Deep learning analysis of drug-induced ECG changes to inform arrhythmia risk and improve diagnosis of congenital long QT syndrome

Edi Prifti (✉ edi.prifti@gmail.com)

French National Research Institute for Sustainable Development <https://orcid.org/0000-0001-8861-1305>

Ahmad Fall

French National Research Institute for Sustainable Development

Giovanni Davogustto

Department of Medicine, Vanderbilt University Medical Center

Alfredo Pulini

Université de Paris, Faculty of Medicine

Isabelle Denjoy

CNMR Maladies Cardiaques Héritaires Rares, Hôpital Bichat

Christian Funck-Brentano

CIC-1421, Pitié-Salpêtrière Hospital, AP-HP <https://orcid.org/0000-0003-0406-5189>

Yasmin Khan

Banook Group

Alexandre Durand-Salmon

Banook Group

Quinn Wells

Department of Medicine, Vanderbilt University Medical Center

Antoine Leenhardt

CNMR Maladies Cardiaques Héritaires Rares, Hôpital Bichat

Jean-Daniel Zucker

Unité Mixte Internationale de Modélisation Mathématique et Informatiques des Systèmes Complexes

Dan Roden

Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, TN

Fabrice Extramiana

CNMR Maladies Cardiaques Héritaires Rares, Hôpital Bichat

Joe-Elie SALEM

APHP

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

Posted Date: February 22nd, 2021

DOI: <https://doi.org/10.21203/rs.3.rs-256040/v1>

License: This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Deep learning analysis of drug-induced ECG changes to inform arrhythmia risk and improve diagnosis of congenital long QT syndrome

Edi Prifti^{1,2,*}, Ahmad Fall¹, Giovanni Davogustto^{3,§}, Alfredo Pulini^{1,4,§}, Isabelle Denjoy⁵, Christian Funck-Brentano⁶, Yasmin Khan⁷, Alexandre Durand-Salmon⁷, Quinn S. Wells³, Antoine Leenhardt⁵, Jean-Daniel Zucker^{1,2}, Dan Roden^{3,8,9}, Fabrice Extramiana⁵ and Joe-Elie Salem^{3,6,*}

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

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

³ Department of Medicine, Vanderbilt University Medical Center, Nashville, TN, USA;

⁴ Université de Paris, Faculty of Medicine

⁵ CNMR Maladies Cardiaques Héritaires Rares, Hôpital Bichat, Paris, France;

⁶ 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 ;

⁷ Banook Group, Nancy, France;

⁸ Department of Pharmacology, Vanderbilt University Medical Center, Nashville, TN, USA

⁹ 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[at]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: 10207 words

Abstract: 113 words

Figures: 6 (and 11 supplemental)

Tables: 0 (and 1 supplemental)

NCT: NCT00773201

Abstract

Congenital or drug-induced long-QT syndromes can cause Torsade-de-Pointes (TdP), a life-threatening ventricular arrhythmia. The current strategy to identify individuals at high risk of TdP consists on measuring the QT duration on the electrocardiogram (ECG), shown to provide limited information. We propose an original method, including training deep neural networks to recognize ECG alterations induced by QT-prolonging drugs, as a comprehensive evaluation of TdP risk. These models accurately detected patients taking QT prolonging drugs during ECGs, while discriminating for the presence and type of congenital long-QT. Moreover, they enhanced prediction of drug-induced TdP events in addition to QT measurement. Analyses of these models revealed footprints of the torsadogenic risk and clinically relevant patient stratification.

Keywords

Torsades de Pointes, machine learning, risk prediction, interpretability, long-QT

Introduction

Torsades de Pointes (TdP) is a ventricular arrhythmia which can degenerate into ventricular tachycardia or ventricular fibrillation, and lead to death.¹ TdP is usually seen in patients with one of the congenital long-QT syndromes (cLQT)² or in association with drug therapy (drug-induced long-QT syndrome, diLQT).

There are three main forms of cLQT: type-1 (cLQT1), caused by loss of function mutations in *KCNQ1*, responsible for the cardiac potassium current I_{Ks} ; type-2 (cLQT2), caused by loss of function mutations in *KCNH2*, responsible for the cardiac potassium current I_{Kr} , and type-3 (cLQT3), caused by mutations in *SCN5A*, encoding the cardiac sodium channel and specifically increasing the non-inactivating (“late”) current I_{Na-L} . On ECG, QTc is prolonged in all three subtypes, but specific ECG waveform patterns including T-wave morphology abnormalities, can be specific for each subtype.³

These conditions have in common abnormalities of the QT interval, an electrocardiographic parameter which represents ventricular repolarization and is measured as the time lapsed between the beginning of the QRS complex and the end of the T-wave of the electrocardiogram (ECG).⁴ Hence, the current risk stratification strategies for TdP are based on the quantification of the QT interval corrected for heart rate (QTc)⁵. Beyond QT duration, the morphology of the QT can also have a characteristic appearance in the three main forms of cLQT.⁶

The incidence of TdP in cLQT is approximately 1-2% per year⁷. In diLQT, the incidence of TdP is variable, according to the drug, often with a dose-effect relationship⁸. However, the consequences of TdP can be catastrophic, leading to sudden death in young, otherwise healthy individuals in cLQT, and can be the reason for withdrawing drugs from the market in diLQT.

However, more than 100 cardiac or non-cardiac drugs (common antimicrobials such as hydroxychloroquine and azithromycin, antipsychotics, antiemetics, anticancer drugs) are approved despite a conditional, possible, or even known TdP risk because they are considered to have a favorable risk-benefit ratio.^{9,10} Most drugs responsible for diLQT and possibly TdP, can be identified by assessing whether they block the ventricular delayed rectifier potassium current (I_{Kr}) translating to prolonged QTc interval, decreased T-wave maximal amplitude and notched T-wave appearance on the ECG.¹¹ Regulatory agencies require new drugs to undergo thorough QT studies, in which only the magnitude of drug-induced QTc prolongation is evaluated as a surrogate for TdP risk.¹² However, it is well-established that limiting ECG evaluation only to the QTc interval duration is poorly predictive of TdP.¹³ QT measurement captures a limited part of the information contained in the ECG signal.¹⁴ An unbiased and

complex examination of the ECG information beyond QT could provide relevant insight into identifying at-risk drugs and patients, and predict future TdP events. Such attempts have not translated to clinical practice for now.^{11,15,16} Furthermore, most physicians prescribing these drugs are unable to correctly quantify QTc or evaluate and manage TdP, and do not have immediate access to QT-expert consultation.¹⁷

Artificial intelligence (AI) is rapidly reaching specific medical expertise.¹⁸ Deep learning (DL) and particularly convolutional neural networks (CNN) have brought a radical change in the field of pattern recognition and machine learning (ML) itself, improving earlier models in learning tasks such as image classification and natural language processing^{19,20}. CNN architectures were applied to detect various types of arrhythmia²¹ including shockable ventricular arrhythmia,²² myocardial infarction^{21,23-26} and coronary artery disease²⁷. However, the use of DL in the context of diagnosis and risk stratification of congenital and drug-induced cardiac arrhythmias need further exploration.

