ORIGINAL ARTICLE

Deep neural networks reveal novel sex-specific electrocardiographic features relevant for mortality risk

Klaske R. Siegersma [1,2†], Rutger R. van de Leur [3,4†],
N. Charlotte Onland-Moret [5], David A. Leon [6,7,8], Ernest Diez-Benavente [2],
Liesbeth Rozendaal[9], Michiel L. Bots [5], Ruben Coronel [10], Yolande Appelman [1],
Leonard Hofstra[1,11], Pim van der Harst [3], Pieter A. Doevendans [3,4],
Rutger J. Hassink [3], Hester M. den Ruijter [2‡], and René van Es [3*‡]

[1]Department of Cardiology, Amsterdam University Medical Centres, VU University Amsterdam, Amsterdam, The Netherlands; [2]Laboratory of Experimental Cardiology, University Medical Center Utrecht, Utrecht University, Utrecht, The Netherlands; [3]Department of Cardiology, University Medical Center Utrecht, Utrecht, The Netherlands; [4]Netherlands Heart Institute, Utrecht, The Netherlands; [5]Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht University, Utrecht, The Netherlands; [6]Faculty of Epidemiology and Population Health, London School of Hygiene & Tropical Medicine, London WC1E 7HT, UK; [7]International Laboratory for Population and Health, National Research University, Higher School of Economics, Moscow 101000, Russian Federation; [8]Department of Community Medicine, UiT The Arctic University of Norway, Tromsø, Norway; [9]Julius Gezondheidscentrum Parkwijk, Utrecht, The Netherlands; [10]Heart Center, Department of Experimental Cardiology, AMC, Amsterdam University Medical Centres, Amsterdam, The Netherlands; and [11]Cardiology Centers of the Netherlands, Amsterdam, The Netherlands

Aims	Incorporation of sex in study design can lead to discoveries in medical research. Deep neural networks (DNNs) accurately predict sex based on the electrocardiogram (ECG) and we hypothesized that misclassification of sex is an important predictor for mortality. Therefore, we first developed and validated a DNN that classified sex based on the ECG and investigated the outcome. Second, we studied ECG drivers of DNN-classified sex and mortality.
Methods and results	A DNN was trained to classify sex based on 131 673 normal ECGs. The algorithm was validated on internal (68 500 ECGs) and external data sets (3303 and 4457 ECGs). The survival of sex (mis)classified groups was investigated using time-to-event analysis and sex-stratified mediation analysis of ECG features. The DNN successfully distinguished female from male ECGs {internal validation: area under the curve (AUC) 0.96 [95% confidence interval (CI): 0.96, 0.97]; external validations: AUC 0.89 (95% CI: 0.88, 0.90), 0.94 (95% CI: 0.93, 0.94)}. Sex-misclassified individuals (11%) had a 1.4 times higher mortality risk compared with correctly classified peers. The ventricular rate was the strongest mediating ECG variable (41%, 95% CI: 31%, 56%) in males, while the maximum amplitude of the ST segment was strongest in females (18%, 95% CI: 11%, 39%). Short QRS duration was associated with higher mortality risk.
Conclusion	Deep neural networks accurately classify sex based on ECGs. While the proportion of ECG-based sex misclassifications is low, it is an interesting biomarker. Investigation of the causal pathway between misclassification and mortality uncovered new ECG features that might be associated with mortality. Increased emphasis on sex as a biological variable in artificial intelligence is warranted.

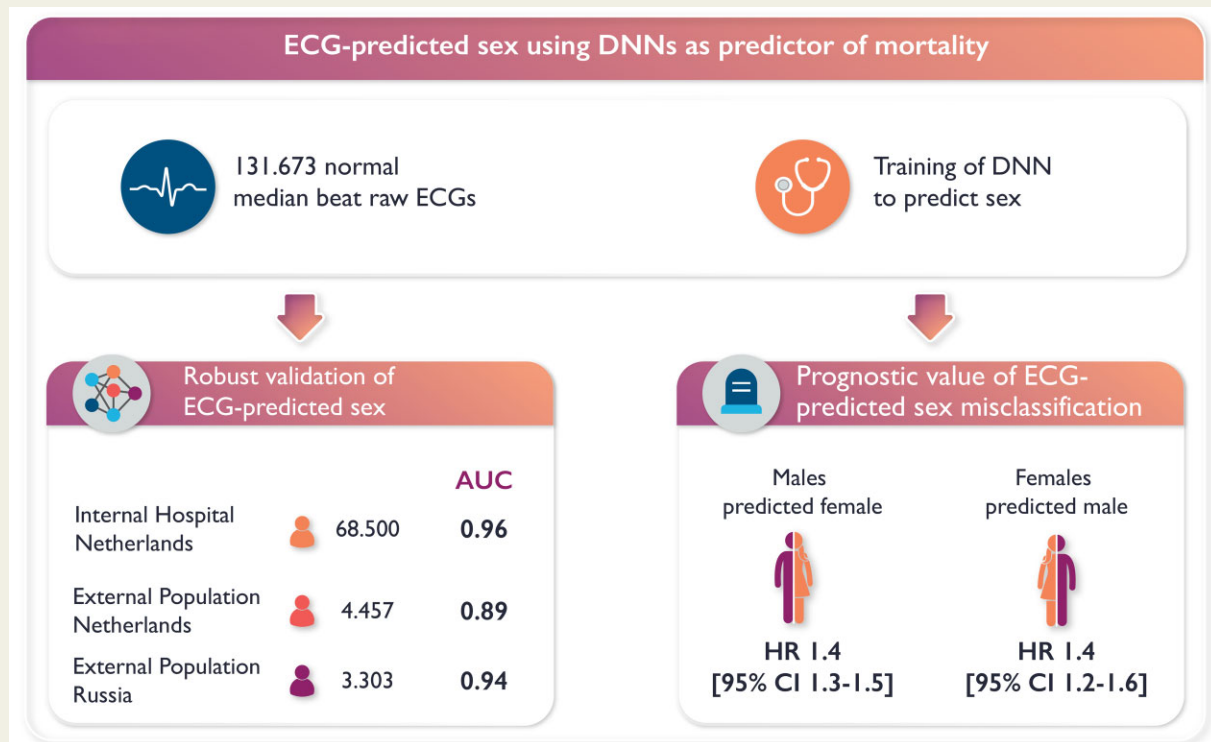* Corresponding author. Tel: +31 622504131, Email: r.vanes@umcutrecht.nl
†These authors contributed equally.
‡These authors contributed equally.

