HeartRhythm

The Official Journal of the Heart Rhythm Society, The Cardiac Electrophysiology Society, and The Pediatric & Congenital Electrophysiology Society

Heart Rhythm Society SM

ECG-only explainable deep learning algorithm predicts the risk for malignant ventricular arrhythmia in phospholamban cardiomyopathy

Rutger R. van de Leur, MD,[1*] Remco de Brouwer, MD,[2,*] Hidde Bleijendaal, MD,[3,4,5] Tom E. Verstraelen, MD, PhD,[3,4] Belend Mahmoud, MD,[2] Ana Perez-Matos, MD,[6] Cathelijne Dickhoff, MD,[7] Bas A. Schoonderwoerd, MD, PhD,[8] Tjeerd Germans, MD, PhD,[9] Arjan Houweling, MD, PhD,[10] Paul A. van der Zwaag, MD, PhD,[11] Moniek G.P.J. Cox, MD, PhD,[2] J. Peter van Tintelen, MD, PhD,[4,12] Anneline S.J.M. te Riele, MD, PhD,[1] Maarten P. van den Berg, MD, PhD,[2] Arthur A.M. Wilde, MD, PhD,[3,4] Pieter A. Doevendans, MD, PhD,[1,4,13,14] Rudolf A. de Boer, MD, PhD,[2,15] René van Es, PhD[1]

ABSTRACT

BACKGROUND Phospholamban (*PLN*) p.(Arg14del) variant carriers are at risk for development of malignant ventricular arrhythmia (MVA). Accurate risk stratification allows timely implantation of intracardiac defibrillators and is currently performed with a multimodality prediction model.

OBJECTIVE This study aimed to investigate whether an explainable deep learning–based approach allows risk prediction with only electrocardiogram (ECG) data.

METHODS A total of 679 *PLN* p.(Arg14del) carriers without MVA at baseline were identified. A deep learning–based variational auto-encoder, trained on 1.1 million ECGs, was used to convert the 12-lead baseline ECG into its FactorECG, a compressed version of the ECG that summarizes it into 32 explainable factors. Prediction models were developed by Cox regression.

RESULTS The deep learning–based ECG-only approach was able to predict MVA with a C statistic of 0.79 (95% CI, 0.76–0.83), comparable to the current prediction model (C statistic, 0.83 [95% CI, 0.79–0.88]; $P = .054$) and outperforming a model based on conventional ECG parameters (low-voltage ECG and negative T waves; C statistic, 0.65 [95% CI, 0.58–0.73]; $P < .001$). Clinical simulations showed that a 2-step approach, with ECG-only screening followed by a full workup, resulted in 60% less additional diagnostics while outperforming the multimodal prediction model in all patients. A visualization tool was created to provide interactive visualizations (https://pln.ecgx.ai).

CONCLUSION Our deep learning–based algorithm based on ECG data only accurately predicts the occurrence of MVA in *PLN* p.(Arg14del) carriers, enabling more efficient stratification of patients who need additional diagnostic testing and follow-up.

KEYWORDS Electrocardiography; Deep learning; Phospholamban; Genetic cardiomyopathy; Explainable artificial intelligence

(Heart Rhythm 2024;■:1–11)

From the [1]Department of Cardiology, University Medical Center Utrecht, Utrecht, The Netherlands, [2]Department of Cardiology, University Medical Center Groningen, Groningen, The Netherlands, [3]Department of Cardiology, Amsterdam UMC location University of Amsterdam, Amsterdam, The Netherlands, [4]European Reference Network for Rare, Low-Prevalence, or Complex Diseases of the Heart (ERN GUARD-Heart), [5]Department of Clinical Epidemiology, Biostatistics and Bioinformatics, Amsterdam University Medical Center, University of Amsterdam, Amsterdam, The Netherlands, [6]Department of Cardiology, St Antonius Hospital, Sneek, The Netherlands, [7]Department of Cardiology, Dijklander Hospital, Hoorn, The Netherlands, [8]Department of Cardiology, Medical Centre Leeuwarden, Leeuwarden, The Netherlands, [9]Department of Cardiology, Noordwest Hospital Group, Alkmaar, The Netherlands, [10]Department of Human Genetics, Amsterdam University Medical Center, Amsterdam, The Netherlands, [11]Department of Genetics, University Medical Center Groningen, Groningen, The Netherlands, [12]Department of Genetics, University Medical Center Utrecht, Utrecht, The Netherlands, [13]Netherlands Heart Institute, Utrecht, The Netherlands, [14]Central Military Hospital, Utrecht, The Netherlands, and [15]Department of Cardiology, Erasmus Medical Center, Rotterdam, The Netherlands.

**These authors contributed equally to this work.*

https://doi.org/10.1016/j.hrthm.2024.02.038

Introduction

Phospholamban (*PLN*) p.(Arg14del) cardiomyopathy is an inherited disease caused by a pathogenic genetic variant in the gene encoding the phospholamban protein.[1,2] This causes this protein to misfold, which in turn causes defects in the regulation of the sarcoplasmic reticulum Ca^{2+} pump.[3] This disturbance in the Ca^{2+} homeostasis of the cardiomyocyte eventually affects the composition of cardiac tissue, resulting in structural abnormalities such as cardiac fibrosis that cause, among others, distinct electrocardiographic changes (low QRS voltage in the extremity leads and negative T waves).[4–6]

The pathogenic *PLN* p.(Arg14del) variant is associated with an arrhythmogenic or dilated cardiomyopathy characterized by progressive heart failure, malignant ventricular arrhythmia (MVA), and sudden cardiac death.[7] All of these characteristics may already have occurred at a young age, but not all carriers of this genetic variant have symptoms because of its incomplete penetrance. The *PLN* p.(Arg14del) genetic variant is a founder mutation in The Netherlands; its prevalence is estimated to be 1:500–1000 in large parts of the country. It has also been identified in several other countries, including Spain, Greece, Vietnam, China, Japan, Canada, and the United States.[8,9] The relatively high prevalence in The Netherlands enables the compilation of uniquely large data sets.

There is no evidence-based disease-modifying therapy available for *PLN* p.(Arg14del) cardiomyopathy, although implantation of an implantable cardioverter-defibrillator (ICD) may improve outcomes. Affected patients are currently treated according to general clinical guidelines, with risk score algorithms being used to identify carriers at particular risk of MVA. The latest validated risk score algorithm uses data from Holter registration, electrocardiography, echocardiography, and cardiac magnetic resonance imaging.[1]

Current prediction models use manual interpretation of the electrocardiogram (ECG), but reports have shown that deep neural networks, a type of artificial intelligence (AI), can be trained to discover more complex patterns in ECGs to diagnose *PLN* p.(Arg14del) cardiomyopathy.[10,11] Although the need for large data sets and the lack of interpretability were former common drawbacks of deep learning, a novel technique that uses a variational auto-encoder (the FactorECG) broadens the applicability of deep neural networks to much smaller data sets while also providing improved explainability (ie, explaining which ECG morphology is associated with the outcome).[12–14] The aim of this study was to evaluate whether this explainable deep learning–based approach could be implemented to assess the risk of MVA using only ECG data, allowing clinicians to make more informed decisions about patient management while simultaneously reducing the total health care burden of this disease.

