



**HAL**  
open science

## Deep learning analysis of ECG for risk prediction of drug-induced arrhythmias and diagnosis of long QT syndrome

Edi Prifti, Ahmad Fall, Giovanni Davogustto, Alfredo Pulini, Isabelle Denjoy, Christian Funck-Brentano, Yasmin Khan, Alexandre Durand-Salmon, Fabio Badilini, Quinn S Wells, et al.

► **To cite this version:**

Edi Prifti, Ahmad Fall, Giovanni Davogustto, Alfredo Pulini, Isabelle Denjoy, et al.. Deep learning analysis of ECG for risk prediction of drug-induced arrhythmias and diagnosis of long QT syndrome. *European Heart Journal*, 2021, 42 (38), pp.3948-3961. 10.1093/eurheartj/ehab588 . hal-03477972

**HAL Id: hal-03477972**

<https://hal.sorbonne-universite.fr/hal-03477972v1>

Submitted on 13 Dec 2021

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# Deep learning analysis of ECG for risk prediction of drug-induced arrhythmias and diagnosis of long QT syndrome

Edi Prifti<sup>1,2,\*</sup>, Ahmad Fall<sup>1</sup>, Giovanni Davogustto<sup>3,§</sup>, Alfredo Pulini<sup>1,4,§</sup>, Isabelle Denjoy<sup>5</sup>, Christian Funck-Brentano<sup>6</sup>, Yasmin Khan<sup>7</sup>, Alexandre Durand-Salmon<sup>7</sup>, Fabio Badilini,<sup>8</sup> Quinn S. Wells<sup>3</sup>, Antoine Leenhardt<sup>5</sup>, Jean-Daniel Zucker<sup>1,2</sup>, Dan M. Roden<sup>3,9,10</sup>, Fabrice Extramiana<sup>5</sup> and Joe-Elie Salem<sup>3,6,\*</sup>

<sup>1</sup> IRD, Sorbonne University, UMMISCO, 32 Avenue Henri Varagnat, F-93143 Bondy, France;

<sup>2</sup> Sorbonne University, INSERM, NutriOmics, 91 Boulevard de l'Hopital, F-75013, Paris, France;

<sup>3</sup> Department of Medicine, Vanderbilt University Medical Center, Nashville, TN, USA;

<sup>4</sup> Université de Paris, Faculty of Medicine

<sup>5</sup> CNMR Maladies Cardiaques Héritaires Rares, Hôpital Bichat, Paris, France;

<sup>6</sup> Clinical Investigation Center Paris-Est, CIC-1901, INSERM, UNICO-GRECO cardio-oncology program, Department of Pharmacology, Pitié-Salpêtrière University Hospital, Sorbonne Université; 7513, Paris, France ;

<sup>7</sup> Banook Group, Nancy, France;

<sup>8</sup> AMPS LLC, NYC, USA

<sup>9</sup> Department of Pharmacology, Vanderbilt University Medical Center, Nashville, TN, USA

<sup>10</sup> Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, TN, USA;

§ These authors contributed equally to the work

\* Corresponding authors

## Correspondence:

Edi Prifti: IRD, Sorbonne University, UMMISCO, 32 Avenue Henri Varagnat, F-93143 Bondy, France. E-mail: [edi.prifti@ird.fr](mailto:edi.prifti@ird.fr) <http://orcid.org/0000-0001-8861-1305>,

Joe-Elie Salem, Center of Clinical Investigation Paris-Est, CIC-1901, Departement of Pharmacology, Pitié-Salpêtrière University Hospital, Sorbonne Université; 47 Boulevard de l'Hopital, 75013 PARIS ; <https://orcid.org/0000-0002-0331-3307>

**Word Counts:** 5000/5000 words

**Abstract:** 250/250 words

**Figures:** 8 (and 6 supplemental)

**Tables:** 0 (and 1 supplemental)

**NCT:** NCT00773201

## Abstract

**Aims:** Congenital (cLQTS) or drug-induced (diLQTS) long-QT syndromes can cause Torsades-de-Pointes (TdP), a life-threatening ventricular arrhythmia. The current strategy for identification of drugs at high-risk of TdP relies on measuring the QT-interval corrected for heart rate (QTc) on ECG. However, QTc has a low positive predictive value.

**Methods:** We used convolutional-neural-network (CNN) models to quantify ECG alterations induced by sotalol, an  $I_{K_r}$ -blocker associated with TdP, aiming to provide new tools (CNN-models) to enhance prediction of diTdP and diagnosis of cLQTS. Tested CNN-models used single or multiple 10sec recordings/patient using 8 leads or single leads in various cohorts: 1029 healthy subjects before and after sotalol intake (n=14135 ECGs); 487 cLQTS patients (n=1083 ECGs: 560 type 1, 456 type 2, 67 type 3); 48 patients with diTdP (n=1105 ECGs, with 147 obtained within 48hours of a diTdP episode).

**Results:** CNN-models outperformed models using QTc to identify exposure to sotalol (ROC-AUC=0.98 vs. 0.72,  $p \leq 0.001$ ). CNN-models had higher ROC-AUC using multiple vs. single 10s-ECG ( $p \leq 0.001$ ). Performances were comparable for 8-lead vs. single lead models. CNN-models predicting sotalol exposure also accurately detected presence and type of cLQTS vs. healthy controls, particularly for cLQT2 (AUC-ROC=0.9), and were greatest shortly after a diTdP event and declining over time ( $p \leq 0.001$ ), after controlling for QTc and intake of culprit drugs. ECG segment analysis identified the J-T<sub>peak</sub> interval as the best discriminator of sotalol intake.

**Conclusion:** CNN-models applied to ECGs outperform QTc measurements to identify exposure to drugs altering the QT-interval, congenital LQTS, and are greatest shortly after a diTdP episode.

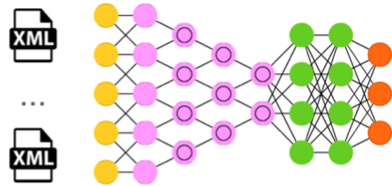
**Keywords:** Torsades de Pointes, machine learning, risk prediction, interpretability, long QT

# Graphical Abstract.

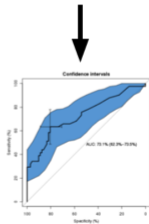
## Training artificial intelligence models



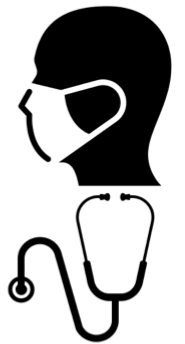
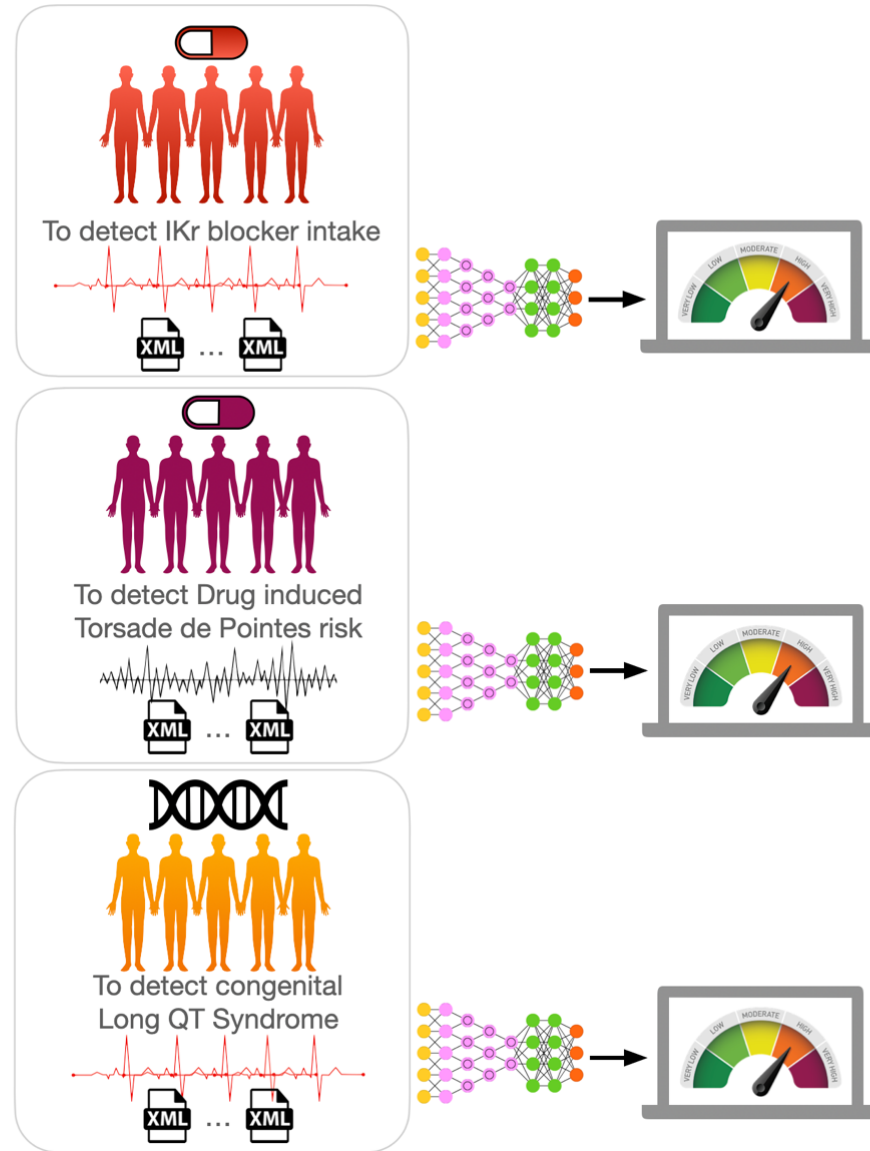
Deep Convolutional Neural Networks models



Models (multilead, unilead) detecting IKr blockade



## Applying artificial intelligence models



## Introduction.

Torsades-de-Pointes (TdP) is a distinctive form of life-threatening polymorphic ventricular arrhythmia associated with prolonged QT interval, corrected for heart rate (QTc), on ECG.<sup>1-3</sup> TdP and QTc prolongation are favored by congenital or drug-induced alterations in potassium and sodium cardiac channels.<sup>4-6</sup> There are three main forms of congenital long QT syndromes (cLQTS): type 1 and type 2 is caused by loss of function mutations in the potassium channels *KCNQ1* (cLQT1,  $I_{Ks}$  current) and *KCNH2* (cLQT2,  $I_{Kr}$  current) respectively; and type 3 is caused by mutations in *SCN5A* increasing the non-inactivating “late” sodium current  $I_{NaL}$  (cLQT3).<sup>6-8</sup> Drug-induced LQTS (diLQTS) is the other main cause of TdP, with almost all culprit drugs blocking  $I_{Kr}$  and the most torsadogenic among them also activating  $I_{NaL}$ .<sup>5</sup> Over 100 cardiac or non-cardiac drugs are currently approved despite favoring TdP risk because these drugs are thought to have a favorable risk-benefit assessment in some patients.<sup>9</sup>

QTc, which reflects ventricular repolarization duration, is the time between the beginning of the QRS complex and the end of the T-wave.<sup>10</sup> QTc is prolonged in cLQTS and diLQTS and is a hallmark of TdP. Specific T-waveform patterns have been described for each subtype of cLQTS and for diLQTS.<sup>11-14</sup> Current individual and population risk stratification strategies for TdP are almost exclusively based on the quantification of QTc.<sup>4</sup> Regulatory agencies require new drugs to undergo thorough QT studies, where the magnitude of drug-induced QTc prolongation is evaluated as a surrogate for TdP risk.<sup>15</sup> However, limiting ECG evaluation to the QTc is poorly predictive of TdP.<sup>16</sup> An unbiased and complex examination of the ECG data beyond simple QTc prolongation could provide relevant insight into identifying drugs and patients at risk of TdP.

Artificial intelligence is being increasingly applied to complex medical problems.<sup>17</sup> Techniques such as deep-learning, including convolutional neural networks (CNN), are bringing a radical change in the field of pattern recognition, improving earlier models in learning tasks such as image classification,

ECG analysis and natural language processing.<sup>18-20</sup> Herein, we tested if such models were able to learn the ECG footprint of sotalol, an IKr-blocker drug inducing TdP, to develop a new tool using ECG to recognize beyond QTc, exposure to IKr-blocker drugs; improve prediction of drug-induced TdP events and classification of cLQTS types particularly cLQT2.

# Methods

## Study cohort datasets and QTc measurement

We studied ECGs from four cohorts (**Figure 1**). The “Generepol cohort” (*NCT00773201*)<sup>14, 21</sup> was conducted at Pitié-Salpêtrière Clinical Investigation Center (start-end: 2008-2012, Paris, France): ECGs from 990 healthy subjects were recorded before and 1, 2, 3 and 4 hours after an 80mg oral sotalol dose (sofT1, sofT2, sofT3 and sofT4). The “Pharmacia’s cohort” was an open-label, nonrandomized study involving healthy controls, (n=39, including 28 males) receiving a fixed oral sotalol sequence administered on three successive days: 24-hour baseline without sotalol (Day 0); 160mg in all participants at 8:00AM Day 1; and 320mg in 21 males at 8:00AM Day 2. The study was conducted at Pharmacia’s Clinical Research Unit (start-end:2002; Kalamazoo, MI).<sup>22, 23</sup> The “cLQTS cohort” was 487 patients confirmed by genetic testing to have one the three main cLQTS followed at the Arrhythmia unit of Bichat Hospital (start-end: 1992-2018, Paris, France; 64% asymptomatic).<sup>24</sup> The “diTdP cohort” was 48 patients prospectively enrolled and followed at Vanderbilt University Medical Center (start-end: 2002-2019, Nashville, TN, USA) who had experienced at least one drug-induced TdP (diTdP) episode; acute cardiac ischaemia at the time of the event and genetically-confirmed underlying cLQTS were exclusion criteria. All cohorts were approved by institutional review boards, and written informed consent was obtained from participants when appropriate.

Recordings from these patients were curated by two expert cardiologists and tracings with ventricular or junctional tachycardia during the 10s acquisition were excluded from the analyses. In all cohorts, QTc was heart-rate corrected with Fridericia’s formula and details concerning the respective inter and intra-observer variability for QTc measurements in the cohorts are detailed elsewhere.<sup>10, 14, 21-24</sup>

## Data preparation

Raw 10s-ECG data (sampling frequency: 250 and 500Hz) were acquired with a variety of devices at the different centers. The 250 Hz signals were up-sampled to 500 Hz using a cubic interpolation. The ECG contained 8 independent leads (LI, LII, V1-V6), allowing for reconstruction of 12 leads (addition of LIII, aVF, aVL, aVR). ECG were provided in .scp or .xml files depending on recording devices (General Electric MAC5500, Marquette MAC15/MACVU, M3700 System, PageWriter Touch/Trim/XL/TC, Mortara ELI200 and Cardionics Cardioplug devices). They were parsed using Biosig software, and Python xmldict library, as appropriate.<sup>25</sup> The data were stored in Python dictionaries and converted onto 3D tensors (8 leads, 5000 time points for each lead, recordings) used to train and test the models. Standardization was performed at the whole ECG level, with each lead signal standardized by the mean of all other lead signals for models including all 8 leads (“multilead”) and at the lead level for “unilead” models. No other transformations, including filtering, were used.

## Sotalol-intake classification with the multilead and unilead models

We used either the eight leads concomitantly (LI, LII, V1-6; termed “multilead”) or each of the 8 leads independently (“unilead”) to train a CNN model to predict *Sot+* (having received sotalol, as a surrogate for  $I_{Kr}$  blockade) and *Sot-* classes (normal ECG before sotalol intake). The Generepol cohort (healthy volunteers before and after sotalol intake) was split into two sets: general training (80% for multilead models, 90% for unilead models) and holdout (20% for multilead models, 10% for unilead models). Ten times 10-fold cross-validation was performed in the general training set for parameter optimization. Each split was performed according to the subjects’ IDs and therefore each training partition had distinct subjects from the testing split. Descriptions of how the multilead and unilead models were constructed are provided in the **supplementary data (Figure S1 and S2)**. After cross-validation, each model was trained on the training set of the Generepol cohort. Then, models were



tested on the holdout Generepol set (completely independent from the training set) and the three other study cohorts.

