Electrocardiogram-based deep learning improves outcome prediction following cardiac resynchronization therapy

Philippe C. Wouters [1]*[†], **Rutger R. van de Leur** [1][†], **Melle B. Vessies** [1], **Antonius M.W. van Stipdonk** [2], **Mohammed A. Ghossein** [3], **Rutger J. Hassink** [1], **Pieter A. Doevendans** [1,4], **Pim van der Harst** [1], **Alexander H. Maass** [5], **Frits W. Prinzen** [3], **Kevin Vernooy** [2], **Mathias Meine** [1], and **René van Es** [1]

[1]Department of Cardiology, University Medical Center Utrecht, Heidelberglaan 100, 3584 CX Utrecht, The Netherlands; [2]Department of Cardiology, Cardiovascular Research Institute Maastricht (CARIM), Maastricht University Medical Centre (MUMC+), Maastricht, The Netherlands; [3]Department of Physiology, Cardiovascular Research Institute Maastricht (CARIM), Maastricht University, Maastricht, The Netherlands; [4]Netherlands Heart Institute, Utrecht, The Netherlands; and [5]Department of Cardiology, Thoraxcentre, University of Groningen, University Medical Center Groningen, Groningen, The Netherlands

See the editorial comment for this article 'Explainable AI for ECG-based prediction of cardiac resynchronization therapy outcomes: learning from machine learning?', by Z. I. Attia and P. A. Friedman, https://doi.org/10.1093/eurheartj/ehac733.

Abstract

Aims	This study aims to identify and visualize electrocardiogram (ECG) features using an explainable deep learning–based algorithm to predict cardiac resynchronization therapy (CRT) outcome. Its performance is compared with current guideline ECG criteria and QRS$_{AREA}$.
Methods and results	A deep learning algorithm, trained on 1.1 million ECGs from 251 473 patients, was used to compress the median beat ECG, thereby summarizing most ECG features into only 21 explainable factors (FactorECG). Pre-implantation ECGs of 1306 CRT patients from three academic centres were converted into their respective FactorECG. FactorECG predicted the combined clinical endpoint of death, left ventricular assist device, or heart transplantation [c-statistic 0.69, 95% confidence interval (CI) 0.66–0.72], significantly outperforming QRS$_{AREA}$ and guideline ECG criteria [c-statistic 0.61 (95% CI 0.58–0.64) and 0.57 (95% CI 0.54–0.60), $P < 0.001$ for both]. The addition of 13 clinical variables was of limited added value for the FactorECG model when compared with QRS$_{AREA}$ (Δ c-statistic 0.03 vs. 0.10). FactorECG identified inferolateral T-wave inversion, smaller right precordial S- and T-wave amplitude, ventricular rate, and increased PR interval and P-wave duration to be important predictors for poor outcome. An online visualization tool was created to provide interactive visualizations (https://crt.ecgx.ai).
Conclusion	Requiring only a standard 12-lead ECG, FactorECG held superior discriminative ability for the prediction of clinical outcome when compared with guideline criteria and QRS$_{AREA}$, without requiring additional clinical variables. End-to-end automated visualization of ECG features allows for an explainable algorithm, which may facilitate rapid uptake of this personalized decision-making tool in CRT.

* Corresponding author. Email: p.wouters@umcutrecht.nl
† These authors shared first authorship.

Structure Graphical Abstract

Key Question

Can an explainable deep learning-based approach (FactorECG) predict outcomes and echocardiographic response after cardiac resynchronization therapy (CRT) using only a standard ECG?

Key Finding

FactorECG significantly outperformed Guideline criteria and QRS_{AREA} for prediction of CRT outcomes. It offset the need for other clinical variables and identified additional predictive ECG features besides the left bundle branch block criteria and QRS_{AREA}.

Take Home Message

FactorECG requires only a standard ECG for superior prediction of outcomes after CRT and is explainable through visualization of key ECG factors. This approach therefore allows for personalized and easy-to-use clinical decision making in CRT.
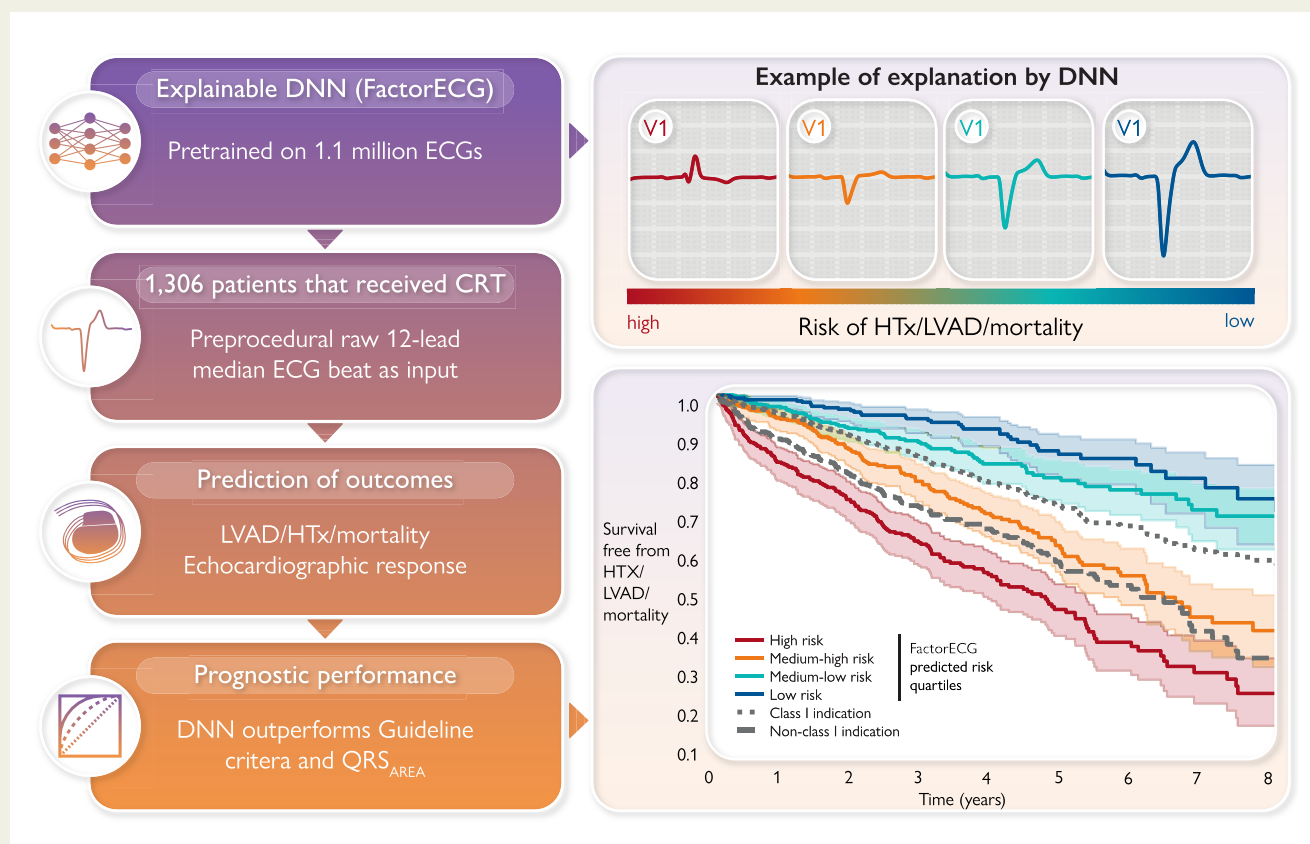
First, an artificial intelligence algorithm (variational auto-encoder) was pretrained on 1.1 million electrocardiograms (ECGs) to learn the underlying continuous factors that generate the ECG (i.e. the FactorECG). In this process, the variational auto-encoder (VAE) learns to reconstruct ECGs as accurately as possible using only 21 continuous factors without any human input. In the training phase, the pre-procedural median beat ECGs of 1306 cardiac resynchronization therapy (CRT) patients were each converted into their FactorECG. These 21 factors were subsequently used as input in a Cox model to predict the primary composite endpoint of left ventricular assist device implantation, heart transplantation and all-cause death, and the secondary endpoint of echocardiographic response. FactorECG significantly improved outcome prediction following CRT when compared with the current guidelines and QRS_{AREA}. The algorithm is explainable by using the decoder to visualize the effect of the ECG factors that significantly predicted outcome on the median beat ECG morphology. Here, for example, the influence of Factor 9 (F_9) is visualized, where higher values represent a more left bundle branch block-like ECG morphology and lower values, a more right bundle branch block-like morphology. CRT, cardiac resynchronization therapy; DNN, deep neural network; ECG, electrocardiogram; HTx, heart transplantation; LVAD, left ventricular assist device.

Introduction

In patients with dyssynchronous heart failure (HF), cardiac resynchronization therapy (CRT) can effectively restore left ventricular (LV) electrical activation and mechanical function, thereby improving clinical outcome.[1] However, for CRT to be beneficial, sufficient LV electrical conduction delay must be present.[2] Currently, patients are selected based on various requirements set out by different guidelines. However, despite indicating the highest level of recommendation, by itself, a Class I indication does not necessarily ensure a sustained response after CRT.[3] Conversely, the effectiveness of CRT is variable and doubted in patients without left bundle branch block (LBBB) morphology or intermediate QRS duration.[2,4,5] Although a substantial portion of these patients will benefit regardless,[6] they are at increased risk of not being considered for treatment.[5,6] Accurate and objective identification of the underlying electrical substrate is therefore crucial to optimize patient selection and ensure optimal treatment.