Graphical Abstract

ECG-predicted sex using DNNs as predictor of mortality

131.673 normal median beat raw ECGs

Training of DNN to predict sex

Robust validation of ECG-predicted sex

		AUC
Internal Hospital Netherlands	68.500	0.96
External Population Netherlands	4.457	0.89
External Population Russia	3.303	0.94

Prognostic value of ECG-predicted sex misclassification

Males predicted female
HR 1.4 [95% CI 1.3-1.5]

Females predicted male
HR 1.4 [95% CI 1.2-1.6]

Keywords Electrocardiography • Sex differences • Neural network • Artificial intelligence

Introduction

Despite increasing awareness of sex differences in cardiology, women remain underrepresented in randomized clinical trials.[1,2] Often, even when enough women are included, a focus on sex stratification is absent, despite pronounced differences between the sexes.[2,3] This also applies to the recent advancement of artificial intelligence (AI) in cardiology.[1] Artificial intelligence and specifically deep neural networks (DNNs) are increasingly used in cardiac research to analyse raw electrocardiogram (ECG) signals for prediction, diagnosis, and prognosis of cardiovascular disease (CVD).[4–7]

Sex differences in the ECGs are well known, as females have a higher heart rate, shorter PR interval and QRS duration, longer corrected QT duration (QTc), different T-wave morphology, and lower precordial QRS and T-wave amplitudes than males.[8–10] Many studies have investigated long-term prognostic information available in the ECG.[11] QRS prolongation, corrected QT prolongation, and T-wave morphology are associated with mortality, with different relations per sex.[11–14] For example, different types of T-wave morphology are associated with all-cause mortality in males and females.[14] For QRS duration and QTc interval, the relation is also different per sex, as healthy males have longer QRS durations and shorter QTc intervals.[13]

Recent advancements in AI and DNN have opened up many new possibilities for clinical applications of the established ECG

technology.[15,16] Deep neural networks classify sex based on ECG with an extremely high accuracy, but it remains unclear where these DNNs base their decisions on (e.g. differences in pathological changes).[17] ECG-based prediction of sex with DNNs could lead to new discoveries and insights. This might provide us with more information on the differences in longevity between the sexes.

We hypothesized that misclassification of sex based on the ECG is associated with survival and that sex-misclassified individuals mirror the survival of their predicted biological sex. Deep neural network–based misclassified men (classified as women) are hypothesized to have similar survival as biological women, and vice versa. Therefore, we externally validate a DNN trained on normal ECGs for classification of sex in large cohort studies and evaluated the outcome. We highlight how sex-specific ECG features affect mortality with mediation analysis.

Methods

Study participants and data acquisition

University Medical Center Utrecht training and internal validation data set

All 10 s 12-lead resting ECGs (*n* = 1 136 113) were acquired in the University Medical Center Utrecht (UMCU) between July 1991 and

Figure 1 Selection of individuals in the different University Medical Center Utrecht data sets; University Medical Center Utrecht training data set, internal validation data set, and data set for time-to-event and mediation analysis.
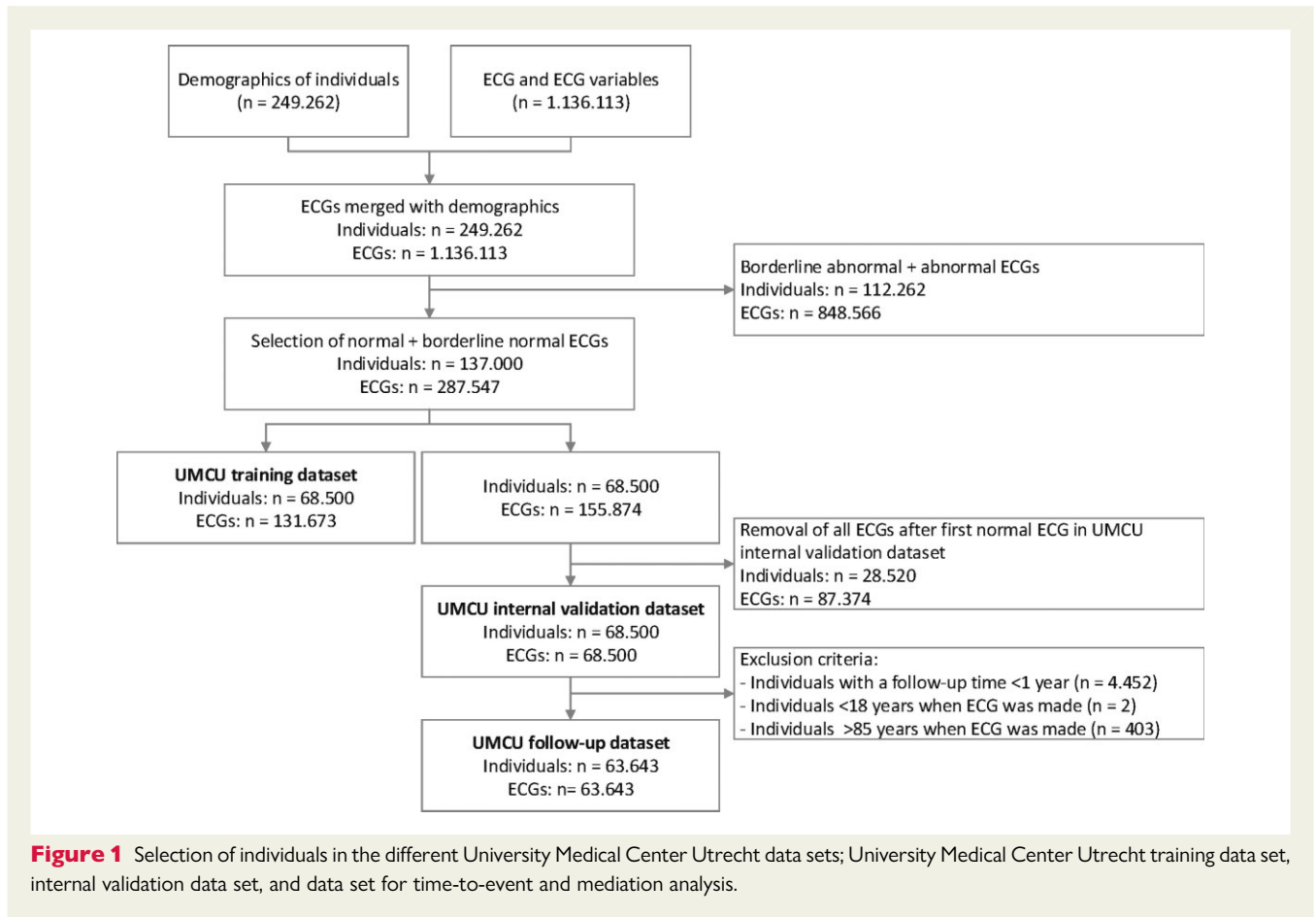
December 2019 from individuals ($n = 249\,262$) aged between 18 and 85 years were selected. Demographic (age, sex, and years of follow-up) and ECG data were extracted from hospital files of these individuals. All individuals ($n = 137\,000$) with at least one normal ECG ($n = 287\,547$) were selected (Figure 1). Only ECGs that were deemed interpretable by the Marquette 12 SL algorithm (GE Healthcare, Chicago, IL, USA) or the annotating physician were included.

The ECGs were recorded using a General Electric MAC V, 5000, or 5500 (GE Healthcare) at 250 or 500 Hz and extracted in raw voltage format. Approximately 31% of the ECGs were recorded at 250 Hz and linear interpolation was used to resample every ECG to 500 Hz. The representative median beat was used in this study and derived from 10 s recordings by aligning all QRS complexes and calculating the median voltage. R peaks were detected using the Stationary Wavelet Transform detector.[18] Extraction of the conventional ECG features, such as PR interval, is described in more detail in the Supplementary material online, Methods. All recordings obtained at non-cardiology departments were systematically annotated by a physician as part of the regular clinical workflow. The other ECGs were annotated by the Marquette 12SL algorithm (GE Healthcare). Diagnostic ECG statements were extracted from these free text annotations using a text mining algorithm described before and were used to determine if an ECG was interpreted as normal (e.g. the classification 'normal ECG' was given by either the physician or the computerized algorithm).[4]

The data set was split randomly 50:50 into a training and internal validation set on the individual level, making sure that there were no overlapping individuals between the training and internal validation data set. For training, all ECGs ($n = 131\,673$) per individual ($n = 68\,500$) were used. For internal validation, only the first normal ECG of each individual was selected, excluding 87 374 ECGs. This resulted in the UMCU internal validation data set of 68 500 ECGs in 68 500 individuals. All data were de-identified during extraction in accordance with the EU General Data Protection Regulation and the written informed consent requirement was waived by the UMCU ethical committee.

