, Blomström-Lundqvist C *et al.* 2020 ESC guidelines for the diagnosis and management of atrial fibrillation developed in collaboration with the European Association for Cardio-Thoracic Surgery (EACTS): the task force for the diagnosis and management of atrial fibrillation of the European Society of Cardiology (ESC) developed with the special contribution of the European Heart Rhythm Association (EHRA) of the ESC. *Eur Heart J* 2021;**42**:373–498.
17. Patton KK, Heckbert SR, Alonso A, Bahrami H, Lima JA, Burke G *et al.* N-terminal pro-B-type natriuretic peptide as a predictor of incident atrial fibrillation in the multiethnic study of atherosclerosis: the effects of age, sex and ethnicity. *Heart* 2013;**99**: 1832–6.