The automatic evaluation of the TdP risk in the cLQT or diLQT setting, based on powerful AI techniques, could prove to be very useful for clinicians and beneficial to patients. TdP events are rare, so we developed and validated an original approach. We trained CNN to recognise ECG alterations induced by sotalol, an anti-arrhythmic betablocker inhibiting I_{Kr} , and known to confer a high risk of TdP, as a proxy of a generalizable torsadogenic risk evaluation. The CNN models could accurately detect sotalol intake on ECG and on independent cohorts, discriminate for the presence and type of cLQT and enhance prediction of drug-induced TdP events in addition to QTc evaluation. Further analyses of these models revealed interpretable footprints of the torsadogenic risk and clinically relevant patient stratification.

Results

Study population characteristics

We used three different cohorts to test our hypothesis that training CNN to recognise specific ECG alterations is accurate and provide additional insight compared to QTc measurement for TdP risk stratification. The first cohort was derived from the “Generepol” study¹¹ and contained 10,292 ECG recordings from 990 healthy subjects (62% women) in normal sinus rhythm before and 1, 2, 3, and 4 hours after the administration of 80mg sotalol (respectively denoted as baseline and sotT1-sotT4; **Figure 1A**). The median number of ECG per participant in the “Generepol” cohort was 15 (IQR=1).

The second cohort included patients evaluated from congenital-LQT patients confirmed by genetic testing and followed over a 26-year period at the Arrhythmia unit of Bichat Hospital (Paris, France). The Bichat cohort included 487 participants with 1,083 ECG recordings (median number of ECG per participant=3, IQR=6). The three most common cLQT forms were represented, with 266 cLQT1 (62% women), 188 cLQT2 (54% women), and 33 cLQT3 (45% women) patients. In the Bichat cohort, 213 participants (44%) had at least one recording performed while on betablocker medications, with 116, 88 and 9 (44%, 47% and 27%) participants for cLQT1, cLQT2, and cLQT3, respectively. ECGs were in normal sinus rhythm, except for 8 (0.7%) with supra-ventricular arrhythmia and 2 (0.2%) with either atrial and/or ventricular pacing.

The third cohort included 48 participants (60% women) with 1,105 ECG recordings (median number of ECG per participant=31, IQR=21) and who experienced at least one drug-induced TdP (diTdP) event without underlying cLQT followed at Vanderbilt University Medical Center (Nashville, TN). Indeed, 66% of the ECG (n=733/1105) were recorded while patients were on drugs with known risk for TdP⁹, with amiodarone (29/48), sotalol (12/48), dofetilide (9/48), fluconazole (7/48) and hydroxychloroquine (4/48), being the most prevalent (>95%). Recordings from these patients were classified into four categories using combination of the time window from ECG to the diTdP event, associated with the presence of premature ventricular contractions (PVC) : <24h, 24-48h, >48h+PVC and >48h-PVC. There were 103, 44, 115 and 843 ECG per group, respectively. 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%) ECG had at least one ventricular and/or an atrial paced complex. At least one PVC was seen in 143 (13%) ECG.

QTc measurements

Serial QTc monitoring, is the main method cardiologists use to evaluate TdP risk in clinical practice.¹³ When ECG exams display values of QTc>480ms or an increase ≥ 60 ms after drug intake compared to the baseline, the patients are considered at significant clinical risk to develop TdP.¹³

In the Generepol healthy volunteer study (cohort 1), the average QTc values at baseline were 14ms higher in women compared to men (391 ± 15 ms vs. 377 ± 16 ms; $p < 2e-16$). The QTc increased by 30 ± 14 ms after sotalol intake compared to baseline. The maximal prolongation in QTc with sotalol intake was pronounced in females (34 ± 14 ms vs 23 ± 12 ms; $p < 2e-16$, **Figure 1C**).

In the Bichat cLQT participants (cohort 2), QTc were measured on the first ECG for each patient, (n=483) as previously described.²⁸ No difference in QTc was detected between the three groups cLQT1 (449 ± 36 ms, 453 ± 40 ms, and 452 ± 35 ms, for cLQT1, cLQT2, and cLQT3 respectively; **Figure 1D**). However, the QTc values of this cohort were 65ms higher when compared with those from the healthy volunteers from the first cohort before sotalol intake (451 ± 38 vs. 386 ± 18 ms, $p < 2e-16$).

In the Vanderbilt participants with diTdP (cohort 3), the QTc values were higher in the immediate 24-hour window from the TdP event (501 ± 70 ms) compared with ECG recorded within the 24-48h (478 ± 45 ms; $p < 0.02$), after 48h with PVC (455 ± 50 ms; $p < 2.14e-8$), and after 48h without PVC (459 ± 45 ms; $p < 8.5e-12$; **Figure 1E**). We added patient ID as a random effect and intake of drugs with known risk of TdP as a fixed effect. The latter was associated with a 7.4ms increase in QTc in average ($p = 0.07$). Similarly, QTc values of this cohort were in average 86ms greater than healthy volunteers before sotalol intake (469 ± 63 ms vs. 386 ± 18 ms; $p < 2e-16$).

Figure 1: Experimental design and QTc distribution across study cohorts.

A: Description of the three study cohorts. The first cohort was composed of healthy volunteers taking 80mg of oral sotalol. This dataset was used to train the models. The second cohort was composed of cLQT patients of type 1, 2 and 3. The third cohort included patients who experienced events of drug-induced TdP with no underlying identified cLQT. **B:** Illustration of ECG use (multilead and unilead) in training and testing the deep CNN models up to the prediction of a risk score and evaluation by the physician. **C:** The distribution of QTc values following the sotalol drug-induced QTc prolongation in the Generepol study¹¹. The X-axis corresponds to the time in minutes since inclusion and data is colour coded by gender. Lines link ECG recordings from the same participant over the duration of the protocol. **D:** Boxplot distribution of the estimated QTc values in the cLQT cohort by subtypes. ECG are grouped and coloured by gender. **E:** Boxplot distribution of QTc values across the di-TdP study cohort grouped by time to TdP event and presence or not of premature ventricular contractions (PVC). ECG are grouped and coloured by gender. The black horizontal lines indicate the 480ms at-risk QTc threshold for the panels C-E.

Sotalol intake is accurately detected from raw ECG signals

To learn the sotalol footprint as a proxy of drug-induced I_{Kr} blockade in the ECG signal, we trained two different CNN models on a subset of cohort 1. The first model used all the leads (LI-II, V1-V6) from raw ECG data (M1:ecg_multilead). The second model used clinical information (age, sex and potassium levels at baseline) in addition to the ECG data (M2:ecg_multilead+clin). Here, we focused on 10s triplicate recordings at baseline, as well as at one, two, and three hours after sotalol intake. This time frame was selected since the absorption peak (C_{max}) of oral sotalol was observed at three hours (**Figure 2A**).

The models provided an output score that was considered as a likelihood of sotalol intake varying between 0 and 1. A score of 0 corresponds to the absence of sotalol whereas 1 its presence (*Sot-*, *Sot+*, respectively). The predicted score at baseline was on average very low (0.06) but increased rapidly for ECGs recorded one hour (*SotT1*, 0.80), two hours (*SotT2*, 0.88) and peaked at three hours after sotalol intake (*SotT3*, 0.95) (**Figure 2B**). Noteworthy, the increase in the predicted score corresponded with the increase in sotalol concentration measured in participants' blood samples two and three hours after intake (**Figure 2A**). No difference was observed in *Sot+* classification score between women and men, indicating the possibility that the CNN integrated potential sex differences.