Abbreviations

AI: artificial intelligence

AUPRC: area under the precision recall curve

AUROC: area under the receiver operating curve

DNN: deep neural network

HR: hazard ratio

ICD: implantable cardioverter-defibrillator

LVEF: left ventricular ejection fraction

MVA: malignant ventricular arrhythmia

NRI: net reclassification improvement

VAE: variational auto-encoder

Methods

Study population and clinical data acquisition

All index patients and relatives carrying the *PLN* p.(Arg14del) variant were identified from a large nationwide registry. Patients who were genetically evaluated in the University Medical Center Utrecht, University Medical Center Groningen, and Amsterdam University Medical Center between 2009 and 2020 were included in the study. Clinical data were collected by chart review from the first clinical contact until last follow-up in both the university and non-university medical centers. Data acquired within 1 year of the first clinical contact and before the first event of MVA were used for training the algorithm. For additional analyses, all ECGs before the first event or end of follow-up were considered. Design and detailed data collection of the nationwide registry have been described in detail before.[15] This study followed the Code of Conduct and the Use of Data in Health Research and was approved by local ethics or institutional review boards.

Electrocardiographic data acquisition

All raw 10-second 12-lead ECGs of the included patients were extracted from the MUSE ECG system (MUSE version 8; GE Healthcare, Chicago, IL) from the 3 university medical centers and resampled to 500 Hz using linear interpolation, if necessary. All ECGs were converted into median beats by aligning all primary QRS complexes (eg, excluding premature ventricular complexes) and taking the median voltage.[16]

Clinical outcomes

The primary outcome of MVA was defined, as previously, as a composite of sustained ventricular tachycardia (>30 seconds or terminated electrically or pharmacologically), ventricular fibrillation, appropriate ICD intervention, or (aborted) sudden cardiac death.[1]

Explainable deep neural network

A recently developed approach that uses a deep neural network to learn explainable features from the 12-lead median beat ECG was employed. These features are explainable in the sense that the clinician obtaining an output from the deep neural network can visualize the ECG morphology that was associated with the outcome.[12] In this approach, a generative deep neural network, called variational auto-encoder (VAE), is used to learn the underlying generative factors of the ECG without any assumptions. This VAE consists of 3 parts—an encoder, the FactorECG (32 continuous factors),

and a decoder—and was pretrained by learning to reconstruct 1,144,331 ECGs of 251,473 patients using only the 32 factors. For training of the VAE, we used all ECGs obtained in the University Medical Center Utrecht across different departments between 1991 and 2000. Overlap in the pretraining cohort and patients included in this study was negligible at 0.01% and could not influence the results because the VAE was trained unsupervised (ie, without any knowledge of the MVA outcome). After training, the pretrained encoder can be used to convert any median beat ECG into its FactorECG, the distinctive set of 32 factors that represent that ECG. In the development study of the FactorECG, we have shown that only 21 factors encode relevant information about the ECG morphology.[12] In this analysis, we used these 21 continuous factors as input to the Cox and logistic regressions models (Figure 1). The VAE tries to reconstruct the 12-lead median beats, and when the reconstruction is very different from the original, this is marker of poor ECG quality or encoding. Therefore, we excluded all ECGs with a Pearson correlation between original and reconstructed ECG below 0.5.

The individual ECG factors can be made explainable on both the model and individual patient level. This was done on the model level by varying the values of the factors individually between −3 and 3 while generating the median beat ECG using the decoder. As the other factors are kept constant, the individual influence of that factor on the ECG morphology can be visualized. Patient-level explanations can be obtained by investigating the FactorECG values of that specific ECG and the coefficients of the prediction model. In this way, we could determine which factors were important in a specific patient to make the prediction. Interactive visualizations of the model are available on https://pln.ecgx.ai. The architecture and training procedures for the FactorECG have been described in detail before.[12]

Predictor variables

Three different sets of predictors were evaluated and compared. Two ECG-only predictor sets, 1 baseline with the accepted conventional ECG criteria (number of leads with negative T waves and presence of low QRS voltage) and 1 with the standardized FactorECG values, were compared with the predictor set used in the multimodal prediction model (the 2 conventional ECG criteria, number of premature ventricular complexes on Holter monitoring and left ventricular ejection fraction [LVEF]).[1] Given the low number of events in this cohort, we selected 12 of the 21 ECG factors most associated with a reduced LVEF in a previous study to achieve at least 5.84 events per predictor.[12] These ECG factors were chosen as reduced LVEF was shown to be a strong predictor for MVA.[1] This number was optimized for the proposed analysis, expected data set, and model fit using an approach as

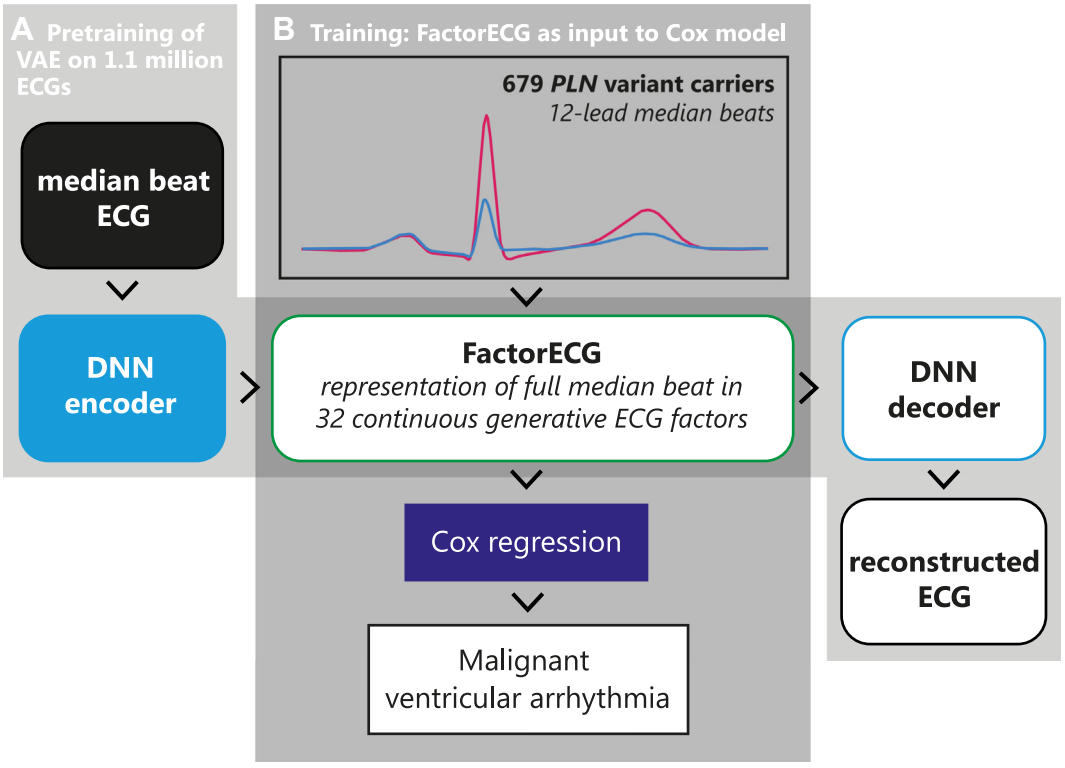

Figure 1
Schematic overview of the applied deep learning–based strategy. In the pretraining phase, the variational auto-encoder (VAE) is trained on a data set of 1.1 million median beat electrocardiograms (ECGs) to learn to reconstruct the ECG as accurately as possible in 32 variables (the FactorECG; **A**). In the training phase, the pretrained VAE is used to convert the *PLN* variant carrier ECGs into their FactorECG (**B**). Of these, 6 ECG factors that were associated with reduced ejection fraction in a previous study were selected and used in a Cox regression model to predict malignant ventricular arrhythmia. The pretrained decoder can be used to visualize which ECG features were important for prediction. DNN = deep neural network.