University Medical Center Utrecht follow-up data set

A subset of the UMCU internal validation data set was used to determine the association between ECG-classified sex and all-cause mortality. Survival data from all individuals were extracted from the Dutch Population Register. Individuals with <1 year of follow-up ($n = 4452$) and individuals aged <18 or >85 years during ECG recording ($n = 405$) were excluded. This enabled the investigation of long-term follow-up and reduced bias caused by individuals who are already in the hospital for a specific reason that reduces life expectancy (e.g. severe trauma or palliative care). These exclusions resulted in a final data set of 63 643 individuals and an equal number of ECGs (Figure 1).

External validation: Know-Your-Heart data set and Utrecht Health Project data set

External validation of the algorithm was performed in two data sets. The Know-Your-Heart (KYH) data set is a cross-sectional population-based study from two Russian cities, Arkhangelsk and Novosibirsk.[19] This cohort consisted of 4647 individuals, aged between 35 and 69 years. Only individuals ($n = 3303$, 1989 females, 60%) with a normal ECG were selected. The full protocol of the KYH study has been described

elsewhere.[19] A detailed description of the ECG acquisition in this data set is provided in the Supplementary material online, Methods.

The Utrecht Health Project (UHP) is an ongoing dynamic population study initiated in a newly developed large residential area in Leidsche Rijn, part of the city of Utrecht.[20] All new inhabitants were invited by their general practitioner to participate in the UHP. Written informed consent was obtained and an individual health profile was made by dedicated research nurses. Survival data were obtained by the general practitioner via the International Classification of Primary Care (ICPC)-codes. The UHP study was approved by the Medical Ethical Committee of the University Medical Center, Utrecht, The Netherlands. The UHP included baseline normal ECGs of 4457 individuals (2469 females, 55.4%), with a median age of 35 years (IQR: 30–43). The full protocol of the UHP cohort has been described elsewhere.[21]

Deep neural network development

A convolutional DNN architecture with several one-dimensional causal dilated convolutional layers was trained to classify sex on the ECG. This network architecture is inspired by van den Oord *et al.* and was described in detail previously.[22,23] The architecture has been previously optimized for use on median beats and no further hyperparameter tuning was performed on the UMCU training data set.[23,24] Training was performed with a binary cross-entropy loss and the Adam optimizer with a learning rate of 0.0001 and batch size of 128.[25,26] Early stopping was performed when the validation did not decrease for 20 epochs. Deep neural network output was a probability indicating the likelihood of an ECG belonging to a female individual. The cut-off value was set to 0.5, i.e. a probability of <0.5 resulted in the classification of the ECG belonging to a male. All algorithm development was performed with the PyTorch package (version 1.7.0).[27]

Statistical analysis

Descriptive statistics of data set and performance evaluation of DNN

The baseline characteristics of the data sets were described as mean ± standard deviation (SD) or median with IQR, where appropriate. The discriminatory performance of the DNN in the UMCU internal validation set and KYH and UHP external validation set was assessed with the area under the receiver operating characteristic (AUC) and accuracy, calculated as the number of correctly classified individuals divided by the total number of individuals. The 95% confidence interval (CI) around the performance measures was obtained using 2000 bootstrap samples. Four groups were identified for subsequent analyses using a predicted probability cut-off of 0.5: correctly classified males and females, biological females classified as male, and biological males classified as female. Conventional ECG features (e.g. PR interval) were compared between these groups. No *P*-values were provided in these comparisons.

Survival analysis in University Medical Center Utrecht follow-up data set

Using data from the UMCU follow-up and the UHP external validation data set, sex-stratified survival analyses with the Kaplan–Meier curves and Cox regression were done to evaluate the differences in survival between the correctly classified and misclassified individuals per sex. All analyses were performed with age as the primary time variable (i.e. correction for late entry or left-truncation), as included individuals had their first ECG at different ages (*Figure 2*).

To investigate underlying mechanisms of the association between DNN-based sex classification and survival, a biological sex-stratified mediation analysis with conventional ECG features (*n* = 57) was performed using the UMCU follow-up data set. Therefore, we used left-truncated

Cox regressions, separately for males and females. These regressions modelled the association between all-cause mortality, (mis)classification, and the conventional ECG parameter. The difference method was applied to quantify the proportion of the association between DNN-based sex classification and mortality that could be explained by conventional ECG features.[28,29] Therefore, the change in the coefficient of the DNN-based classification was first calculated when an ECG feature was added to the Cox regression, compared with the model that did not include that feature. Second, the mediation was determined by calculation of the proportion effect explained (PEE), i.e. the change in coefficient as a percentage of the total effect (the coefficient of ECG-based classification without the ECG feature). Bootstrap resampling with replacement (*n* = 500) was implemented to obtain 95% CI around the PEE.[28]

To investigate possible non-linear relationships between ECG variables and survival, we included a sex-stratified *post hoc* evaluation of the association between conventional ECG features that mediated >10% of the association between classification and all-cause mortality or where the difference in PEE in males and females was >10%. These cut-off points were chosen as we assume these are clinically relevant and as these might explain sex differences in ECG features. In this analysis, the selected ECG features were added one-by-one to a Cox regression model with age as the underlying time scale. A natural cubic spline was modelled for the relation between the ECG feature and all-cause mortality as the outcome. Models were developed for males and females separately. Hazard ratios (HRs) relative to the median value of the ECG variable in all samples were used to visualize the non-linear relationship.

All statistical analyses were executed using R version 3.5 (R Foundation for Statistical Computing). The Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis Statement for the reporting of diagnostic models was followed, where appropriate.[30]

Results

Characteristics of the study population

The median age of included individuals at the time of their ECG acquisition in the UMCU data set was 57.2 [interquartile range (IQR): 44.7–67.6] years. Supplementary material online, Table S1 shows the baseline characteristics of the ECGs used in the UMCU training and internal validation set separated for males and females. No significant differences in patient characteristics and ECG features were present between the data set used for training and internal validation. Follow-up was available for 104 848 (76.5%) of individuals and the median follow-up time for the UMCU follow-up data set was 8.7 (IQR: 4.4–14.6) and 8.9 (IQR 4.6–15.0) years for males and females, respectively. *Table 1* shows the distributions of the ECG features, stratified by sex and classification of sex, for the UMCU follow-up, the KYH external validation, and the UHP external validation data sets. Follow-up was available in the UHP data set [median follow-up: 16.9 years (IQR: 15.3–18.0)], but not for the KYH cohort.