The output scores were then converted into a binary variable *Sot-* or *Sot+* based on a threshold of ≥ 0.5 (*Sot+*; **Figure 2B**). Performance indicators (ROC AUC, accuracy, precision, recall (all ranging in [0,1]) and F1 score (ranging in [0,0.5])) were evaluated for each model in each ECG recording individually or averaged for multiple ECG of the same participant (i.e. voting strategy on triplicates) at any of the different time points, in the training, cross-validation and holdout sets (**Figures 2C; Figure S2-S5**). For comparison with current practice electrocardiographic markers, we also tested the performance of QTc measurements for the same recordings (M3:QTcF) and with the same additional clinical information (M4:QTcF+clin) to discriminate the presence of sotalol intake on ECG.

The average cross-validation test ROC AUC of the model M1:ecg_multilead for discriminating ECG of patients before or after sotalol intake was 0.948 when computed for each ECG and 0.98 with the voting approach (see methods). Similarly, for the model M2:ecg_multilead+clin, the average test accuracy was 0.948 (ECG) and 0.98 with voting (**Figure 2C, Figure S1**). No statistical difference was, however, observed between M1 and M2. This indicated that the information contained in age, sex, and potassium level was likely embedded in the ECG

footprint captured by the CNN model. Therefore, the model trained solely with raw ECG data was deemed sufficient to use in the rest of the study.

The linear regression model based on the QTc alone (M3:QTcF), displayed a lower ROC AUC 0.695 (ECG) and 0.720 (voting) than the CNN model (M1:ecg_multilead; ROC AUC: 0.948 and 0.98; respectively, $p < 1.5 \times 10^{-141}$). After integrating additional clinical information in the QTc model (M4:QTcF+clin), its performance increased significantly ($p < 3.3 \times 10^{-16}$) to 0.717 (ECG) and 0.750 (voting) compared with M3:QTcF (**Figure 2C; Figure S1**). QTc based models, were less performant than CNN models, even after integration of relevant clinical covariates. This observation reinforces our hypothesis that the sotalol footprint is more complex than just a QT prolongation. All four models introduced above, displayed significantly higher ROC AUC scores in the test subset (average CV values) with the voting (triplicates of 10s ECG) predictions as compared with the single 10s ECG ($p < 1.2 \times 10^{-20}$ for M1:ecg_multilead). This demonstrates the importance of multiple short signal acquisitions or longer recordings (>10s). The precision of the M1 model (voting) was high, 0.955 [0,1] as was the recall 0.927 [0,1] and F1-score 0.470 [0,0.5]. Results were similar in the holdout set (**Figure 2C; Figure S1**). Other performance indicators are detailed for all the models in **Figure S2-S5**).

Finally, we tested the hypothesis that the sotalol intake footprint could be even picked up by single channels alone during ECG acquisition. For this, we trained eight different models – one for each channel (LI, LII, V1-V5, V6; see methods). Their performance were comparable with the multilead model (**Figure 2C**). 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). Surprisingly, when this model trained on one lead was tested on the rest of the leads, it performed well, with average holdout AUC of 0.883 (10s ECG) and 0.96 (voting). However, while the average 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 (see **Table S1** for detailed data).

Figure 2: Classification performance of CNN and linear regression (QT) models in discriminating baseline ECG before sotalol intake from those after sotalol intake (SotT1, SotT2, SotT3).

A) Boxplots, illustrating the distribution of circulating sotalol concentration in (ng/ml) in the Generepol cohort two and three hours after 80mg oral sotalol intake. Data is displayed separated and coloured by gender. **B)** Scatterplot illustrating the evolution of the classification score for the *Sot+* class (y-axis) across time from inclusion (x-axis). All points (averaged ECGs) of a study participant are linked together as trajectories and are coloured by gender. Summarized loess distribution of the data $\pm$ standard error is overlaid on top and grouped by gender. The red horizontal line corresponds to the *Sot+*/*Sot-* classification threshold. **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 (on triplicates of 10s) 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 ≥ 0.5 , the ECG was classified as *Sot+*. With the voting approach, an average *Sot+* classification score was computed. The same threshold was applied to predict the *Sot-*/*Sot+* class. Blue, orange and brown colours 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.

Sotalol-intake classification models discriminate congenital LQT profiles

Since drug-induced and cLQT both display prolonged QTc, we hypothesized that the models trained to recognize the sotalol footprint would also be able to recognize ECG from cLQT patients from healthy controls; and discriminate particularly well cLQT2. Indeed, this form of cLQT shares a similar pathophysiological mechanism with diLQT, (i.e genetic I_{Kr} loss of function and pharmacological I_{Kr} blockade, respectively). We used the M1 multilead model trained on cohort 1 (“Generepol” training set) and applied to the ECG from the “Generepol” holdout subset and of the cLQT patients (cohort 2). The model’s prediction results confirmed our hypothesis (**Figure 3A**). First, we noticed that the majority (97%) of ECG recordings on sotalol from cohort 1 holdout subset were classified as *Sot+* using the voting strategy. Similarly, the majority of ECGs recorded before sotalol intake (control) were classified as *Sot-* (95%; **Figure 3A left**). Second, most ECGs from cLQT patients (66%) were classified as *Sot+*. Of them, cLQT2 displayed the strongest proportion (80%, 74%) of *Sot+*, followed by cLQT3 (64%, 67%) and cLQT1 (55%, 51%), respectively at the individual 10s ECG level and after voting (**Figure 3A right**).

For an improved understanding of the predictive power of the CNN model, we performed ROC analyses to discriminate groups from one another using single 10s ECG. **Figure 3B** displays results comparing ECG from patients on sotalol versus controls, cLQT1, cLQT2 and cLQT3 groups, with ROC AUCs of 0.99, 0.78, 0.58 and 0.67, respectively. Similarly, we compared ECG from cLQT2 with those from cLQT1 and cLQT3, patients on sotalol and control groups and observed ROC AUCs of 0.66, 0.57, 0.58 and 0.9, respectively (**Figure 3C**). The low AUC (0.58) of ECG from cLQT2 patients vs. those from individuals on sotalol indicated that the CNN model could not discriminate well these two groups. Indeed, the footprint detected by the model in the ECGs of both conditions were very similar. Similarly, when we applied the *Sot+* classification score to discriminate cLQT2 patients from healthy controls, we obtained a high AUC score of 0.9.

Noteworthy, the CNN model trained to recognise sotalol intake moderately separated cLQT2 from cLQT1 and cLQT3, with AUCs of 0.66 and 0.57, respectively. Although the mechanisms underlying cLQT1 and cLQT3 mutations are different from those of cLQT2, these data were not used to train the *Sot+*/*Sot-* model. cLQT1, and to a lesser extent cLQT3 footprints on the ECG were different from that left after sotalol intake (AUC of 0.78, and 0.67, respectively; **Figure 3B**). This is in accordance with sotalol intake also activating I_{Na-L} (the mechanism of cLQT3)²⁹ on top of strongly inhibiting I_{Kr} (the mechanism of cLQT2).

We then tested with a mixed linear model the *Sot*+ classification score in relation with the type of cLQT controlled for variation in betablocker intake and possibly different ECG acquisition devices for a same patient (ID). The average classification score for cLQT2 ECG was 0.53, significantly higher than the score for cLQT1 (0.34, $p < 7.4e-7$) and cLQT3 (0.43, $p < 0.14$), after adjustment for betablockers having a significant interaction with *Sot*+ classification for cLQT2 ($p < 7.3e-6$, effect size 0.19; no significant effect on cLQT1 and cLQT3). Individual *Sot*+ classification scores are displayed in **Figure S8**. Of note, age and sex were not significantly associated with *Sot*+ classification score in this cohort.

