POST-HOC UNCERTAINTY QUANTIFICATION FOR QT INTERVAL MEASUREMENTS WITH ENSEMBLES OF ELECTROCARDIOGRAPHIC LEADS AND DEEP MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Deep neural networks (DNNs) have achieved state-of-the-art performance in QT interval measurement but are missing uncertainty quantification, which is necessary for safer decision making. Uncertainty is usually encoded in DNNs through probability distributions over model weights. In this paper, we combine this approach with notions of multisensory integration whereby neural systems account for uncertainty by optimally integrating all available sensory inputs. We thus approximate the posterior predictive distribution of the QT interval given a multi-lead ECG as a weighted average across leads (lead integration) and models (deep ensembling) and derive $100(1-\alpha)\%$ Bayesian prediction intervals (PIs). We apply this method to QT-based cardiac drug safety monitoring and compare it to an adapted version of conformal prediction. The Bayesian and conformal approaches yield comparable empirical coverage (77%-82% for mean PI widths of $\sim$28 milliseconds, α = 0.1). The former is more straightforward and shows better error-based calibration. Data and code implementation are available at https://anonymous.4open.science/r/qt-uncertainty-FD50/.

1 INTRODUCTION

The prolongation of the QT interval is a surrogate biomarker for the risk of torsades de pointes (TdP), a life-threatening arrhythmia. Electrocardiography (ECG) monitoring is therefore essential to the prevention of TdP induced by otherwise useful medications. The American Heart Association provided guidelines for such ECG monitoring (Drew et al., 2004; 2010; Tisdale et al., 2020) but their use is yet to be widespread (Putnikovic et al., 2022). Given the difficulty and unreliability of manual QT measurements (Malik, 2004; Lester et al., 2019), robust algorithms could allow accurate, continuous and real-time drug safety monitoring. Deep Learning (DL) has achieved state-of-the-art ECG interval measurement performance (Giudicessi et al., 2021; Hicks et al., 2021; Diaw et al., 2022). However, uncertainty quantification (UQ) of DL-based ECG interval measurements, important for safe decision making and model explainability, remains under-explored.

In DL, predictive uncertainty has mostly been quantified through the posterior over model weights using Bayesian neural networks (BNNs) or approximates (Gal & Ghahramani, 2016). Deep ensembling (Lakshminarayanan et al., 2017), which also fits within approximate Bayesian inference (Wilson & Izmailov, 2020), is fast becoming the gold standard for UQ in DNNs as it has outperformed approximate BNNs (Ovadia et al., 2019; Ashukha et al., 2020). Less common UQ approaches in DL include test-time data augmentation, mostly used in medical image processing (Ayhan & Berens, 2018; Wang et al., 2019). Conformal prediction (Vovk et al., 2005) has been of great interest for the machine learning (ML) community at large (Angelopoulos et al., 2020; Ndiaye, 2022) as it is distribution-free and model-agnostic.

In cognitive science (e.g. study of perception), uncertainty is usually encoded at input level of neural systems rather than in synaptic weights (equivalent of DNN weights). Pearce (2020) suggests that incorporating this paradigm in current UQ methods for DL could push the field forward. This is relevant to QT interval measurement that starts, in common practice, with an efficient visualization of all leads of a given ECG recording, similar to how the visual system optimally perceives an object by combining all available sensory signals (Jacobs, 1999). Here, uncertainty depends more on the object being observed (external or aleatoric uncertainty) than on the observer (internal or epistemic).

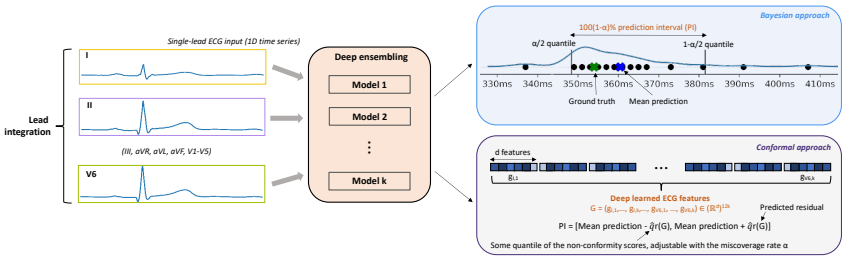

Figure 1: Construction of QT prediction intervals for a standard 12-lead ECG using a deep ensemble

In this paper, we consider approximate Bayesian and conformal approaches to generate QT prediction intervals (PIs). Our contributions are: *(i)* We draw insights from how the visual system accounts for uncertainty to propose a probabilistic model of QT measurement on multi-lead ECG recordings (lead integration). We combine lead integration with deep ensembling to approximate the posterior predictive distribution from which a PI is derived (cf. Figure 1). *(ii)* We build adaptative conformal PIs by leveraging the ECG features learned by a deep ensemble *(iii)* We demonstrate the Bayesian and conformal approaches on real-world clinical data and show that approximate Bayesian inference is more practical.