Currently, electrical characteristics derived from the electrocardiogram (ECG), such as LBBB morphology and QRS duration, are used to determine eligibility for CRT.[2] Multiple ECG criteria for LBBB have been defined[7] and inter-observer variability is high.[8] Moreover, a variety of LV electrical activation patterns are concealed in the ECG, further complicating clinical decision-making.[9] Recently, QRS_{AREA} has emerged as a new and objective computerized measure.[3,10] QRS_{AREA} is independently associated with survival and echocardiographic response, outperforming LBBB morphology and QRS duration.[3,10] As such, QRS_{AREA} partly overcomes the challenges of subjective ECG interpretation, but (subtle) ECG characteristics, besides the QRS complex, are still not considered.

Machine learning has gained interest as a means of integrating large number of variables, thereby producing advanced clinical decision models. The SEMMELWEIS-CRT score, for example, outperforms many already existing risk scores, but relies on 33 clinical variables.[11] Besides being laborious to use, such models also rely on human interpretation of input variables such as left ventricular ejection fraction (LVEF), the New York Heart Association (NYHA), LBBB morphology and QRS duration, which are all subjectively assessed. Hence, although such models may predict response to CRT, large amounts of clinical variables will still need to be acquired, extracted, and entered in such models.[11–14]

A recent development in the field of machine learning, called deep learning, can learn features from the raw ECG signal without the necessity for any human interpretation.[15] Deep learning algorithms may therefore be used to automatically detect, identify, and classify ECG abnormalities that are associated with non-response or poor outcome after CRT. Although the need for very large data sets and the lack of interpretability were deemed common drawbacks of deep learning, a novel technique that uses a variational auto-encoder (VAE; the FactorECG) was recently introduced.[16] This approach enables physicians to better understand and verify the learned ECG features of deep learning algorithms, and makes the technique available to much smaller data sets.

The present study seeks to compare contemporary guideline ECG criteria for CRT implantation and QRS_{AREA} with the FactorECG for the prediction of a combined clinical endpoint and echocardiographic response. In addition, we aim to identify and visualize ECG features associated with these outcome measures.

Methods

Study design

All data were acquired for routine patient care and handled anonymously, and were collected as part of the multicentre Maastricht–Utrecht–Groningen (MUG) registry.[10] Under these circumstances, informed consent was waived by the Institutional Review Board at the time of the study. All study procedures were performed in compliance with the Declaration of Helsinki.

Only patients who received a *de novo* CRT device with a transvenous LV lead were considered for the present study (*Figure 1*). A baseline ECG (within 3 months before implantation) was required for the primary endpoint analysis, whereas a paired echocardiographic examination at baseline and follow up (6–12 months) was required for the secondary endpoint. Echocardiographic examinations from various vendors were used to determine LV end-systolic volume (LVESV), and LVEF was calculated using the Simpson's modified biplane method (IntelliSpace Cardiovascular, Philips, Eindhoven).

The primary endpoint was a combined clinical endpoint consisting of left ventricular assist device (LVAD) implantation, heart transplantation (HTx), and all-cause mortality. The secondary endpoint was echocardiographic non-response, defined as a relative decrease in LVESV of <15%.[17] In addition, three tertiary endpoints were investigated: (i) a composite of HF hospitalization and the primary endpoint, (ii) HF hospitalization alone, and (iii) ≥1 point of NYHA functional class improvement.

Electrocardiographic data

For all patients, standard 12-lead ECGs were exported and converted into median heart beats using the MUSE ECG system (MUSE version 8; GE Healthcare, Chicago, IL, USA). The median beat data were constructed by aligning all QRS complexes in the 10-second ECG of the same shape (e.g. excluding premature ventricular complexes), and generating a representative QRS complex by taking the median voltage.[18] Automated ECG readings were used to derive QRS duration and other typical ECG parameters. LBBB morphology was defined according to the 2013 European Society of Cardiology (ESC) and 2013 the American Heart Association(AHA) criteria at the time (see Supplementary material online, *Table S1*), as previously reported.[7] Using these morphological definitions, indications for CRT implantation were determined according to the current 2021 ESC guidelines.[2] Strauss criteria provide similar risk stratification when compared with the 2013 ESC criteria,[7] and were therefore not evaluated. Without exception, all digitally available ECGs were selected for analysis.

To calculate QRS_{AREA}, first all ECGs were semi-automatically recoded into vectorcardiograms, consisting of three orthogonal leads (X, Y, and Z). To this end, the Kors conversion matrix was used in custom Matlab software (MathWorks Inc).[19] The three orthogonal leads from the vectorcardiogram form a 3D-vector loop, from which QRS_{AREA} was calculated as the sum of the area under the QRS complex as $\sqrt{X_{AREA}^2 + Y_{AREA}^2 + Z_{AREA}^2}$.

Deep learning approach

A recently developed approach to use deep neural networks in an explainable method, referred to as the FactorECG, was used. Here, the complete median beat ECG is analysed using a VAE, which divides the ECG into morphological features without any assumptions (e.g. an agnostic approach). For this approach, the VAE was pretrained to learn these morphological features (or underlying generative factors) of the ECG, using a data set of 1 144 331 ECGs from 251 473 consecutive patients that underwent ECG recording in the University Medical Centre Utrecht between July 1991 and August 2020.[16] Overlap in the pretraining cohort and patients included in this study was negligible at 0.04% and could not influence the results since the VAE was trained unsupervised (i.e. without any knowledge of CRT outcome).

The VAE is a generative artificial intelligence algorithm that consists of three parts: (i) an encoder neural network, (ii) the FactorECG (a compressed version of the ECG in only 32 disentangled continuous factors), and (iii) the decoder neural network (*Figure 2A*). The goal of the VAE is to learn to 'compress' the ECG, without human interference, into a reduced number of continuous and independent variables that are presumably related to the
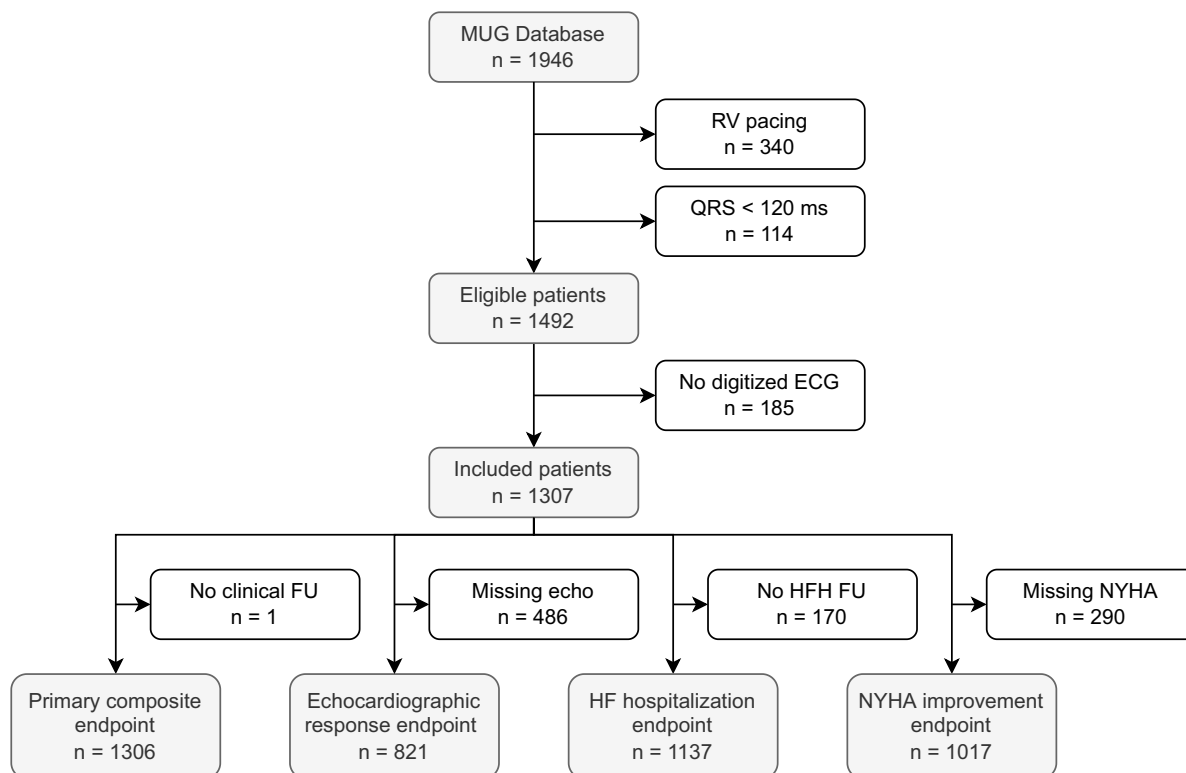
Figure 1 Flow chart for the inclusion of patients in this study. ECG, electrocardiogram; FU, follow up; HF, heart failure; HFH, heart failure hospitalization; NYHA, New York Heart Association; RV, right ventricular.

underlying (patho)physiological generative processes of the ECG. Pretraining of the VAE was performed unsupervised by entering the median beat ECGs into the algorithm and reconstructing the same ECG, while calculating the difference between the original and reconstructed ECG to optimize the network. After training, the first part of the VAE (encoder) can be used to convert any median beat ECG into its FactorECG, the distinctive set of 32 factors that represent that ECG. Importantly, it has been shown before that only 21 of the 32 factors encode significant information.[16] Hence, only these 21 factors were used in subsequent models. In the training step of the current analysis, the 21 continuous FactorECG values for every ECG, as calculated by the encoder, are used in Cox and logistic regression models to perform prediction of the different endpoints (*Figure 2B*).