Figure 3: CNN model performance in classifying study participants as *Sot*+/*Sot*-.

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.

Sotalol-intake classification models display increased risk for TdP events

We evaluated the usefulness of the M1 multilead CNN model to predict the risk of diTdP events in the third cohort. We quantified the association between *Sot+* classification and the TdP footprint on ECG from patients having had at least one diTdP event. The TdP footprint was coded as a four-class variable combining the time window since the diTdP event (<24h, 24-48h, >48h) along with the existence or absence of PVC for the >48h timeframe. We therefore fitted a mixed linear model to describe the TdP footprint phenotype as a function of QTc, intake of drugs with known risk for TdP, the *Sot+* classification score and patient (ID).

As expected, TdP footprint was associated with QTc ($p < 1.87e-10$), as well as intake of drugs at risk of TdP ($p < 3.17e-7$). Additionally, the TdP footprint was also associated with *Sot+* classification score (24h from the event compared with >48h without PVC ($p < 0.0018$; **Figure 4**; adjusting for QTc, TdP risky drug intake as fixed effects and patient ID as random effects). In other words, ECG closest to the TdP event had greater *Sot+* score. Interestingly, this association persisted after adjustment for QTc and the intake of drugs with known risk of TdP. These observations indicate that the CNN model learned to recognize additional features in the signal of sotalol exposed ECG, which allowed discriminating diTdP risk, beyond QTc duration and intake of QT prolonging drugs.

Figure 4: Model's classification *Sot+* 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).

CNN models uncover novel clinically relevant representation of the ECG data

Besides being among the most accurate models in ML, providing there is enough data to train, deep neural networks offer other advantages. One of them is their ability to automatically discover novel features and complex representations from the data. A way to access such knowledge is by extracting the output of the inner layers of the networks, also known as embeddings.³⁰ After being processed in the various layers of the CNN to distinguish *Sot+* from *Sot-*, ECG were transformed by the model's functions and weights in embeddings, consisting on vectors of 512 values representing learned abstractions. These were extracted for all ECG from the three study cohorts (see methods). A non-linear dimension reduction was applied for easier visualization using t-distributed Stochastic Network Embedding (t-SNE) (**Figure 5**; see methods). When annotating all ECG from the first study cohort as a function of the predicted risk score, we noticed a gradient pattern of the ECG as a function of the *Sot+* risk prediction score (**Figure 5A**). Interestingly, this corresponded closely to the timing of ECG from sotalol intake (**Figure 5B**), demonstrating the relevance of what the model “learned” from the data in recognizing the sotalol footprint on the ECG.

Next, we focused on the second cohort and analyzed all ECG obtained from the cLQT participants using the same t-SNE mapping. Interestingly, most of the ECG from the cLQT2 group were located in the high *Sot+* classification score zone (top part of the map), while those from the cLQT1 and cLQT3 were uniformly distributed in the map (**Figure 5C**). In this experiment, we retrieved these similarities between cLQT2 and sotalol intake even in the model's “perception” of the ECG, indicating that the CNN model automatically learned clinically relevant characteristics.

We then visualized the ECG from the third diTdP study cohort. Most of these ECG were located near the average to high-risk areas of the embedding map, especially those ECG that were recorded within the 24h of the diTdP event (**Figure 5D**). Noteworthy, as introduced above, most of these patients were heavily medicated, some of whom with drugs known to prolong the QT and block I_{Kr} . All these factors, each adding specific footprints to ECG signals, influenced the model's ability to recognize the sotalol footprint.

Altogether, these results indicate that besides the classification accuracy in recognizing sotalol footprint, the CNN model provided us with embeddings, which condense important information. Aside from being clinically relevant, such novel representations of the data open perspectives for novel stratification of the ECG and consequently patients (**Figure S9**).

Figure 5: Risk prediction model's deep embeddings reveal clinically relevant structure of the data
All panels illustrate 2 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 first study cohort. Grayscale indicates the *Sot*+ classification score ranging in [0:1]. **B:** Same t-SNE map as A) where ECG are coloured 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 second cohort of cLQT patients. ECG are annotated by cLQT type. **D:** Same t-SNE map as A) with ECG from the third cohort of patients having experienced at least one diTdP event. ECG are coloured by the four groups of TdP intensity footprint (timeframe from the diTdP event and presence/absence of PVCs on the ECG).

Interpretability analyses of the models uncover sotalol intake footprint.

Machine learning research has focused on building accurate models for large data collections, often at the expense of interpretability. However, it is critical to understand why and how a decision is made, especially for healthcare providers³¹. We explored the different CNN models trained to recognize sotalol intake and attempted to uncover the drug's footprint. For this, we adapted an *attribution interpretability algorithm*³² based on input perturbation using a sliding window of 100ms (50 points). We used segmented ECG signals by beats in order to summarize the importance of each point (i.e. feature) of the signal in the model's classification decision using the same occlusion approach. **Figure 6** displays averaged standardized beat ECG signal for leads LII, V2 and V3 for each time point of the first cohort. Overlaid behind the ECG profile is displayed the standardized feature importance profile (i.e. FIP) as quantified with the occlusion approach. The negative part of the FIP indicates features, which the model uses to provide a *Sot-* score, while the positive ones a *Sot+* score. Its amplitude is proportional to the importance of the features.

We noticed that the FIP changed with sotalol concentration in the blood (maximum at 3 hours, **Figure 2A**). 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 as well as on the T-wave onset and offset. These features are used by the model to recognize an ECG without sotalol footprint —the QRS complex indicating a regular occurring attribute of the ECG signal potentially used to calibrate the data input. One hour after sotalol intake, the FIP distribution started to change. The intensity of the QRS lowered and the importance of the signal after the T-wave and before P-onset increased for *Sot+* classification. This region corresponds to the RR distance, which is the determinant of heart rate. Indeed, sotalol has beta-blocking properties known to slow the sinus rate, which seems to be captured by the model and used to classify these ECG as *Sot+* at first. Two hours after taking the drug, we noticed the FIP increasing at T-wave onset (corresponding to the J-T_{peak} interval), which reached maximum intensity three hours after intake. At that time, the I_{Kr}-blockade mechanism was active and strongly apparent in the ECG footprint. Moreover, we noticed that the model focused more on this marker rather than the RR seen at sotT1 and sotT2. Simultaneously the importance of the QRS diminished while shifting towards the J-T_{peak}. We performed the same experiment in the unilead models trained in ECG's leads V2 and V3 and noticed similar behaviour of the FIP (**Figure 5**). When overlapping ECG signals summarized at each time point, we retrieved a decrease in maximal amplitude of the T-wave of lead V2 and V3 as well as a prolongation of the QT distance (**Figure S12**) as previously described.¹¹

Figure 5: 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 analysed together. Similarly, the standardized feature importance profile (FIP) is summarized behind the ECG profile. Colours for both the ECG and FIP indicate intensity of the FIP.