Sex classification and feature detection

The AUC for sex classification with DNN in the UMCU internal validation data set (*n* = 68 500) was 0.96 (95% CI: 0.96–0.97), with an accuracy of 0.89 (95% CI: 0.89–0.89). Overall, misclassified individuals had ECG characteristics resembling those of their biological counterparts (*Table 1* and see Supplementary material online, Table S2). Thus, males classified as females had higher ventricular
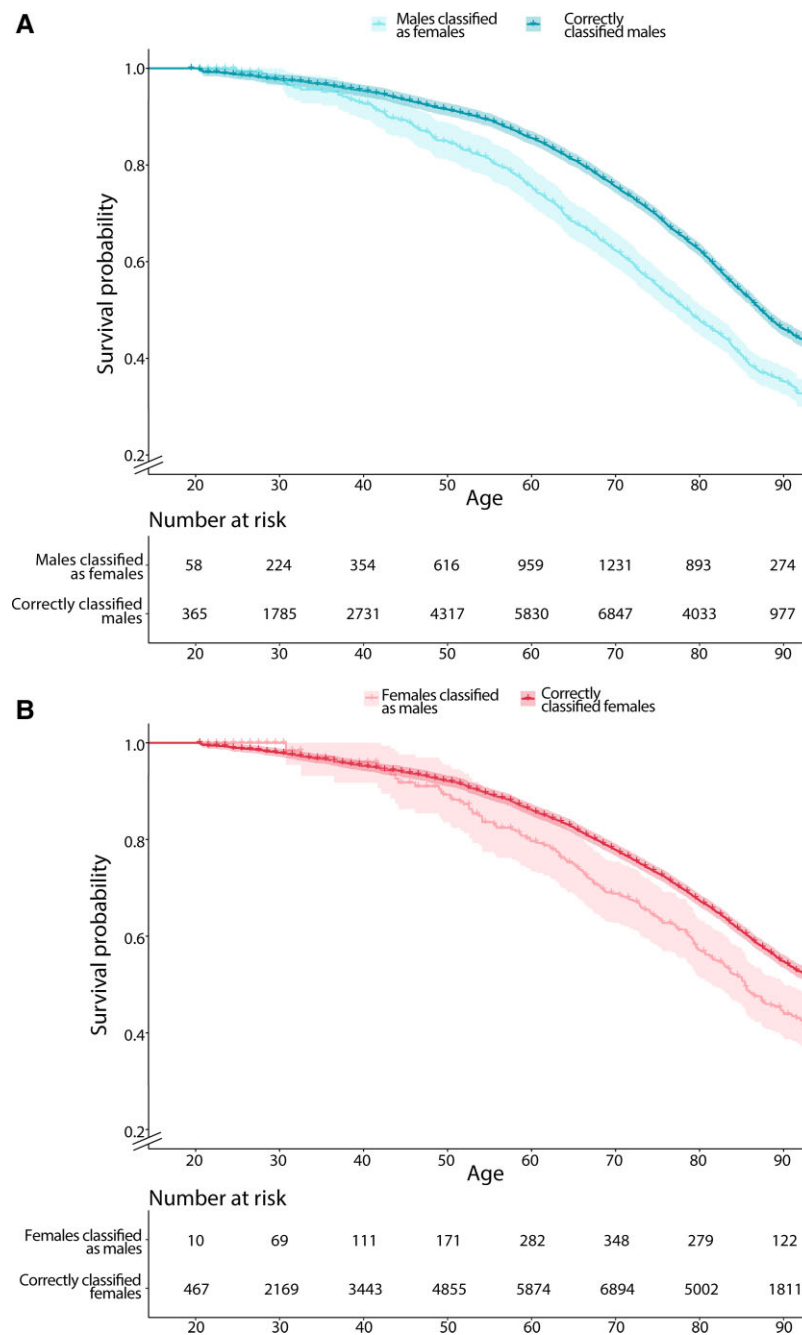
Figure 2 Kaplan–Meier curves (left-truncated) of males (*A*) and females (*B*), separately plotted for correctly classified and misclassified groups, corrected for late entry and stratified by classification of sex as was the output of the deep neural networks.

rate, shorter PR and QRS duration, longer QTc, and lower Sokolow–Lyon and Cornell voltages than correctly classified males. Females classified as males, when compared with correctly classified females, had a longer PR and QRS duration, but differences were less evident.

Evaluation of the individuals that were misclassified on sex in the UMCU follow-up data set (*Table 1*) showed that they were older than their correctly classified biological peers. The median age of misclassified females was 61 (IQR: 49–72) vs. 56 (IQR: 42–67) years for correctly classified females, and 61 (IQR: 49–70) and 57 (IQR: 44–66) years for misclassified and correctly classified males, respectively.

External validation using Know-Your-Heart and Utrecht Health Project data set

The AUC and accuracy for the DNN in the KYH external validation data set were, respectively, 0.89 (95% CI: 0.88–0.90) and 0.81 (95%

Table 1 Overview of baseline and electrocardiographic features in the University Medical Center Utrecht follow-up data set, Know-Your-Heart, and Utrecht Health Project external validation data set and the distribution between correctly classified males and females and their misclassified biological peers (IQR, interquartile range)

| | UMCU follow-up | | | | Know-Your-Heart | | | | Utrecht Health Project | | | |
| | Males | | Females | | Males | | Females | | Males | | Females | |
	Correctly classified	Misclassified	Correctly classified	Misclassified	Correctly classified	Misclassified	Correctly classified	Misclassified	Correctly classified	Misclassified	Correctly classified	Misclassified
Individuals, n	26914	5108	30339	1282	902	412	1787	202	1872	116	1764	705
Deceased, n (%)	3313 (12.3)	880 (17.2)	3041 (10.0)	210 (16.4)	—	—	—	—	40 (2.1)	2 (1.7)	17 (1.0)	11 (1.6)
Age at ECG in years [median (IQR)]	57 (44, 66)	61 (49, 70)	56 (42, 67)	61 (49, 72)	53 (45, 61)	56 (47, 64)	54 (45, 62)	57 (49, 64)	36 (31, 43)	40 (34, 51)	34 (29, 43)	33 (29, 41)
Follow-up duration in years [median (IQR)]	9.0 (4.6, 15.0)	7.5 (3.7, 12.8)	8.9 (4.6, 15)	10.0 (4.9, 15.6)	—	—	—	—	16.9 (15.3, 18.0)	15.9 (15.2, 17.2)	16.7 (15.2, 17.9)	17.4 (15.6, 18.1)
Ventricular rate in beats per minute [median (IQR)]	67 (59, 77)	74 (64, 86)	71 (63, 81)	69 (61, 80)	63 (57, 70)	62 (56, 70)	64 (59, 71)	63 (58, 69)	60 (54, 66)	60 (54, 72)	66 (60, 72)	60 (54, 72)
PR interval in ms [median (IQR)]	158 (144, 174)	154 (138, 172)	150 (136, 166)	158 (144, 176)	158 (146, 174)	156 (140, 174)	152 (138, 168)	154 (142, 168)	—	—	—	—
QRS duration in ms [median (IQR)]	96 (90, 104)	90 (84, 98)	86 (80, 92)	94 (86, 100)	96 (90, 102)	94 (88, 102)	90 (84, 96)	92 (86, 98)	98 (92, 104)	94 (86, 102)	86 (80, 92)	88 (82, 96)
QT interval in ms [median (IQR)]	388 (368, 412)	382 (356, 408)	386 (364, 408)	388 (366, 412)	402 (384, 422)	406 (386, 426)	410 (392, 430)	410 (392, 428)	394 (376, 414)	395 (370, 425)	394 (376, 412)	394 (376, 414)
Corrected QT in ms [median (IQR)]	411 (400, 427)	422 (407, 439)	419 (406, 436)	417 (405, 436)	411 (398, 426)	416 (402, 430)	425 (412, 438)	421 (409, 435)	396 (381, 412)	408 (390, 417)	406 (392, 422)	405 (389, 420)
SL-voltage in mV [median (IQR)]	2.02 (1.63, 2.44)	1.82 (1.47, 2.23)	1.91 (1.55, 2.29)	1.97 (1.59, 2.41)	2.2 (1.8, 2.6)	2 (1.6, 2.4)	2 (1.6, 2.3)	2 (1.6, 2.4)	2.3 (1.9, 2.8)	2.1 (1.6, 2.5)	1.9 (1.6, 2.3)	2.0 (1.6, 2.4)
Cornell-voltage in mV [median (IQR)]	1.41 (1.08, 1.77)	1.30 (0.98, 1.66)	1.15 (0.85, 1.48)	1.32 (0.98, 1.70)	1.5 (1.2, 1.8)	1.3 (1, 1.7)	1.2 ((0.9, 1.5)	1.5 (1.1, 1.8)	1.1 (0.8, 1.5)	1.1 (0.8, 1.4)	0.8 (0.5, 1.1)	0.8 (0.5, 1.1)
SL-product in mV [median (IQR)]	193 (153, 238)	163 (130, 202)	165 (132, 202)	182 (146, 227)	210 (168, 248)	189 (154, 229)	177 (145, 212)	183 (147, 227)	222 (180, 274)	193 (155, 239)	164 (136, 197)	172 (139, 208)
Cornell-product in mV (median (IQR))	136 (101, 175)	117 (86, 153)	100 (72, 131)	122 (89, 159)	142 (108, 176)	128 (95, 162)	105 (78, 137)	141 (103, 169)	111 (79, 147)	105 (70, 129)	64.8 (44, 91.3)	69.8 (43.9, 101.5)