proposed by Riley and coworkers[17] with the R package *pmsampsize*. Detailed definitions of all predictor variables have been described before.[1]

Clinical utility

Potential consequences of using the different prediction models to determine ICD implantation with different thresholds for 5-year risk of MVA were explored. For each model and threshold, the model was used to determine which patients would receive an ICD implantation. Afterward, we labeled carriers who had an ICD implanted and experienced MVA as true positives as these carriers might have suffered from sudden cardiac death without an ICD. On the other hand, carriers who had no ICD implanted and also did not experience MVA were labeled true negatives; in these carriers, we correctly refrained from implanting an ICD, given the risk of adverse effects such as inappropriate shock, infection, collapsed lung, and others. Following from this, carriers who had an ICD but did not experience MVA or had no ICD but did experience MVA were labeled false positives and false negatives.

In addition to the 3 predictor sets, we evaluated a 2-step approach whereby only patients with a high predicted risk by the ECG-only FactorECG model were referred for additional diagnostics. With this approach, fewer additional diagnostics might be needed as not all diagnostic tests need to be performed on all patients at every follow-up appointment. In that subgroup, we simulated that echocardiography and 24-hour Holter monitoring were performed, and if a carrier had an LVEF <50% or >500 premature ventricular complexes per 24 hours on Holter monitoring, an ICD was implanted. The risk threshold for additional diagnostics was chosen at the best tradeoff of positive and negative predictive value in the current cohort.

Statistical analysis

Multivariable Cox proportional hazards models were used to evaluate the effect of the 3 different predictor sets on the risk of MVA while taking the time-to-event into account. For all models, the proportional hazards assumption was verified, and nonlinear relationships were investigated by natural cubic splines. Multivariable hazard ratios (HRs) were reported to investigate the effect of the different predictors on MVA. As the ECG factors were standardized, the HR was also used as a measure of importance for the individual ECG factors. Backward selection by the Akaike information criterion was used to achieve the sparsest model for the FactorECG predictor set.

As a result of the retrospective design, there were missing values in some predictor variables. Missing data were considered missing at random, and multiple imputation using chained equations was performed (with all characteristics from Table 1 and the ECG factors). Given a mean proportion of missing values of approximately 30%, we generated 30 imputed data sets.[18] Results on the imputed data sets were pooled with Rubin's rules.

Table 1 Baseline characteristics of the study population

Characteristic	Missing	Overall (N = 679)
Patient demographics		
Age, y	0 (0)	42 (27–55)
Male sex	0 (0)	294 (43)
Proband	6 (1)	113 (17)
History		
First-degree family member with MVA	0 (0)	91 (13)
NYHA class >I	0 (0)	62 (9.1)
Electrocardiography		
Ventricular rate, beats/min	228 (34)	71 (63–81)
PR duration, ms	239 (35)	148 (132–164)
QRS duration, ms	228 (34)	86 (80–98)
Corrected QT duration, ms	228 (34)	411 (398–430)
No. of leads with negative T waves	120 (18)	1 (0–2)
Low-voltage ECG	61 (9)	95 (15)
NSVT on Holter monitoring	0 (0)	67 (10)
24-hour PVC count >500	273 (40)	125 (31)
Imaging		
LVEF	224 (33)	54 (48–60)
RVEF	146 (22)	65 (50–65)
MRI LGE	417 (61)	77 (29)
Outcomes		
MVA	0 (0)	72 (10)
Duration of follow-up, y	0 (0)	4.3 (1.7–7.4)

Categorical variables are presented as number (percentage). Continuous variables are presented as median (interquartile range).
ECG = electrocardiogram; LGE = late gadolinium enhancement; LVEF = left ventricular ejection fraction; MRI = magnetic resonance imaging; MVA = malignant ventricular arrhythmia; NSVT = nonsustained ventricular arrhythmia; NYHA = New York Heart Association; proband = first member of a family in whom the *PLN* p.(Arg14del) variant was found; PVC = premature ventricular complex; RVEF = right ventricular ejection fraction.

Internal validation of the discriminatory performance (as measured by Harrell's C statistic) was performed by a bootstrap-based optimism estimation technique. Here, all model development steps (including multiple imputation and pooling using Rubin's rules) were repeated on 500 bootstrap samples.[19] Each new pooled model was tested on the original data, and the optimism was defined as the mean difference in the C statistic between the original and bootstrapped data sets. This value is subtracted from the apparent performance measure (ie, the C statistic in original data from a model fitted on the original data).[20] These optimism-corrected measures have been shown to be an unbiased estimate of the generalizability of the model, without losing any data for training.[21] The bootstrap samples were also used to determine the 95% confidence intervals (CIs) around the C statistic. Permutation tests were used to compare the C statistic from the different predictors sets.

In addition, we computed a range of performance metrics at the clinically accepted risk time frame of 5 years. At this time point, the predicted 5-year risk of MVA was derived for the different models. The area under the receiver operating curve (AUROC) and area under the precision recall curve (AUPRC)

were computed for this predicted risk. Moreover, net reclassification improvement (NRI), sensitivity, specificity, and positive and negative predictive values were derived at 3 different prespecified clinically used probability cutoffs: 5%, 7.5%, and 10%.[22] Finally, we investigated the robustness of the algorithm for different ECGs of the same patient by calculating the median of the predicted 5-year risk of MVA's SD per patient in a single year.

Baseline characteristics were expressed as mean ± SD or median with interquartile range (IQR), where applicable. All statistical analyses were performed with Python version 3.9. The Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis statement for the reporting of diagnostic models was followed where applicable.[23]

Results

Study population

The total cohort consisted of 1067 *PLN* p.(Arg14del) variant carriers. After exclusion of patients with MVA at baseline (n = 65 [6%]), patients without follow-up data (n = 221 [21%]), and patients without any baseline test in the participating centers (n = 102 [9.6%]), 679 *PLN* carriers were included in the analysis. Raw 12-lead ECG waveforms within 1 year of first presentation were available for 472 (70%) patients, and of these, 451 (96%) were of adequate quality. Performance of the pretrained VAE for the included ECGs was good, with a Pearson correlation coefficient between original and reconstructed ECG of 0.89. A total of 72 patients (10%) reached the primary outcome of MVA during a follow-up of 4.3 years (IQR, 1.7–7.4 years). The composite consisted of appropriate ICD therapy, sustained ventricular tachycardia/ventricular fibrillation, and sudden cardiac death in 37, 26, and 9 patients, respectively. Additional baseline characteristics are shown in Table 1. Differences between patients with and without a raw 12-lead ECG are shown in Supplemental Table 1.