Discussion

QT prolongation in ECG, although not very reliable, was shown to be associated with TdP and it is therefore considered a surrogate risk factor for TdP being largely used in clinical practice¹⁴. Here, we proposed an original approach to improve TdP risk prediction. We hypothesized that it would be possible to use cutting edge AI models to learn the footprint of drugs at high-risk of TdP in healthy volunteers. We could use these models to quantify a novel risk score on other participants exposed to these drugs or with cLQT forms. 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 paradigm of I_{Kr} blockade, the mechanism by which drugs cause QTc prolongation and predispose for TdP. The CNN models detected accurately on ECG the intake of drugs at risk of TdP, and discriminated the presence and type of cLQT, being particularly accurate for cLQT2. Moreover, these models enhanced 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-T_{peak} interval to recognize the sotalol induced ECG footprint alteration. Models based on a single lead performed in general as well as those using 8 leads.

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 AI model to learn robustly. The rationale of using a cohort exposed to sotalol is that this drug is known to prolong ventricular repolarization through I_{Kr}-inhibition^{33,34}, that rarely but-dose dependently can lead to TdP. Then, the CNN models developed here were able to accurately classify if a patient did or did not take sotalol, regardless of the time after drug consumption. Furthermore, multiple acquisitions taken together with a voting approach improved the classification. This demonstrated the presence of rich information contained within the electrocardiograms, exceeding sole measurement of QT duration along 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} current¹¹. Here, we demonstrated that the similarities of the sotalol ECG footprint with cLQT2 allowed us to accurately classify 80% of the ECG from cLQT2 patients. This result offers the potential for clinical applications such as screening incoming patients for cLQT and discrimination of types, with very low cost, before using more expensive genetic tests or scarce expert ECG repolarization evaluation. Although QTc interval is prolonged in all cLQT, the ECG waveforms carry specificities including T-wave morphology abnormalities that are specific of each type of cLQT.⁶ However, the models developed herein were not trained to distinguish the different cLQT groups, particularly cLQT type 1 and 3, for which more data is needed.

The use of betablockers was a significant factor influencing the classification score in cLQT participants. Sotalol, has betablocker properties and this was learned by the model. However, cLQT2 patients still displayed a higher *Sot+* score compared with cLQT1 and cLQT3, after adjusting for betablockers intake. Therefore, we do expect that ECG CNN based models will be able to correctly discriminate the different types of cLQT, provided that they are trained with enough ECG data.

When applied to an independent study cohort of patients who experience diTdP events, the *Sot+* score was higher within the 24 hours of TdP events compared to ECG from same

individuals more than 24 hours apart from the event. These results demonstrate that it is possible to discriminate patients who have a higher TdP footprint, despite a prolonged QTc induced by medication intake for most participants. Our models could be helpful to diagnose patients who experienced an out of hospital TdP event or risk stratify patients on short-term TdP events.

To the best of our knowledge, this is the first study which successfully deploys the original approach of learning drug footprints to predict heart pathology risk based on ECG. A prior study was able to correlate drug concentrations on ECG using CNN.³⁵ The authors analysed 10-s 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 repository³⁶, and found that CNN model was superior to QT interval measurement alone in predicting plasma dofetilide concentration. However, the database used in their study was relatively small and they did not use cross-validation in training, with the risk of overfitting. Furthermore, they could not conclude with regards to the occurrence of arrhythmic risk, application to cLQT, and interpretability of their findings (**Figure 5**) 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,²⁵ the 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 prolapse³⁷. The ROC AUCs were in the 77%-94% range for the 4 classes. Similar technology was also used by Attia et al³⁸, 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, developed a low ejection fraction in the years to come, suggesting that the model was able to detect features of this condition before it became clinically diagnosed. Unfortunately, the healthy volunteers from the first study cohort misclassified by our model as *Sot+*, before any sotalol intake, were not followed, precluding any evaluation of their subsequent risk for TdP and sudden-death.

Noteworthy, 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 original and unexpected

result, indicating that sotalol footprint is detected by all leads and in similar ways, with the exception of lead V1. Moreover, the ECG data were recorded with a multitude of different acquisition devices and some ECG, recorded in 250Hz, were upsized using interpolation techniques. But still, the results were robust, regardless of the technical variability of the data. 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 analysed 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 footprint in the ECG, which changed with time as the sotalol concentration in the blood 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 receptors. 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 of the prediction process is increasingly requested³⁹ when not mandatory⁴⁰, especially for CNN, which trains millions of parameters and who are considered as “black boxes”. J-T_{peak} features emerged as the main attribute allowing for discrimination between *Sot+* and *Sot-* signals. This is concordant with the emerging literature on the importance of this segment when predicting for drug-induced TdP pro-arrhythmic risk beyond QTc.^{41,42}

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 ECG from the three study cohorts according to their clinical relevance. These embeddings can be used to automatically stratify ECG and ultimately patients, in novel classes that are yet to be characterized (**Figure S10**). However, identifying the best embeddings can be challenging since the exploration space of model architectures can be very large. More research and training data are needed in the context of translational clinical applications of CNN models in TdP for an improved representation, understanding and finally patient care.

Methods

Study cohort datasets

Three cohorts were used in this study. The first was derived from the “Generepol” study (clinical trials.gov: *NCT00773201*)¹¹ conducted at the Clinical Investigation Center Paris Est (Paris, Pitié-Salpêtrière hospital, France), in which 10,292 ECG from 990 healthy subjects were recorded before and 1, 2, 3 and 4 hours after the administration of 80mg of sotalol (sotT1, sotT2, sotT3 and sotT4, respectively). Multiple ECG recordings were taken at baseline and sotT3 with a median total recording number of 15 per patient, (IQR=1, min=11, max=19), acquired during the one-day intervention. The study participants were on majority women (n=614, 62%) with an average age of 28 years old (median=24; IQR=12; min=18; max=60). The study protocol was approved by the Committee for the Protection of Human Subjects of Paris Ile-de-France and written informed consent was obtained from all subjects prior to study enrolment. The study was performed in accordance with the principles of the international conference on harmonization guidelines on good clinical practice and the world medical association declaration of Helsinki.

The second cohort was obtained from congenital-LQT patients confirmed by genetic testing followed over a 26-year period at the Arrhythmia unit of Bichat Hospital (Paris, France). This cohort included 266 participants with cLQT1 (ECGs=560), 188 with cLQT2 (ECGs=456), and 33 with cLQT3 (ECGs=67). Out of the 487 cLQT patients, 282 (60%) were female, and 199 (41%) had longitudinal follow-up with an average of 5 ECG recordings (median=3, IQR=6, min=1, max=15). The longest follow-up was for 23 years (median=0, IQR=1). Median age of participants in this cohort was 28 years (IQR: 29, min=0, max=84). 213 (44%) participants in this cohort had at least one ECG recording under betablocker medication and 16 (3%) had a cardiac implantable electronic device. The use of these confidential, electronically processed patient data was approved by the French National Commission for Data Protection and Liberties (Commission Nationale de l’Informatique et des Libertés).

The third cohort was composed of participants prospectively enrolled and followed at Vanderbilt University Medical Center (Nashville, TN, USA) who had experienced at least one TdP event when exposed to drugs with known risk of TdP⁹, in the absence of acute cardiac ischemia at the time of the event (n=48 and ECG=1,105). Most ECGs (n=733, 66%) were recorded while patients were taking drugs with well-established IKr blocking activity and known risk for TdP.^{9,43} Amiodarone, sotalol, hydroxychloroquine, dofetilide and fluconazole, were the most prevalent (>95%). Some patients took multiple drugs with known risk for TdP

simultaneously (69% with only one drug, 24% with 2 drugs, and 5% with 3 drugs). None of the participants were diagnosed with genetically confirmed cLQT. The median follow-up of these participants was 4 years (IQR=10, min=0, max=17), and had a median of 31 ECG recordings (IQR=21; min=3; max=55) . Median age at the time of the first ECG recording was 60 years, (IQR=26, min=18, max=85), and 29 (60%) were female. Recordings from these patients were curated by two expert cardiologists and those patients displaying ventricular tachycardia (n=1) and junctional tachycardia (n=22) during the 10-second acquisition were excluded from the analyses. The ECGs were classified into four categories-based time windows from the TdP event (<24h, 24-48h, >48h+PVC and >48h-PVC) and the presence of premature ventricular contractions (PVC) with 103, 44, 115, and 843 ECGs, respectively. This study was approved by Vanderbilt University Medical Center review committee (IRB#040551), written informed consent was obtained from all subjects prior to study enrolment.