### **Voting vs. single ECG analysis**

Performance indicators were computed using both 10s single-ECG signal analysis (ECG level in figures) and by averaging risk scores from multiple recordings acquired within minutes of a same timepoint (multiple 10s recordings; patients' level in figure; ie. The voting analysis) for a given patient and condition. The output provided by the models was a score ranging from 0 to 1 indicating a likelihood of being *Sot+* (having ingested sotalol). In order to classify a patient into *Sot+* or *Sot-* classes, ECG from the same patient were processed by the models and the patient was affected as being *Sot+* (versus *Sot-*) based on the mean classification score of the different 10s-ECG; on which a threshold of 0.5 was applied (*Sot+* if  $\text{score} \geq 0.5$ ). The performance metrics of all tested models can be found in **Figures 3-5**, **Supplementary Figures 3-4** and **Supplementary Table 1**.

### **Embedding analyses**

CNN models generate outputs, such as *Sot+* or *Sot-*, by analyzing raw input through a series of intermediate “layers” termed embeddings.<sup>26</sup> A distinctive feature of CNN models is their ability to discover novel representations of complex data, and one way to access such knowledge is by extracting the embeddings (transformation of the input data by the neural network). In this study, ECGs were transformed in the CNN embeddings by deriving vectors of 512 values. To represent these complex datasets in two dimensions for human interpretation, a nonlinear dimension reduction technique was applied based on the t-SNE algorithm<sup>27</sup> (perplexity=100, iteration=1000) using the Rtsne package. ECG data (vectors of 512 values) were thus visualized annotated as points on these maps. All dimensions of the embeddings were used to identify partitions with the k-means method with default

parameters implemented in base R. Details concerning embedding analyses are in the supplementary data (**Figure S5**).

### **ECG segment occlusion analysis (interpretability)**

We sought to identify which parts of the ECG signals were most useful in our classification models to classify an ECG as *Sot+*. To accomplish this goal, we iteratively dropped (“occluded”) a predefined portion of the data (in this case, a segment of the ECG signal) and reperformed the prediction. Here, we used a window of 50 points (corresponding to 100ms in 500Hz recordings) that was iteratively moved across the signal to identify which parts of the ECG signal were the most useful for the classification of ECG as *Sot+*. Feature Importance profile (FIP) was generated for each segment and provided us with a relevant score for identifying which ECG segments were more or less important for predicting *Sot+*. Details concerning occlusion methods are in the supplementary data (**Figure S6**). We implemented the occlusion method in Python with Tensorflow-2.

### **Statistical analyses**

Data are presented as count and frequencies, or median and IQR for categorical and continuous variables, respectively. We used mixed-effects linear models to best describe the data and their relations while controlling for random effects such as patient ID (multiple recordings per patient). Models were compared using ANOVA and the best models were selected based on AIC. Accuracy, recall, precision, f1-score and ROC-AUC were used to evaluate the different models generated. The Chi<sup>2</sup>-test was used for comparing proportions. Statistics and graphics were performed using R-packages (lme445, lmerTest, ggplot2, pROC). A  $p \leq 0.05$  was deemed significant; all tests were two-tailed.

**Data sharing.** Data and models used herein are available from the corresponding authors, upon reasonable request.

# Results

## Study population characteristics

The main characteristics of the 4 study cohorts are summarized in **Figure 1**.

The Generepol cohort contained 10,292 10s-ECG recordings from 990 healthy subjects (62% women, median [range] age=24[18-60]) in sinus rhythm before and 1, 2, 3, and 4 hours after the administration of 80mg sotalol (respectively denoted as baseline and sotT1-sotT4). The median number of 10s-ECG/ participant in this cohort was 15 [range=12-18].

The Pharmacia's cohort contained 3,843 10s-ECG recordings from 39 healthy subjects (46% women, median [range] age=25[18-45]) in sinus rhythm before and up to 12 hours after the intake of 160mg sotalol on Day 1 and 320mg sotalol on Day 2. The median number of 10s-ECG/ participant in this cohort was 114 [range=42-117].

The cLQTS cohort included 487 participants (median [range] age=28[0-84]; confirmed by genetic testing) with 1,083 10s-ECG recordings (median number of ECG/patient=3, IQR=6, longest follow-up=23 years). The three cLQTS types were represented, with 266 cLQT1 (62% women), 188 cLQT2 (54% women), and 33 cLQT3 (45% women) patients. A total of 213 participants (44%) had at least one recording performed while on beta-blocker, with 116, 88 and 9 (44%, 47% and 27%) participants for cLQT1, cLQT2, and cLQT3, respectively. ECGs were in sinus rhythm, except for 8 (0.7%) with supra-ventricular arrhythmia and 2 (0.2%) with either atrial and/or ventricular pacing.

The diTdP cohort included 48 participants (60% women; median [range] age at the time of the first ECG=60[18-85] years) with 1,105 10s-ECG recordings (median number of ECG/patient=31). The median follow-up was 4 years [range=0-17]. Sixty-six percent of the 10s-ECG (n=733/1105) were recorded while patients were on  $I_{Kr}$ -blocker drugs with known-risk for TdP,<sup>9</sup> with amiodarone (29/48), sotalol (12/48), dofetilide (9/48), fluconazole (7/48) and hydroxychloroquine (4/48), being the most prevalent.<sup>4, 14, 28</sup> Some patients took multiple drugs with TdP known-risk (one drug: 69%, 2 drugs:

24%, 3 drugs: 5%). Recordings from these patients were classified into four categories using combination of delay between ECG intake and the diTdP event, associated with the presence/absence of premature ventricular contractions (PVC): <24h, 24-48h, >48h+PVC and >48h-PVC. Of these 1,105 ECG recordings, 930 were obtained in sinus rhythm (84%), 171 (15%) in supraventricular arrhythmia, and 4 in junctional rhythm. A total of 162 (15%) and 183 (17%) 10s-ECG had at least one ventricular and/or an atrial paced complex. At least one PVC was seen in 143 (13%) ECGs.

## QTc evaluation

Serial QTc surveillance is the method cardiologists use to evaluate TdP risk in clinical practice.<sup>15</sup> When QTc>480ms or is increased  $\geq 60$ ms after drug intake compared to baseline, patients are considered at potential TdP risk.<sup>15</sup> In Generepol, the mean QTc at baseline was 14ms higher in women versus men ( $391\pm 15$ ms vs.  $377\pm 16$ ms;  $p<2e-16$ ), as is well-recognized.<sup>8, 29</sup> The maximal QTc prolongation after sotalol was more pronounced in women versus men ( $34\pm 14$ ms vs.  $23\pm 12$ ms;  $p<2e-16$ ). Similar results were obtained in Pharmacia's study when comparing QTc before and after sotalol intake (**Figures 1 and 2**).

In cLQTS, no difference in QTc was detected among the three types of cLQTS on the first ECG available for each patient ( $449\pm 36$ ms,  $453\pm 40$ ms, and  $452\pm 35$ ms; for cLQT1, cLQT2, and cLQT3 respectively,  $n=483$ ; **Figure 2**). Mean QTc in the cLQTS cohort was 65ms greater than the pre-sotalol values from Generepol ( $451\pm 38$  vs.  $386\pm 18$ ms,  $p<2e-16$ ).

In the diTdP cohort, QTc values were higher within 24-hour of diTdP events ( $501\pm 70$ ms) versus within 24-48h ( $478\pm 45$ ms;  $p<0.02$ ), or versus >48h with and without PVC ( $455\pm 50$ ms,  $p<2.14e-8$  and  $459\pm 45$ ms,  $p<8.5e-12$ , respectively; **Figure 2**). The mean QTc in the diTdP cohort was 86ms longer than the pre-sotalol values from Generepol ( $469\pm 63$ ms vs.  $386\pm 18$ ms,  $p<2e-16$ ).

## CNN models and Sotalol intake on ECG

To learn the sotalol footprint as a proxy of drug-induced  $I_{Kr}$ -blockade on ECG, we trained different CNN models (M) on a subset of Generepol. The first model used all leads (LI-II, V1-V6) from raw ECG data (M1:ecg\_multilead). A second model used clinical information (age, sex and serum potassium) in addition to the ECG data (M2:ecg\_multilead +clin). In this study, we first focused on 10s-ECG recordings at baseline before sotalol, and one, two, and three hours after sotalol intake.

The models provided an output score indicating a likelihood of sotalol intake in the [0-1] range. A score of 0 predicted the absence of sotalol intake whereas 1 corresponded to the highest probability for sotalol exposure. The mean predicted score at baseline was low (0.06) but increased rapidly for ECG recorded at one (*SotT1*, 0.80), two (*SotT2*, 0.88) and peaked at three hours after sotalol (*SotT3*, 0.95) (**Figure 3**); there were no sex differences. Notably, this increase in model score predictions tracked the increase in sotalol blood concentration (**Figure 3**). The output score was then converted into a binary variable based on a threshold (*Sot-* if model's derived score < 0.5, *Sot+* if  $\geq 0.5$ ). Performance indicators (ROC-AUC, accuracy, precision, recall (all ranging within [0-1]) and F1 score (ranging within [0-0.5]) were evaluated for each model in 10s-ECG recording individually or on the mean of multiple 10s-ECG of the same participant at a given time-point ("voting strategy"), in the training, cross-validation, and holdout sets (**Figures 3-4, Supp3-4**). The mean cross-validation ROC-AUC of M1:ecg\_multilead for discriminating ECG of patients before versus after sotalol intake was 0.948 when computed on single 10s-ECG and 0.98 with the voting approach. Similarly, for M2:ecg\_multilead +clin, the mean test accuracy was 0.948 (ECG) and 0.98 with voting (**Figures 3, Supp3**). No difference was observed between M1 and M2. This indicated that the information contained in age, sex, and serum potassium was likely embedded in the ECG footprint captured by the CNN model. Therefore, M1:ecg\_multilead model was deemed sufficient to be used thereafter. Its precision (voting), recall and F1-score were very high (0.955, 0.927, 0.470, respectively).

For comparison with current practice, we also tested the performance of QTc (M3:QTcF) alone, and with the same additional clinical information as above (M4:QTcF+clin) to discriminate on the presence/absence of sotalol intake. The linear regression model based on QTc alone (M3:QTcF), displayed a lower ROC-AUC of 0.695 (10s-ECG) and 0.720 (voting) versus M1:ecg\_multilead (ROC-AUC: 0.948 and 0.98, respectively;  $p < 1.5e-141$ ). After integration of clinical data to QTc (M4:QTcF+clin), models performance increased significantly ( $p < 3.3e-16$ ) to 0.717 (ECG) and 0.750 (voting) versus M3:QTcF (**Figures 3, Supp3**). Overall, QTc models were less effective than CNN models, even after integration of relevant clinical covariates. All four models (M1-4) displayed significantly higher ROC-AUC with the voting versus individual 10s-ECG strategy ( $p < 1.2e-20$  for M1:ecg\_multilead, **Table Supp1**). This demonstrates the importance of having longer recordings of at least 30s (mainly triplicates of 10s-ECG in our study). Results were similar in the holdout set (**Figures 3-4, Supp3**). All performance indicators for all these models are in **Figures Supp3-4 and Table Supp1**.

Thereafter, we tested the hypothesis that the ECG footprint for sotalol exposure could also be detected by analysis of single leads. For this, we trained eight different models – one for each lead (LI, LII, V1-V6; see **methods**). Their performances were comparable to the multilead models (**Figures 3, Supp4**). The best scores were obtained with the model trained and tested on lead LII (M5:ecg\_unilead\_LII; ROC-AUC=0.958 (10s-ECG) and 0.992 (voting) in the holdout set). When this model trained on one lead was tested on the rest of the leads, it performed well, with mean holdout AUC-ROC of 0.883 (10s-ECG) and 0.96 (voting). However, while the mean recall was high 0.913 (10s-ECG) and 0.963 (voting), the precision was lower 0.597 (10s-ECG) and 0.605 (voting). Similar results were obtained with other unilead models, except for the one trained on V1, which did not generalize well on the other leads (**Figure Supp4, Table Supp1**).

Finally, we validated M1:ecg\_multilead and M5:ecg\_unilead\_LII models (trained in the training subset of Generepol) in Pharmacia's cohort, an independent dataset of healthy controls before and after sotalol intake. Both M1 and M5 models performed very well to discriminate sotalol intake using ECGs (ROC-AUC=0.94-0.98 depending on the models, 10s-ECG vs. voting; **Figure 5**).

## **CNN models and cLQTS types**

Since diLQTS and cLQTS are both characterized by prolonged QTc, we hypothesized that the models (M1:ecg\_multilead) trained to recognize the sotalol ECG footprint would also be able to discriminate ECG from cLQTS subjects compared to Generepol baseline data, particularly for cLQT2, which shares the same pathophysiological mechanism of  $I_{Kr}$  blockade with sotalol-induced LQTS. We used M1:ecg\_multilead trained on a subset of Generepol (80%) and applied it to evaluate its potential in discriminating ECG from the healthy volunteers before and after sotalol intake (20% holdout from Generepol, never used for training) and cLQTS patients. The model's prediction results confirmed our hypothesis (**Figure 4**). First, we showed that the vast majority of ECGs before and after sotalol intake (95%, 97%, voting, respectively) from holdout Generepol cohort were correctly classified as *Sot+* and *Sot-*, respectively (**Figure 4**). Second, most ECGs from cLQTS (66%) were classified as *Sot+*. cLQT2 displayed the strongest proportion (80%, 74%) of *Sot+*, followed by cLQT3 (64%, 67%) and cLQT1 (55%, 51%) at the individual 10s-ECG level and after voting, respectively. **Figure 4** displays AUC-ROC results comparing ECGs from healthy participants on sotalol versus their baseline ECG before sotalol (controls), cLQT1, cLQT2 and cLQT3. M1:ecg\_multilead was highly efficient (AUC-ROC=0.9) in discriminating cLQT2 from healthy controls contrasting with low AUC-ROC (0.58) of ECG analysis from cLQT2 versus healthy subjects having received sotalol. These results indicate that M1:ecg\_multilead could not discriminate well between these latter two groups, supporting the hypothesis of shared ECG footprint alterations between cLQT2 and sotalol intake ( $I_{Kr}$  blockade).

Notably, M1:ecg\_multilead moderately separated cLQT2 from cLQT1 and cLQT3 (**Figure 4C**). The mean M1:ecg\_multilead ECG-derived score in cLQT2 was 0.53, significantly higher than cLQT1 (0.34,  $p < 7.4e-7$ ) and cLQT3 (0.43,  $p < 0.14$ ), after adjustment for beta-blockers (accounting for significant interaction between beta-blockers intake and cLQT2, effect-size=0.19,  $p < 7.3e-6$ ; but not for other cLQTS types). Of note, age and sex were not significantly associated with M1:ecg\_multilead score in cLQTS.

### **CNN models and diTdP events**

We evaluated M1:ecg\_multilead model to predict the risk of diTdP events in the diTdP cohort. We quantified the association between M1:ecg\_multilead score and the TdP footprint on ECG from patients who had had a diTdP event. The TdP footprint was coded as a four-class variable combining the delay from the diTdP event (<24h, 24-48h, >48h) and the existence or absence of PVCs in the >48h subgroup. Using a mixed linear model, we showed that TdP footprint was associated with the M1:ecg\_multilead score (highest within 24h from diTdP vs. >48h from diTdP without PVC (mean:0.68 vs. 0.56,  $p < 0.0018$ , respectively; **Figure 6**) after adjusting for a significant association with QTc ( $p < 1.87e-10$ ) and intake of drugs with a known risk for TdP ( $p < 3.17e-7$ ).

### **CNN and novel representation of ECG data**

The complex representation of an ECG, learned by the layers of the M1:ecg\_multilead CNN model, is contextual to the presence or absence of the sotalol footprint. We extracted these representations (embeddings) of all the ECGs of the studied cohorts by accessing the output of the last convolutional layers (see **supplementary methods**). When annotating all ECGs from Generepol as a function of the M1:ecg\_multilead predicted risk score (**Figure 7A**), we noticed a gradient pattern corresponding closely to the time between ECG acquisition and sotalol intake (**Figure 7B**). This demonstrated the



relevance of what the model “learned” from the ECG data in recognizing sotalol exposure. In cLQTS, most of cLQT2 ECG were located in the high-level score zone of the T-SNE map (top part of the map), indicating ECG features resembling those of sotalol induced  $I_{Kr}$ -blockade as seen previously. This, contrasts with those from cLQT1 and cLQT3, which were uniformly distributed in the t-SNE map (**Figure 7C**). In the diTdP cohort, most ECGs were located near the average to high-risk zones of the t-SNE map, being particularly high when recorded within 24h of the diTdP (**Figure 7D**), at a time when residual  $I_{Kr}$  blockade was most likely to be present. Taken together, these results indicate that the classification accuracy in recognizing the sotalol footprint also extends to CNN M1-model-identified embeddings, which condense clinically relevant information. Such novel representations of the data open perspectives for novel TdP risk stratification of ECG and patients (**Supp Figure S5**).