For KYH, no follow-up information was available, while for UHP, the PR interval was not available.

ECG, electrocardiogram; IQR, interquartile range.

CI: 0.80–0.82). Similar trends regarding the distribution of ECG features as in the UMCU internal validation data set were seen when different classifications were compared (Table 1). Moreover, misclassified individuals were overall older (median age misclassified females: 57, IQR: 49–64, vs. correctly classified females: 54, IQR: 45–62, median age misclassified males: 56.0, IQR: 47–64, vs. correctly classified males: 53, IQR: 45–61).

The AUC and accuracy for the DNN in the UHP data set were, respectively, 0.94 (95% CI: 0.93–0.94) and 0.82 (95% CI: 0.80–0.83). In this data set, the individuals were overall younger than in the KYH data set and UMCU internal validation data set with a median age of 36 (IQR: 31–44) for males and 34 (IQR: 29–42) for females. Yet, only misclassified males were older than their correctly classified biological peers (median age: 40, IQR: 34–51 vs. 36, IQR: 31–43). The median age between misclassified and correctly classified females did not differ (misclassified females: 33, IQR: 29–41 vs. correctly classified females: 34, IQR: 29–43).

Sex-specific survival analysis

In the UMCU follow-up data set, 3251 (10%) of included females and 4193 (13%) of included males died during follow-up. Mortality risk in this data set was higher for biological males compared with biological females (HR: 1.33, 95% CI: 1.27–1.39). In both sexes, a higher proportion of misclassified individuals died compared with their correctly classified biological peers: 16.4% ($n = 210$) vs. 10% ($n = 3041$) of misclassified and correctly classified females, 17.2% ($n = 880$) vs. 12.3% ($n = 3313$) of misclassified and correctly classified males. This was also shown in the Kaplan–Meier curves of both sexes (Figure 2) and confirmed by the Cox regression that showed misclassified individuals had a higher mortality risk referenced to their correctly classified biological peers (HR misclassified females compared with correctly classified females: 1.37, 95% CI: 1.19–1.57 and HR misclassified males compared with correctly classified males: 1.38, 95% CI: 1.28–1.49).

Follow-up was also available for the individuals in the UHP external validation data set. In this data set, mortality was low, with 28 (1.1%) females and 42 (2.1%) males who died. The mortality risk for males was higher compared with females (HR: 1.62, 95% CI: 1.01–2.62). Misclassified females, when compared with their correctly classified peers, had a higher mortality risk, although not significantly (HR: 1.61, 95% CI: 0.76–3.46), while misclassified males had a lower mortality risk when compared with their correctly classified peers (HR: 0.39, 95% CI: 0.09–1.64).
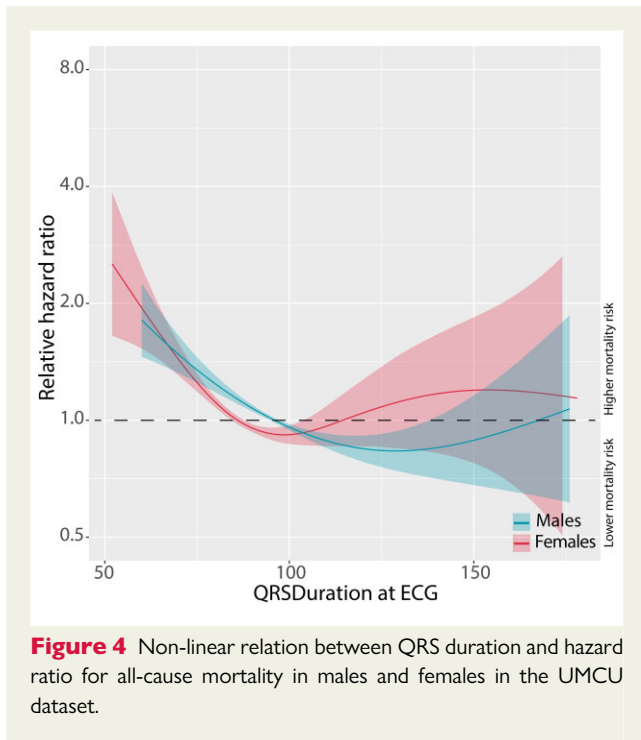
Sex-stratified analysis of mediation by electrocardiographic features on survival