Model performance

The baseline ECG-only model (consisting of the number of negative T waves and low QRS voltage as predictors) predicted MVA with an optimism-corrected C statistic of 0.65 (95% CI, 0.58–0.73). The FactorECG model (consisting of 7 ECG factors) outperformed the baseline model with an optimism-corrected C statistic of 0.79 (95% CI, 0.76–0.83; $P < .001$) and was comparable to the multimodal prediction model (optimism-corrected C statistic of 0.83 [95% CI, 0.79–0.88]; $P = .054$).

On investigating performance by the clinically accepted 5-year predicted risk of MVA, the AUROC and AUPRC were 0.67, 0.86, and 0.89 and 0.12, 0.27, and 0.30 for the baseline ECG-only, FactorECG, and multimodal prediction models, respectively. The overall NRI for the FactorECG model compared with the baseline ECG-only was 30 (95% CI, 13–49), with 42% (95% CI, 26–59) more patients with MVA correctly moved upward to the group with a risk >7.5%. On comparing the FactorECG model with the multimodal prediction model, the NRI was 6.3% (95% CI, −5.3 to 18), with 6.4% (95% CI, 0.0–14.5) more patients with MVA moved upward to the group with a risk >7.5%. This indicates that the FactorECG model identifies more patients with MVA than the baseline ECG-only model, without missing cases compared with the multimodal model. An overview of the AUROC, AUPRC, NRI, sensitivity, specificity, and positive and negative predictive values at different probability thresholds for all predictor sets can be found in Table 2. With use of the 3 prediction models to stratify carriers in 4 quartiles by their predicted 5-year risk of MVA, a clear distinction in risk between the groups can be observed for the FactorECG and multimodal model (Figure 2A–C). In the lowest 3 risk groups, almost no events are observed for these models, whereas the baseline ECG-only model is not able to distinguish groups without events.

On taking multiple ECGs per patient into account, we can predict the 5-year risk of MVA for every ECG. In total, 3849 ECGs were available of 514 individual patients. The median SD of these predicted probabilities within an individual patient, grouped by year, is 0.02 (IQR, 0.008–0.05). A histogram of the predicted probabilities and an overview of the predicted probabilities over time can be found in Supplemental Figures 1 and 2.

The most important predictors in the FactorECG model were F_1 (inferolateral ST-segment and T-wave morphology; HR, 0.56 [0.39–0.81]) and F_5 (inferolateral negative T waves; HR, 2.48 [1.70–3.61]); in the multimodal model, LVEF (HR, 0.96 per 1% increase [95% CI, 0.94–0.98]) and 24-hour premature ventricular complex count (HR, 1.33 per 1 log increase [95% CI, 1.16–1.55]) were most predictive. Multivariable HRs and CIs for all prediction models are shown in Table 3. Similar HRs were found in predicting only appropriate ICD therapy and only ventricular tachycardia/ventricular fibrillation and sudden cardiac death. Univariable HRs and HRs for the separate end points in the composite can be found in Supplemental Tables 2 and 3.

Clinical applicability

Different scenarios with varying thresholds for the 5-year predicted risk of MVA to determine which patients should receive an ICD implantation were investigated (Figure 3). At a clinically accepted 5-year risk threshold of 5% (1% risk per year), the baseline ECG-only model performed the worst with a sensitivity of 80% and specificity of only 36%. The FactorECG model outperformed the baseline model with a sensitivity of 92% and specificity of 52%, whereas the multimodal model had a higher specificity of 62% and higher sensitivity of 95%. This indicates that when the FactorECG model is used at a 5% threshold, a similar number of patients that will experience MVA without having an ICD implanted are missed, at the cost of implanting more ICDs. A similar trend is observed at the higher 5-year risk thresholds of 7.5% and 10%, in which significantly fewer ICDs are implanted but with more false negatives.

Next to the implementation of the models alone, a more clinically applicable 2-step approach was investigated

Table 2 Prognostic performance measures for the predicted risk of malignant ventricular arrhythmia at 5 years for the different predictor sets (A–C) at 3 different probability cutoffs and the 2-step approach (D)

	A. Baseline ECG-only			B. FactorECG			C. Multimodal			
	5%	7.5%	10%	5%	7.5%	10%	5%	7.5%	10%	D. 2-step
AUROC	0.68			0.86			0.89			NA
AUPRC	0.12			0.27			0.30			NA
Sensitivity	80	52	45	92	90	82	95	90	78	90
Specificity	36	75	83	52	64	74	62	71	78	75
PPV	7	12	14	11	14	16	13	16	18	18
NPV	97	96	96	99	99	98	100	99	98	99
NRI	Ref	Ref	Ref	34[a]	30[a]	28[a]	13[a]	6	7	NA
NRI_e	Ref	Ref	Ref	16	38[a]	36[a]	4	1	3	NA
NRI_{ne}	Ref	Ref	Ref	19[a]	−8[a]	−8[a]	8[a]	5[a]	4[a]	NA

The net reclassification improvement (NRI) was computed in comparison to the predictor set A for predictor set B and in comparison to the predictor set B for predictor set C.

AUPRC = area under the precision recall curve; AUROC = area under the receiver operating curve; NA = not applicable; NPV = negative predictive value; NRI_e = net reclassification improvement for patients with an event; NRI_{ne} = net reclassification improvement for patients without an event; PPV = positive predictive value.

[a]Statistically significant.

(Figure 4). With this simulated approach, all patients first have an ECG; and then, only in the high-risk patients as predicted by the FactorECG model, echocardiography and Holter monitoring data are needed. A threshold to determine which patients were high risk of 7.5% was used as this provided the best tradeoff of positive and negative