## **Interpretability analyses of CNN**

**Figure 8** displays the results of the “occlusion analysis” designed to identify ECG sub-segments (i.e. features) most important for the models. In lead II, we found that the standardized feature importance profile (FIP) changed with increased sotalol blood concentration (maximum at 3h in Generepol). Initially, at inclusion (before sotalol intake), the FIP was highly negative over the QRS and positive, although with low amplitude, on the P-wave offset and T-wave onset and offset. These features are used by the model to recognize normal ECG complexes without a sotalol footprint — the QRS complex indicating a regularly occurring attribute used to calibrate the data input. One hour after sotalol intake, the FIP distribution started to change. The FIP intensity of the QRS decreased and the importance of the signal after the T-wave and before P-onset increased. This region corresponds to the RR time, i.e., the cardiac heart rate. Indeed, sotalol has beta-blocking properties known to slow the sinus rate, which were captured by the model. Two hours after sotalol, the FIP increased in the first part of the T-wave (corresponding to the J-Tpeak interval), which reached maximum intensity 3h after

sotalol. At that time,  $I_{Kr}$ -blockade was active and strongly apparent on ECG. We performed the same experiment in unilead models trained on V2 and V3 and FIP behaved similarly (**Figures 8, Supp6**).

## Discussion

QTc prolongation, although imperfect, has been shown to be associated with TdP and is currently used in clinical practice as a surrogate for evaluating the risk of TdP.<sup>30</sup> Here, we propose a new approach to improve TdP risk prediction. We hypothesized that it would be possible to use cutting edge artificial intelligence models to learn the footprint of drugs at high-risk of TdP in healthy volunteers. We then used these models to quantify a novel risk score in other participants exposed to these drugs or in patients with cLQTS. The main finding of our study is that training deep CNN models using raw digital ECG data allows for an automated and comprehensive TdP risk stratification that complements QTc measurement. The CNN was trained to recognize ECG alterations induced by sotalol as a model of  $I_{Kr}$ -blockade, the major mechanism by which drugs cause QTc prolongation and predispose to TdP.<sup>14</sup> The CNN models accurately detected ECG associated with the intake of drugs at risk of TdP, and discriminated the presence and type of cLQTS, being particularly accurate for cLQT2. Moreover, these models improved prediction of diTdP event, even after controlling for QTc and intake of drugs at known-risk of TdP. Analyses of the CNN models highlighted specific interpretable ECG features, particularly the J-Tpeak interval to recognize the sotalol-induced ECG footprint. Models based on a single lead performed in general as well as those using 8 leads, except for V1.

Because TdP is a relatively rare event, we first used a population of healthy volunteers exposed to sotalol so we could generate enough labeled data for the CNN model to be robust. The rationale for using a cohort exposed to sotalol is that this drug is known to prolong ventricular repolarization through  $I_{Kr}$ -inhibition, that rarely but-dose dependently can lead to TdP.<sup>31, 32</sup> The CNN models developed here were able to accurately classify if a patient was or not exposed to sotalol, regardless of the time after drug intake. Furthermore, multiple acquisitions taken together with a voting approach improved the classification. This demonstrated the presence of rich information contained within the electrocardiograms, exceeding the sole measurement of QTc including with relevant clinical

information. Classification from ECG features learned in the CNN models could become a useful approach in compliance ascertainment and drug adjustment; eventually more practical, less costly and faster than standard blood analysis.

Similar molecular and physiological mechanisms to sotalol action are known to be involved in cLQT2 patients with *KCNH2* mutations, which also leads to decrease  $I_{Kr}$  current.<sup>14</sup> Here, we demonstrated that the similarities of the sotalol ECG footprint with cLQT2 allowed to accurately classify 80% of the ECG from cLQT2 patients. This result has potential clinical applications such as screening incoming patients for cLQTS and discrimination of types, with very low cost, before using more expensive genetic tests or scarce expert ECG repolarization evaluation. Although QTc is prolonged in all cLQTS, the ECG waveforms carry specificities including T-wave morphology abnormalities that are specific to each type of cLQTS.<sup>33</sup> However, the models developed herein were not trained to distinguish the different cLQTS groups, particularly cLQT1 and 3, for which more data are needed.

When applied to an independent study cohort of patients who experienced diTdP events, our CNN-model-derived scores were higher within 24 hours of the diTdP events vs. ECGs from same individuals more than 24 hours (and even more 48 hours) after or before the event. These results indicate that such models could be helpful to diagnose patients who experienced an out-of-hospital TdP event or even risk stratify patients with continuous surveillance for emerging diTdP events.

To the best of our knowledge, this is the first study, which successfully deploys the original approach of learning drug footprints to predict drug-induced heart pathology risk based on ECG. A prior study was able to correlate drug concentrations on ECG using CNN.<sup>34</sup> The authors analyzed 10s-ECG recordings of 42 patients receiving dofetilide, another  $I_{Kr}$ -blocker antiarrhythmic drug, or placebo. In their experiments, they used the data from two distinct prospective randomized controlled trials available in the Physionet repository,<sup>35</sup> and found that their CNN model was superior to QTc

alone in predicting plasma dofetilide concentration. However, the database used in their study was relatively small (dozens of patients) and they did not use cross-validation in training, with the well-known risk of overfitting. Furthermore, they could not assess the capacity of their AI model to detect an arrhythmic risk, or cLQTS, and interpretability of their findings was not performed (**Figure 8**) as done in the present study.

Other studies have focused on CNN modeling of other cardiac diseases using multilead ECG input. For the detection of anterior myocardial infarction,<sup>36</sup> Liu et al. used a 4-lead approach that led to accuracies >90% with a 5-fold cross-validation. Tison et al. created a CNN-hidden Markov model that took 12-lead input in order to detect pulmonary arterial hypertension, hypertrophic cardiomyopathy, cardiac amyloid and mitral valve prolapse.<sup>37</sup> The ROC-AUC were in the 77%-94% range for the 4 conditions. Similar technology was also used by Attia et al,<sup>34</sup> who applied a CNN model on a large database (n >97,000 patients) to detect left ventricular dysfunction. They used a large holdout set (n>52,000 patients) and achieved an overall accuracy of 86%. Moreover, a subset of the patients, which were erroneously classified as ventricular dysfunction, later developed a low ejection fractions, suggesting that the model was able to detect features of this condition before it became clinically diagnosed. Unfortunately, the healthy volunteers from Generepol misclassified by our model as taking sotalol before any intake were not followed, precluding any evaluation of their subsequent risk for TdP and sudden-death.

We introduced CNN models trained with data obtained from one lead only. They were as accurate as the multilead model not only when classifying holdout data from the same leads but also from leads on which they were not trained. This is an unexpected result, and indicates that the sotalol footprint is detected by all leads and in similar ways, with the exception of lead V1. Moreover, the ECG data were recorded with different acquisition devices and some ECGs, recorded in 250Hz, were upsized using interpolation techniques. Still, the results were robust, regardless of the recording device. This paves

the way to clinical applications where the patients or physicians could record a single electrode ECG, which could then be sent to a centralized server and analyzed by the CNN models, with the goal of stratifying the risk for the patient to develop a TdP.

We also explored the CNN models to understand how the decision process was made and what was the model looking for in the ECG to provide a prediction. The occlusion-based interpretability algorithm uncovered the sotalol ECG footprint, which changed with time as the blood sotalol concentration increased. The analysis of the footprint was consistent with existing knowledge on how sotalol influences cardiomyocyte action potential, mainly through blockade of  $I_{Kr}$  and beta-adrenergic receptor blockade. This approach opens novel avenues of research and applications in the context of drug monitoring for the pharmaceutical industry, but also plays an important role in the acceptability of AI in clinics. Providing an explanation for the prediction process is increasingly requested when not mandatory,<sup>38,39</sup> especially for “black boxes” such as deep neural networks, which train millions of parameters. J-Tpeak features emerged as the main attribute allowing for discrimination of sotalol intake. This is concordant with the emerging literature on the importance of this specific segment when predicting for diTdP beyond QTc.<sup>40</sup>

Lastly, we demonstrated that besides risk prediction, the CNN models learn clinically relevant knowledge. A Post-Hoc analysis of the network’s deep embeddings grouped ECGs from the studied cohorts according to their clinical relevance (**Figure 7**). These embeddings can be used to automatically stratify ECG and ultimately patients, in novel classes that are yet to be characterized. However, identifying the best embeddings can be challenging since the number of model architectures to explore can be very large. More research and training data are needed in the context of translational clinical applications of CNN models for diagnosis of the different types of cLQTS and prediction of diTdP.

**Funding.** This study was partly funded by the French Research Agency (2021-2024): Agence nationale de la recherche (ANR) DeepECG4U project. Further validation of the algorithm developed herein is planned in real-life patients prospectively included with various age, gender, ethnicity, geographic location and cardiovascular risk factors within the ANR funded DeepECG4U project.

## REFERENCES

1. Dessertenne F. [Ventricular tachycardia with 2 variable opposing foci]. Arch Mal Coeur Vaiss 1966;**59**(2):263-72.
2. Rosso R, Hochstadt A, Viskin D, Chorin E, Schwartz AL, Tovia-Brodie O, Laish-Farkash A, Havakuk O, Gepstein L, Banai S, Viskin S. Polymorphic ventricular tachycardia, ischaemic ventricular fibrillation, and torsade de pointes: importance of the QT and the coupling interval in the differential diagnosis. Eur Heart J 2021.
3. Stramba-Badiale M, Karnad DR, Goulene KM, Panicker GK, Dagradi F, Spazzolini C, Kothari S, Lokhandwala YY, Schwartz PJ. For neonatal ECG screening there is no reason to relinquish old Bazett's correction. Eur Heart J 2018;**39**(31):2888-2895.
4. Roden DM. Drug-induced prolongation of the QT interval. N Engl J Med 2004;**350**(10):1013-22.
5. Yang T, Chun YW, Stroud DM, Mosley JD, Knollmann BC, Hong C, Roden DM. Screening for acute IKr block is insufficient to detect torsades de pointes liability: role of late sodium current. Circulation 2014;**130**(3):224-34.
6. Schwartz PJ, Ackerman MJ, Antzelevitch C, Bezzina CR, Borggrefe M, Cuneo BF, Wilde AAM. Inherited cardiac arrhythmias. Nat Rev Dis Primers 2020;**6**(1):58.
7. Itoh H, Crotti L, Aiba T, Spazzolini C, Denjoy I, Fressart V, Hayashi K, Nakajima T, Ohno S, Makiyama T, Wu J, Hasegawa K, Mastantuono E, Dagradi F, Pedrazzini M, Yamagishi M, Berthet M, Murakami Y, Shimizu W, Guicheney P, Schwartz PJ, Horie M. The genetics underlying acquired long QT syndrome: impact for genetic screening. Eur Heart J 2016;**37**(18):1456-64.
8. Salem JE, Yang T, Moslehi JJ, Waintraub X, Gandjbakhch E, Bachelot A, Hidden-Lucet F, Hulot JS, Knollmann BC, Lebrun-Vignes B, Funck-Brentano C, Glazer AM, Roden DM. Androgenic Effects on Ventricular Repolarization: A Translational Study From the International Pharmacovigilance Database to iPSC-Cardiomyocytes. Circulation 2019;**140**(13):1070-1080.
9. Schwartz PJ, Woosley RL. Predicting the Unpredictable: Drug-Induced QT Prolongation and Torsades de Pointes. J Am Coll Cardiol 2016;**67**(13):1639-1650.
10. Saque V, Vaglio M, Funck-Brentano C, Kilani M, Bourron O, Hartemann A, Badilini F, Salem JE. Fast, accurate and easy-to-teach QT interval assessment: The triplicate concatenation method. Arch Cardiovasc Dis 2017;**110**(8-9):475-481.
11. Moss AJ, Zareba W, Benhorin J, Locati EH, Hall WJ, Robinson JL, Schwartz PJ, Towbin JA, Vincent GM, Lehmann MH. ECG T-wave patterns in genetically distinct forms of the hereditary long QT syndrome. Circulation 1995;**92**(10):2929-34.
12. Schwartz PJ, Ackerman MJ. The long QT syndrome: a transatlantic clinical approach to diagnosis and therapy. Eur Heart J 2013;**34**(40):3109-16.
13. Salem JE, Bretagne M, Lebrun-Vignes B, Waintraub X, Gandjbakhch E, Hidden-Lucet F, Gougis P, Bachelot A, Funck-Brentano C, French Network of Regional Pharmacovigilance C. Clinical characterization of men with long QT syndrome and torsades de pointes associated with hypogonadism: A review and pharmacovigilance study. Arch Cardiovasc Dis 2019;**112**(11):699-712.
14. Salem JE, Germain M, Hulot JS, Voiriot P, Lebourgeois B, Waldura J, Tregouet DA, Charbit B, Funck-Brentano C. GENomE wide analysis of sotalol-induced IKr inhibition during ventricular REPOLarization, "GENEREPOl study": Lack of common variants with large effect sizes. PLoS One 2017;**12**(8):e0181875.
15. Administration USDoHaHSFaD. *E14 Clinical Evaluation of QT/QTc Interval Prolongation and Proarrhythmic Potential for Non-Antiarrhythmic Drugs — Questions and Answers (R3) from the FDA, although* <https://www.fda.gov/files/drugs/published/E14-Clinical-Evaluation-of-QT-QTc->



[Interval-Prolongation-and-Proarrhythmic-Potential-for-Non-Antiarrhythmic-Drugs-Questions-and-Answers-%28R3%29-Guidance-for-Industry.pdf](#)

16. Drew BJ, Ackerman MJ, Funk M, Gibler WB, Kligfield P, Menon V, Philippides GJ, Roden DM, Zareba W, American Heart Association Acute Cardiac Care Committee of the Council on Clinical Cardiology tCoCN, the American College of Cardiology F. Prevention of torsade de pointes in hospital settings: a scientific statement from the American Heart Association and the American College of Cardiology Foundation. *Circulation* 2010;**121**(8):1047-60.
17. Obermeyer Z, Emanuel EJ. Predicting the Future - Big Data, Machine Learning, and Clinical Medicine. *N Engl J Med* 2016;**375**(13):1216-9.
18. LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature* 2015;**521**(7553):436-44.
19. Hinton G. Deep Learning-A Technology With the Potential to Transform Health Care. *JAMA* 2018;**320**(11):1101-1102.
20. Bos JM, Attia ZI, Albert DE, Noseworthy PA, Friedman PA, Ackerman MJ. Use of Artificial Intelligence and Deep Neural Networks in Evaluation of Patients With Electrocardiographically Concealed Long QT Syndrome From the Surface 12-Lead Electrocardiogram. *JAMA Cardiol* 2021;**6**(5):532-538.
21. Salem JE, Dureau P, Bachelot A, Germain M, Voiriot P, Lebourgeois B, Tregouet DA, Hulot JS, Funck-Brentano C. Association of Oral Contraceptives With Drug-Induced QT Interval Prolongation in Healthy Nonmenopausal Women. *JAMA Cardiol* 2018;**3**(9):877-882.
22. Extramiana F, Badilini F, Sarapa N, Leenhardt A, Maison-Blanche P. Contrasting time- and rate-based approaches for the assessment of drug-induced QT changes. *J Clin Pharmacol* 2007;**47**(9):1129-37.
23. Sarapa N, Morganroth J, Couderc JP, Francom SF, Darpo B, Fleishaker JC, McEnroe JD, Chen WT, Zareba W, Moss AJ. Electrocardiographic identification of drug-induced QT prolongation: assessment by different recording and measurement methods. *Ann Noninvasive Electrocardiol* 2004;**9**(1):48-57.
24. Extramiana F, Badilini F, Denjoy I, Vaglio M, Green CL, Kligfield P, Leenhardt A, Maison-Blanche P. Sex influences on ventricular repolarization duration in normal subjects and in type 1, 2 and 3 long QT syndrome patients: Different effect in acquired and congenital type 2 LQTS. *J Electrocardiol* 2020;**62**:148-154.
25. Baillet S, Friston K, Oostenveld R. Academic software applications for electromagnetic brain mapping using MEG and EEG. *Comput Intell Neurosci* 2011;**2011**:972050.
26. Su C, Tong J, Zhu Y, Cui P, Wang F. Network embedding in biomedical data science. *Brief Bioinform* 2018.
27. Van der Maaten L, Hinton G. Visualizing data using t-SNE. *Journal of machine learning research* 2008;**9**(11).
28. Nguyen LS, Dolladille C, Drici MD, Fenioux C, Alexandre J, Mira JP, Moslehi JJ, Roden DM, Funck-Brentano C, Salem JE. Cardiovascular Toxicities Associated With Hydroxychloroquine and Azithromycin: An Analysis of the World Health Organization Pharmacovigilance Database. *Circulation* 2020;**142**(3):303-305.
29. Salem JE, Alexandre J, Bachelot A, Funck-Brentano C. Influence of steroid hormones on ventricular repolarization. *Pharmacol Ther* 2016;**167**:38-47.
30. Hondeghem LM. Drug-Induced QT Prolongation and Torsades de Pointes: An All-Exclusive Relationship or Time for an Amicable Separation? *Drug Saf* 2018;**41**(1):11-17.
31. Funck-Brentano C. Pharmacokinetic and pharmacodynamic profiles of d-sotalol and d,l-sotalol. *Eur Heart J* 1993;**14 Suppl H**:30-5.
32. Haverkamp W, Breithardt G, Camm AJ, Janse MJ, Rosen MR, Antzelevitch C, Escande D, Franz M, Malik M, Moss A, Shah R. The potential for QT prolongation and pro-arrhythmia by non-