Explainability of the individual ECG factors was achieved by visualizing their influence on the median beat ECG morphology. This was done at the model level by varying the values of the individual ECG factors between −3 and 3, while reconstructing the ECG using the decoder. The other factors were kept constant, which allows for visualization of the distinct median beat ECG morphology that every factor entails. Moreover, patient-level explanations can be obtained by investigating the FactorECG values of that specific ECG, in combination with the coefficients of the model. This way, we can determine which factors were important in a specific patient to make the prediction. Interactive visualizations of the model are available at https://crt.ecgx.ai. The architecture and training procedures for the FactorECG have been described in detail before.[16]

Statistical analysis

Baseline characteristics were expressed as mean ± standard deviation (SD), or median with interquartile range (IQR), where applicable. Depending on the normality of data, differences in continuous variables were assessed using the Student's *t*-test or Mann–Whitney *U* test. Conversely, categorical variables were tested using the χ^2 test or Fisher's exact test.

Models using different guideline criteria, QRS_{AREA} and the per-patient 21 significant standardized FactorECG values, were compared. For the primary endpoint, multivariable Cox proportional hazard models were fitted to take time-to-event into account. For the secondary endpoint, a similar approach was applied, with multivariable logistic regression to predict the binary endpoint of LVESV non-response <15%. In a second step, the added value of the models to a combination of 13 standard clinical parameters was assessed. Clinical parameters known to be associated with CRT outcome were entered into the multivariable models (i.e. Cox regression for the primary endpoint and logistic regression for secondary endpoint): sex, age, and aetiology [i.e. ischaemic cardiomyopathy (ICM) or non-ICM], weight, height, baseline NYHA class, rhythm (sinus rhythm or atrial fibrillation), baseline LVEF, baseline end diastolic volume, baseline interventricular mechanical delay (IVMD), haemoglobin, creatinine levels, and presence of diabetes. As there were missing values of some parameters, multivariate imputation using chained equations was performed using only these clinical parameters as input.

For all models, non-linear relationships were investigated using natural cubic splines, and for the Cox models, the proportional hazards assumption was verified. Hazard ratios (HRs) and odds ratios (ORs) were reported to investigate the importance of individual predictors, such as the standardized FactorECG values. Model fit for all models was assessed using Akaike's Information Criterion, discrimination using Harell's C-statistic, and calibration using the calibration slope. The apparent C-statistic and calibration slope were obtained by applying the model to the original data. Internal validation was performed by using a bootstrap-based optimism estimation technique, where all model development steps were repeated on the 500 bootstrap samples and the model was tested on the original data.[20] The 'optimism', which is the mean difference between the performance
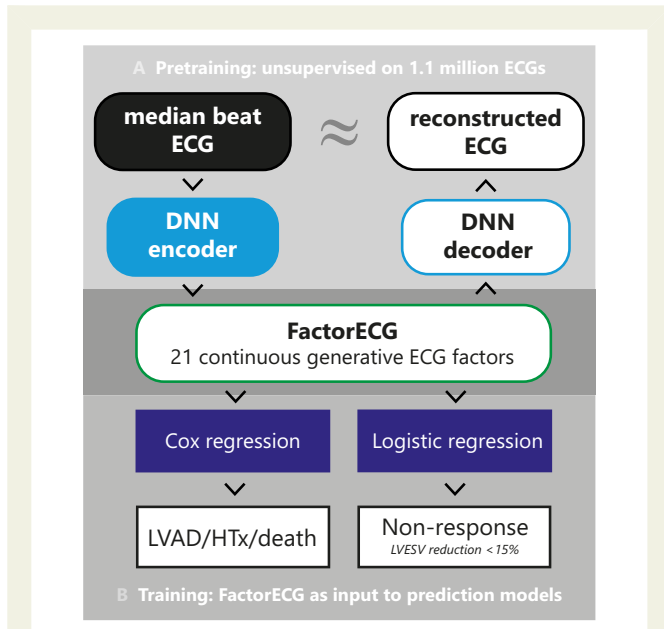
Figure 2 Schematic representation of the series of algorithms and processes: a variational auto-encoder, the FactorECG and reconstructions. (*A*) In the pretraining phase, the variational auto-encoder is trained on a data set of 1.1 million median beat electrocardiograms from the University Medical Center Utrecht to learn the underlying factors that generate the electrocardiogram. In this process, the variational auto-encoder learns to reconstruct electrocardiograms as accurately as possible using only the FactorECG continuous factors. (*B*) In the training phase, the 21 significant electrocardiogram factors for every median beat electrocardiogram in the cardiac resynchronization therapy population are obtained using the encoder. These factors are used as input in Cox and logistic regression models to predict outcome (composite of left ventricular assist device implantation, heart transplantation, and death) or non-response (left ventricular end-systolic volume reduction <15% after cardiac resynchronization therapy implantation). DNN, deep neural network; ECG, electrocardiogram; HTx, heart transplantation; LVAD, left ventricular assist device; LVESV, left ventricular end-systolic volume; VAE, variational auto-encoder.

measure in the original and bootstrapped data set, was subtracted from the apparent performance measures. 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] Confidence intervals (CIs) around the performance measures were obtained using 2000 bootstrap samples. All statistical analyses were performed using Python version 3.8. The Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis Statement for the reporting of diagnostic models was followed, where applicable.[22]

Results

Baseline characteristics

A real-world CRT population was gathered from three Dutch academic hospitals (*n* = 1946), of which 1492 were eligible after exclusion for RV pacing and QRS duration <120 ms. Of the 1492 patients, 1307 had a digital ECG available in the 90 days before implantation, and 1306 were included in the analysis for the primary endpoint (*Figure 1*,

Table 1 Baseline characteristics

Variable	Missing, *n* (%)	Overall (*n* = 1306)
Patient demographics		
Age (years)	0 (0)	68.3 (60.0–74.7)
Male sex, *n* (%)	0 (0)	919 (70.4)
Length (cm), mean (SD)	66 (5.1)	174 (8.9)
Weight (kg), mean (SD)	60 (4.6)	81.9 (16.1)
ICM, *n* (%)	0 (0)	649 (49.7)
DM, *n* (%)	2 (0.1)	328 (25.2)
Pre-procedural NYHA, *n* (%)	29 (2.2)	
I		28 (2.2)
II		513 (40.2)
III		672 (52.6)
IV		64 (5.0)
ICD, *n* (%)	0 (0)	1226 (93.9)
Laboratory measurements		
NT-proBNP (pmoL/L), median (IQR)	605 (46.3)	1379 (587–2845)
Haemoglobin (mmoL/L), median (IQR)	436 (33.3)	8.5 (7.8–9.1)
Creatinine (mmoL/L), median (IQR)	48 (3.7)	102 (83–130)
Electrocardiography		
Sinus rhythm, *n* (%)	9 (0.7)	1096 (84.5)
PR duration (ms), median (IQR)	216 (16.5)	184.0 (164.0–213.5)
QRS duration (ms), median (IQR)	0 (0)	158.0 (146.0–172.0)
QTc duration (ms), median (IQR)	0 (0)	486.0 (463.0–510.0)
LBBB (ESC 2013), *n* (%)	0 (0)	1028 (78.7)
LBBB (AHA), *n* (%)	0 (0)	173 (13.3)
QRS$_{AREA}$ (µVs), median (IQR)	0 (0)	108.2 (76.0–151.0)
Echocardiography		
LVEDV (mL), median (IQR)	355 (27.2)	205.0 (157.1–271.0)
LVESV (L), median (IQR)	349 (26.7)	151.0 (113.0–209.0)
LVEF (%), median (IQR)	321 (24.5)	24.0 (18.9–30.0)
IVMD (ms), median (IQR)	522 (40.0)	45.0 (22.0–64.0)
Procedure characteristics		
CRT-P, *n* (%)	0 (0)	80 (6.1)
LV lead position, *n* (%)	38 (2.9)	
Anterior		135 (10.6)
Lateral		466 (36.8)
Posterior		667 (52.6)
Outcomes		
Duration of follow up (years), median (IQR)	0 (0)	3.48 (2.08–5.24)

Continued

Table 1 *Continued*

Variable	Missing, n (%)	Overall (n = 1306)
Primary endpoint (LVAD, HTx or death), n (%)	0 (0)	385 (30)
LVESV reduction (%), median (IQR)	485 (37.1)	20.9 (0.5–41.4)
LVESV non-responder endpoint, n (%)	485 (37.1)	355 (43)
Composite of primary endpoint and heart failure hospitalization, n (%)	169 (12.9)	406 (35.7)
Heart failure hospitalization endpoint, n (%)	169 (12.9)	133 (11.7)
Post-procedural NYHA, n (%)	249 (19.1)	
I		178 (16.8)
II		650 (61.5)
III		216 (20.4)
IV		13 (1.2)
NYHA improvement endpoint	289 (22.1)	509 (50)

CRT-P, cardiac resynchronization therapy pacemaker; DM, diabetes mellitus; ICD, implantable cardioverter defibrillator; ICM, ischaemic cardiomyopathy; IVMD, interventricular mechanical delay; IQR, interquartile range; LBBB, left bundle branch block; LV, left ventricular; LVEDV, left ventricular end diastolic volume; LVEF, left ventricular ejection fraction; LVESV, left ventricular end-systolic volume; NT-proBNP, N-terminal pro-B-type natriuretic peptide; NYHA, New York Heart Association; SD, standard deviation.

Table 1). The median time between ECG and implantation was 1 day (IQR 1–6 days). The pretrained VAE performed well in the current population, with a Pearson correlation between the original and reconstructed ECG of 0.86. The ESC guideline CRT indications, using the 2013 ESC criteria for LBBB, were as follows: Class I 737 (56%), Class IIa 401 (31%), Class IIb, and Class III 168 (13%; Table 1). When applying the AHA criteria for LBBB, the indications were as follows: Class I 134 (10%), Class IIa 787 (60%), Class IIb, and Class III 385 (30%).