Sex-stratified mediation analyses in the UMCU follow-up data set showed that the relationship between misclassification of sex and mortality was mediated, at least in part, by conventional ECG features in both sexes. Figure 3 shows the ECG features that were selected according to clinically relevant criteria ($n = 17$). These features either mediated for >10% or suppressed less than −10%. Features with >10% difference in PEE between males and females were also selected. The PEE of some of the features was clinically relevant in both sexes, although for some features, the PEE was negative in one of the sexes. Features that were selected in both males
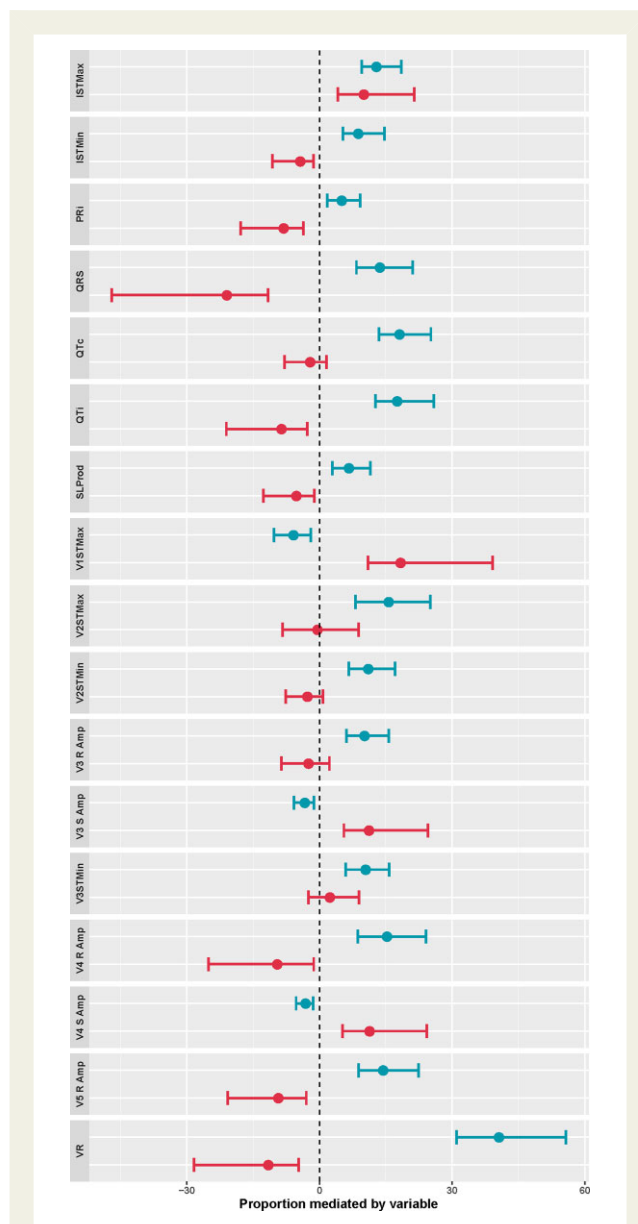
Figure 3 Proportion of the relation between classification and survival that is mediated by a selection of standard electrocardiographic features, stratified for females (red, lower point in each subplot) and males (blue, upper point in each subplot). Only electrocardiographic features with more than 10% mediation or with a more than 10% difference between mediation in males and females are displayed. I/V2/V3 ST Max/Min, maximum and minimum amplitude of ST segment in Leads I, V2, and V3; PRi, PR interval; QRS, duration of QRS complex; QTc, corrected QT interval; QTi, QT interval; SLProd, Sokolow–Lyon product; V3/V4/V5 R/S Amp, R- and S-wave amplitude in Leads V3, V4, and V5; VR, ventricular rate.

and females, i.e. features in which the PEE was less than −10% of >10%, were ventricular rate (male: 41%, female: 12%), QRS duration (male: 13%, female: −21%), and maximum T-wave amplitude in Lead I (male: 13%, female: 10%). In males, the ventricular rate was the strongest mediator between DNN-based sex classification and survival

Figure 4 Non-linear relation between QRS duration and hazard ratio for all-cause mortality in males and females in the UMCU dataset.

(41%). In females, this was the maximum amplitude of the ST segment in V1 (18%). Moreover, a strongly negative PEE was found for QRS duration in females. A complete overview of all ECG features included in the mediation analyses is shown in *Supplementary material online, Table S3*.

Figure 3 shows the features used as input for non-linear modelling of all-cause mortality, for males and females separately. The QRS duration was associated with a lower mortality risk in both sexes. Restricted cubic spline analysis showed a non-linear relationship between QRS duration and mortality (*Figure 4*): in both males and females, the HR decreased with elongation of the QRS duration until 100 ms. After this threshold, QRS duration was no longer associated with increased mortality risk in both males and females. The non-linear relation between short QRS duration and higher mortality risk was confirmed in males in the UHP external validation data set, but not in females (see *Supplementary material online, Figure S1*). Figures of other selected ECG features and mortality risk are shown in *Supplementary material online, Figure S2*.

Discussion

We show excellent performance of DNN in the classification of sex based on ECGs across different populations and ECG devices. Misclassification of sex on the ECG is associated with a higher mortality risk, independent of age. We hypothesized that the survival curve of misclassified individuals would copy the survival curve of their biological counterparts. However, this hypothesis was not fully confirmed as the survival of males classified as females was indeed worse compared with the correctly classified males in the internal validation data set. Yet, the UHP external validation data set showed a non-significant trend towards confirmation of the hypothesis.

Subsequent mediation analyses on survival showed that conventional ECG features, i.e. ventricular rate, QRS duration, and several features that relate to the amplitude, are mainly important in misclassified males, but less so in females. Unexpectedly, we observed that shortened QRS duration increased mortality risk which has not previously been described. Our study highlights the importance of studying sex differences with AI to uncover new biology.

Our study shows a similar performance of the DNN on median beats for classification of sex as was described on full 10 s ECG, while also validating these results in two external test sets with completely different patient groups and ECG devices.[17] This is one of the first studies to show that DNNs trained on median beats are easily transferrable to other data sets. Moreover, the study by Attia *et al.* stated the necessity to understand the relevance of discordance between ECG-classified and true biological sex for the individual, as this remained unknown until now. Our analysis of ECG features focused on this knowledge gap and showed that misclassification occurs when the ECGs become more alike, i.e. females who have ECG features similar to males are more often misclassified and the same holds for males who have ECG features similar to females, which was confirmed in external validation. Furthermore, discordance between ECG-classified and true biological sex was associated with a reduced survival in both sexes.

A higher mortality risk in sex-misclassified males (classified as female) was mediated by different ECG features such as ventricular rate and QT interval. It is known that increased ventricular rate and corrected QT interval are associated with mortality.[31–35] It is assumed that a higher resting heart rate is an indicator of an overall worse physical condition.[32,34] A longer QTc has been described before to predicting a high risk of dying, even when the QTc is within normal boundaries. This was confirmed in our *post hoc* analysis (see *Supplementary material online, Figure S2*).[11]

In contrast to males, the association between misclassified females (classified as males) and survival was only minimally mediated by conventional ECG parameters. The feature that mediated the relation to the largest part, yet only 18%, was the positive T-wave amplitude in V1. Deep neural network therefore picked up subtle ECG differences between males and females that seem to have prognostic value.

Surprisingly, the mediation analysis showed a negative mediation for QRS duration in females. This means that QRS duration might act as a suppressor. Indeed, the non-linear *post hoc* analysis showed increased mortality risk in those with a short QRS duration (<80 ms). This was validated in the UHP external validation data set. After an optimum at 100 ms for females and 125 ms for males, a higher QRS duration is associated with worse survival, which is confirmed by previous studies.[13] The relationship between a shorter QRS duration and increased mortality risk has, to our knowledge, not been previously identified. However, it has been hypothesized that increasing extension of the Purkinje system into the walls of the ventricular system is associated with a shorter QRS duration, which makes these individuals more at risk for idiopathic ventricular re-entrant arrhythmias.[36]

Strengths in this study are, first, the large amount of annotated ECG data that have been used, which gave the opportunity to only select normal ECGs.[26] A large number of ECG data from regular care are also important in the light of the structural underrepresentation of women in cardiovascular research.[37–39] Our study included many females, which gave us the opportunity to specifically study sex

differences and perform all analyses in a sex-stratified manner. Secondly, this study was the first study to externally validate a sex classification algorithm. Despite more awareness, sex stratification is often not performed, which also applies to the validation of AI algorithms.[1] Thirdly, this study is unique in that it provided new insights into sex-specific ECG features that are associated with mortality, through the classification of sex with ECG-based AI. The presented methodology and results feed future research into sex-specific conductivity mechanisms that influence survival, unravelling the conundrum of sex differences in longevity.