- anti-arrhythmic drugs: clinical and regulatory implications. Report on a Policy Conference of the European Society of Cardiology. *Cardiovasc Res* 2000;**47**(2):219-33.
33. Porta-Sanchez A, Spillane DR, Harris L, Xue J, Dorsey P, Care M, Chauhan V, Gollob MH, Spears DA. T-Wave Morphology Analysis in Congenital Long QT Syndrome Discriminates Patients From Healthy Individuals. *JACC Clin Electrophysiol* 2017;**3**(4):374-381.
34. Attia ZI, Sugrue A, Asirvatham SJ, Ackerman MJ, Kapa S, Friedman PA, Noseworthy PA. Noninvasive assessment of dofetilide plasma concentration using a deep learning (neural network) analysis of the surface electrocardiogram: A proof of concept study. *PLoS One* 2018;**13**(8):e0201059.
35. Moody GB, Mark RG, Goldberger AL. PhysioNet: a research resource for studies of complex physiologic and biomedical signals. *Comput Cardiol* 2000;**27**:179-82.
36. Liu W, Zhang M, Zhang Y, Liao Y, Huang Q, Chang S, Wang H, He J. Real-Time Multilead Convolutional Neural Network for Myocardial Infarction Detection. *IEEE Journal of Biomedical and Health Informatics* 2018;**22**(5):1434-1444.
37. Tison GH, Zhang J, Delling FN, Deo RC. Automated and Interpretable Patient ECG Profiles for Disease Detection, Tracking, and Discovery. *arXiv:1807.02569 [cs]* 2018.
38. Martens D, Vanthienen J, Verbeke W, Baesens B. Performance of classification models from a user perspective. *Decision Support Systems* 2011;**51**(4):782-793.
39. Bryce G, Seth F. European Union Regulations on Algorithmic Decision-Making and a “Right to Explanation”. *AI Magazine* 2017;**38**(3).
40. Vicente J, Zusterzeel R, Johannesen L, Mason J, Sager P, Patel V, Matta MK, Li Z, Liu J, Garnett C, Stockbridge N, Zineh I, Strauss DG. Mechanistic Model-Informed Proarrhythmic Risk Assessment of Drugs: Review of the "CiPA" Initiative and Design of a Prospective Clinical Validation Study. *Clin Pharmacol Ther* 2018;**103**(1):54-66.

## Figures legend

### **Figure 1: Experimental design and main characteristics of study cohorts.**

Description of the main characteristics of the four study cohorts. The Generepol cohort was composed of healthy volunteers given a single 80mg dose of oral sotalol. This dataset was used to train and test the models. The Pharmacia's cohort was composed of 39 healthy volunteers before (Day 0) and after a single 160mg dose of oral sotalol (Day1), followed in some men by a single 320mg dose of oral sotalol (day2). This cohort was only used to test the models. The cLQTS cohort was composed of cLQTS patients of type 1, 2 and 3. The diTdP cohort included patients who experienced events of drug-induced TdP with no underlying identified cLQTS.

### **Figure 2. QTc distribution across study cohorts**

The distribution of QTc values following the sotalol-induced QTc prolongation in the Generepol (**A**) and Pharmacia's cohort (**D**). The X-axis represents the time (minutes) following sotalol administration and lines link ECG recordings from the same participant over the duration of the protocol. Summarized loess (local regression) distribution of the data +/- standard-error is overlaid on top and grouped by gender (males in blue and females in red). **B**) Boxplots of the estimated QTc values in the cLQTS cohort by subtypes. **C**) Boxplots of QTc values across the diTdP cohort grouped by time to TdP event and presence or not of premature ventricular contractions (PVC). ECG are grouped and colored by gender and the black horizontal lines indicate the 480ms at-risk QTc threshold (**A-D**).

**Figure 3: Classification performance of CNN and linear regression (QT) models in discriminating baseline ECG before sotalol from those after sotalol intake (SotT1, SotT2, SotT3) in Generepol.**

**A)** Boxplots, illustrating the distribution of circulating sotalol concentration (ng/ml) in Generepol cohort two and three hours after 80mg oral sotalol intake. Data are displayed separated and colored by gender. **B)** Scatterplot illustrating the evolution of the M1:ecg\_multilead classification score for the *Sot+* class (y-axis) across time from inclusion (x-axis) in the Generepol cohort. All points (averaged ECGs) of a study participant are linked together as trajectories and are colored by gender. Summarized loess (local regression) distribution of the data +/- standard-error is overlaid on top and grouped by gender. The red horizontal line corresponds to the *Sot+/Sot-* classification threshold ( $=0.5$ ). **C)** ROC-AUC for the CNN multilead models (M1, M2), non-CNN standard QT-based linear regression models (M3, M4) as well as all CNN unilead M5 models in classifying each individual 10s-ECG recording (top) or using a voting strategy (in triplicates of 10s-ECG per study participant and time point, bottom). Multiple 10s-ECG recorded at each time point were assigned a *Sot+* classification score. When the risk score was  $\geq 0.5$ , the ECG was classified as *Sot+*. With the voting approach, a mean *Sot+* classification score was computed. The same threshold was applied to predict the *Sot-/Sot+* class. Blue, orange and brown colors respectively depict the training, test and holdout subsets of the first study cohort (see **Figure 1**). Each model tested on the same lead as trained is annotated by a red star. For the multilead models, all leads are used to train and test.

**Figure 4: CNN model performance in classifying study participants as *Sot+/Sot-* in Generepol holdout dataset and cLQTS.**

**A)** Left: Percentage of all ECG for study participants, which are classified as *Sot+* in the holdout Generepol dataset (healthy volunteers before (Control) and one to three hours after sotalol intake (Sotalol)) as well as the cLQT1, cLQT2 and cLQT3 groups. Right: Similar to the left panel, with the exception that groups of ECG were classified as *Sot+* using the patient voting strategy instead of individual 10s-ECG. **B)** ROC curves indicating the separation between patients on sotalol (Sotalol)

and each of the control, cLQT1, cLQT2, cLQT3 groups. C) ROC curves indicating the separation between cLQT2 and cLQT1, cLQT3, sotalol exposed and control groups.

**Figure 5. CNN model performance in classifying study participants as *Sot+*/*Sot-* in Pharmacia's cohort.**

Scatterplot illustrating the evolution of the M1:ecg\_multilead (A) and M5:ecg\_unilead\_II (B) classification score for the *Sot+* class (y-axis) across time from inclusion (x-axis) in Pharmacia's cohort. All points, single ECGs of a study participant, are linked together as trajectories and are colored by gender. Summarized loess (local regression) distribution of the data +/- standard-error is overlaid on top and grouped by gender. The horizontal dotted line corresponds to the *Sot+*/*Sot-* classification threshold. C) ROC curves indicating the separation between subjects on sotalol (Sotalol; 2 to 4 hours post 160mg intake on Day 1, and 320mg Day 2) versus before (Day0 and before intake of Day 1) in Pharmacia's study.

**Figure 6: M1:ecg\_multilead *Sot+*/*Sot-* model's classification score in relation to diTdP imprint intensity in ECG.**

Boxplots indicating the distribution of CNN M1 model's classification score in patients' ECG as a function of TdP imprint intensity groups. Shape indicates the intake of drugs with known risk for TdP (triangles) vs. none (circles).

**Figure 7: Risk prediction model's deep embeddings reveal clinically relevant data structure.**





All panels illustrate two dimensions after a t-SNE transformation of the 512 dimensions of the multilead M1 model's embeddings (see methods). A) T-SNE map where each point represents an ECG from the Generepol cohort. Grayscale indicates the M1:ecg\_multilead classification score

ranging in [0:1]. **B)** Same t-SNE map as A) where ECG are colored by the experimental setup, from inclusion before and 1, 2, 3 or 4 hours after sotalol intake. **C)** Same t-SNE map with ECG from the cLQTS cohort. ECG are annotated by cLQTS type. **D)** Same t-SNE map as A) with ECG from the diTdP cohort of patients having experienced at least one diTdP event. ECG are colored by the four groups of TdP intensity footprint (timeframe from the diTdP event and presence/absence of PVCs on the ECG).

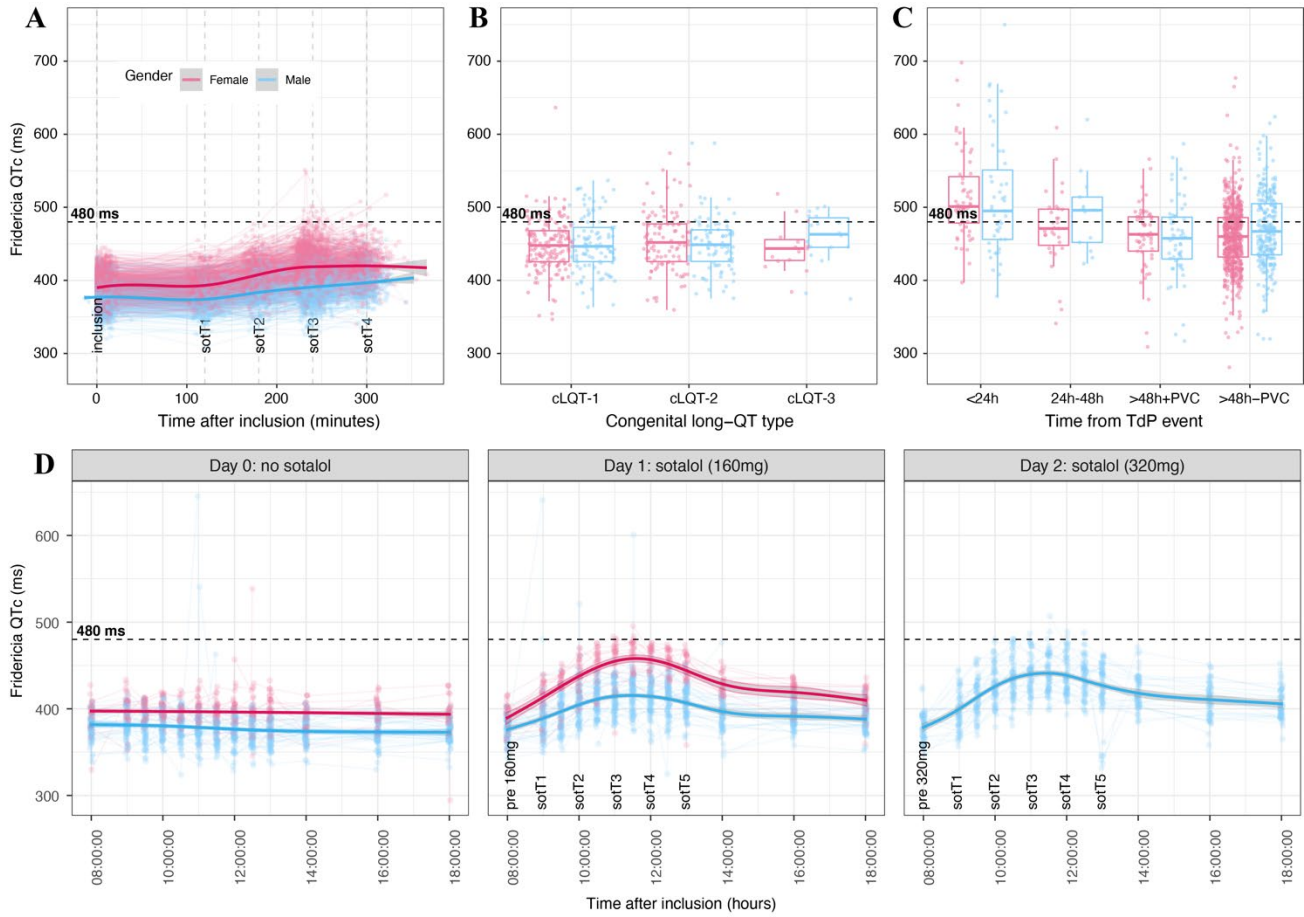
### **Figure 8: Interpretability of the sotalol footprint on the ECG signal**

This figure displays an averaged signal of the standardized ECG for each segmented beat for leads LII, V2 and V3. All signals from the same time points were analyzed together. Similarly, the standardized feature importance profile (FIP) is summarized and laid behind the ECG profile. Colors for both the ECG and FIP indicate intensity of the FIP.