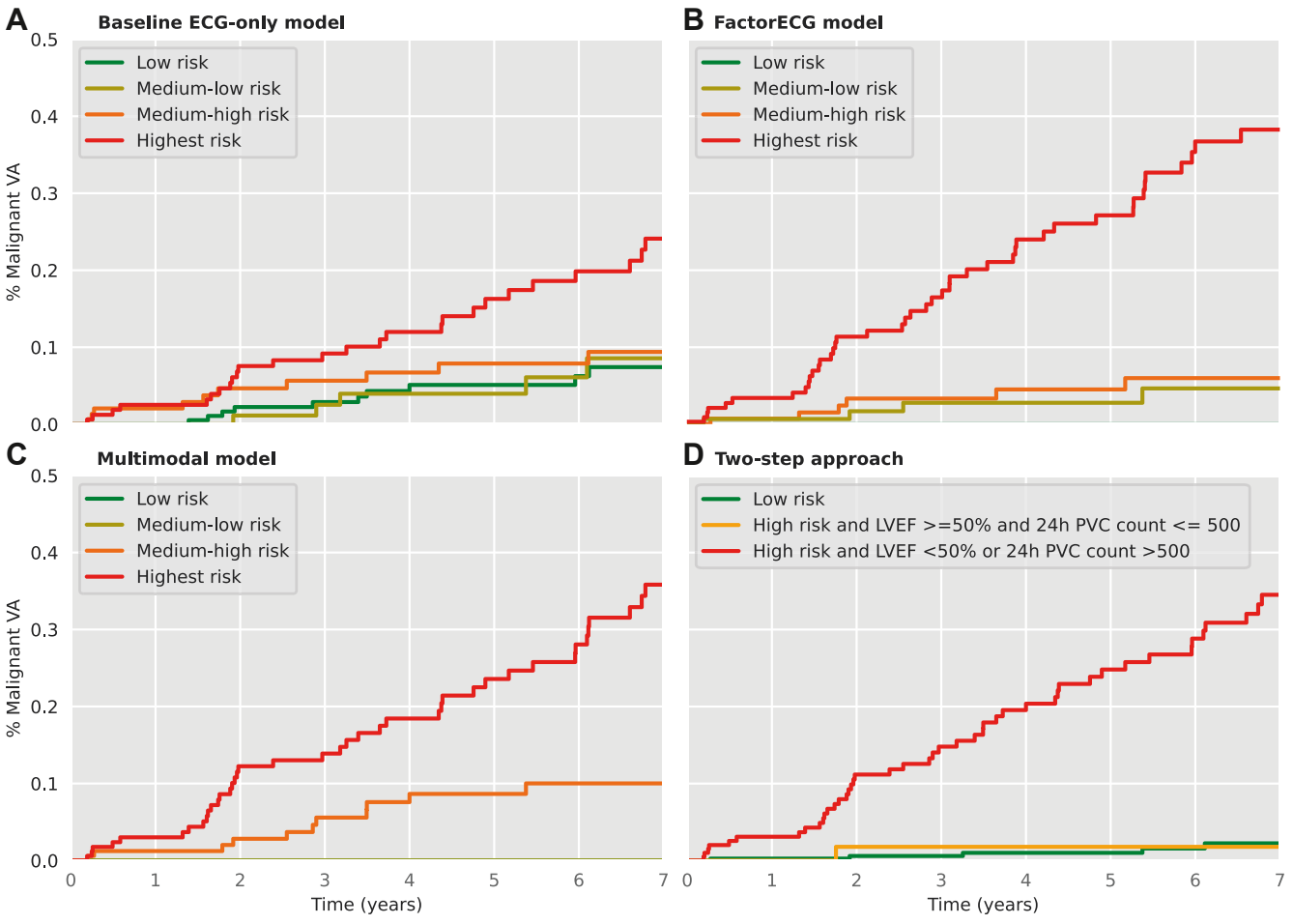

Figure 2

Kaplan-Meier plots for the different predictor sets (**A–C**) and the 2-step approach (**D**). For the prediction models (**A–C**), the 5-year predicted risk of malignant ventricular arrhythmia (VA) is split in 4 quartiles (risk groups). For the 2-step approach, an approach was simulated whereby only patients with a high predicted risk (>7.5%) using the FactorECG model were referred for additional diagnostics. In that subgroup, we simulated that echocardiography and 24-hour Holter monitoring were performed. ECG = electrocardiogram; LVEF = left ventricular ejection fraction; PVC = premature ventricular complex.

Table 3 Hazard ratios, confidence intervals, and *P* values for the different predictor sets evaluated in multivariable Cox proportional hazards models

Predictor	HR (95% CI)	P value
Baseline ECG-only model		
No. of leads with negative T waves	1.12 (1.00–1.24)	.03
Low QRS voltage	3.52 (2.07–5.97)	<.001
FactorECG model		
Factor 1	0.56 (0.39–0.81)	<.001
Factor 5	2.48 (1.70–3.61)	<.001
Factor 8	1.24 (0.93–1.65)	.14
Factor 12	1.50 (1.10–2.04)	.01
Factor 25	0.73 (0.54–1.01)	.05
Factor 26	0.74 (0.55–0.99)	.043
Factor 30	0.76 (0.58–0.99)	.046
Multimodal model		
No. of leads with negative T waves	1.10 (0.89–1.23)	.10
Low QRS voltage	1.76 (0.98–3.20)	.06
LVEF (per % increase)	0.96 (0.94–0.98)	<.001
24-hour PVC count (per 1 log increase)	1.34 (1.16–1.54)	<.001

CI = confidence interval; HR = hazard ratio; other abbreviations as in Table 1.

predictive value (ie, referring the least amount of patients without missing too many patients with MVA). In applying this risk threshold, only 39% of patients need to be referred. In this referred group, we simulated an ICD implantation when either the LVEF was <50% or >500 premature ventricular complexes per 24 hours on Holter monitoring were recorded. This 2-step approach outperformed all other models with a sensitivity of 88% and specificity of 76% (Figure 3).

Model explainability

F_1 (inferolateral ST-segment and T-wave morphology) and F_5 (inferolateral negative T waves) were significantly associated with the risk of MVA during follow-up, with more negative values corresponding to a higher risk for F_1 and more positive values for F_5. Both these factors represent the shape of the inferolateral ST segment and T wave and are significantly correlated with each other in this population (Pearson $r = -0.39$; $P < .001$). The factor traversals of a combined change in F_1 and F_5 showed that this combination represents a change in ECG morphology from normal QRS voltage and repolarization toward lower QRS voltage and inferolateral symmetrical negative T waves without any ST deviation (Figure 5). Interestingly, the effect of this morphologic change was nonlinearly related with the predicted 5-year risk of MVA, and the risk already exceeded 5% when the T waves are still positive (Figure 5). Other factors were not significantly correlated (Pearson $r < 0.22$ for all) in this population, and their factor traversals are therefore shown for each factor individually in Supplemental Figures 3–7. The visualizations show that increased PR interval (factor 8), reduced R-wave height in V_2 through V_4 (factor 12), right bundle branch block–like QRS morphology (factor 25), inferolateral T-wave morphology (factor 26), and increased QT interval (factor 30) are also associated with increased risk of MVA.