Primary endpoint: combined clinical outcome

A total of 385 patients (30%) reached the primary endpoint of LVAD implantation (n = 11), HTx (n = 4), or all-cause mortality (n = 370). The median follow-up time was 3.5 years (IQR 2.1–5.2 years). Optimism-corrected C-statistics were derived for the different predictor sets in predicting the occurrence of the primary endpoint (Table 2, Supplementary material online, Tables S2, S4–S8). According to current guideline criteria for CRT implantation, a Class I indication was significantly associated with freedom of the primary endpoint, when compared with a non-Class I indication. However, this association was only seen when using the ESC [c-statistic 0.57 (95% CI 0.54–0.60)], but not the AHA definition [c-statistic 0.50 (95% CI 0.47–0.53)] of LBBB morphology (Table 2). A stronger association with outcome was seen using FactorECG [c-statistic 0.69 (95% CI 0.66–0.72), P <

0.001 for both AHA and ESC definitions]. Moreover, FactorECG had a significantly stronger association with outcome than QRS$_{AREA}$ [c-statistic 0.61 (95% CI 0.58–0.64), P < 0.001].

When subdividing QRS$_{AREA}$ and FactorECG into four quartiles, better discriminative performance for the occurrence of the primary endpoint was achieved using FactorECG (Figure 3). A significantly higher event-free survival at three years was seen in the lowest risk FactorECG group when compared with QRS$_{AREA} \geq$ 150 µVs (94% vs. 89%; log rank P = 0.01). Additionally, 3-year event-free survival for the highest risk FactorECG quartile was significantly worse than in patients with QRS$_{AREA} <$ 75 µVs (63% vs. 73%; log rank P < 0.005).

Secondary endpoint: echocardiographic non-response

Pre- and post-procedural echocardiograms were available in 821 patients. Long-term echocardiographic non-response was observed in 355 patients (43%). All evaluated models were significantly associated with echocardiographic non-response (Table 2, Supplementary material online, Tables S3, S9–S13). However, guideline classifications performed the worst, using either the ESC [c-statistic 0.61 (95% CI 0.57–0.64)] or AHA definition [c-statistic 0.56 (95% CI 0.53–0.60)] of LBBB morphology. FactorECG [c-statistic 0.69 (95% CI 0.65–0.72)] and QRS$_{AREA}$ [c-statistic 0.70 (95% CI 0.67–0.74)] had similar associations with non-response (P = 0.12), but were both significantly stronger associated with response than either guideline recommendation (P < 0.001, Figure 3). Differences in the extent of reverse remodelling, stratified according to four groups of FactorECG and QRS$_{AREA}$, were similar (Figure 3).

Tertiary endpoints

The availability of tertiary endpoints is summarized in Figure 1 and Table 1. FactorECG was significantly associated with the composite of the primary endpoint combined with HF hospitalization [c-statistic = 0.68 (95% CI 0.65–0.70)], and HF hospitalization alone [c-statistic = 0.70 (95% CI 0.66–0.74)], outperforming QRS$_{AREA}$ and the guideline criteria [P < 0.001 for all comparisons (see Supplementary material online, Tables S14–S24]. None of the models showed additional predictive value for prediction ≥1 point NYHA improvement when compared with a baseline model that only consisted of pre-procedural NYHA class (see Supplementary material online, Tables S14, S25–S29).

Subgroup analysis

Performance of FactorECG and QRS$_{AREA}$ were compared, stratified by known subgroups associated with clinical outcome (Table 3). The strongest association of FactorECG was observed in patients with non-ICM [c-statistic 0.77 (95%CI 0.73–0.81)], which was significantly higher when compared with QRS$_{AREA}$ [c-statistic 0.62 (95%CI 0.57–0.67)]. Using the ESC definition of LBBB morphology, FactorECG outperformed QRS$_{AREA}$ in patients with LBBB [c-statistic 0.71 (95%CI 0.68–0.74) vs. c-statistic 0.61 (95% CI 0.58–0.65)], and non-LBBB [c-statistic 0.66 (95% CI 0.60–0.71) vs. c-statistic 0.52 (95% CI 0.46–0.58)]. The same observation was made when evaluating patients with an intermediate QRS duration (<150 ms) and patients with ICM. Importantly, FactorECG and QRS$_{AREA}$ demonstrated comparable associations with echocardiographic response regardless of the subgroup analysed (Table 3).

Table 2 Optimism-corrected C-statistic for outcome and response

Predictors	Outcome		Response	
	C-statistic	95% CI	C-statistic	95% CI
2013 AHA criteria	0.50	(0.47–0.53)	0.56	(0.53–0.60)
2013 ESC criteria	0.57	(0.54–0.60)	0.61	(0.57–0.64)
QRS$_{AREA}$	0.61	(0.58–0.64)	0.70	(0.67–0.74)
FactorECG	0.69	(0.66–0.72)	0.69	(0.65–0.72)
Clinical	0.69	(0.67–0.72)	0.67	(0.64–0.71)
QRS$_{AREA}$/clinical	0.71	(0.68–0.74)	0.72	(0.68–0.75)
FactorECG/clinical	0.72	(0.69–0.75)	0.70	(0.67–0.74)

AHA, American Heart Association; ESC, European Society of Cardiology.

Additional value of clinical model

Readily available patient characteristics, known to be associated with CRT outcome, were entered into a clinical model.[11] The clinical model was significantly associated with outcome [c-statistic 0.69 (95% CI 0.67–0.72)] and response [c-statistic 0.60 (95%CI 0.56–0.64)] (*Table 2*, Supplementary material online, *Tables S4–S13*). However, for both endpoints, the ECG-only FactorECG model demonstrated similar associations when compared with the clinical model ($P = 0.48$ and $P = 0.10$, respectively). For outcome, the addition of a 13-variable clinical model significantly improved upon QRS$_{AREA}$ (Δ c-statistic 0.10, $P < 0.001$), whereas its addition to FactorECG was of limited added value (Δ c-statistic 0.03, $P < 0.001$). By contrast, concerning echocardiographic non-response, the added value of the clinical model was negligible (Δ c-statistic 0.01, $P = 0.002$).

Explainable deep learning through factor visualization

Electrocardiogram factors that were significantly associated with outcome and non-response are summarized in *Figure 4*. Exact HRs for outcome and the ORs for non-response are summarized in Supplementary material online, *Tables S5 and S10*, respectively. Visualizations of the most important ECG factors, using factor traversals, are shown in *Figure 5*, whereas Supplementary material online, *Figure S1* displays a complete 12-lead visualization of all factors. Factors associated with 'both' non-response and poor outcome were interpreted as follows: F$_1$ (absent QRS notching and ST-segment deviation, but lateral T-wave inversion), F$_9$ (transition from LBBB morphology to more right bundle branch block morphology with smaller right precordial S-wave amplitudes), F$_{10}$ (increased ventricular rate), and F$_{19}$ (decreased anterior QS amplitude and lateral notched R). Importantly, F$_8$ and F$_{15}$ (increased PR interval and P-wave duration) were only associated with worse outcomes, whereas F$_5$ (decreased QRS duration and JTc interval) and F$_{26}$ (decreased QRS duration and amplitude of LBBB morphology) were only associated with non-response. Similar factors (F$_1$, F$_9$, and F$_{19}$), mostly representing reduced QRS and T-wave voltage with increased QT duration, were found to be predictive for HF hospitalization when compared with the model for the primary outcome alone (see Supplementary material online, *Table S21*). However, F$_{25}$, which represents reduced QRS duration, was also predictive for HF hospitalization.

Clinical applicability using risk groups

Using a combination of predictions of the FactorECG algorithm for both echocardiographic non-response and 3-year clinical outcome, four distinct groups could be identified to assist patient selection (see Supplementary material online, *Table S30*). Here, QRS$_{AREA}$ could not differentiate between good and poor outcomes in echocardiographic responders (median QRS$_{AREA}$ 151 vs. 152 μVs, respectively) or non-responders (median QRS$_{AREA}$ 84 vs. 83 μVs, respectively).

In the first group, with both predicted response and good outcome ($n = 338$), 76% of the patients were responders, and only 14% experienced the primary endpoint during follow up. In the second group of poor 3-year outcomes despite an echocardiographic response ($n = 72$), patients were more frequently male, had ICM, higher NT-proBNP, high QRS duration, and the worst ESV and LVEF. Conversely, in the third group, CRT non-responders with good clinical outcome regardless ($n = 96$) were predominantly characterized by shorter QRS duration, lowest LVESV, and highest LVEF. In the fourth group of patients, with both poor outcome and non-response, significantly more ICM, NYHA Class III, and non-LBBB was observed when compared with the other subgroups. In this worst performing subgroup ($n = 314$), the primary endpoint occurred in 46% of the patients during follow up, and response occurred in only 36% of the patients.

In contrast, when using current ESC guidelines for selection of patients eligible for CRT, in Class I patients ($n = 499$) response occurred in 65% and the primary outcome endpoint in 26% during follow up. In patients with Class IIa ($n = 226$) or IIb/III ($n = 96$) indications, response occurred in 50% and 33%, and the primary outcome endpoint in 35% and 37%, respectively. A comparison of the classification in the four FactorECG groups and the guideline-based groups can be found in *Figure 2C*.