**Figure 1. Experimental design and main characteristics of study cohorts.**

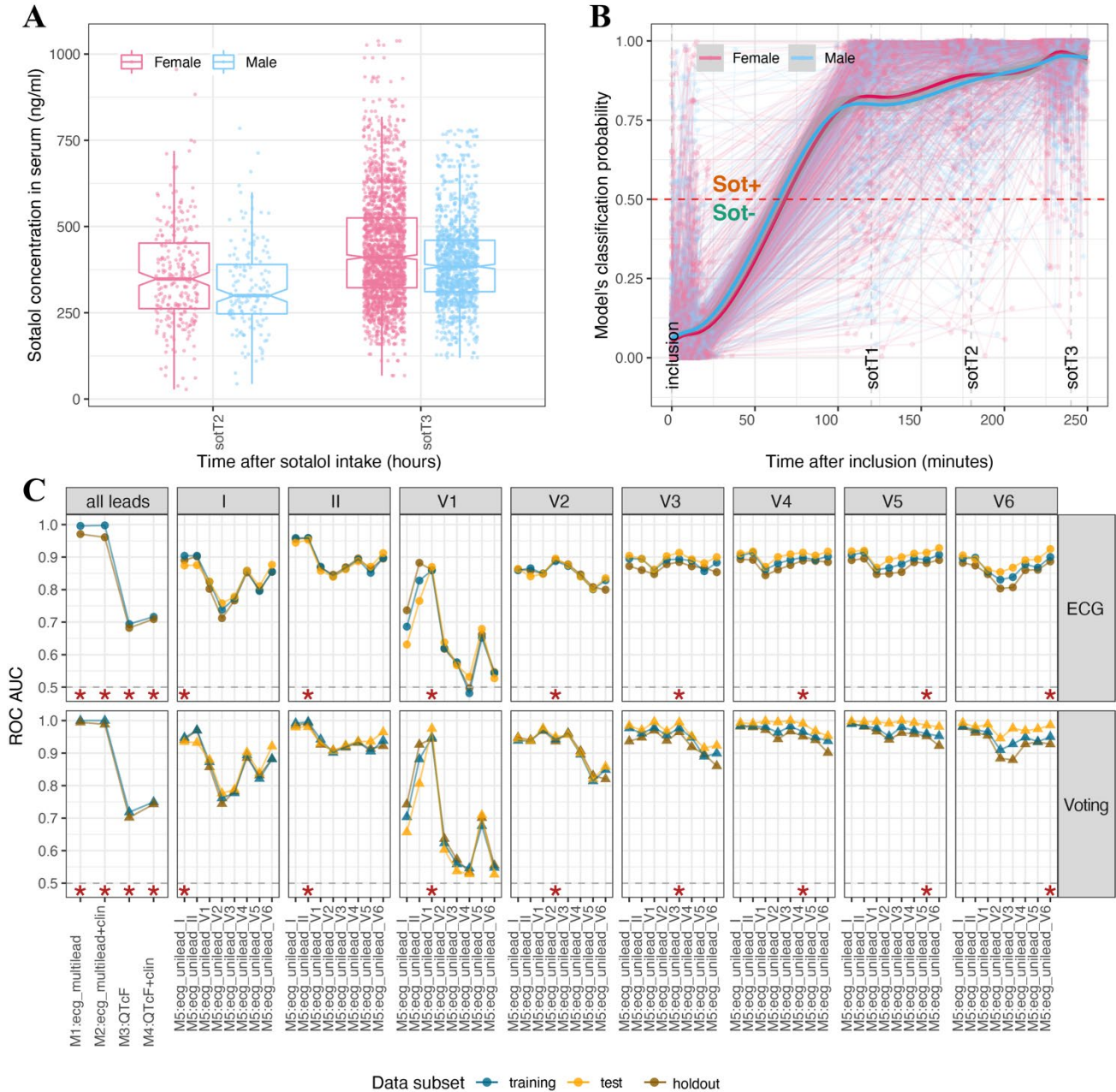
<p><b>Generepol</b> (n=990; 10292 ECG) <b>Training (80%)</b> (n=792; 8245 ECG) baseline (4014 ECG), sotalolol (4231 ECG) <b>Holdout (20%)</b> (n=198; 2047 ECG) baseline (2047 ECG), sotalolol (1048 ECG)</p> 	<p><b>Female</b> Age (28±11 years) (n = 614; 62%) QTc baseline = 391±15 ms QTc maximal (80 mg sotalol) = 425±21 ms Delta QTc max = 34±14 ms</p>	<p><b>Male</b> Age (28±10 years) (n = 376; 38%) QTc baseline = 377±16 ms QTc maximal (80 mg sotalol) = 400±20 ms Delta QTc max = 22±12 ms</p>
<p><b>Pharmacia</b> Total (n=39; 3843 ECG) Day 0 (n=39; 1542 ECG) Day 1 (n=39; 1482 ECG) Day 2 (n=21; 819 ECG)</p> 	<p><b>Female</b> Age (26±8 years) (n = 18; 46%) QTc baseline = 390±18 ms QTc maximal day 1 (160 mg sotalol) = 459±22ms QTc maximal day 2 (320 mg sotalol) = NA Delta QTc max day 1 (160 mg sotalol) = 74±16 ms Delta QTc max day 2 (320 mg sotalol) = NA</p>	<p><b>Male</b> Age (27±8 years) (n = 21; 54%) QTc baseline = 376±13 ms QTc maximal day 1 (160 mg sotalol) = 424±25 ms QTc maximal day 2 (320 mg sotalol) = 450±22 ms Delta QTc max day 1 (160 mg sotalol) = 48±21 ms Delta QTc max day 2 (320 mg sotalol) = 76±13 ms</p>
<p><b>Congenital LQT (cLQTS)</b> Total (n=487; 1083 ECG) LQT1 (n=266; 560 ECG) LQT2 (n=188; 456 ECG) LQT3 (n=33; 67 ECG)</p> 	<p><b>Female</b> Age (33±17 years) (n = 282; 58%) QTc cLQT1 = 448±35 ms QTc cLQT2 = 455±42 ms QTc cLQT3 = 446±33 ms</p>	<p><b>Male</b> Age (30±19 years) (n = 205; 42%) QTc cLQT1 = 451±38 ms QTc cLQT2 = 450±38 ms QTc cLQT3 = 458±36 ms</p>
<p><b>Drug-induced TdP (diTdP)</b> Total (n=48; 1105 ECG) 24h (n=38; 103 ECG) 48h (n=31; 44 ECG) &gt;48h+PVC (n=28; 115 ECG) &gt;48h-PVC (n=48; 843 ECG)</p> 	<p><b>Female</b> Age (56±16 years) (n = 29; 60%) QTc &lt;24h = 515±60 ms QTc &lt;48h = 471±54 ms QTc &gt;48h+PVC = 460±46 ms QTc &gt;48h-PVC = 460±43 ms</p>	<p><b>Male</b> Age (64±16 years) (n = 19; 40%) QTc &lt;24h = 514±81 ms QTc &lt;48h = 493±54 ms QTc &gt;48h+PVC = 457±54 ms QTc &gt;48h-PVC = 469±50 ms</p>

**Figure 2. QTc distribution across study cohorts.**

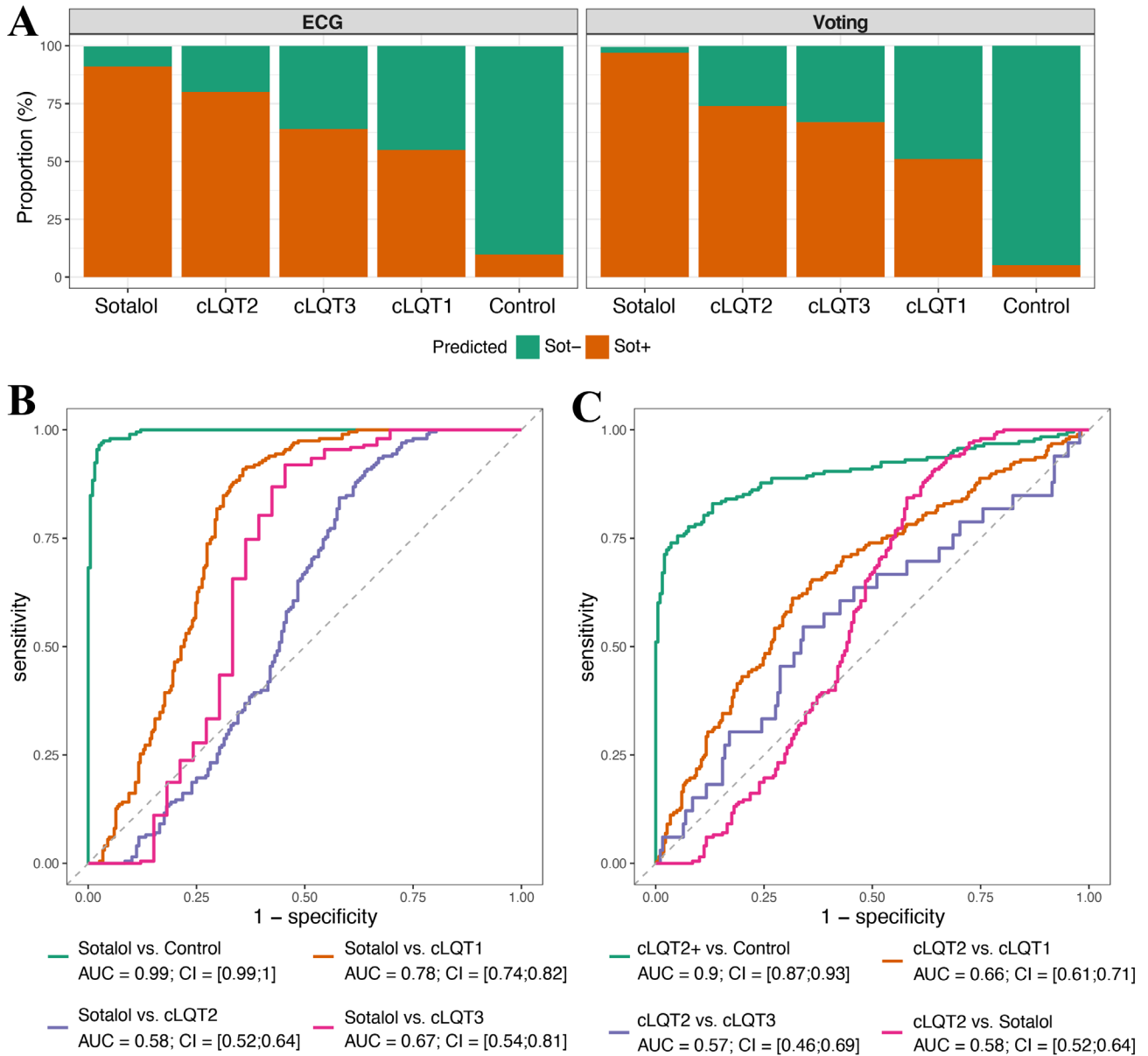




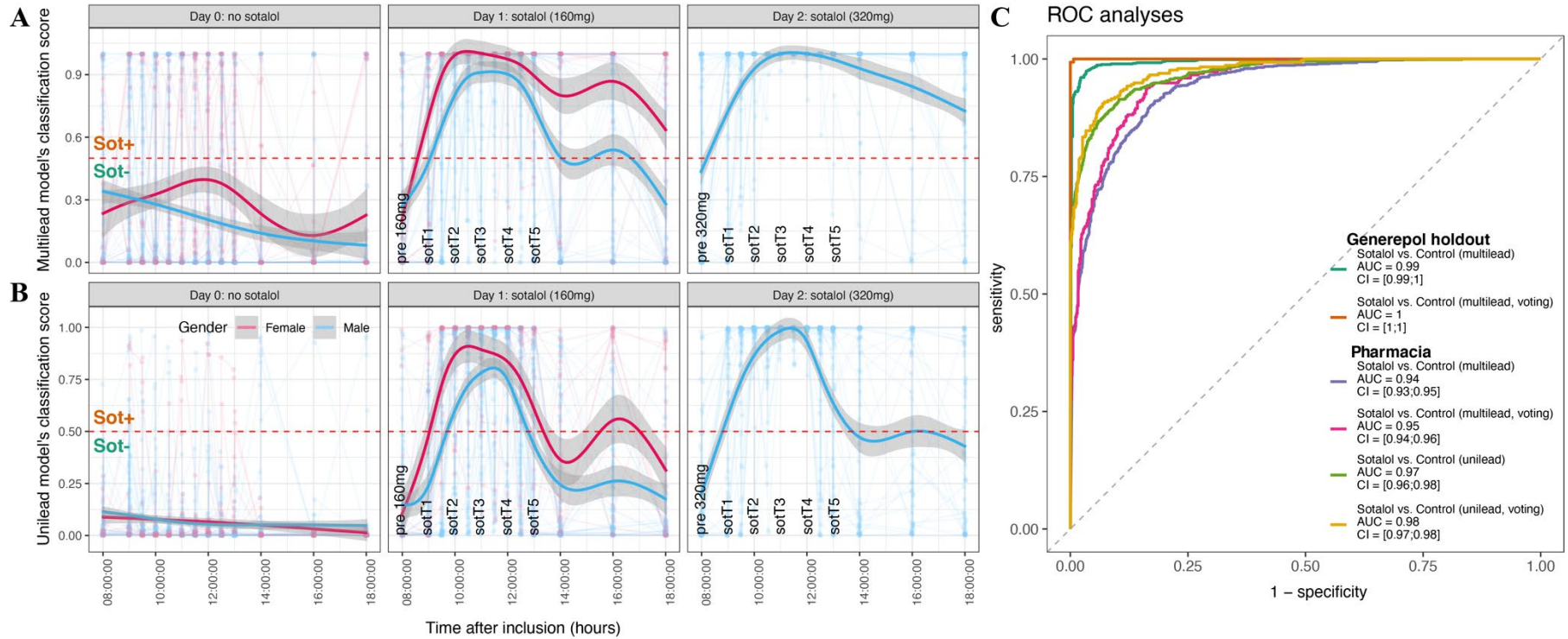
**Figure 3. Classification performance of CNN and linear regression (QT) models in discriminating baseline ECG before sotalol from those after sotalol intake (SotT1, SotT2, SotT3) in Generepol.**



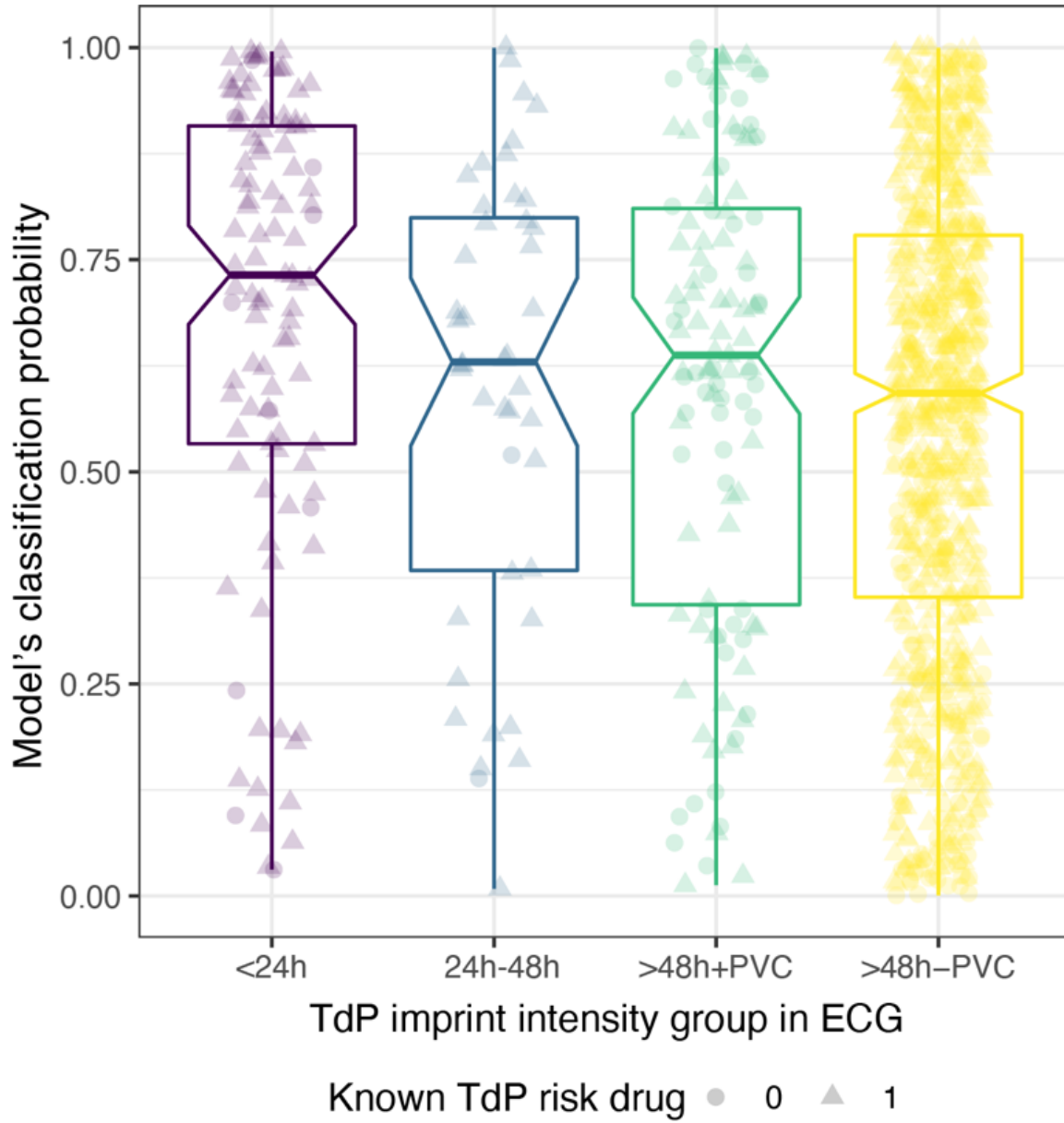
**Figure 4. CNN model performance in classifying study participants as Sot+/Sot- in Generepol holdout dataset and cLQTS.**