This study has some limitations. First, the hospital-visiting population that was used in this study had an ECG for a specific reason, although we selected only ECGs classified as 'normal' by either the ECG recording software or the examining physicians. Underlying pathophysiology that does not directly affect the ECG, but also stress and anxiety related to an out-patient clinic or hospital visit, might induce subtle changes on the ECG, including an increase in heart rate. This could make the DNN less generalizable to non-hospital populations. However, our external validation analyses showed the decrease in performance to be limited.[19,21] Secondly, no causes of death were known for the UMCU follow-up data set, which hampered the analysis of cardiovascular mortality. Thirdly, the UHP external validation data set was significantly different from the UMCU internal validation data set. In general, individuals in the UHP data set were younger. This, in combination with a low number of events, might have caused our inability to show a significantly reduced survival risk in misclassified individuals. Yet, we were able to validate the relation between shortening of the QRS duration in the UHP external validation data set for males, which is a promising feature for future studies. Fourthly, survival information in the KYH external validation cohort was unavailable. Therefore, our findings regarding survival and the influence of different ECG features of misclassified individuals warrant further validation. Nonetheless, the pattern of differences in the ECG characteristics according to classification status was replicated in the KYH data set.

Conclusion

Deep neural networks accurately classify sex based on the ECG. While the proportion of misclassified individuals is low, ECG-based sex misclassification is an interesting biomarker for mortality, with increased mortality for misclassification in both sexes. Using mediation analysis to investigate the causal pathway between misclassification and mortality, new ECG features that might be associated with mortality have been discovered. An increased emphasis on sex as a biological variable in AI is warranted.

Supplementary material

Supplementary material is available at *European Heart Journal – Digital Health*.

Authors' contributions

K.R.S., R.R.v.d.L., N.C.O.-M., H.M.d.R. and R.v.E. designed the study and the analysis plan. K.R.S. and R.R.v.d.L. executed the analysis plan and did all data analysis. K.R.S. and R.R.v.d.L. drafted the initial manuscript and this was critically edited by N.C.O.-M., H.M.d.R. and R.v.E. D.A.L. and E.D-B. helped with analysis of the KYH external validation data set, L.R. and M.L.B. with analysis of the UHP external validation data set. R.C., Y.A., L.H., P.v.d.H., P.D., and R.J.H. helped with interpretation of results and critically edited the manuscript. All authors approved the final work and agree to be accountable for the accuracy and integrity of the work.

Data availability

The UMCU data cannot be shared publicly due to the privacy of individuals who participated in the study. Programming code is available upon request to the corresponding author. When interested in the data for scientific open access use, contact the corresponding author.

References

1. Tannenbaum C, Ellis RP, Eyssel F, Zou J, Schiebinger L. Sex and gender analysis improves science and engineering. *Nature* 2019;**575**:137–146.
2. Bots SH, den Ruijter HM. Recommended heart failure medications and adverse drug reactions in women call for sex-specific data reporting. *Circulation* 2019;**139**: 1469–1471.
3. Vogel B, Acevedo M, Appelman Y, Bairey Merz CN, Chieffo A, Figtree GA, Guerrero M, Kunadian V, Lam CSP, Maas AHEM, Mihailidou AS, Olszanecka A, Poole JE, Saldarriaga C, Saw J, Zühlke L, Mehran R. The Lancet women and cardiovascular disease Commission: reducing the global burden by 2030. *Lancet* 2021;**397**:2385–2438.
4. van de Leur RR, Blom LJ, Gavves E, Hof IE, van der Heijden JF, Clappers NC, Doevendans PA, Hassink RJ, van Es R. Automatic triage of 12-lead ECGs using deep convolutional neural networks. *J Am Heart Assoc* 2020;**9**:e015138.
5. Ko WY, Siontis KC, Attia ZI, Carter RE, Kapa S, Ommen SR, Demuth SJ, Ackerman MJ, Gersh BJ, Arruda-Olson AM, Geske JB, Asirvatham SJ, Lopez-Jimenez F, Nishimura RA, Friedman PA, Noseworthy PA. Detection of hypertrophic cardiomyopathy using a convolutional neural network-enabled electrocardiogram. *J Am Coll Cardiol* 2020;**75**:722–733.

6. Attia ZI, Noseworthy PA, Lopez-Jimenez F, Asirvatham SJ, Deshmukh AJ, Gersh BJ, Carter RE, Yao X, Rabinstein AA, Erickson BJ, Kapa S, Friedman PA. An artificial intelligence-enabled ECG algorithm for the identification of patients with atrial fibrillation during sinus rhythm: a retrospective analysis of outcome prediction. *Lancet* 2019;**394**:861–867.

7. Attia ZI, Kapa S, Lopez-Jimenez F, McKie PM, Ladewig DJ, Satam G, Pellikka PA, Enriquez-Sarano M, Noseworthy PA, Munger TM, Asirvatham SJ, Scott CG, Carter RE, Friedman PA. Screening for cardiac contractile dysfunction using an artificial intelligence–enabled electrocardiogram. *Nat Med* 2019;**25**:70–74.

8. Rijnbeek PR, Van Herpen G, Bots ML, Man S, Verweij N, Hofman A, Hillege H, Numans ME, Swenne CA, Witteman JCM, Kors JA. Normal values of the electrocardiogram for ages 16-90 years. *J Electrocardiol* 2014;**47**:914–921.

9. van der Ende MY, Siland JE, Snieder H, van der Harst P, Rienstra M. Population-based values and abnormalities of the electrocardiogram in the general Dutch population: the LifeLines Cohort Study. *Clin Cardiol* 2017;**40**:865–872.

10. Simonson E, Blackburn H, Puchner TC, Eisenberg P, Ribeiro F, Meja M. Sex differences in the electrocardiogram. *Circulation* 1960;**22**:598–601.

11. Zhang Y, Post WS, Blasco-Colmenares E, Dalal D, Tomaselli GF, Guallara E. Electrocardiographic QT interval and mortality: a meta-analysis. *Epidemiology* 2011;**22**:660–670.

12. Noseworthy PA, Peloso GM, Hwang SJ, Larson MG, Levy D, O'Donnell CJ, Newton-Cheh C. QT interval and long-term mortality risk in the Framingham heart study. *Ann Noninvasive Electrocardiol* 2012;**17**:340–348.

13. Badheka AO, Singh V, Patel NJ, Deshmukh A, Shah N, Chothani A, Mehta K, Grover P, Savani GT, Gupta S, Rathod A, Marzouka GR, Mitrani RD, Moscucci M, Cohen MG. QRS duration on electrocardiography and cardiovascular mortality (from the national health and nutrition examination survey - III). *Am J Cardiol* 2013;**112**:671–677.