Discussion

In this large, multicentre, real-world data set, an explainable deep learning–based algorithm (FactorECG) was predictive for long-term clinical outcome, HF hospitalization, and echocardiographic non-response after CRT implantation. FactorECG outperformed contemporary guideline criteria and vectorcardiographic QRS$_{AREA}$ for clinical outcome and HF hospitalization (*Structured Graphical Abstract*).
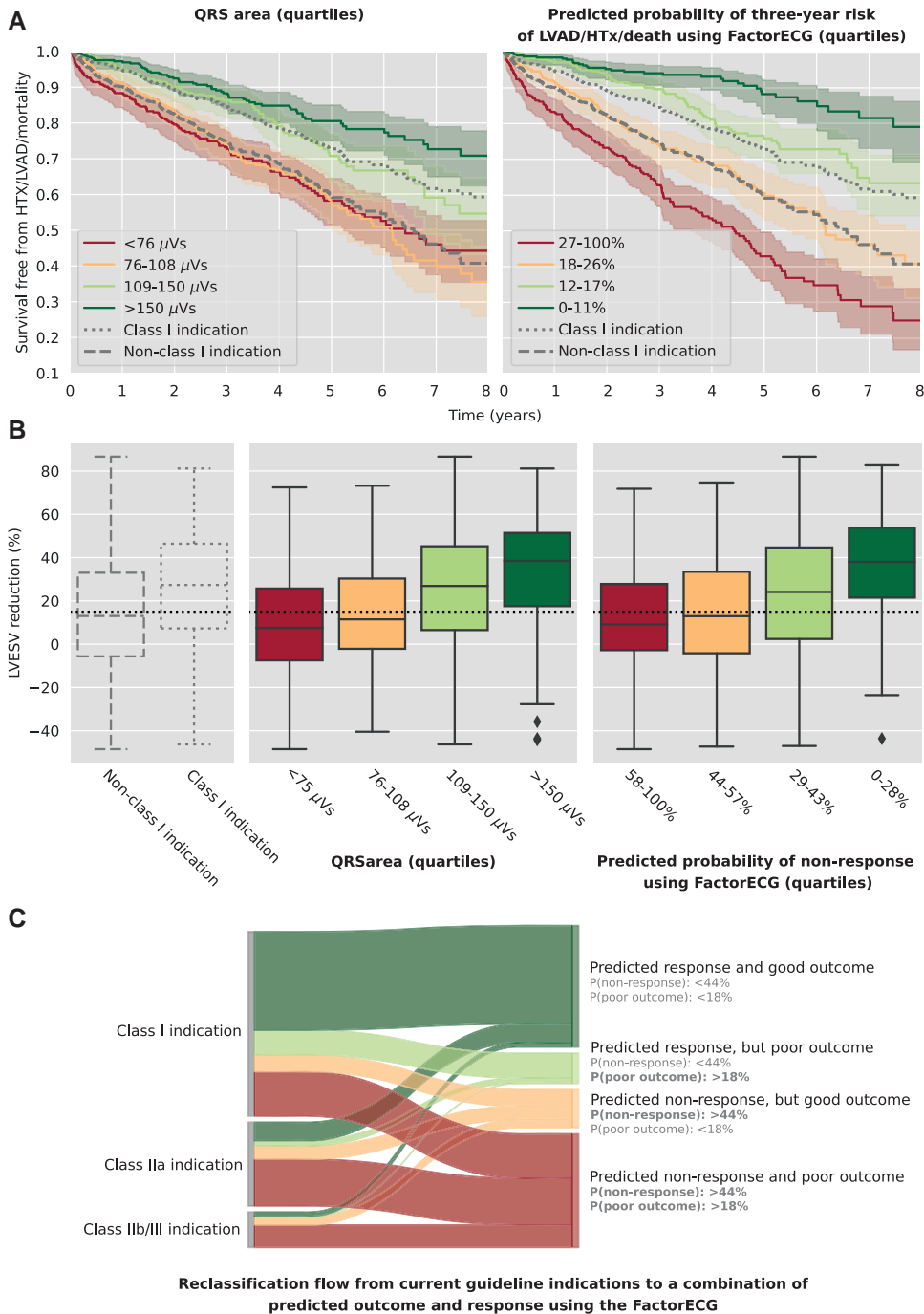
Figure 3 Clinical utility of FactorECG and QRS$_{AREA}$ in cardiac resynchronization therapy. QRS$_{AREA}$ and FactorECG predicted probabilities were divided into four quartiles of equal size. Quartiles of FactorECG better differentiate clinical outcomes when compared with QRS$_{AREA}$ and guidelines using the European Society of Cardiology criteria of left bundle branch block (*A*). Similar associations with echocardiographic response were seen when compared with QRS$_{AREA}$, while still outperforming guideline criteria (*B*). The reclassification flow from the guidelines to the FactorECG predictions is shown in *C*. Here, a combination of predicted clinical outcome and response is assessed by setting the probability cut-off at 50% of the data. The probability cut-offs in *C* therefore correspond to the upper two and lower two quartiles in *A* and *B* combined. ECG, electrocardiogram; HTx, heart transplantation; LVAD, left ventricular assist device; LVESV, left ventricular end-systolic volume.

Importantly, only a readily available 12-lead ECG is required since little added value was obtained using additional clinical input variables. The user-independent analysis and automated visualization of key ECG features allows for patient-specific interpretation of the algorithm (*Figure 6*), which may facilitate its adoption into clinical practice as a valuable alternative for the selection of CRT candidates. Lastly, an online visualization tool was created to provide interactive visualizations (https://crt.ecgx.ai).

Table 3 Optimism-corrected C-statistic in various subgroups

Subgroup	Outcome [c-statistic (95% CI)]		Response [c-statistic (95% CI)]	
	QRS$_{AREA}$	FactorECG	QRS$_{AREA}$	FactorECG
Male	0.60 (0.57–0.63)	0.67 (0.64–0.70)	0.69 (0.65–0.73)	0.70 (0.66–0.74)
Female	0.61 (0.53–0.69)	0.77 (0.71–0.83)	0.70 (0.63–0.77)	0.73 (0.66–0.79)
ICM	0.58 (0.54–0.62)	0.63 (0.60–0.67)	0.65 (0.59–0.70)	0.67 (0.61–0.72)
Non-ICM	0.62 (0.57–0.67)	0.77 (0.73–0.81)	0.72 (0.67–0.77)	0.74 (0.70–0.79)
LBBB[a]	0.61 (0.58–0.65)	0.71 (0.68–0.74)	0.71 (0.67–0.75)	0.73 (0.69–0.76)
Non-LBBB[a]	0.52 (0.46–0.58)	0.66 (0.60–0.71)	0.53 (0.43–0.63)	0.55 (0.46–0.65)
QRS ≥150 ms	0.62 (0.58–0.66)	0.70 (0.66–0.73)	0.71 (0.67–0.75)	0.73 (0.69–0.77)
QRS <150 ms	0.58 (0.53–0.63)	0.72 (0.67–0.76)	0.62 (0.55–0.70)	0.67 (0.60–0.73)

ICM, ischaemic cardiomyopathy; LBBB, left bundle branch block.
[a]Morphology evaluated according to ESC 2013 criteria.

Deep learning–based prediction of outcome

For the first time, deep learning has been used to predict clinical outcome after CRT using only the raw pre-procedural ECG [c-statistic 0.69 (95% CI 0.66–0.72)]. In contrast, previous studies aimed to predict CRT outcome using machine learning to unify a vast number of clinical variables into a single model. The SEMMELWEIS-CRT score combined 33 clinical variables for the prediction of all-cause mortality, reporting a mean internally calculated c-statistic of 0.69, derived from 1510 patients in a single centre.[11] Similarly, three other studies combined a plethora of pre-implantation characteristics, including ECG and complex echocardiography data, totalling 19, 45, or even 77 variables.[12–14] Another study compared an unsupervised principal component analysis model with QRS$_{AREA}$.[23] Here, similar results for QRS$_{AREA}$ [HR = 0.46 (95% CI 0.39–0.55)] and their model [HR = 0.45 (95% CI 0.38–0.53)] were seen for the composite endpoints of death, LVAD, or HTx.

Differences in primary clinical endpoints in the aforementioned studies complicate a direct comparison with the present study. However, similar or better performance was observed with respect to predicting clinical outcomes without relying on complex 'statistical' models.[11,13] Moreover, our approach outperformed QRS$_{AREA}$ with respect to clinical outcome, whereas unsupervised machine learning of baseline QRS waveforms previously failed to do so.[23] Most importantly, all previously proposed models require collection and calculation of many clinical variables, which are highly operator dependent, cumbersome, and likely to dissuade clinicians from rapidly adopting such an approach.[11–14] Although significant added benefit was obtained upon the addition of a clinical model to QRS$_{AREA}$, the increase in model performance was three-fold smaller for FactorECG. Rather, our proposed approach requires only a standard 12-lead ECG, without heavily depending on additional clinical input variables, or manual selection of the QRS complex. It is therefore conceivable that the clinical practicality of our ECG-only approach outweighs the limited benefit of increasing the c-statistic by 0.03 by using 13 clinical variables. For research purposes, an online tool has been developed where the ECG can be uploaded and predictions for CRT outcome can be made (https://crt.ecgx.ai and https://encoder.ecgx.ai).