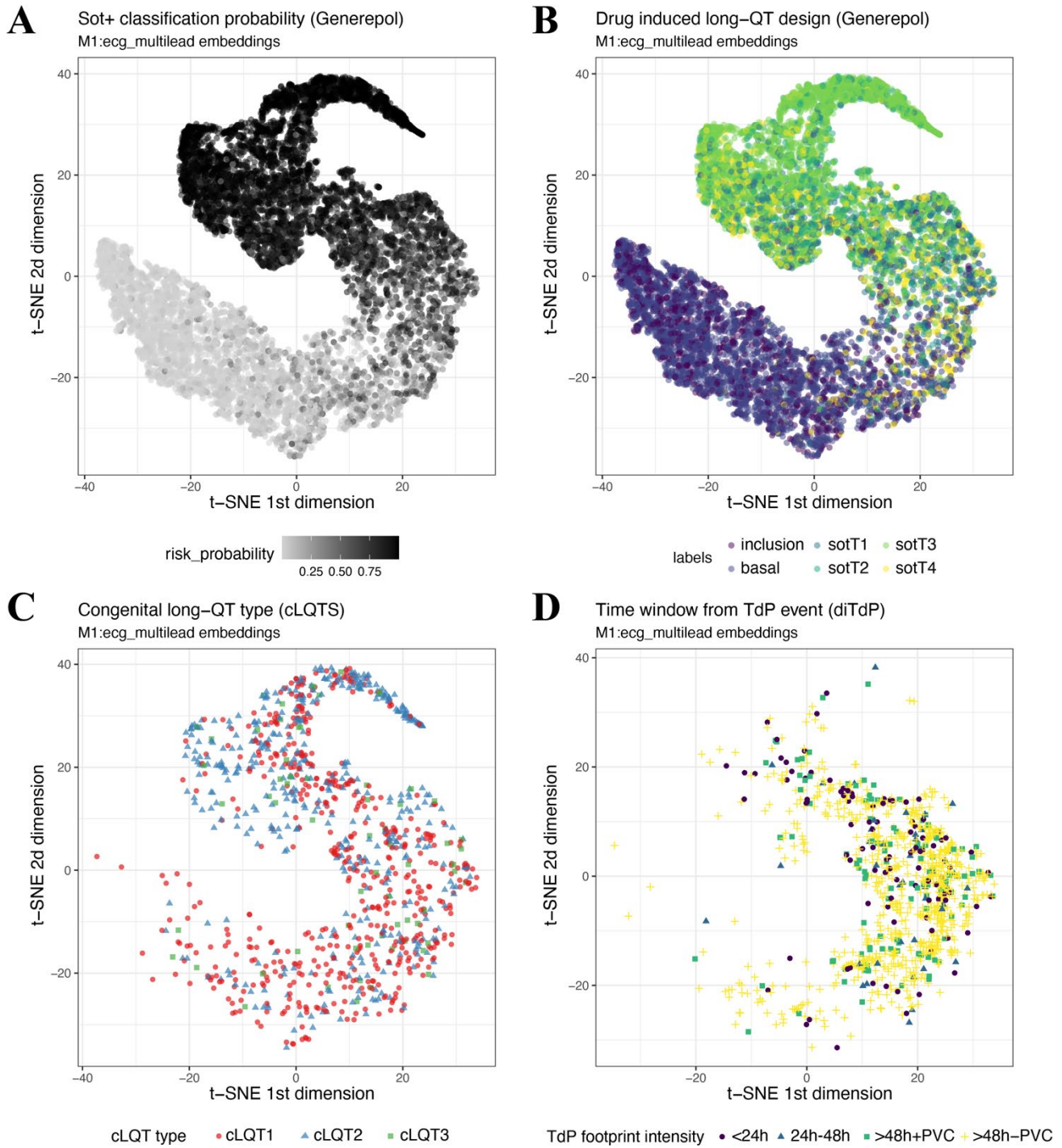
**Figure 5. CNN model performance in classifying study participants as Sot+/Sot- in Pharmacia's cohort.**



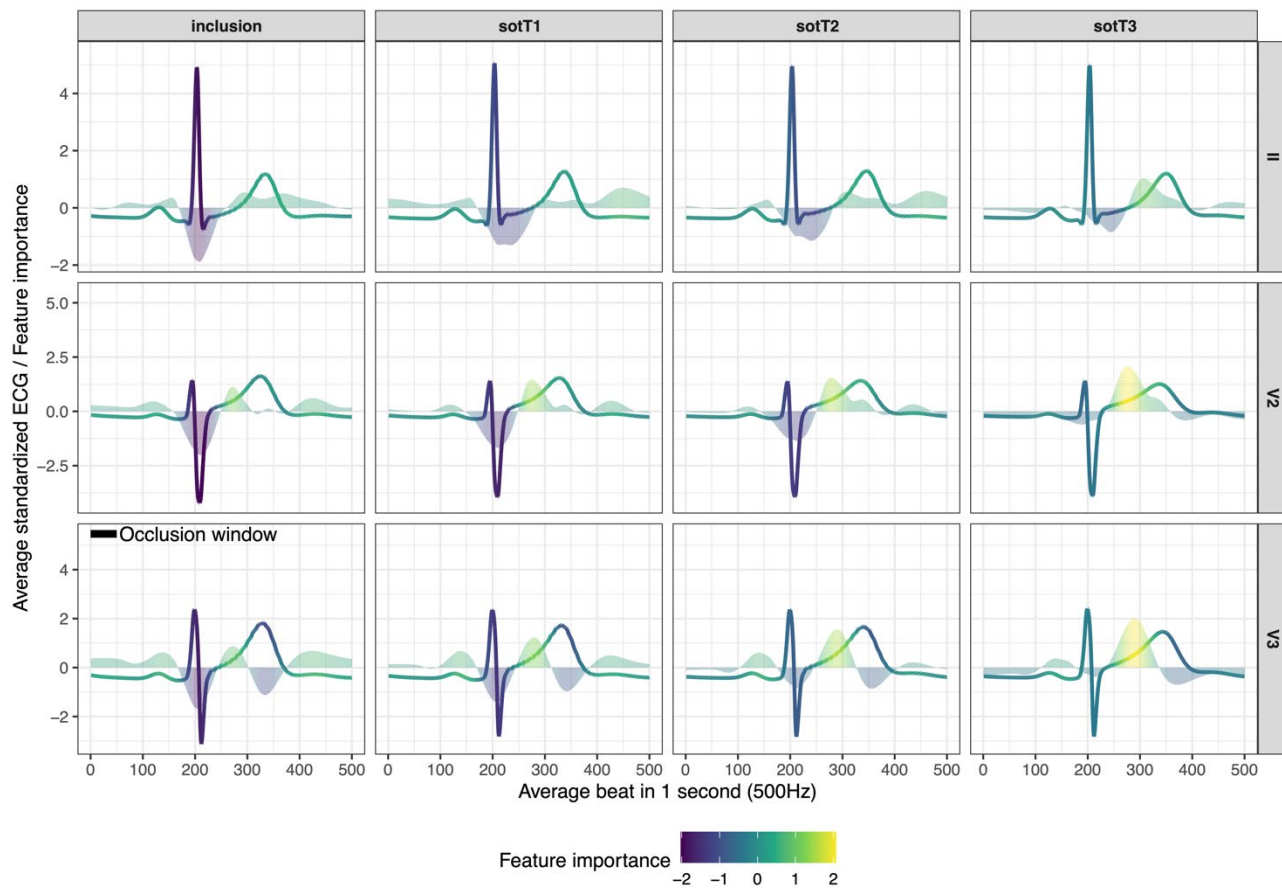
**Figure 6.** M1:ecg\_multilead model's score in relation to diTdP imprint intensity in ECG.



**Figure 7. Risk prediction model's deep embeddings reveal clinically relevant data structure.**



**Figure 8. Interpretability of the sotolol footprint on the ECG signal**



# Deep learning analysis of I<sub>Kr</sub>-blocker drug-induced ECG changes to inform arrhythmia risk and improve diagnosis of congenital long QT syndrome

Edi Prifti<sup>1,2,#</sup>, Ahmad Fall<sup>1</sup>, Giovanni Davogusto<sup>3,§</sup>, Alfredo Pulini<sup>1,4,§</sup>, Isabelle Denjoy<sup>5</sup>, Christian Funck-Brentano<sup>6</sup>, Yasmin Khan<sup>7</sup>, Alexandre Durand-Salmon<sup>7</sup>, Fabio Badilini<sup>8</sup>, Quinn S. Wells<sup>3</sup>, Antoine Leenhardt<sup>5</sup>, Jean-Daniel Zucker<sup>1,2</sup>, Dan Roden<sup>3,8,9</sup>, Fabrice Extramiana<sup>5</sup> and Joe-Elie Salem<sup>3,6,#</sup>

<sup>1</sup> IRD, Sorbonne University, UMMISCO, 32 Avenue Henri Varagnat, F-93143 Bondy, France;

<sup>2</sup> Sorbonne University, INSERM, NutriOmics, 91 Boulevard de l'Hopital, F-75013, Paris, France;

<sup>3</sup> Department of Medicine, Vanderbilt university medical center, Nashville, TN, USA;

<sup>4</sup> Université de Paris, Faculty of Medicine

<sup>5</sup> CNMR Maladies Cardiaques Héritaires Rares, Hôpital Bichat, Paris, France;

<sup>6</sup> Clinical Investigation Center Paris-Est, CIC-1901, INSERM, UNICO-GRECO cardio-oncology program, Department of Pharmacology, Pitié-Salpêtrière University Hospital, Sorbonne Université; 7513, Paris, France ;

<sup>7</sup> Banook Group, Nancy, France;

<sup>8</sup> AMPS LLC, NYC, USA

<sup>9</sup> Department of Pharmacology, Vanderbilt University Medical Center, Nashville, TN, USA

<sup>10</sup> Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, TN, USA;

§ These authors contributed equally to the work

# Corresponding authors

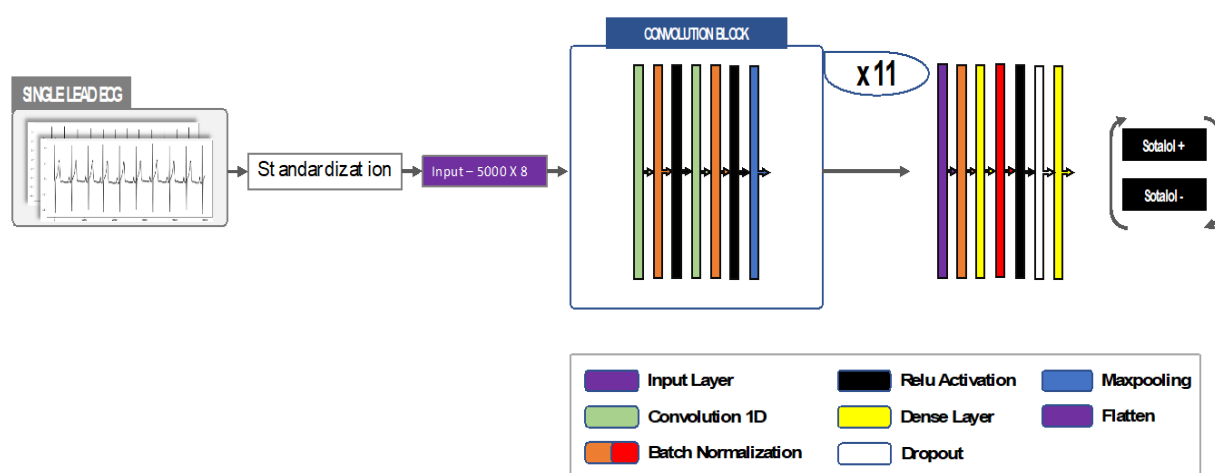
## Methods.

### Sotalol-intake classification with the multilead and unilead models

We used either the eight leads concomitantly (LI, LII, V1-6; i.e multilead hereafter) or each of these 8 leads independently (i.e unilead hereafter) to train a CNN model to predict *Sot+* and *Sot-* classes. The Generepol cohort was split in two sets: general training (80% for multilead models, 90% for unilead models) and holdout (20% for multilead models, 10% for unilead models). Ten times 10-fold cross-validation was performed in the general training set for parameter optimization. Each split was performed according to the subjects' IDs and therefore each training partition had distinct subjects from the testing split. The details concerning the multilead and unilead model's construction can be found below.

#### *The multi-lead model*

The model was composed of 11 blocks of convolution and each block contained two Conv1D (kernel=3) with the same number of filters and a maxpooling1D layer (pool size=2). The number of filters for each block were 8, 8, 8, 16, 32, 64, 128, 256, 512, 1024, and 2048, respectively (**Figure S1**). Zero padding was used for each Conv1D in order to have an output with the same length as the input (option "same"). The remaining model parameters were the default. After the convolutional blocks, the data were fed to a dense layer (512 nodes), a 'Relu' non-linear activation function, a dropout layer (70%) before final classification. Adam optimizer, binary cross-entropy loss, early stopping (patience=50) and a method to reduce the learning rate ('ReduceLRonPlateau' function) were used. The class weights were computed in the training set to balance the output classes before training.



**Figure S1 : CNN model architecture for Sotalol intake classification using multilead signal**

Model composed of 11 blocks of convolution, each containing two Conv1D (kernel=3) with the same number of filters and a maxpooling1D layer, followed by a ReLU activation, a dropout layer (70%) before final classification. The red Batch Normalization layer was used to extract the embeddings (see embedding analyses section in methods and Supp methods).

#### *The single-lead models*

We explored the capacity of any single ECG lead to provide information on the footprint of the sotalol intake. We designed eight CNN models based on the same architecture that were trained



on each single lead (LI-LII, V1-6), and tested them on the same lead they were trained as well as on other leads. The ECG were not denoised nor pre-processed (**Figure S2**)— they were only standardized. We used the Generepol cohort, split in two sub-datasets: global training (90%), which was used for any training, validation, and evaluation tasks; and holdout (10%). The holdout was not used for training or hyper-parameter tuning but was solely used for evaluating performance of the final trained model. We subdivided the “global training” in small datasets using the subjects’ IDs as follows: (i) sub-training (75%), used for training; (ii) sub-validation (10%), used during training validation to monitor performance and apply early stop if necessary; and (iii) sub-evaluation (15%), used for selecting the best model among them.

Traditional linear convolutional architectures such as VGG have many the layers, which are redundant thus making the network heavy and difficult to optimize. We designed a DenseNet-like model architecture for each lead to overcome those limitations. The DenseNet architecture connects all the layers densely together. Therefore, each layer receives inputs from all the preceding ones and passes its own output to all subsequent layers. The consequence is that the final output layer has direct information from every signal layer from the input.

The architecture of our model starts with an input layer followed by a convolutional layer, batch-normalization and Leaky ReLu activation. During this first step, higher features representations were extracted and transmitted to the next step. The second step consisted of 8 successive dense convolutional blocks (*DenseBlocks*). A dense convolutional block was composed of several convolutional sub-blocks (*DenseLayers*), each sub-block began with 4 layers of bottleneck composed of a batch normalization followed by a Leaky ReLu activation, a 1x1 convolution and a dropout layer. This bottleneck reduced the filter space dimensionality, decreasing the number of feature maps whilst retaining their salient features, and allowed to reduce the number of parameters while improving computational needs. After the bottleneck followed a batch normalization layer, a Leaky ReLu activation, a convolution layer and another dropout layer. These convolutional sub-blocks were densely connected, meaning that each sub-block took the output of all previous sub-blocks, in a feed-forward manner.

The dense connections are illustrated with the multiple arrows in **Figure S2**. The use of dense connections reduced the number of convolutional layers and filters, and therefore the number of parameters by reusing the knowledge of previous layers similarly to ResNet. Between each dense convolutional block, there was a transition block (TransitionBlock) with a bottleneck, followed by a Leaky ReLu activation and an average pooling 1D layer, which used the average of points reducing the dimensionality from the previous output. The transition block aimed to down sample the data flow through the network and interconnect the dense convolutional blocks. The third step of the network was a fully connected classifier, the final output of the dense convolutional blocks was flattened through a global average pooling 1D, and then fed to successive dense layers and Leaky ReLu activation. All the dropout layers had a common rate of 0.2. The final output activation was *Softmax*, which provided a posterior probability for each *Sot* class (*Sot-*, *Sot+*). We used the Adam optimizer and binary cross entropy as loss function.