14. Porthan K, Viitasalo M, Jula A, Reunanen A, Rapola J, Väänänen H, Nieminen MS, Toivonen L, Salomaa V, Oikarinen L. Predictive value of electrocardiographic QT interval and T-wave morphology parameters for all-cause and cardiovascular mortality in a general population sample. *Heart Rhythm* 2009;**6**:1202–1208.e1.

15. van de Leur RR, Boonstra MJ, Bagheri A, Roudijk RW, Sammani A, Taha K, Doevendans PAFM, van der Harst P, van Dam Peter M, Hassink RJ, van Es R, Asselbergs FW. Big data and artificial intelligence: opportunities and threats in electrophysiology. *Arrhythmia Electrophysiol Rev* 2020;**9**:146–154.

16. Siontis KC, Noseworthy PA, Attia ZI, Friedman PA. Artificial intelligence-enhanced electrocardiography in cardiovascular disease management. *Nat Rev Cardiol* 2021;**18**:465–478.

17. Attia ZI, Friedman PA, Noseworthy PA, Lopez-Jimenez F, Ladewig DJ, Satam G, Pellikka PA, Munger TM, Asirvatham SJ, Scott CG, Carter RE, Kapa S. Age and sex estimation using artificial intelligence from standard 12-lead ECGs. *Circ Arrhythmia Electrophysiol* 2019;**12**:e007284.

18. Kalidas V, Tamil L. Real-time QRS detector using stationary wavelet transform for automated ECG analysis. In: Proceedings - 2017 IEEE 17th International Conference Bioinformatics and Bioengineering (BIBE), Washington, DC, 2017. p457–461.

19. Cook S, Malyutina S, Kudryavtsev AV, Averina M, Bobrova N, Boytsov S, Brage S, Clark TG, Diez Benavente E, Eggen AE, Hopstock LA, Hughes A, Johansen H, Kholmatova K, Kichigina A, Kontsevaya A, Kornev M, Leong D, Magnus P, Mathiesen E, McKee M, Morgan K, Nilssen O, Plakhov I, Quint JK, Rapala A, Ryabikov A, Saburova L, Schirmer H, Shapkina M, Shiekh S, Shkolnikov VM, Stylidis M, Voevoda M, Westgate K, Leon DA. Know your heart: Rationale, design and conduct of a cross-sectional study of cardiovascular structure, function and risk factors in 4500 men and women aged 35-69 years from two russian cities, 2015-18 [version 3; referees: 3 approved]. *Wellcome Open Res* 2018;**3**:67.

20. Scheltens T, De Beus MF, Hoes AW, Rutten FH, Numans ME, Mosterd A, Kors JA, Grobbee DE, Bots ML. The potential yield of ECG screening of hypertensive patients: the Utrecht Health Project. *J Hypertens* 2010;**28**:1527–1533.

21. Grobbee DE, Hoes AW, Verheij TJM, Schrijvers AJP, Van Ameijden EJC, Numans ME. The Utrecht Health Project: optimization of routine healthcare data for research. *Eur J Epidemiol* 2005;**20**:285–290.

22. van den Oord A, Dieleman S, Zen H, Simonyan K, Vinyals O, Graves A, Kalchbrenner N, Senior A, Kavukcuoglu K. WaveNet: a generative model for raw audio. arXiv preprint arXiv:1609.03499. 12 September 2016.

23. Van De Leur RR, Taha K, Bos MN, van der Heijden JF, Gupta D, Cramer MJ, Hassink RJ, van der Harst P, Doevendans PA, Asselbergs FW, van Es R. Discovering and visualizing disease-specific electrocardiogram features using deep learning: proof-of-concept in phospholamban gene mutation carriers. *Circ Arrhythmia Electrophysiol* 2021;**14**:e009056.

24. Bos MN, Van De Leur RR, Vranken JF, Gupta D, van der Harst P, Doevendans P, van Es R. Automated comprehensive interpretation of 12-lead electrocardiograms using pre-trained exponentially dilated causal convolutional neural networks. 2020 Computing in Cardiology, Rimini, 2020. p10–13.

25. Lin T-Y, Goyal P, Girshick R, He K, Dollár P. Focal loss for dense object detection. In: IEEE International Conference on Computer Vision, Venice, 2017: 2999–3007.

26. Kingma DP, Ba JL. Adam: A method for stochastic optimization. In: 3rd International Conference of Learning Representations (ICLR) 2015 - Conf Track Proc, San Diego, CA. 1–15.

27. Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, Killeen T, Lin Z, Gimelshein N, Antiga L, Desmaison A. PyTorch: an imperative style, high-performance deep learning library. *Adv Neural Inform Proc Syst* 2019;**32**. https://www.apa.org/pubs/journals/psp.

28. Judd CM, Kenny DA. Process analysis: estimating mediation in treatment evaluations. *Eval Rev* 1981;**5**:602–619.

29. Baron RM, Kenny DA. The moderator-mediator variable distinction in social psychological research. Conceptual, strategic, and statistical considerations. *J Pers Soc Psychol* 1986;**51**:1173–1182.

30. Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD) the TRIPOD statement. *Circulation* 2015;**131**:211–219.

31. Reunanen A, Karjalainen J, Ristola P, Heliovaara M, Knekt P, Aromaa A. Heart rate and mortality. *J Intern Med* 2000;**247**:231–239.

32. Alhalabi L, Singleton MJ, Oseni AO, Shah AJ, Zhang ZM, Soliman EZ. Relation of higher resting heart rate to risk of cardiovascular versus noncardiovascular death. *Am J Cardiol* 2017;**119**:1003–1007.

33. Raisi-Estabragh Z, Cooper J, Judge R, Khanji MY, Munroe PB, Cooper C, Harvey NC, Petersen SE, Li Y. Age, sex and disease-specific associations between resting heart rate and cardiovascular mortality in the UK BIOBANK. *PLoS One* 2020;**15**:e0233898.

34. Kannel WB, Kannel C, Paffenbarger RS, Cupples LA. Heart rate and cardiovascular mortality: the Framingham study. *Am Heart J* 1987;**113**:1489–1494.

35. Seccareccia F, Pannozzo F, Dima F, Minoprio A, Menditto A, Lo Noce C, Giampaoli S. Heart rate as a predictor of mortality: the MATISS project. *Am J Public Health* 2001;**91**:1258–1263.

36. Coronel R, Potse M, Haïssaguerre M, Derval N, Rivaud MR, Meijborg VMF, Cluitmans M, Hocini M, Boukens BJ. Why ablation of sites with Purkinje activation is antiarrhythmic: the interplay between fast activation and arrhythmogenesis. *Front Physiol* 2021;**12**:648396.

37. Scott PE, Unger EF, Jenkins MR, Southworth MR, McDowell T-Y, Geller RJ, Elahi M, Temple RJ, Woodcock J. Participation of women in clinical trials supporting FDA approval of cardiovascular drugs. *J Am Coll Cardiol* 2018;**71**:1960–1969.

38. Vitale C, Fini M, Spoletini I, Lainscak M, Seferovic P, Rosano GM. Under-representation of elderly and women in clinical trials. *Int J Cardiol* 2017;**232**:216–221.

39. Pilote L, Raparelli V. Participation of women in clinical trials: not yet time to rest on our laurels. *J Am Coll Cardiol* 2018;**71**:1970–1972.