Data preparation

Raw ECG data were acquired with a variety of different devices in the different centres (Pitié-Salpêtrière, Bichat and Vanderbilt University Medical Center hospitals). The ECG were associated to 8 independent leads (LI-LII, V1-V6), allowing for reconstruction of 12 leads (namely adding LIII, aVF, aVL, aVR). Each recording lasted 10s, with a sampling frequency varying from 250 or 500 Hz. The Pitié-Salpêtrière cohort recruitment, started in 2008 and ended in 2012). ECG recordings were provided in the .scp files format (Cardioplug device; Cardionics Inc, Brussels, Belgium) and were parsed using the Biosig software⁴⁴. The ECG data used corresponded to a segment of 10 consecutive seconds sampled at 500 Hz. Similarly, raw ECG data from the Bichat study were acquired from 1992 to 2018. Raw .xml files were parsed using the python xmldict library. In total, 875 files were acquired at 500 Hz, and 208 were acquired at 250 Hz, for a total of 1,083 recordings. The 250 Hz signals were up-sampled to 500 Hz using a cubic interpolation (interp1d function from the scipy python library). Devices used for ECG signal acquisition were General Electric MAC 5500 (n=830), Marquette MAC15 (n=189), Marquette MAC VU (n=42), Cardionics Cardioplug (n=3). Finally, data from the Vanderbilt cohort were acquired from 2002 to 2019. ECG were also raw .xml files sampled at 500Hz for most of the files (n=971). The remaining ones sampled in 250Hz were interpolated (n=134) as described above. Devices used for ECG signal acquisition were M3700 System (n=128), PageWriter Touch (n=517), PageWriter Trim (n=10), PageWriter XL (n=6), PageWriter TC (n=444).

No filtering, nor any other transformation was applied to the data considered in these studies, except for standardization, which was performed at the whole ECG level, i.e. each lead signal was standardized by the mean of all other lead signals for a given ECG for the multilead models and at the lead level for the unilead model. The data were stored in python dictionaries that were used for training the models. For model training, the data was converted onto 3D tensor (8 leads, 5000 time points for each lead, recordings).

QTc measurement and QTc calculation

For all ECG recorded in the Generopol cohort, the QTc values, i.e. heart-rate corrected QT using the Fridericia formula⁵ were measured by expert technicians and controlled by expert cardiologists (Cardibase, Banook group, Nancy, France). The method consisted in measuring three consecutive QT values with a tangent-based approach mainly on lead LII. Details concerning the inter and intra-observer variability are detailed in our previously published work¹¹. For the Bichat cohort, QTc measurement was based on the overlap method on the averaged QT as previously described²⁸. For the Vanderbilt cohort, QTc was automatically provided by the acquisition machines before being manually validated by an arrhythmia specialist.

Statistical analyses

Data are presented as count and frequencies for categorical variables and median and IQR for continuous variables, unless otherwise specified. Multiple ECG recordings were acquired for most of the study participants in the three different cohorts. We used mixed effects linear models to best describe the data and their relations while controlling for random effects such as patient ID. Summary statistics were extracted from the models along with standard error estimates and p-values. We used the R packages lme4⁴⁵ and lmerTest for this. We tested multiple combinations in the models with increasing complexity. Models were compared using ANOVA and the best models were selected on AIC. When age was introduced in the model, it was centred and the estimate of the intercept describe the average age. When gender was added to the model, interaction was searched to compare effects. A model example is $qtcf \sim 1 + condition*sex + age.c + (1|patient)$. Chi2 test was used for comparing proportions. All model results were processed and were analysed in R. All graphics were coded in R with the ggplot2 package. ROC analyses were performed using the pROC package. Accuracy, recall, precision, f1-score and ROC AUC were the different metrics used to evaluate the data. A p-value less than 0.05 and a FDR less than 0.1 were considered to indicate statistical significance; all tests were two tailed.

Embedding analyses

Embeddings from the multilead classification model were obtained for each ECG by accessing the output of the last group of convolutional, dense and batch normalization layers. A nonlinear dimension reduction technique was applied based on the t-SNE algorithm (perplexity=100, iteration=1000) using the Rtsne package. ECG were visualized onto these maps annotated by available information. Unsupervised learning was applied to the embedding data in order to explore for clinically relevant structure. All dimensions of the embeddings were used to identify partitions with the k-means method with default parameters implemented in base R. The relation between unsupervised embedding partitions and phenotypic conditions were tested using Chi2.

Voting strategy

Performances were computed using both single-signal analysis and by averaging risk scores from multiple recordings 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 affect the patient to one of the *Sot+* or *Sot-* classes, multiple ECG from the same patient were processed by the models and the patient was affected as being *Sot+* (versus *Sot-*) based on the average classification score of the different ECG; on which a threshold of 0.5 was applied (*Sot+* if score ≥ 0.5).

Sotalol-intake classification with the multilead model

We used the eight channels (LI, LII, V1-6; allowing for the reconstruction of the 12 leads standard ECG) to train a CNN model to predict *Sot+* and *Sot-* classes. The Generepol cohort was split in two sets: general training (80%) and holdout (20%). 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 model (**Figure S6**), was composed of 11 blocks of convolution: each block containing 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. Zero padding was used for each Conv1D to keep the same output dimensions (option "same"). The remaining model parameters were left as default. After the convolutional blocks, the data were fed to a dense layer (512 nodes), a 'Relu' non-linear activation function and 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. After cross-validation, the model was trained on the whole general training set and then tested on the holdout database and the two other study cohorts (cLQT and drug-induced TdP patients). The code was written in python 3.8.5 and Keras 1.1.2.

Sotalol-intake classification with the unilead models

We designed 8 CNN models based on the same architecture that were trained on each single channel (LI, LII, V1-6) and tested on each of the leads independently. We split the Generopol cohort in two sub-datasets: global train (90%), which was used for any training, validation and evaluation tasks and a holdout set (10%). The holdout was not used for training or hyperparameter tuning but was solely used for evaluating performance of the final trained model. We subdivided the “global training” sub-dataset in small datasets using the subjects’ IDs: (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. Specifically, we designed a DenseNet-like model architecture, which mitigates the problem that many of the layers are redundant thus making the network heavy and difficult to optimize.