Besides training the model’s weights, it is crucial to find the optimal hyperparameters for both the model as well as the training process in order to reach high performance in generalization. We identified the following variable key parameters:

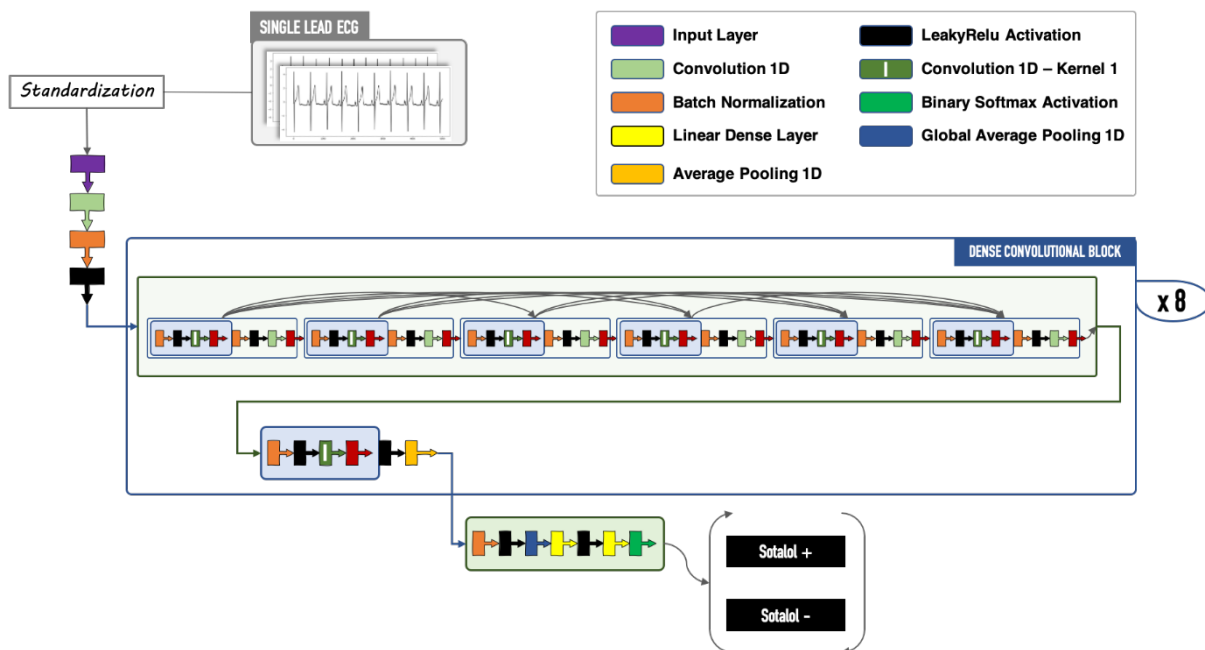
- Training process parameters:
  - Learning rate of optimizer function: 0.01 to 0.001

- Batch size: 32, 64, 128, 256
- Model specific parameters:
  - Bottleneck: TRUE or FALSE
  - Compression rate: 1.0, 0.3
  - Dense layers: 6, 7, 8
  - Dense blocks: 6, 7, 8
  - Dropout rate: 0, 0.2, 0.5
  - Main activation function: ReLu LeakyReLu

The optimal Learning Rate was found by progressively reducing the learning rate by a factor of 0.5 after 5 consecutive iteration without performance improvement. Beyond these variable hyper parameters, we used the following static hyper parameters for all training processes:

- Batch size: 64 (64 x num\_of\_available\_gpus)
- Early training patience: 30 (this parameter denotes after how many consecutive iterations without performance improvement to stop training)
- Epochs: 400 (maximum number of iterations, usually early stop around 50 with small dataset)
- Optimizer function: Adam

The process resulted in numerous models, for each lead we chose the best model by comparing their performance on the evaluation subset, at the end of this process we obtained 8 models, one for each ECG lead. All these models were trained one more time on the global training set with 400 iterations resulting in more robust models.



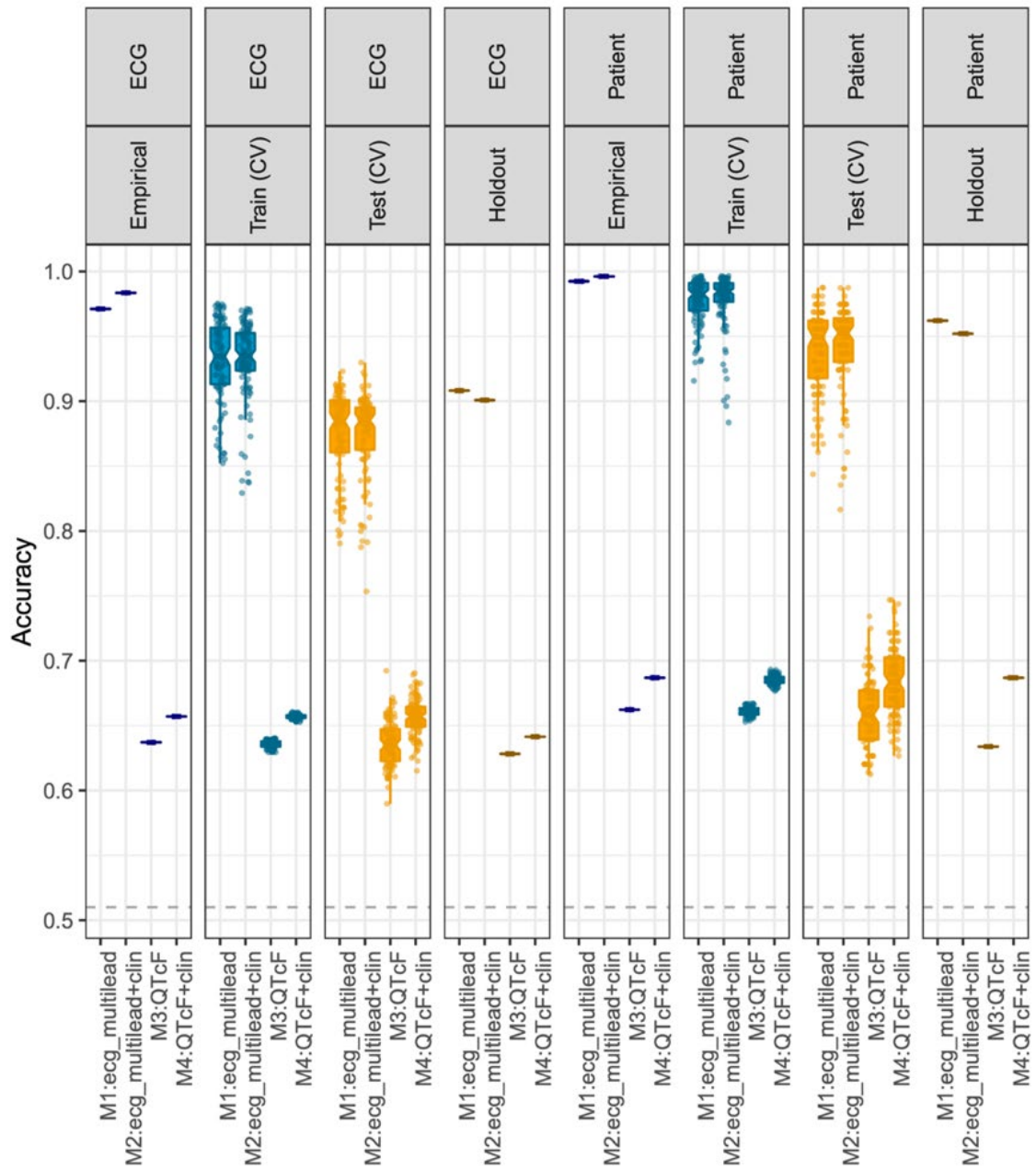
**Figure S2 : CNN model architecture for Sotalol intake classification using single lead signal**

After hyper-optimization process, we identified a common best hyperparameters combination for every ECG lead. Each model is made of 8 DenseBlocks each with 6 DenseLayers. The growth rate is set to 12 and the global convolution kernel size is set to 3. The initial convolution filters number is set to 24. The bottleneck is definitely used, and compression rate is set to 1.0 which means no compression is used. Each model has around 3millions trained variables.

## Results

### Sotalol intake is accurately detected from raw ECG signals

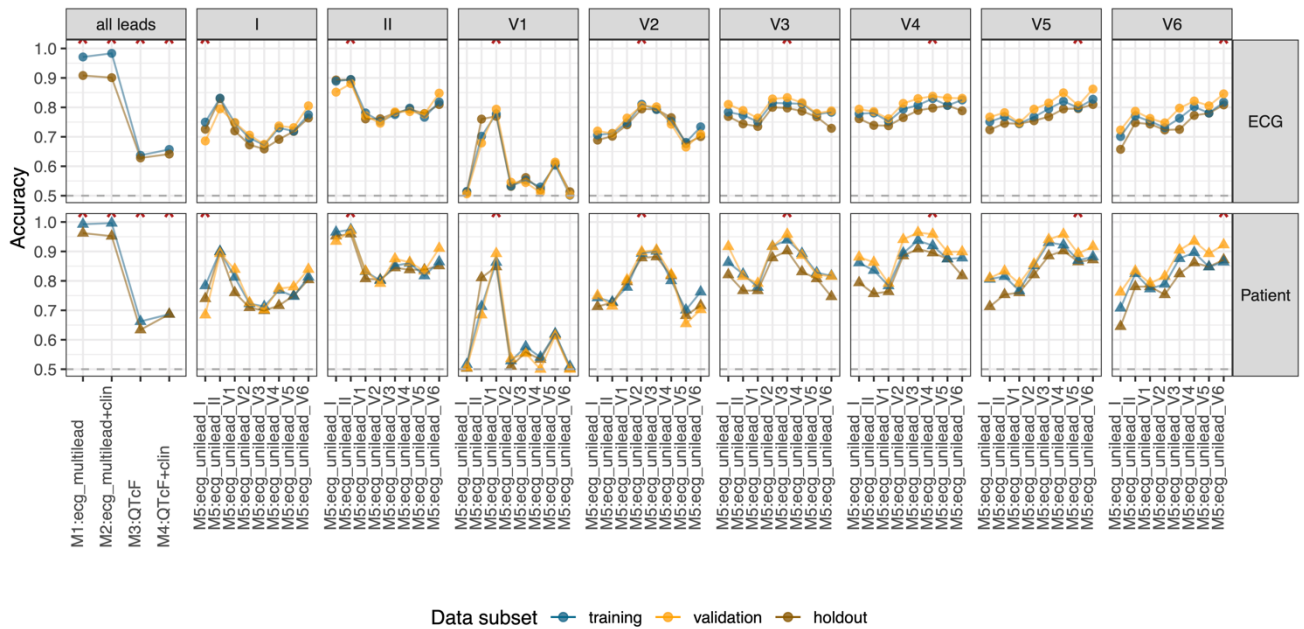
ECG data were recorded at baseline (*Sot-*) as well as at 1, 2, 3, and 4 hours after sotalol intake (*Sot+*; denoted sotT1, sotT2, sotT3 and sotT4, respectively). Triplicates were recorded at baseline and at sotT3. Sotalol concentration was also measured in participants' serum samples at two and three hours after intake (**Figure 3A**). To learn the sotalol footprint in the ECG signal, several convolutional deep neural networks (CNN) were built and tested. The first group of models was based on the multilead data, i.e., all data from the 10s ECG were used simultaneously as input to the model. A second set of models were trained with unilead data, i.e., using only data from a single lead (LI-II, V1-6). For the multilead model, we split the dataset in training and holdout sets. Furthermore, the training dataset was further divided into Train and Test following a 10-times 10-fold cross validation design. Detailed results for the accuracy performance score are displayed in **Figure S3**.



**Figure S3 : Accuracy of the different multilead models in classifying ECG of patients before and after sotalol intake in Generepol cohort**

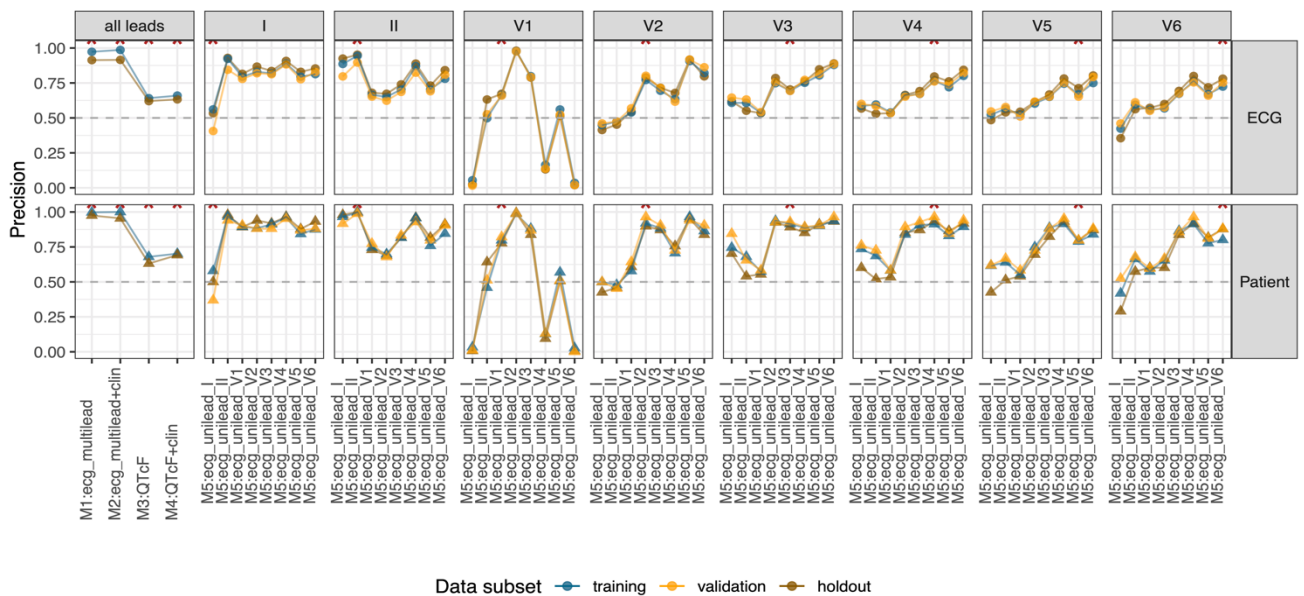
Accuracy for the multilead models (M1, M2), and QT-based logistic regression models (M3, M4) in classifying ECG of patients before and after sotalol intake by using each ECG recording (ECG) or using a voting strategy (using 10s triplicates) per patient (patient). Dark blue, blue, orange, and brown colors depict respectively the training, training Cross Validation (CV), test Cross Validation, and holdout subsets of the first study cohort (Generepol study).

Other performance metrics are described in **Figure S4** and **Table S1** (excel file).



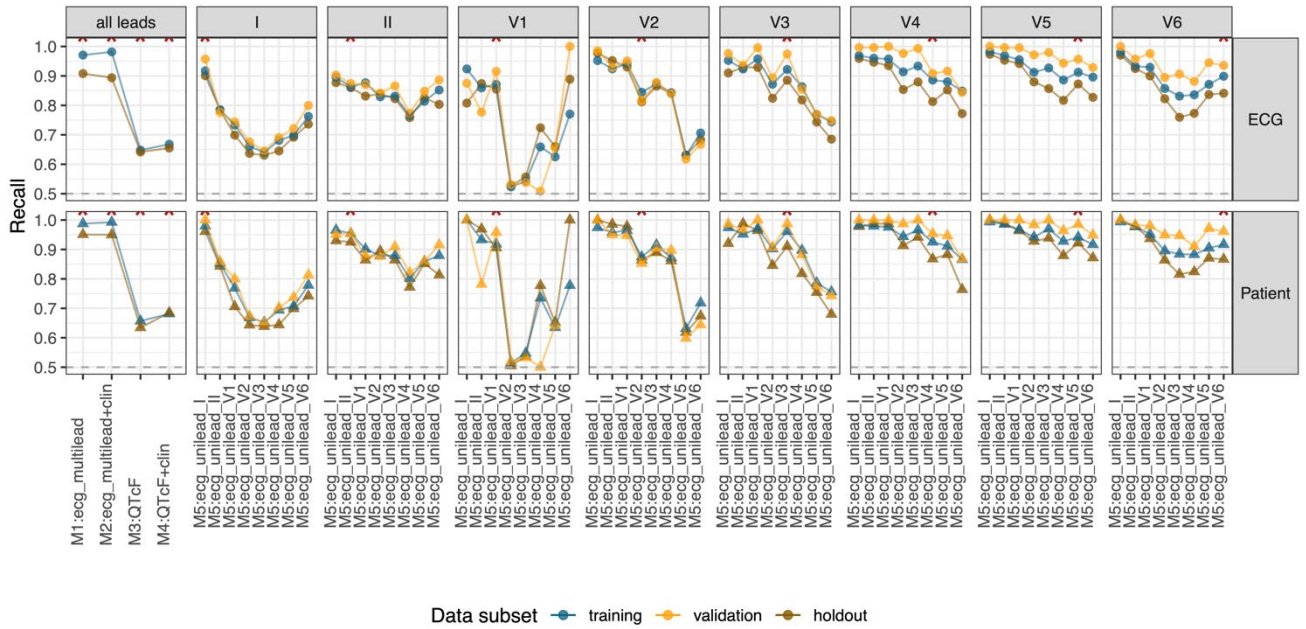
**Figure S4A : Accuracy of the different models in classifying ECG of patients before and after sotalol intake in Generepol cohort**

Accuracy for the multilead models (M1, M2), QT-based logistic regression models (M3, M4) as well as all unilead M5 models in classifying each ECG recording (top) or using a voting strategy (Patient) per patient (bottom). Blue, orange and brown colors depict respectively the training, validation, and holdout subsets of the first study cohort. Each model trained and tested in the same lead is annotated by red \*.