Echocardiographic and functional response

The proportion of 43% non-responders is in accordance with previous literature and highlights the need for better patient selection.[3,17] In our study, a head-to-head comparison of FactorECG and QRS$_{AREA}$ provided similar results for the prediction of echocardiographic non-response [c-statistic 0.69 (95% CI 0.65–0.72) and 0.70 (95% CI 0.67–0.74), $P = 0.12$]. However, next to identifying the electrical substrate on the ECG, characterization of the extent of mechanical impairment is of importance as well, especially in patients with ICM. In fact, adding strain-based parameters of mechanical dyssynchrony to QRS$_{AREA}$ improves prediction of 6-month response (c-statistic 0.76), and is therefore also likely to add value to FactorECG.[3] Simple multivariate logistic regression models, consisting of only four variables, have also been shown to be associated with sustained echocardiographic response (c-statistic 0.774), a surrogate marker of stable disease remission.[3] None of the described models provided added value to predict NYHA class improvement, likely because NYHA class is nonspecific and its assessment is subjective and prone to bias.[24]

Identifying electrocardiogram features beyond the QRS complex

FactorECG improves upon heatmap-based attempts to make deep learning explainable, as such approaches merely highlight 'where' on the ECG significant features are detected but provide no information on which morphological change explains the prediction.[16] Rather, FactorECG allows for 'quantifiable' identification of specific ECG features, rendering physicians able to evaluate and confirm the clinical rationale of said features. This is reflected by our results that confirm the known importance of LBBB morphology and QRS$_{AREA}$ for the prediction of echocardiographic response.[3,10] Using FactorECG, all types of LV conduction delay, as reflected in the QRS complex, can be represented by combining ECG factors 5, 9, 19, and 27. Interestingly, although QRS$_{AREA}$ was associated with outcome, ECG factors that incorporate QRS duration were not associated with outcome (*Figure 5*). This may be because, in the presence of sufficient electrical substrate, a subset of patients with moderate QRS prolongation are still
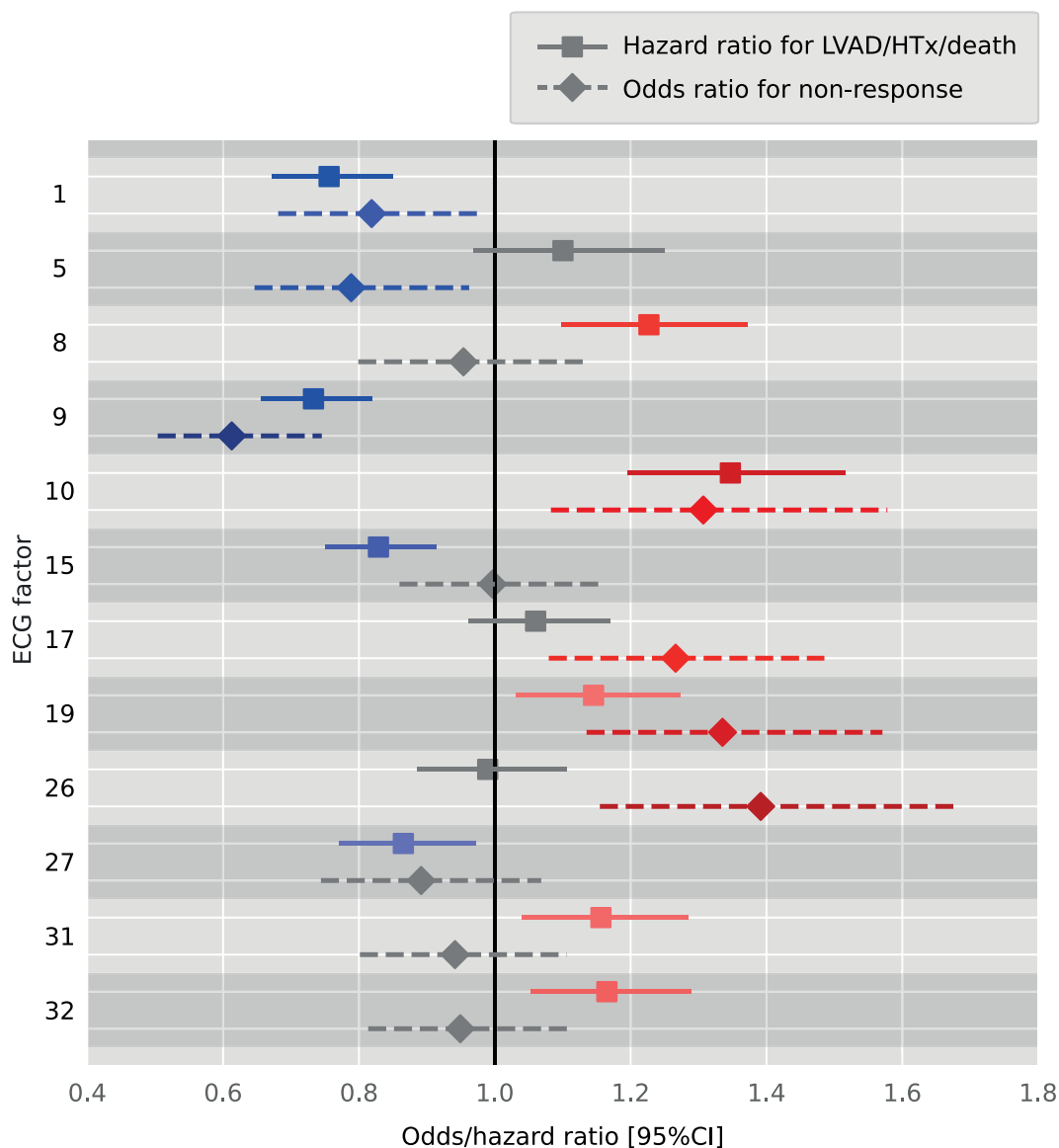
Figure 4 Hazard and odds ratios for the models predicting either the clinical endpoint or echocardiographic non-response (left ventricular end-systolic volume reduction < 15%) using the electrocardiogram factors as the only input for the model. Colours correspond with factor traversal reconstructions in *Figure 5*. All electrocardiogram factors were standardized and hazard and odds ratios can be interpreted as importance scores. ECG, electrocardiogram; HTx, heart transplantation; LVAD, left ventricular assist device; LVESV, left ventricular end-systolic volume.

likely to respond.[2,3,5,25] This is also underscored by our results, since FactorECG also predicted outcome in patients with QRS duration <150 ms [c-statistic 0.72 (0.67–0.76)]. Likewise, when corrected for various other ECG features, no significant association with QRS duration and outcome remains, as also reported previously.[26]

Visualization of ECG factors also identified various other ECG characteristics known to be associated with outcome and/or response, including the PR interval and P-wave duration (F_8 and F_{15}). The fact that correction of atrioventricular dromotropathy increases LV filling and LV pump function may explain the increased risk of poor outcome in the present study.[27] Similarly, prolonged P-wave duration >120 ms, indicating interatrial myopathy, has been linked to supraventricular arrhythmias, stroke, and mortality.[28] In addition, the QRS-T angle,[29] JTc

interval,[30] and T-wave area[31] have been raised as potentially important determinants of response or outcome. However, various other subtle markers of ischaemia, dyssynchrony, or risk of arrhythmia may be represented by FactorECG.

Indeed, when evaluated by itself, a large number of other factors can be identified from the ECG.[7] Unfortunately, accurately identifying these factors, and interpreting their interrelated meaning, is highly complex. In the first place because there is lack of consensus[7] and inter-observer disagreement[8] as to what truly defines LBBB morphology. Matters are further complicated when septal and LV activation patterns are concealed, or wrongly mimicked.[9] Finally, various unknown ECG criteria may have remained undetected. Interpretation of the LBBB ECG is therefore complex and misleading. In this regard, FactorECG
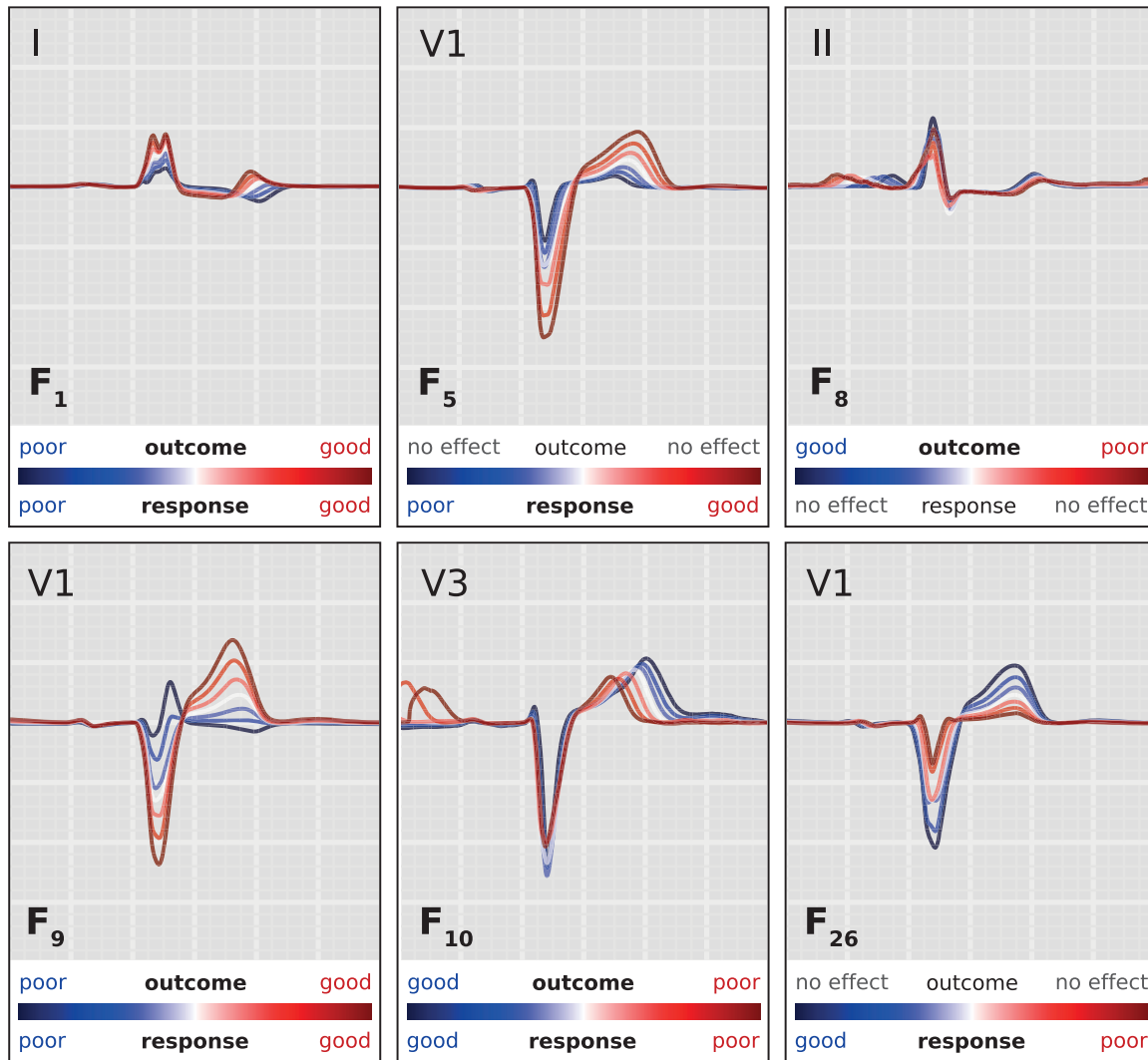
Figure 5 Factor traversals of a subset of the electrocardiogram factors associated with both clinical outcome (composite endpoint of left ventricular assist device/HTx/death) and echocardiographic response (left ventricular end-systolic volume reduction >15%). In each graph, the corresponding factor is varied from −3 (blue) to 3 (red) standard deviations from the mean of 0 (white line), which represents the mean electrocardiogram in the cardiac resynchronization therapy population. For each factor, the lead showing the most easily interpretable effect is shown in the upper left corner. A complete 12-lead electrocardiogram of all factors can be found in Supplementary material online, *Figure S1*. ECG, electrocardiogram; HTx, heart transplantation; LVAD, left ventricular assist device; LVESV, left ventricular end-systolic volume.

allows for a unified and agnostic approach, is user independent, and is inherently explainable.