The model’s architecture (**Figure S7**) started 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, consisting of eight successive dense convolutional blocks (*DenseBlocks*). A dense convolutional block was composed of several convolutional sub-blocks (*DenseLayers*), each sub-block begins with four layers of bottlenecks composed of a batch normalization followed by a Leaky ReLu activation, a 1x1 convolution and a dropout layer. These bottlenecks reduced the filter space dimensionality, decreasing the number of feature maps whilst retaining their salient features. The bottleneck technique allowed reducing 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 in a feedforwarding. Finally, 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 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 likelihood for each class *Sot-* and *Sot+*. We used the Adam optimizer and

binary cross entropy as loss function. In order to find an optimal combination of the hyperparameters to obtain better precision, we ran a hyper-optimization process. For each ECG lead and for each hyperparameters combination, we trained a model on the small training set with 50 iterations. The following hyperparameters were used to optimize the models:

- Learning rate for the optimizer function (Adam): 0.01 to 0.001
- Bottleneck : True or False
- Compression rate : 1.0, 0.3
- Dense layers : 6, 8
- Dense blocks : 6, 8
- Dropout rate : 0, 0.2, 0.5

The process resulted in numerous models and for each ECG lead we chose the best model by comparing their performance on the evaluation subset. Eventually, at the end of this process we obtained eight 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. The models were written and trained in Python with Tensorflow 2.

Interpretability of the CNN models

We used the occlusion method implemented in python with Tensorflow 2. This method consists in iteratively occluding a predefined portion of the data and making a prediction. This perturbation of the input data allows assessing the importance of features in the final classification. Here, we used a window of 50 points (corresponding to 100ms in 500Hz recordings) that was iteratively slid across the signals in an attempt to identify which parts of the signal were the most useful for the classification of ECG as *Sot+*. Eventually, for each hidden feature, we could quantify its importance in the resulting classification result. We first tested the interpretability on the multi-lead model. However, although this model was highly performant, it was difficult to interpret. The signals 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). This meant that a positive contribution for the classification in lead LII, for instance, should be considered with all the positive contributions in all leads for any beat. Therefore, it was difficult to assess what lead and beat were relevant given the fact the occlusion method considered all time and space variables as a single entity (**Figure S10**). To achieve better human interpretability, we then used a single-lead model.

Bibliography

- 1 Roden, D. M. Drug-induced prolongation of the QT interval. *N Engl J Med* **350**, 1013-1022, doi:10.1056/NEJMra032426 (2004).
- 2 Nakano, Y. & Shimizu, W. Genetics of long-QT syndrome. *J Hum Genet* **61**, 51-55, doi:10.1038/jhg.2015.74 (2016).
- 3 Moss, A. J. *et al.* ECG T-wave patterns in genetically distinct forms of the hereditary long QT syndrome. *Circulation* **92**, 2929-2934, doi:10.1161/01.cir.92.10.2929 (1995).
- 4 Yap, Y. G. & Camm, A. J. Drug induced QT prolongation and torsades de pointes. *Heart* **89**, 1363-1372, doi:10.1136/heart.89.11.1363 (2003).
- 5 Saque, V. *et al.* Fast, accurate and easy-to-teach QT interval assessment: The triplicate concatenation method. *Arch Cardiovasc Dis* **110**, 475-481, doi:10.1016/j.acvd.2016.12.011 (2017).
- 6 Porta-Sanchez, A. *et al.* T-Wave Morphology Analysis in Congenital Long QT Syndrome Discriminates Patients From Healthy Individuals. *JACC Clin Electrophysiol* **3**, 374-381, doi:10.1016/j.jacep.2016.10.013 (2017).
- 7 Mazzanti, A. *et al.* Interplay Between Genetic Substrate, QTc Duration, and Arrhythmia Risk in Patients With Long QT Syndrome. *J Am Coll Cardiol* **71**, 1663-1671, doi:10.1016/j.jacc.2018.01.078 (2018).
- 8 Medicine, N. L. o. *DailyMed: SOTALOL- sotalol hydrochloride tablet*, <<https://dailymed.nlm.nih.gov/dailymed/drugInfo.cfm?setid=56879738-8662-4f5c-8386-761ab2b5e46f>> (
- 9 Woosley, R. L., Black, K., Heise, C. W. & Romero, K. CredibleMeds.org: What does it offer? *Trends Cardiovasc Med* **28**, 94-99, doi:10.1016/j.tcm.2017.07.010 (2018).
- 10 Nguyen, L. S. *et al.* Cardiovascular Toxicities Associated With Hydroxychloroquine and Azithromycin: An Analysis of the World Health Organization Pharmacovigilance Database. *Circulation* **142**, 303-305, doi:10.1161/CIRCULATIONAHA.120.048238 (2020).
- 11 Salem, J. E. *et al.* GENomE wide analysis of sotalol-induced IKr inhibition during ventricular REPOLarization, "GENEREPOl study": Lack of common variants with large effect sizes. *PLoS One* **12**, e0181875, doi:10.1371/journal.pone.0181875 (2017).
- 12 Administration, U. S. D. o. H. a. H. S. F. a. D. *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>> (
- 13 Drew, B. J. *et al.* Prevention of torsade de pointes in hospital settings: a scientific statement from the American Heart Association and the American College of Cardiology Foundation. *Circulation* **121**, 1047-1060, doi:10.1161/CIRCULATIONAHA.109.192704 (2010).
- 14 Hondeghem, L. M. Drug-Induced QT Prolongation and Torsades de Pointes: An All-Exclusive Relationship or Time for an Amicable Separation? *Drug Saf* **41**, 11-17, doi:10.1007/s40264-017-0584-4 (2018).
- 15 Graff, C. *et al.* Covariate analysis of QTc and T-wave morphology: new possibilities in the evaluation of drugs that affect cardiac repolarization. *Clin Pharmacol Ther* **88**, 88-94, doi:10.1038/clpt.2010.51 (2010).
- 16 Johannesen, L. *et al.* Improving the assessment of heart toxicity for all new drugs through translational regulatory science. *Clin Pharmacol Ther* **95**, 501-508, doi:10.1038/clpt.2013.238 (2014).
- 17 Viskin, S. *et al.* Inaccurate electrocardiographic interpretation of long QT: the majority of physicians cannot recognize a long QT when they see one. *Heart Rhythm* **2**, 569-574, doi:10.1016/j.hrthm.2005.02.011 (2005).
- 18 Obermeyer, Z. & Emanuel, E. J. Predicting the Future - Big Data, Machine Learning, and Clinical Medicine. *N Engl J Med* **375**, 1216-1219, doi:10.1056/NEJMp1606181 (2016).
- 19 LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436-444, doi:10.1038/nature14539 (2015).
- 20 Hinton, G. Deep Learning-A Technology With the Potential to Transform Health Care. *JAMA* **320**, 1101-1102, doi:10.1001/jama.2018.11100 (2018).
- 21 Acharya, U. R. *et al.* Application of deep convolutional neural network for automated detection of myocardial infarction using ECG signals. *Information Sciences* **415**, 190-198 (2017).