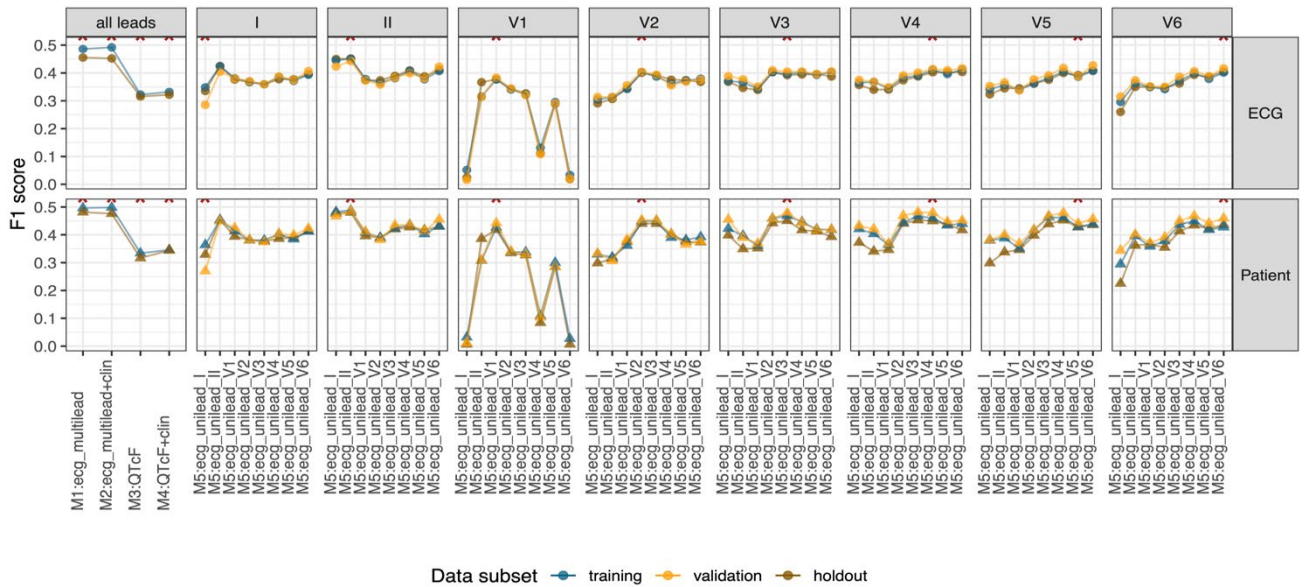
**Figure S4B : Precision of the different models in classifying ECG of patients before and after sotalol intake in Generepol cohort**

Precision for the multilead models (M1, M2), QT-based logistic regression models (M3, M4), as well as all unilead M5 models in classifying each ECG recording (top) or using a voting strategy per patient (Patient). Blue, orange and brown colors depict respectively the training, validation and holdout subsets of the first study cohort. Each model trained and tested in the same lead is annotated by red \*.



**Figure S4C : Recall of the different models in classifying ECG of patients before and after sotalol intake in Generepol cohort**

Recall for the multilead models (M1, M2), QT-based logistic regression models (M3, M4) as well as all unilead M5 models in classifying each ECG recording (top) or using a voting strategy per patient (Patient). Blue, orange and brown colors depict respectively the training, validation and holdout subsets of the first study cohort. Each model trained and tested in the same lead is annotated by red \*.

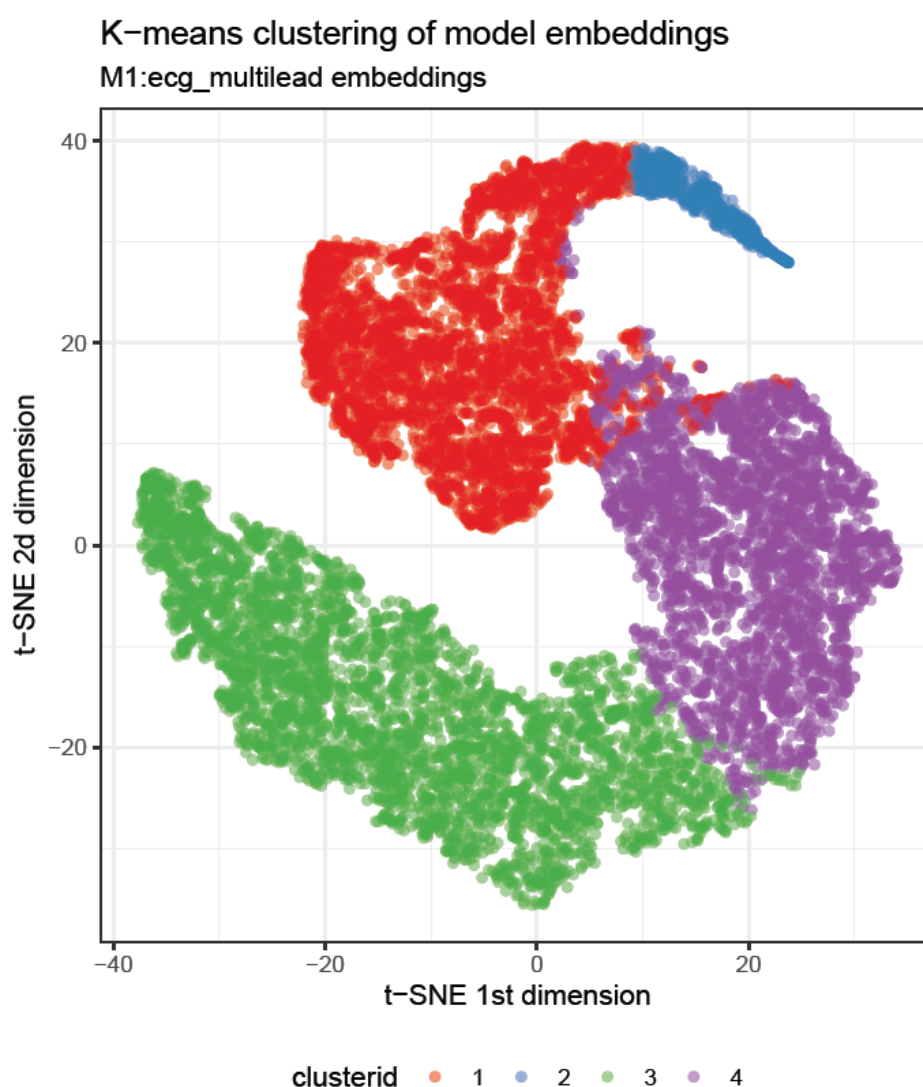


**Figure S4D : F1-score of the different models in classifying ECG of patients before and after sotalol intake in Generepol cohort**

F1-score for the multilead models (M1, M2), QT-based logistic regression models (M3, M4) as well as all unilead M5 models in classifying each ECG recording (top) or using a voting strategy per patient (Patient). Blue, orange and brown colors depict respectively the training, validation and holdout subsets of the first study cohort. Each model trained and tested in the same lead is annotated by red \*.

## Unsupervised learning of the total classification embeddings

The complex representations learned by the deep layers of the multilead CNN model were obtained as embeddings by accessing the output after the last group of convolutional, dense and batch normalization layers. After batch normalization, the output was fed as input to a ReLU activation function and a dropout layer before the final classification. This embedding layer, composed of 512 nodes, allows transforming the initial ECG (8 lead \* 5000 signal points) to 512 values. We applied this strategy to all ECG from the Generepol, cLQTS, and diTDP cohorts and analyzed them altogether. To allow for easier visualization we applied a dimension reduction technique (from 512 to 2 dimensions) based on the t-SNE algorithm (perplexity=100, iteration=1000, using the Rtsne implementation; **Figure 7**).



**Figure S5 : Unsupervised k-means clustering of the embeddings (k=4).**

Map illustrating 2 dimensions after multiple dimension reduction of 512 dimensions of the classification model's embedding, using t-SNE algorithm. Each point is an ECG from all three study cohorts. ECG are colored with the cluster id resulting from the k-means unsupervised learning.

Besides visualization analyses described in the main text where clinically relevant data distribution was observed, we applied an unsupervised learning approach to identify structure in this novel representation space. We used k-means with default parameters and  $k=4$  to cluster the ECG using all 512 dimensions of the embeddings and obtained four clusters (1 to 4) with 28, 4, 40, and 28% of the ECGs, respectively. The obtained partition is visualized on the same t-SNE map as for the previous analyses in the main text and concurs with the two-dimensional reduction of the t-SNE map (**Figure 7; Figure S5**). We then analyzed the distribution of the different study groups and conditions within these novel clusters. 90% of all the baseline ECG from the Generepol cohort are grouped in cluster #3 while 90% of the ECG after Sotalol intake are grouped in the clusters #1 and #4. The relation between the clustering and the simplified drug intake protocol was significant ( $p < e-16$ , Chi2-test).

Similarly, we tested the relation between the clusters obtained with unsupervised learning and the number of ECG from the different types of congenital long-QT patients. 43% of the cLQT-1 ECG were grouped in cluster #4 and 29% in cluster #3. 44% of the cLQT-2 ECG were found in cluster #1 and 27% in cluster #4. A similar distribution was also observed for ECG from the cLQT-3 participants, with 34% in cluster #1 and 43% in cluster #4. The relation between the clustering and the cLQTS types was highly significant ( $p < 8e-24$ , Chi2-test).

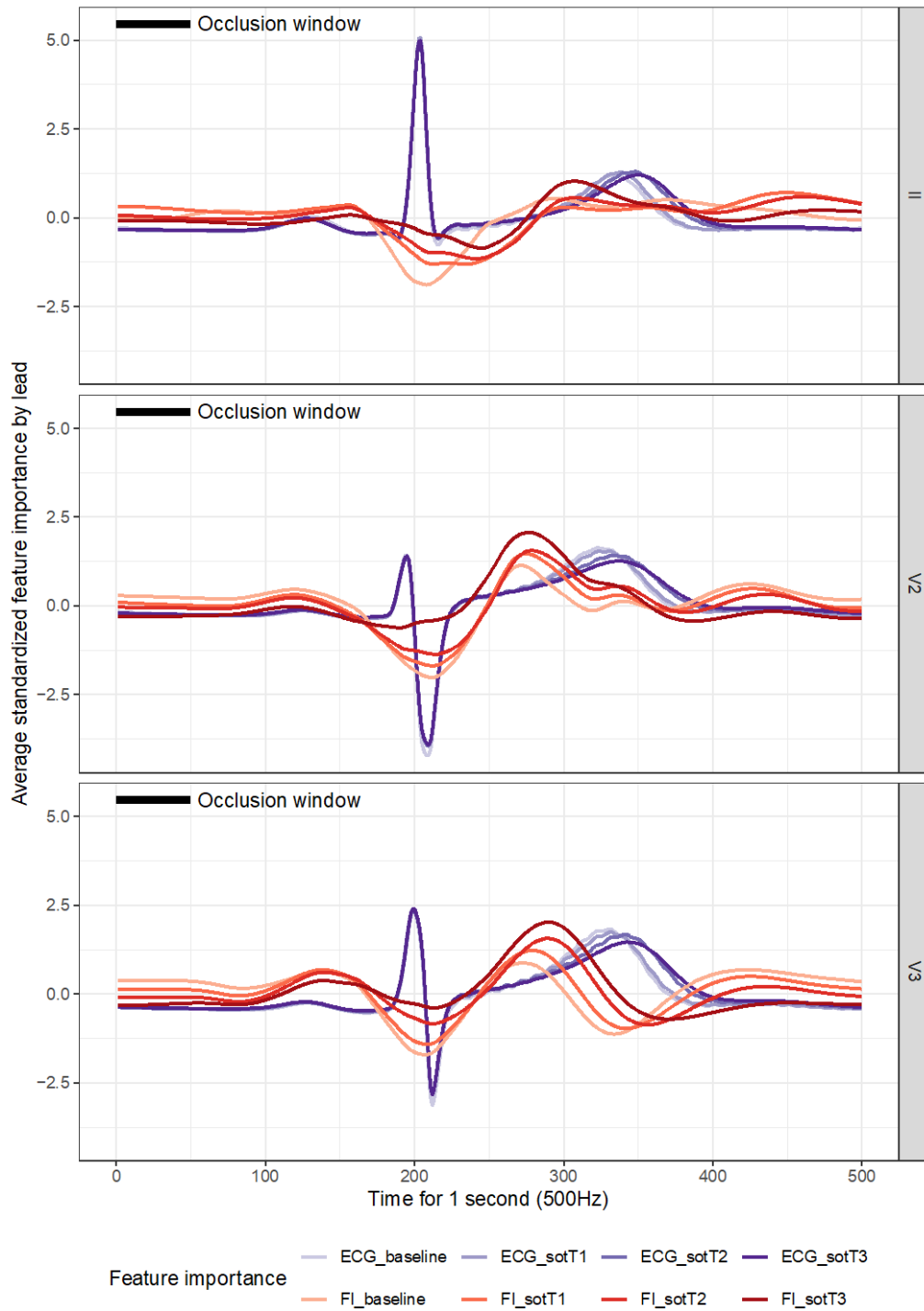
We then tested the relation between the clusters and the number of ECG in the different time-frame from TdP events from the diTdP study cohort. 63% of the ECG within 24h from the TdP event were assigned to cluster #4 and 25% in cluster #1. ECG within 48h of the TdP event were assigned predominantly to cluster #4 (77%). Similar trends were seen for the other periods from TdP. The relation between clustering and the TdP time-window types was significant ( $p < 0.0017$ , Chi2-test).



## Interpretability of the CNN models

We attempted to identify which features from the ECGs were the most useful for the classification as *Sot+* and *Sot-*. There are different approaches when considering interpretability of neural networks. Among the different algorithms for interpretability, we used the occlusion method, which permits to explore the contribution of each part of the signal to the final prediction. This method consists in iteratively occluding a given chunk of data of a predefined size and to make a prediction. Eventually, for each hidden feature we could quantify its importance in the resulting classification result. In our analyses, we used a window of 50 points (100ms) that was iteratively sliding across the signals.

We first tested the interpretability on the multi-lead model. However, although this model was demonstrated to be highly accurate, it was difficult to interpret. Indeed, the signal from different ECG leads were mixed together in increasingly complex abstractions throughout the neural network. The signal of each lead at each instant was considered at the same time during the occlusion process, leading the interpretability to a fusion of data coming both from space (leads) and time (10s). To achieve a more human comprehensible interpretation, we used the single-lead models. The single-lead models allowed us to better identify patterns in ECG that drove interpretation by the model as shown in **Figure 8** of the main text. We used segmented ECG signals by beats in order to summarize the importance of each point of the signal in the model's classification decision using the same occlusion approach. **Figure S6** displays standardized ECG signals averaged at the beat level for each time point of the Generepol cohort for leads LII, V2, and V3, (blue colors). Overlaid behind the ECG profile is displayed the standardized feature importance profile (i.e. FIP) quantified with the occlusion approach for each time point (red colors). There is lengthening of the QT duration and decrease in maximal T wave-amplitude (blue colors in **Figure S6**) after sotalol intake, as well as the change in the FIP footprint (red colors in **Figure S6**), with incremental importance of J-T<sub>peak</sub> segment in discrimination of ECG alteration induced by sotalol.



**Figure S6 : Interpretability of the single lead CNN classification model**

This figure displays an averaged signal of the standardized ECG for each segmented beat for leads LII, V2 and V3. All signals from the same time points were grouped together (blue colors). Similarly, the standardized feature importance profile (i.e. FIP) is summarized on top the ECG profiles (red colors).