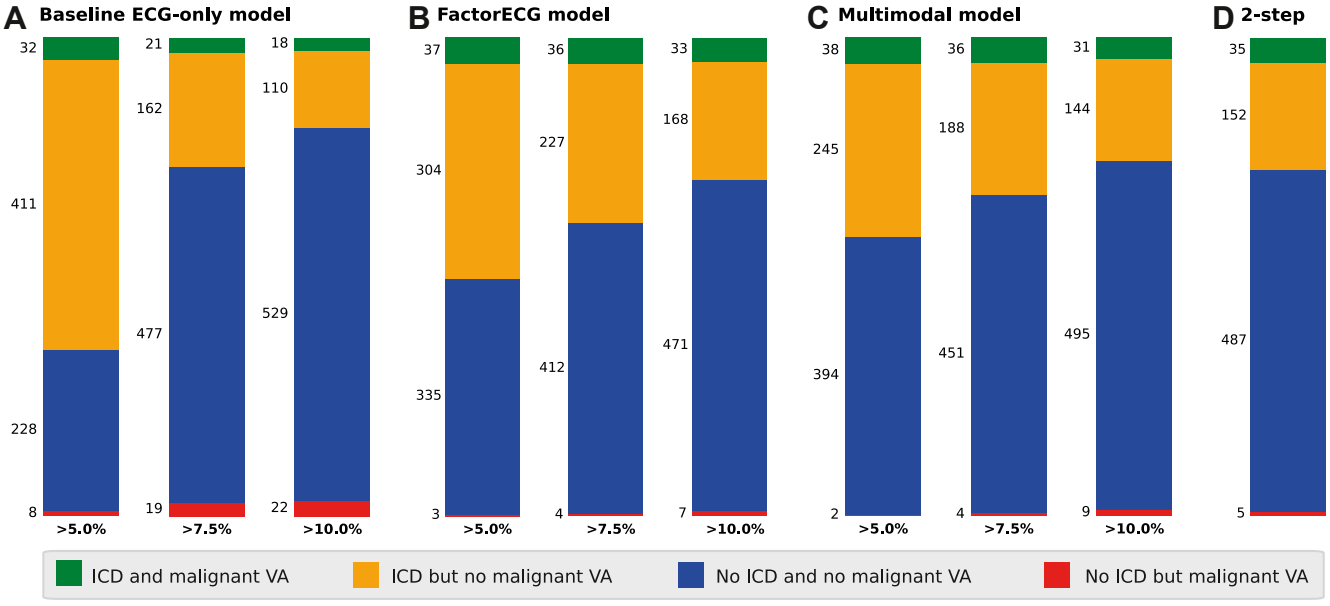

Figure 3
Clinical utility plots for the different predictor sets (**A–C**) and the 2-step approach (**D**). The bars represent the clinical implications of using different 5-year risk of malignant ventricular arrhythmia (VA) thresholds for the decision to implant an implantable cardioverter-defibrillator (ICD). For the 2-step approach, an approach was simulated whereby only patients with a high predicted risk (>7.5%) using the FactorECG model were referred for additional diagnostics. In that subgroup, we simulated that echocardiography and 24-hour Holter monitoring were performed, and if a carrier had a left ventricular ejection fraction <50% or >500 premature ventricular complexes per 24 hours on Holter monitoring, an ICD was implanted. ECG = electrocardiogram.

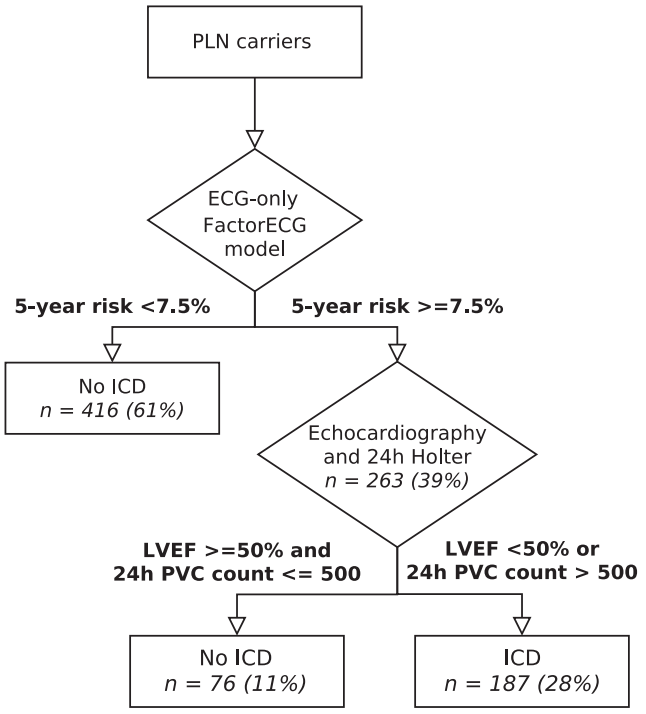

Figure 4
Overview of the 2-step approach. An approach was simulated in which all *PLN* variant carriers first have an electrocardiogram (ECG) only. This ECG is evaluated by the FactorECG prediction model, and only the high-risk patients are referred for additional diagnostics (echocardiography and 24-hour Holter monitoring). When carriers had a left ventricular ejection fraction (LVEF) <50% or a premature ventricular complex (PVC) count >500 per 24 hours on Holter monitoring, an implantable cardioverter-defibrillator (ICD) was implanted.

Discussion

This study shows that an explainable deep learning–based approach using only ECG data is able to predict the risk of MVA with an optimism-corrected C statistic 0.79 (95% CI, 0.75–0.85) in a large cohort of *PLN* p.(Arg14del) carriers. Addition of echocardiographic and Holter monitoring data in the group with high predicted risk based on the FactorECG improved predictive ability further (ie, a 2-step approach), outperforming the current multimodal model in all patients. Such a 2-step approach could allow more efficient risk stratification of *PLN* p.(Arg14del) carriers; reduce the burden of monitoring visits for these carriers; and lead to a significant decrease in costs by reducing the number of visits, diagnostics, and ICD implantations. Deep learning–based ECG analysis may enhance the possibilities for remote monitoring of genetic variant carriers. An online tool to convert any ECG into its FactorECG and to predict prognosis in *PLN* patients is available through https://pln.ecgx.ai.

Clinical applicability and prior studies

This is the first study attempting risk stratification in carriers of the *PLN* p.(Arg14del) genetic variant using only ECG data. The current best practice in risk stratification of known *PLN* p.(Arg14del) carriers involves the use of a risk score combining structural, electrophysiologic, and functional parameters.[1] This multimodal algorithm has an optimism-corrected C statistic of 0.83 (95% CI, 0.79–0.88) in the current analysis. Whereas an ECG-only model containing conventional ECG features of *PLN* cardiomyopathy (low QRS voltage and negative T waves) was not able to reach similar predictive performance (optimism-corrected C statistic of 0.65 [95% CI, 0.58–0.73]), the deep learning–based ECG-only model did perform comparably (optimism-corrected C statistic of 0.79 [95% CI, 0.76–0.83]). Net reclassification analysis confirmed that the FactorECG algorithm outperformed the baseline ECG-only model at all risk thresholds, without missing patients with MVA within 5 years compared with the multimodal algorithm (Table 2).

Clinically, such an ECG-only algorithm could be used in a 2-step approach involving a first pass with the ECG model alone, followed by additional diagnostics in patients deemed at risk of MVA. The multimodal algorithm would not be feasible for such a 2-step approach because it requires many different modalities of diagnostic data. If acceptable negative predictive values can be achieved with only ECG (possibly at home or by the general practitioner), the large burden of monitoring visits could be reduced, especially for asymptomatic carriers. Whereas the conventional ECG-only model did not reach adequate negative predictive values to be usable in such an approach, the FactorECG model was able to reach a negative predictive value of 99% in 61% of the patients at a 5-year risk threshold of 7.5%. As visualized in Figure 3, this results in more accurate risk prediction than either method alone as well as being more accurate than the multimodal model in all patients.

The presence of the *PLN* p.(Arg14del) genetic variant is established by genotyping of potentially affected index patients presenting with related signs and symptoms, followed by genetic cascade screening in close family members. Both Bleijendaal and coworkers[10] and van de Leur and coworkers[11] have shown that a deep learning–based approach may aid in the diagnosis of the genetic variant in the general population as well, aiding in the identification of the index patients. This study builds on their results by providing the risk stratification required for optimal management after initial diagnosis.