Clinical implications

The FactorECG algorithm can be used in every patient that is considered for CRT. When provided with the baseline ECG, the patient-specific ECG factors that are associated with response and outcome are identified and combined into an individual risk score, and a patient-specific visualization of these factors is given (*Figure 6*). Hence, assessment of the electrical substrate as a 'continuum', rather than the current binary classification of LBBB morphology, is achieved. While similar in size, the CRT non-response *and* poor outcome subgroups, as predicted by the FactorECG, performed worse than patients without a Class I indication for CRT according to the ESC guideline

criteria. Importantly, 39% of patients in this worst performing subgroup had a Class I indication (*Figure 2C*). FactorECG therefore enables better classification of patient eligibility, without compromising the total proportion of patients deemed suitable for CRT implantation.

Future perspectives

Our self-contained ECG-based model was especially effective in females and the non-ICM population (c-statistic = 0.77 for both), but additional clinical variables are required to improve performance in patients with ICM. A future study will address the importance of adding strain-based mechanical dyssynchrony to FactorECG.[3] In addition, optimal placement of the LV lead is of importance to enhance response in CRT patients. This is particularly important in patients with scars, but also in patients with heterogeneous LV electrical activation.[9] In the future, FactorECG may
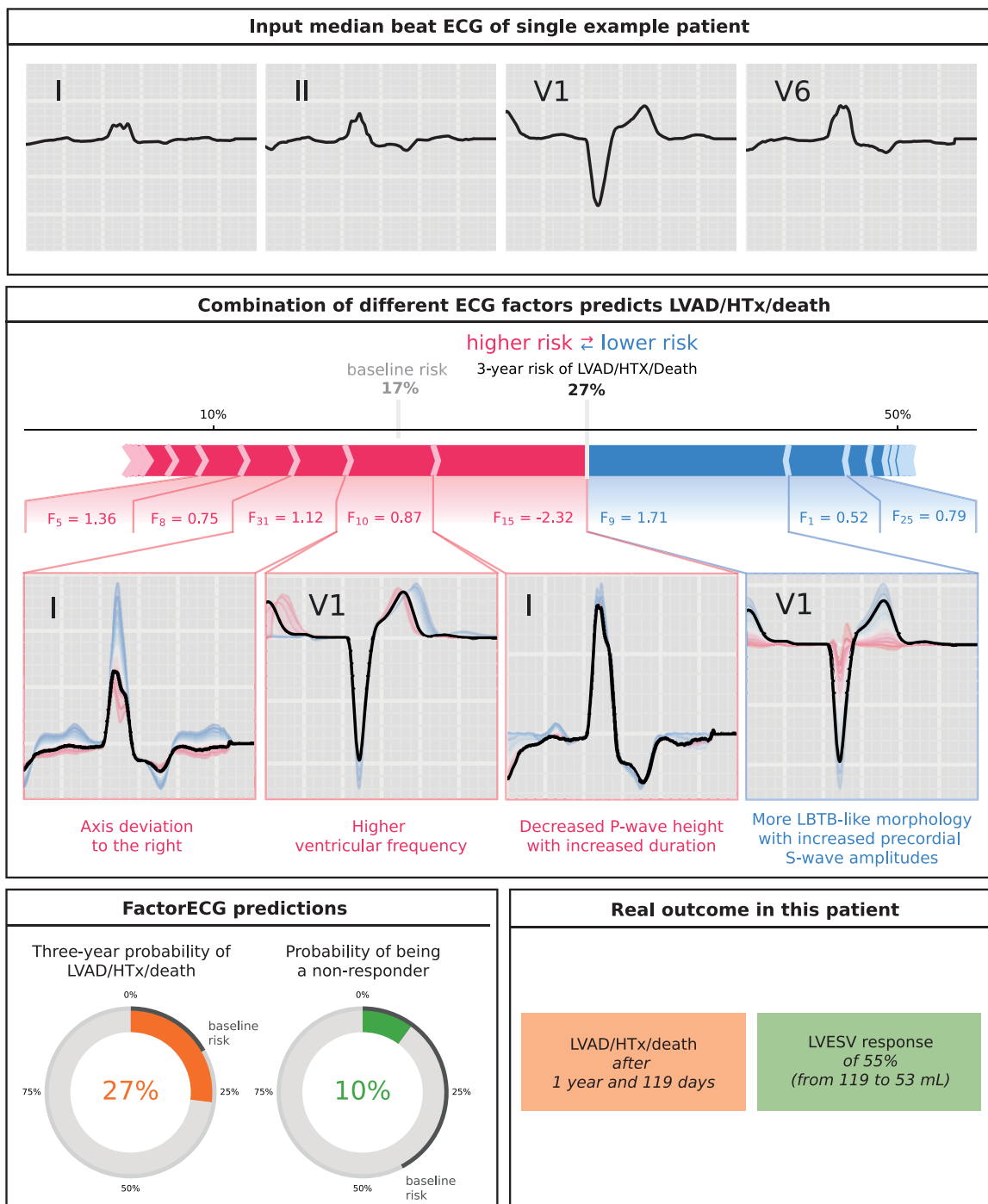
Figure 6 Patient-level example of a prediction with the FactorECG explanation. A standard 12-lead electrocardiogram is entered into a deep learning model, which automatically translates this electrocardiogram into its FactorECG containing all distinct features. These factors are entered into the Cox and logistic regression models, and predicted probabilities for both left ventricular assist device/HTx/death and non-response are shown to the user. This patient responded well to cardiac resynchronization therapy but died within 3 years regardless. Despite the presence of a 'typical' left bundle branch block morphology (F_9), FactorECG demonstrates that this prediction of high probability of poor outcome is driven by increased ventricular frequency (F_{10}), long PR interval with broad P-wave (F_{15}), and axis deviation to the right (F_{31}). ECG, electrocardiogram; HTx, heart transplantation; LVAD, left ventricular assist device; LVESV, left ventricular end-systolic volume.

use ECG-derived data to identify the site of latest electrical activation, thereby guiding LV lead implantation.[9] Moreover, the results need to be validated in a patient group that received a CRT-P device, as recent reports have shown similar survival between patients with a CRT-P and a CRT-D.[32] Lastly, prospective studies with FactorECG are warranted to acquire CE certification, allowing its use as a medical device.

Strengths and limitations

Our data were derived from a large multicentre database, and thereby represent a real-world population. Internal validation by means of bootstrapping was performed, which allows for unbiased validation of the complete data set and is therefore considered the recommended approach for internal validation of any prediction model.[20,21,32] As a result, performance was not assessed in a single train-test split because this approach only validates an example model in an arbitrarily chosen and small data subset and produces a poorer model by default.[33] We acknowledge that external validation of data sets with a different patient population remains important to investigate the generalizability of our results. However, by using regular prediction models (i.e. logistic regression and Cox regression) with a limited number of predicting variables as input (only the 21 factors), the risk of overfitting is low. Although ECG data were derived from a single vendor, previous studies have shown that ECG-based deep learning results generalize well to other cohorts with different ECG manufacturers.[34,35] Despite QRS_{AREA} being calculated manually, performance is identical to that of automated calculation.[36] Although measurement of LVESV is user dependent, excellent intra- and inter-observer reliabilities were previously demonstrated in a subpopulation of this study.[3]

Many clinicians regard deep learning as a 'black box', which limits trust in such algorithms.[16] However, our approach to make the model inherently explainable may abate this concern and increase willingness to facilitate clinical adoption of the FactorECG. Although an overall c-statistic of 0.69 leaves room for improvement, our approach is unique in its clinical practicality, with better risk stratification than QRS_{AREA}. The addition of a few important clinical values might further increase the predictive value of FactorECG. Especially use of strain parameters has been shown to be highly predictive, also in addition to QRS_{AREA},[3] or when used in machine-learning models.[12] As a result, no direct comparison with pre-existing scores could be performed.[11] Conversely, our approach only requires a standard 12-lead ECG, and no advanced and highly user-dependent measurements are needed. Lastly, ethnicity and cause of death were not systemically gathered, and our results cannot be generalized to patients receiving an upgrade to CRT.