- 22 Acharya, U. R. *et al.* Automated identification of shockable and non-shockable life-threatening ventricular arrhythmias using convolutional neural network. *Future Generation Computer Systems* **79**, 952-959, doi:10.1016/j.future.2017.08.039 (2018).
- 23 Li, D., Zhang, J., Zhang, Q. & Wei, X. in *2017 IEEE 19th International Conference on e-Health Networking, Applications and Services (Healthcom)*. 1-6 (IEEE).
- 24 Acharya, U. R. *et al.* Automated detection of arrhythmias using different intervals of tachycardia ECG segments with convolutional neural network. *Information Sciences* **405**, 81-90, doi:10.1016/j.ins.2017.04.012 (2017).
- 25 Liu, W. *et al.* Real-Time Multilead Convolutional Neural Network for Myocardial Infarction Detection. *IEEE Journal of Biomedical and Health Informatics* **22**, 1434-1444, doi:10.1109/JBHI.2017.2771768 (2018).
- 26 Chen, T. M., Huang, C. H., Shih, E. S. C., Hu, Y. F. & Hwang, M. J. Detection and Classification of Cardiac Arrhythmias by a Challenge-Best Deep Learning Neural Network Model. *iScience* **23**, 100886, doi:10.1016/j.isci.2020.100886 (2020).
- 27 Acharya, U. R. *et al.* Automated detection of coronary artery disease using different durations of ECG segments with convolutional neural network. *Knowledge-Based Systems* **132**, 62-71, doi:10.1016/j.knosys.2017.06.003 (2017).
- 28 Extramiana, F. *et al.* 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* **62**, 148-154, doi:10.1016/j.jelectrocard.2020.08.016 (2020).
- 29 Yang, T. *et al.* Screening for acute IKr block is insufficient to detect torsades de pointes liability: role of late sodium current. *Circulation* **130**, 224-234, doi:10.1161/CIRCULATIONAHA.113.007765 (2014).
- 30 Su, C., Tong, J., Zhu, Y., Cui, P. & Wang, F. Network embedding in biomedical data science. *Brief Bioinform*, doi:10.1093/bib/bby117 (2018).
- 31 Rudin, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat Mach Intell* **1**, 206-215 (2019).
- 32 Zeiler, M. D. & Fergus, R. Visualizing and Understanding Convolutional Networks. *arXiv* (2013).
- 33 Funck-Brentano, C. Pharmacokinetic and pharmacodynamic profiles of d-sotalol and d,l-sotalol. *Eur Heart J* **14 Suppl H**, 30-35 (1993).
- 34 Haverkamp, W. *et al.* 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* **47**, 219-233 (2000).
- 35 Attia, Z. I. *et al.* Noninvasive assessment of dofetilide plasma concentration using a deep learning (neural network) analysis of the surface electrocardiogram: A proof of concept study. *PLoS One* **13**, e0201059, doi:10.1371/journal.pone.0201059 (2018).
- 36 Moody, G. B., Mark, R. G. & Goldberger, A. L. PhysioNet: a research resource for studies of complex physiologic and biomedical signals. *Comput Cardiol* **27**, 179-182 (2000).
- 37 Tison, G. H., Zhang, J., Delling, F. N. & Deo, R. C. Automated and Interpretable Patient ECG Profiles for Disease Detection, Tracking, and Discovery. *arXiv:1807.02569 [cs]* (2018).
- 38 Attia, Z. I. *et al.* Screening for cardiac contractile dysfunction using an artificial intelligence-enabled electrocardiogram. *Nat Med* **25**, 70-74, doi:10.1038/s41591-018-0240-2 (2019).
- 39 Martens, D., Vanthienen, J., Verbeke, W. & Baesens, B. Performance of classification models from a user perspective. *Decision Support Systems* **51**, 782-793, doi:10.1016/j.dss.2011.01.013 (2011).
- 40 Bryce, G. & Seth, F. European Union Regulations on Algorithmic Decision-Making and a "Right to Explanation". *AI Magazine* **38**, doi:10.1609/aimag.v38i3.2741 (2017).
- 41 Vicente, J. *et al.* Mechanistic Model-Informed Proarrhythmic Risk Assessment of Drugs: Review of the "CiPA" Initiative and Design of a Prospective Clinical Validation Study. *Clin Pharmacol Ther* **103**, 54-66, doi:10.1002/cpt.896 (2018).
- 42 Couderc, J. P. *et al.* T-wave morphology abnormalities in benign, potent, and arrhythmogenic I(kr) inhibition. *Heart Rhythm* **8**, 1036-1043, doi:10.1016/j.hrthm.2011.02.005 (2011).
- 43 De Bruin, M. L., Pettersson, M., Meyboom, R. H., Hoes, A. W. & Leufkens, H. G. Anti-HERG activity and the risk of drug-induced arrhythmias and sudden death. *Eur Heart J* **26**, 590-597, doi:10.1093/eurheartj/ehi092 (2005).
- 44 Baillet, S., Friston, K. & Oostenveld, R. Academic software applications for electromagnetic brain mapping using MEG and EEG. *Comput Intell Neurosci* **2011**, 972050, doi:10.1155/2011/972050 (2011).
- 45 Bates, D., Maechler, M., Bolker, B. & Walker, S. Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software* **67**, 1-48, doi:10.18637/jss.v067.i01. (2015).

Figures

Figure 1

Experimental design and QTc distribution across study cohorts. A: Description of the three study cohorts. The first cohort was composed of healthy volunteers taking 80mg of oral sotalol. This dataset was used to train the models. The second cohort was composed of cLQT patients of type 1, 2 and 3. The third cohort included patients who experienced events of drug-induced TdP with no underlying identified cLQT. B: Illustration of ECG use (multilead and unilead) in training and testing the deep CNN models up to the prediction of a risk score and evaluation by the physician. C) The distribution of QTc values following the sotalol drug-induced QTc prolongation in the Generepol study¹¹. The X-axis corresponds to the time in minutes since inclusion and data is colour coded by gender. Lines link ECG recordings from the same participant over the duration of the protocol. D) Boxplot distribution of the estimated QTc values in the cLQT cohort by subtypes. ECG are grouped and coloured by gender. E) Boxplot distribution of QTc values across the di-TdP study cohort grouped by time to TdP event and presence or not of premature ventricular contractions (PVC). ECG are grouped and coloured by gender. The black horizontal lines indicate the 480ms at-risk QTc threshold for the panels C-E.

Figure 2

Classification performance of CNN and linear regression (QT) models in discriminating baseline ECG before sotalol intake from those after sotalol intake (SotT1, SotT2, SotT3). A) Boxplots, illustrating the distribution of circulating sotalol concentration in (ng/ml) in the Generepol cohort two and three hours after 80mg oral sotalol intake. Data is displayed separated and coloured by gender. B) Scatterplot illustrating the evolution of the classification score for the Sot+ class (y-axis) across time from inclusion (x-axis). All points (averaged ECGs) of a study participant are linked together as trajectories and are coloured by gender. Summarized loess 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. 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 (on triplicates of 10s) 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 ≥ 0.5 , the ECG was classified as Sot+. With the voting approach, an average Sot+ classification score was computed. The same threshold was applied to predict the Sot-/Sot+ class. Blue, orange and brown colours 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 3

CNN model performance in classifying study participants as Sot+/Sot-. 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 4

Model's classification Sot+ 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 5

Risk prediction model's deep embeddings reveal clinically relevant structure of the data. All panels illustrate 2 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 first study cohort. Grayscale indicates the Sot+ classification score ranging in [0:1]. B: Same t-SNE map as A) where ECG are coloured by the experimental setup, from inclusion before and 1, 2, 3 or 4 hours after sotalolol

intake. C: Same t-SNE map with ECG from the second cohort of cLQT patients. ECG are annotated by cLQT type. D: Same t-SNE map as A) with ECG from the third cohort of patients having experienced at least one diTdP event. ECG are coloured by the four groups of TdP intensity footprint (timeframe from the diTdP event and presence/absence of PVCs on the ECG).

Figure 6

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 analysed together. Similarly, the standardized feature importance profile (FIP) is summarized behind the ECG profile. Colours for both the ECG and FIP indicate intensity of the FIP.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [PRIFTIdeepcgtdpsupplementarytext.pdf](#)