Explaining the AI algorithm

The term *black box* is often used to describe models resulting from the extensive training of a machine learning algorithm.[24] These models may become too complex to be interpreted by humans using them to reach an output from a given input, which in turn may cause a level of distrust in the output.[25] Our approach provides improved explainability by allowing clinicians to visualize the influence of specific median beat ECG morphology on the predictions.[12,26] Previous studies have shown that a similar approach with the FactorECG can be used to predict risk of MVA in patients with dilated cardiomyopathy and outcomes in cardiac resynchronization therapy recipients.[13,14]

Our visualizations confirm that the FactorECG prediction is mostly based on known *PLN* cardiomyopathy ECG features

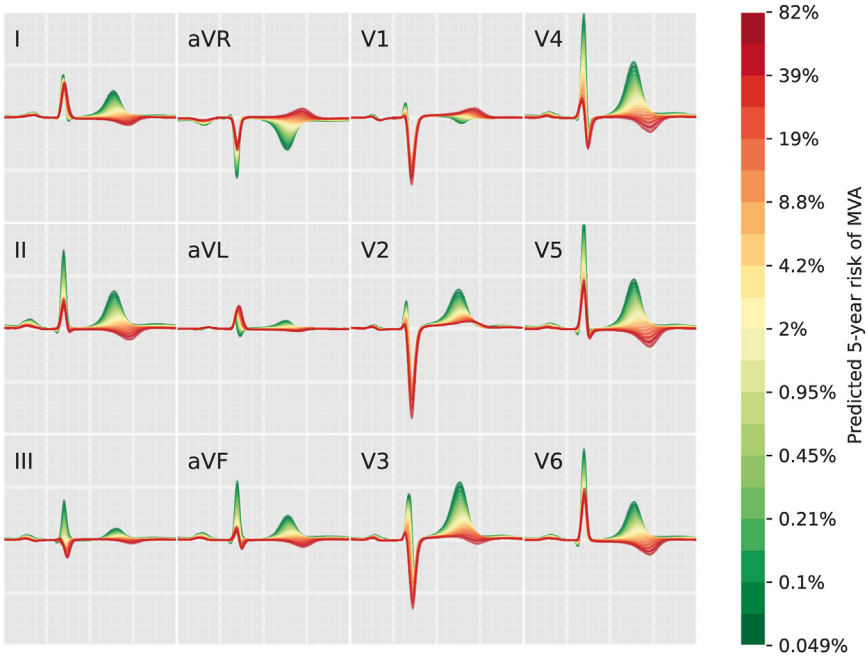

Figure 5
Factor traversals for the 2 most important electrocardiogram factors to visualize electrocardiographic features that the model used to predict malignant ventricular arrhythmia (MVA). In the current plot, we varied the values for factor 1 and factor 5 simultaneously as these factors are strongly correlated in the current population while keeping the other factors constant at their mean value. For each combination, the 5-year risk of MVA is derived by the Cox regression model and visualized.

(eg, reduced QRS voltage and inferolateral symmetrical negative T waves as represented by the combinations of F_1 and F_5), as shown in Figure 5. Interestingly, it uses these features as a continuous spectrum and already predicts a risk higher than the threshold of 5% before the appearance of negative T waves, but only with a reduced R- and T-wave height. This might explain why the model outperforms the baseline ECG-only model, as this uses binary cutoff points for QRS voltage and negative T waves. Other ECG features shown by the visualizations are an increased PR interval (F_8; Supplemental Figure 3), rSR′ in V_1 with slurred S waves inferolaterally (F_{25}; Supplemental Figure 5), and reduced lateral T-wave height (F_{26}; Supplemental Figure 6), although all borderline significant (Table 2). We expect that this direct input-output relationship makes using the algorithm a more attractive option to clinicians by increasing trust in the outcome. An interactive tool for explainability is available through https://pln.ecgx.ai.

Strengths and limitations

The main strength of this study is that the *PLN* registry allowed leveraging of a uniquely large cohort of deeply phenotyped *PLN* p.(Arg14del) carriers.[15] However, there are several limitations to this study. First, no external validation for the prediction models or the risk thresholds in the 2-step scenario analysis could be performed as there are currently no other cohorts of *PLN* p.(Arg14del) carriers available. To minimize the risk of overoptimism, we prespecified our predictor sets before the analysis, selected a limited number of predictors in every model, and performed a rigorous internal validation using a bootstrap-based resampling technique.[21] Second, the retrospective nature of the data comes with missing values; sufficient data for analysis were available for 679 of 1002 eligible patients, 451 with a baseline raw ECG of adequate quality (45% of those eligible). Missingness is mostly due to the long time span of the cohort; the first medical contact for some patients occurred in the 1990s in many different hospitals across The Netherlands and preceded the genetic diagnosis. This meant that many patients did not have any diagnostic tests or follow-up available. Moreover, we were only able to access the raw 12-lead ECGs of the patients with their first medical contact in the 3 university medical centers, leading to additional missingness. Most important, there was no difference in the occurrence of MVA between patients with and without a raw 12-lead ECG. Third, the primary outcome of MVA was defined as a composite of several end points, one of which was appropriate ICD intervention. Thus, appropriate ICD intervention was given the same weight as sudden cardiac death or ventricular fibrillation, similar to the current prediction model in *PLN* p.(Arg14del) variant carriers. This may result in an overestimation of the true 5-year risk of sudden cardiac death because not all appropriate ICD interventions equate with cardiac arrest. Removing ICD recipients, however, would not be a valid alternative; this would mean all high-risk carriers are removed, leading to selection bias and underestimation of the true sudden cardiac death risk. Reassuringly, HRs for the ECG factors were similar in evaluating the end points in the composite (appropriate ICD therapy, ventricular tachycardia/ventricular fibrillation, and sudden cardiac death) separately (Supplemental Figure 3).

Finally, we were not able to provide metrics for the 10-year predicted risk of MVA as we have a median follow-up duration of only 4.3 years. Given that the ICD battery life is around 10 years, false positives and true negatives (ie, patients who did not experience MVA) are relative as they can have MVA after 5 years.

Future perspectives

A machine learning–based approach could aid in both diagnosis of the cardiomyopathy-associated variant and risk stratification to help clinicians more efficiently organize their health care system. Currently, the *PLN* p.(Arg14del) genetic variant is mainly prevalent in The Netherlands. As more affected families and relatives are identified, both in The Netherlands and abroad, the health care burden of diagnosis and risk assessment will rise. This is of importance because with rising health care costs and the high barriers to accessing health care in some nations, deep learning–based ECG analysis can provide a remote solution to manage this group of patients. Moreover, the approach in this study may also be of use for researchers studying other uncommon types of (genetic) cardiomyopathy.

Conclusion

An ECG-only explainable deep learning–based algorithm is able to predict the occurrence of MVA in *PLN* p.(Arg14del) carriers with an optimism-corrected C statistic of 0.79 (95% CI, 0.75–0.85), which could allow an alternative stratification relying on the ECG only, precluding additional diagnostics and follow-up. Such a 2-step approach could reduce the burden of monitoring visits for *PLN* p.(Arg14del) carriers and lead to a significant decrease in costs by reducing the number of visits, diagnostics, and ICD implantations.

Appendix

Supplementary data

Supplementary data associated with this article can be found in the online version at https://doi.org/10.1016/j.hrthm.2024.02.038.

Acknowledgments

We would like to thank Anton Oomen, MD, cardiologist, and the St Antonius Medical Centre in Sneek, The Netherlands, for their hospitality and assistance with data collection.

Funding Sources: This work was supported by the Dutch Heart Foundation and co-financed by The Netherlands Organisation for Health Research and Development (ZonMw, No. 104021004) and the Dutch Heart Foundation (No. 2019B011). Support from The Netherlands CardioVascular Research Initiative, an initiative supported by the Dutch Heart Foundation [CardioVasculair Onderzoek Nederland (CVON) projects: PREDICT2 2018-30, eDETECT 2015-12, and Double Dose 2020B005], is acknowledged. Additional financial support was received from CURE-PLaN, a network funded by the Leducq Foundation. Dr te Riele was supported by the ZonMW Off Road grant 2021.