Conclusions

FactorECG, an inherently explainable and end-to-end automated deep learning model, can accurately predict long-term clinical outcome, HF hospitalization, and echocardiographic non-response in patients eligible for CRT. Moreover, it outperformed contemporary guideline ECG criteria and QRS_{AREA} with superior discriminative ability. This approach is based solely on a standard 12-lead ECG, without heavily relying on additional clinical parameters, and visualizes patient-specific key features associated with outcome and response. Besides QRS morphology, T-wave amplitude and inversion, ventricular rate, PR interval, and P-wave duration were identified as important ECG factors. The FactorECG thereby facilitates personalized decision-making in CRT while being easy to use, allowing rapid uptake for everyday clinical practice.

Supplementary material

Supplementary material is available at *European Heart Journal* online.

Data availability

Data sharing requests will be considered upon 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. 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 predict the outcome and response to CRT, is available through: https://encoder.ecgx.ai.

References

1. Vernooy K, van Deursen CJM, Strik M, Prinzen FW. Strategies to improve cardiac resynchronization therapy. *Nat Rev Cardiol* 2014;**11**:481–493.
2. Glikson M, Nielsen JC, Kronborg MB, Michowitz Y, Auricchio A, Barbash IM, *et al.* 2021 ESC Guidelines on cardiac pacing and cardiac resynchronization therapy: developed by the task force on cardiac pacing and cardiac resynchronization therapy of the European Society of Cardiology (ESC) With the special contribution of the European Hear. *Eur Heart J* 2021;**42**:3427–3520.
3. Wouters PC, van Everdingen WM, Vernooy K, Geelhoed B, Allaart CP, Rienstra M, *et al.* Does mechanical dyssynchrony in addition to QRS area ensure sustained response to cardiac resynchronization therapy? *Eur Heart J Cardiovasc Imaging* 2021:jeab264 .
4. Sipahi I, Chou JC, Hyden M, Rowland DY, Simon DI, Fang JC. Effect of QRS morphology on clinical event reduction with cardiac resynchronization therapy: meta-analysis of randomized controlled trials. *Am Heart J* 2012;**163**:260–267.e3.
5. Sipahi I, Carrigan TP, Rowland DY, Stambler BS, Fang JC. Impact of QRS duration on clinical event reduction with cardiac resynchronization therapy: meta-analysis of randomized controlled trials. *Arch Intern Med* 2011;**171**:1454–1462.
6. Salden OAE, Vernooy K, van Stipdonk AMW, Cramer MJ, Prinzen FW, Meine M. Strategies to improve selection of patients without typical left bundle branch block for cardiac resynchronization therapy. *JACC Clin Electrophysiol* 2020;**6**:129–142.
7. van Stipdonk AMW, Hoogland R, ter Horst I, Kloosterman M, Vanbelle S, Crijns HJGM, *et al.* Evaluating electrocardiography-based identification of cardiac resynchronization therapy responders beyond current left bundle branch block definitions. *JACC Clin Electrophysiol* 2020;**6**:193–203.
8. van Stipdonk AMW, Vanbelle S, Ter Horst IAH, Luermans JG, Meine M, Maass AH, *et al.* Large variability in clinical judgement and definitions of left bundle branch block to identify candidates for cardiac resynchronisation therapy. *Int J Cardiol* 2019;**286**:61–65.
9. Wouters PC, Vernooy K, Cramer MJ, Prinzen FW, Meine M. Optimizing lead placement for pacing in dyssynchronous heart failure: the patient in the lead. *Heart Rhythm* 2021;**18**:1024–1032.
10. Ghossein MA, van Stipdonk AMW, Plesinger F, Kloosterman M, Wouters PC, Salden OAE, *et al.* Reduction in the QRS area after cardiac resynchronization therapy is associated with survival and echocardiographic response. *J Cardiovasc Electrophysiol* 2021;**32**:813–822.
11. Tokodi M, Schwertner WR, Kovács A, Tősér Z, Staub L, Sárkány A, *et al.* Machine learning-based mortality prediction of patients undergoing cardiac resynchronization therapy: the SEMMELWEIS-CRT score. *Eur Heart J* 2020;**41**:1747–1756.
12. Cikes M, Sanchez-Martinez S, Claggett B, Duchateau N, Piella G, Butakoff C, *et al.* Machine learning-based phenogrouping in heart failure to identify responders to cardiac resynchronization therapy. *Eur J Heart Fail* 2019;**21**:74–85.
13. Kalscheur MM, Kipp RT, Tattersall MC, Mei C, Buhr KA, DeMets DL, *et al.* Machine learning algorithm predicts cardiac resynchronization therapy outcomes. *Circ Arrhythm Electrophysiol* 2018;**11**:e005499.
14. Liang Y, Ding R, Wang J, Gong X, Yu Z, Pan L, *et al.* Prediction of response after cardiac resynchronization therapy with machine learning. *Int J Cardiol* 2021;**344**:120–126.
15. van de Leur RR, Boonstra MJ, Bagheri A, Roudijk RW, Sammani A, Taha K, *et al.* Big data and artificial intelligence: opportunities and threats in electrophysiology. *Arrhythm Electrophysiol Rev* 2020;**9**:146–154.
16. van de Leur RR, Bos MN, Taha K, Sammani A, Yeung MW, van Duijvenboden S, *et al.* Improving explainability of deep neural network-based electrocardiogram interpretation using variational auto-encoders. *Eur Heart J Digit Heatlh* 2022;**3**:390–404.

17. Foley PWX, Chalil S, Khadjooi K, Irwin N, Smith REA, Leyva F. Left ventricular reverse remodelling, long-term clinical outcome, and mode of death after cardiac resynchronization therapy. *Eur J Heart Fail* 2011;**13**:43–51.

18. GE Healthcare. Marquette 12SL ECG Analysis Program Physician's Guide. 2012. Chicago, IL. https://www.gehealthcare.com/products/diagnostic-cardiology/marquette-12slhttps://www.gehealthcare.com/products/diagnostic-cardiology/marquette-12sl (30 April 2020).

19. Kors JA, van Herpen G, Sittig AC, van Bemmel JH. Reconstruction of the Frank vector-cardiogram from standard electrocardiographic leads: diagnostic comparison of different methods. *Eur Heart J* 1990;**11**:1083–1092.

20. Harrell FE Jr, Lee KL, Mark DB. 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 FEJ, 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. Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *Br J Surg* 2015;**102**:148–158.

23. Feeny AK, Rickard J, Trulock KM, Patel D, Toro S, Moennich LA, *et al.* Machine learning of 12-lead QRS waveforms to identify cardiac resynchronization therapy patients with differential outcomes. *Circ Arrhythm Electrophysiol* 2020;**13**:e008210.

24. Raphael C, Briscoe C, Davies J, Ian Whinnett Z, Manisty C, Sutton R, *et al.* Limitations of the New York Heart Association functional classification system and self-reported walking distances in chronic heart failure. *Heart* 2007;**93**:476–482.

25. van Stipdonk AMW, Ter Horst I, Kloosterman M, Engels EB, Rienstra M, Crijns HJGM, *et al.* QRS area is a strong determinant of outcome in cardiac resynchronization therapy. *Circ Arrhythm Electrophysiol* 2018;**11**:e006497.

26. Khidir MJH, Delgado V, Ajmone Marsan N, Schalij MJ, Bax JJ. QRS duration versus morphology and survival after cardiac resynchronization therapy. *ESC Heart Fail* 2017;**4**:23–30.

27. Salden FCWM, Huntjens PR, Schreurs R, Willemen E, Kuiper M, Wouters P, *et al.* Pacing therapy for atrioventricular dromotropathy: a combined computational-experimental-clinical study. *Europace* 2022;**24**:784–795.

28. Martínez-Sellés M, Elosua R, Ibarrola M, de Andrés M, Díez-Villanueva P, Bayés-Genis A, *et al.* Advanced interatrial block and P-wave duration are associated with atrial fibrillation and stroke in older adults with heart disease: the BAYES registry. *Europace* 2020;**22**:1001–1008.

29. Sweda R, Sabti Z, Strebel I, Kozhuharov N, Wussler D, Shrestha S, *et al.* Diagnostic and prognostic values of the QRS-T angle in patients with suspected acute decompensated heart failure. *ESC Heart Fail* 2020;**7**:1817–1829.

30. Maass AH, Vernooy K, Wijers SC, van't Sant J, Cramer MJ, Meine M, *et al.* Refining success of cardiac resynchronization therapy using a simple score predicting the amount of reverse ventricular remodelling: results from the markers and response to CRT (MARC) study. *Europace* 2018;**20**:393.

31. Engels EB, Végh EM, van Deursen CJM, Vernooy K, Singh JP, Prinzen FW. T-wave area predicts response to cardiac resynchronization therapy in patients with left bundle branch block. *J Cardiovasc Electrophysiol* 2015;**26**:176–183.

32. Hadwiger M, Dagres N, Haug J, Wolf M, Marschall U, Tijssen J, *et al.* Survival of patients undergoing cardiac resynchronization therapy with or without defibrillator: the RESET-CRT project. *Eur Heart J* 2022;**43**:2591–2599.

33. Steyerberg EW, Harrell FEJ. Prediction models need appropriate internal, internal-external, and external validation. *J Clin Epidemiol* 2016;**69**:245–247.

34. Attia IZ, Tseng AS, Benavente ED, Medina-Inojosa JR, Clark TG, Malyutina S, *et al.* External validation of a deep learning electrocardiogram algorithm to detect ventricular dysfunction. *Int J Cardiol* 2021;**329**:130–135.

35. van de Leur RR, Bleijendaal H, Taha K, Mast T, Gho JMIH, Linschoten M, *et al.* Electrocardiogram-based mortality prediction in patients with COVID-19 using machine learning. *Neth Heart J* 2022;**30**:312–318.

36. Plesinger F, van Stipdonk AMW, Smisek R, Halamek J, Jurak P, Maass AH, *et al.* Fully automated QRS area measurement for predicting response to cardiac resynchronization therapy. *J Electrocardiol* 2020;**63**:159–163.