Disclosures: The UMC Groningen, which employs several of the authors, received research grants and/or fees from AstraZeneca, Abbott, Boehringer Ingelheim, Cardior Pharmaceuticals GmbH, Ionis Pharmaceuticals, Inc, Novo Nordisk, and Roche (outside the submitted work). Rudolf A. de Boer has had speaker engagements with Abbott, AstraZeneca, Bayer, Bristol Myers Squibb, Novartis, and Roche (outside the submitted work). Rutger R. van de Leur and René van Es are cofounders, shareholders, and board members of Cordys Analytics B.V., a spin-off of the UMC Utrecht that has licensed AI-ECG algorithms, not including the algorithm studied in the current manuscript. The UMC Utrecht receives royalties from Cordys Analytics for potential future revenues. Pieter A. Doevendans is founder and shareholder of HeartEye B.V., an ECG-device company. The other authors declare that there is no conflict of interest.

Authorship: All authors attest they meet the current ICMJE criteria for authorship.

Code Availability: Programming code to train and use the FactorECG model is available through https://github.com/rutgervandeleur/ecgxai. An online tool to convert any ECG into its FactorECG and to predict prognosis in *PLN* patients is available through https://pln.ecgx.ai.

Data Availability: Data sharing requests will be considered on reasonable request to the corresponding author if accompanied by clear research objectives, a statistical analysis plan, and data requirements. If approved, information will be provided under the terms of a data sharing agreement.

Address reprint requests and correspondence: Dr R.R. van de Leur, Department of Cardiology, University Medical Center Utrecht, Heidelberglaan 100, 3584 CX Utrecht, The Netherlands. E-mail address: r.r.vandeleur@umcutrecht.nl

References

1. Verstraelen TE, van Lint FH, Bosman LP, et al. Prediction of ventricular arrhythmia in phospholamban p.Arg14del mutation carriers—reaching the frontiers of individual risk prediction. Eur Heart J 2021;42:2842–2850.
2. Zwaag PA, Rijsingen IA, Asimaki A, et al. Phospholamban R14del mutation in patients diagnosed with dilated cardiomyopathy or arrhythmogenic right ventricular cardiomyopathy: evidence supporting the concept of arrhythmogenic cardiomyopathy. Eur J Heart Fail 2012;14:1199–1207.
3. Rijdt WP, Tintelen JP, Vink A, et al. Phospholamban p.Arg14del cardiomyopathy is characterized by phospholamban aggregates, aggresomes, and autophagic degradation. Histopathology 2016;69:542–550.
4. de Brouwer R, Meems LM, Verstraelen TE, et al. Sex-specific aspects of phospholamban cardiomyopathy: the importance and prognostic value of low-voltage electrocardiograms. Heart Rhythm 2022;19:427–434.
5. Hof IE, van der Heijden JF, Kranias EG, et al. Prevalence and cardiac phenotype of patients with a phospholamban mutation. Neth Heart J 2019;27:64–69.
6. Haghighi K, Gardner G, Vafiadaki E, et al. Impaired right ventricular calcium cycling is an early risk factor in R14del-phospholamban arrhythmias. J Pers Med 2021;11:502.
7. van Rijsingen IA, van der Zwaag PA, Groeneweg JA, et al. Outcome in phospholamban R14del carriers: results of a large multicentre cohort study. Circ Cardiovasc Genet 2014;7:455–465.
8. Cheung CC, Healey JS, Hamilton R, et al. Phospholamban cardiomyopathy: a Canadian perspective on a unique population. Neth Heart J 2019;27:208–213.
9. Jiang X, Xu Y, Sun J, Wang L, Guo X, Chen Y. The phenotypic characteristic observed by cardiac magnetic resonance in a PLN-R14del family. Sci Rep 2020;10:16478.
10. Bleijendaal H, Ramos LA, Lopes RR, et al. Computer versus cardiologist: is a machine learning algorithm able to outperform an expert in diagnosing a

phospholamban p.Arg14del mutation on the electrocardiogram? Heart Rhythm 2021;18:79–87.
11. van de Leur RR, Taha K, Bos MN, et al. Discovering and visualizing disease-specific electrocardiogram features using deep learning. Circ Arrhythm Electrophysiol 2021;14:e009056.
12. van de Leur RR, Bos MN, Taha K, et al. Improving explainability of deep neural network–based electrocardiogram interpretation using variational auto-encoders. Eur Heart J Digit Health 2022;3:390–404.
13. Wouters PC, van de Leur RR, Vessies MB, et al. Electrocardiogram-based deep learning improves outcome prediction following cardiac resynchronization therapy. Eur Heart J 2022;44:680–692.
14. Sammani A, van de Leur RR, Henkens MT, et al. Life-threatening ventricular arrhythmia prediction in patients with dilated cardiomyopathy using explainable electrocardiogram-based deep neural networks. Europace 2022;24:1645–1654.
15. Bosman LP, Verstraelen TE, van Lint FH, et al. The Netherlands Arrhythmogenic Cardiomyopathy Registry: design and status update. Neth Heart J 2019;27:480–486.
16. GE Healthcare. Marquette 12SL ECG Analysis Program. Physician's Guide. https://landing1.gehealthcare.com/rs/005-SHS-767/images/45351-MUSE-17Nov2022-6-1-Quick-Reference-Guide-LP-Diagnostic-Cardiology.pdf. Accessed November 26, 2023.
17. Riley RD, Ensor J, Snell KI, et al. Calculating the sample size required for developing a clinical prediction model. BMJ 2020;368:m441.
18. White IR, Royston P, Wood AM. Multiple imputation using chained equations: issues and guidance for practice. Stat Med 2011;30:377–399.
19. Schomaker M, Heumann C. Bootstrap inference when using multiple imputation. Stat Med 2018;37:2252–2266.
20. Harrell FE, Lee KL, Mark DB. Tutorial in biostatistics multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Stat Med 1996;15:361–387.
21. Steyerberg EW, Harrell FE, Borsboom GJ, Eijkemans MJ, Vergouwe Y, Habbema JD. Internal validation of predictive models efficiency of some procedures for logistic regression analysis. J Clin Epidemiol 2001;54:774–781.
22. Leening MJ, Vedder MM, Witteman JC, Pencina MJ, Steyerberg EW. Net reclassification improvement: computation, interpretation, and controversies: a literature review and clinician's guide. Ann Intern Med 2014;160:122–131.
23. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): the TRIPOD statement. Ann Intern Med 2015;162:55.
24. Durán JM, Jongsma KR. Who is afraid of black box algorithms? On the epistemological and ethical basis of trust in medical AI. J Med Ethics 2021;47:329–335.
25. Petch J, Di S, Nelson W. Opening the black box: the promise and limitations of explainable machine learning in cardiology. Can J Cardiol 2021;38:204–213.
26. van de Leur RR, Hassink RJ, van Es R. Variational auto-encoders improve explainability over currently employed heatmap methods for deep learning–based interpretation of the electrocardiogram. Eur Heart J Digit Health 2022;3:502–504.