2 Uncertainty Quantification

2.1 Approximate Bayesian inference

Standard ECG recordings have multiple leads, each sensing the cardiac electrical activity from a different spatial viewpoint. Denote $D = \{(X_t^i, y_i)\}_{i=\{1,\dots,n\}}$ a dataset of n ECG samples where $X_t = \{x_t^l\}_{l=\{1,\dots,L\}}$ represents a L-lead ensemble of ECG beats (typically, L = 12) and y the corresponding QT interval. Denote $f_\theta : x_t \rightarrow y$ a QT estimator parameterized by θ, optimized on D and capable of analyzing all L types of lead. We propose to approximate the posterior predictive distribution $Y = p_\theta(y|X_t, D)$ by generating multiple QT estimates $f_\theta(x_t^l)$. We subsequently derive a PI of level $1-\alpha$, $\alpha \in [0, 1]$, following the equal-tailed method, $PI = [y^{(\alpha/2)}, y^{(1-\alpha/2)}]$ where $y^{(\alpha/2)}$ and $y^{(1-\alpha/2)}$ are the $\alpha/2$ and 1-$\alpha/2$ quantiles of the predictive distribution Y.

Lead integration. In its simplistic Bayesian model, the human visual system encodes uncertainty by linearly combining all available cues or sensory signals, each weighed in proportion to their reliability (cf. Appendix A.1). Human ECG annotators usually solve the task of finding the QT interval y by superimposing all lead components of X_t to better define the beginning of the QRS complex and end of the T wave. This resembles an optimal cue integration system as the annotator aggregates all information provided by the different leads, which can have various degrees of reliability. For instance, some leads might have lower amplitude T waves or be noisier. We propose to encode this notion of lead reliability in automated measurement using a linear pooling model (Stone, 1961),

$$Y = p_\theta(y|X_t, D) = p_\theta(y|x_t^1, \dots, x_t^L, D) = \sum_{l=1}^{L} w_l p_\theta(y|x_t^l, D) \tag{1}$$

where the weight w_l denotes the reliability of lead x_t^l relative to the other leads, all weights summing to 1. This formulation requires prior knowledge on lead reliability.

Implicit combination with deep ensembling (UQ-EL). Denote $\{f_{\theta_1}, \dots, f_{\theta_k}\}$ an ensemble of k single-lead DL models parameterized by $\theta_1, \dots, \theta_k$ respectively. We can define the mean QT estimator $f_\theta = \frac{1}{k}\sum_{j=1}^{k} f_{\theta_j}$. Assuming that, a priori, all leads have fairly equal reliability ($w_l \approx 1/L$), we then approximate Y (Equation 1) with the L estimates $f_\theta(x_t^l)$.

Explicit combination with deep ensembling (UQ-ELM[1]). The posterior distribution of the deep ensemble is approximatively a uniformly weighted mixture model (Lakshminarayanan et al., 2017),

[1]Stands for UQ with an Ensemble of Leads and Models

$$p_\theta(y|x_t, D) \approx \frac{1}{k}\sum_{j=1}^{k} p(y|x_t, \theta_j) \tag{2}$$

With $w_l \approx 1/L$, we can rewrite the predictive distribution $p_\theta(y|X_t, D)$ using Equations 1 and 2 as

$$Y = p_\theta(y|X_t, D) \approx \frac{1}{L}\sum_{l=1}^{L} p_\theta(y|x_t^l, D) \approx \frac{1}{k \times L}\sum_{l=1}^{L}\sum_{j=1}^{k} p(y|x_t^l, \theta_j) \tag{3}$$

Here, we approximate Y using the $k \times L$ estimates $\hat{y}_{l,j} = f_{\theta_j}(x_t^l)$.

2.2 Locally adaptative split conformal prediction (LASCP)

Given n past observations and a prespecified miscoverage rate α, CP consists in fitting a model f_θ on the n samples and building for a new observation Y_{n+1} a predictive interval PI_α such that $P(Y_{n+1} \in PI_\alpha) \geq 1 - \alpha$. In split CP (SCP) (Papadopoulos et al., 2002), the n samples are split in 2 sets for model fitting (performed on, say, D_{train}) and nonconformity calibration (performed on I_2 of size n_2). Nonconformity scores are computed as absolute residuals $\epsilon = |y - \hat{y}|$ and the $\lceil (n_2 + 1)(1 - \alpha) \rceil / n_2$ quantile of the scores ($\hat{q}_{n_2}$) calibrated on I_2 is used to define constant-length PIs $[\hat{y} - \hat{q}_{n_2}, \hat{y} + \hat{q}_{n_2}]$. LASCP (Papadopoulos et al., 2008; 2011; Lei et al., 2018) creates adaptative intervals by first fitting a model $r : x \rightarrow \epsilon$ (residual fitting) on another calibration dataset I_1. In the traditional LASCP algorithm, f and r have the same explicit input features. In this paper, we use the ECG features learned by a deep ensemble to train a shallow ML model for residual fitting. The proposed framework for LASCP calibration on L-lead ECG data using a deep ensemble is as follows:

1. For each L-lead ensemble in the first calibration set $I_1 = \{(X_t^i, y_i)\}_{i=\{1,\ldots,n_1\}}$ of size n_1, compute the absolute residuals $\epsilon_i = |y_i - \bar{\hat{y}}_i|$ with $\bar{\hat{y}}_i = \frac{1}{k \times L}\sum_{l=1}^{L}\sum_{j=1}^{k} \hat{y}_{l,j}$
2. Train a model $r : (\mathbb{R}^d)^{k \times L} \rightarrow \mathbb{R}$ on $\{(G_i, \epsilon_i)\}_{i=\{1,\ldots,n_1\}}$ with G_i the set of d-dimensional single-lead ECG features extracted by the ensemble of k models.
3. For every sample in the second calibration set $I_2 = \{(X_t^i, y_i)\}_{i=\{1,\ldots,n_2\}}$, compute the nonconformity score $s_i = \frac{|y_i - \bar{\hat{y}}_i|}{r(G_i)}$. Then, compute the aforementioned quantile $\hat{q}_{n_2}$ of $S = \{s_i\}_{i=\{1,\ldots,n_2\}}$. For each new ECG sample, derive $PI = [\bar{\hat{y}} - \hat{q}_{n_2} r(G), \bar{\hat{y}} + \hat{q}_{n_2} r(G)]$.

3 Experiments and results

Clinical data[2]. Study 1 (**S1**) stems from a prospective randomized placebo-controlled clinical trial including 22 healthy subjects (Johannesen et al., 2014). Each subject was followed during 5 periods during which they received a placebo (*P*) or one of the following drugs: Dofetilide (*D*), Quinidine (*Q*), Ranolazine (*R*), and Verapamil (*V*). For each period, 12-lead ECGs were recorded for 24 hours and 3 10-second ECGs were extracted at 16 timepoints (1 point pre-dose and 15 points post-dose) leading to a total of 5219 10s ECG recordings. Similarly, Study 2 (**S2**) aimed at analyzing the ECG effects of *D*, Lidocaine + *D*, Mexiletine + *D* and Moxifloxacin + Diltiazem (Johannesen et al., 2016). ECGs were extracted at 14 timepoints, leading to a total of 4211 10s ECG recordings. Semi-automated QT annotations made on representative median beats are available for each 12-lead ECG in S1 and S2.

Deep ensembling with 5-fold cross-validation. Resampling techniques can be used to generate multiple distinct models from a single training set. We constitute a training set using ECGs recorded before/after administration of *V* (non QT-prolonging drug) and *Q* (QT-prolonging) so as to learn a wide range of ECG morphologies. We refer to this set as **S1a** and the remaining data in S1 as **S1b**. We split in 5 groups the 22 subjects in S1a for patient-stratified cross-validation. We generate 5 versions of the single-lead residual neural network (ResNet) proposed by Diaw et al. (2022). Data

[2] Available at https://physionet.org/content/ecgrdvq/1.0.0/ and https://physionet.org/content/ecgdmmld/1.0.0/

preparation and DNN optimization are conducted as in the original paper (Diaw et al., 2022). For a single-lead ECG input (1.2 second average beat sampled at 500 Hz), the ResNet yields d = 32 global features by average pooling the last feature maps before the regression head.

LASCP calibration. We split S1b in two distinct sets, S1b = $I_1 \cup I_2$ (cf. Appendix A.2 for a summary of the subsets). We train a gradient boosting regressor on I_1 for residual fitting. Model parameters and residual RMSE scores are detailed in Appendix A.3.

Evaluation metrics. We compute the prediction interval coverage probability (CP), defined as the proportion of target QT intervals that fall within the PI, the mean prediction interval width (MW), and the mean absolute deviation (MAD), which measures how far the target QT intervals not covered by the PIs are from the closest PI bounds (lower or upper). We also evaluate the error-based calibration of the proposed methods by dividing the set of absolute differences between ground-truth and mean prediction (errors) into bins of 100 samples and computing the MW and mean absolute error (MAE) within each bin.

Data	Method	CP	MW (ms)	MAD (ms)
S1b (N = 3163)	UQ-EL	82%	**29.49**	**7.20**
	UQ-ELM	**95%**	43.91	9.81
S2 (N = 4211)	UQ-EL	77%	**28.27**	3.95
	UQ-ELM	**90%**	40.47	**3.26**
	LASCP	82%	28.67	3.50

Table 1: Evaluation of the 90% PIs (α = 0.1) on N 12-lead ECG ensembles

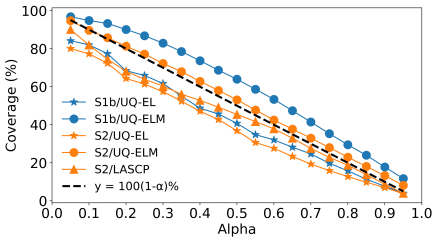

Figure 2: CP versus α

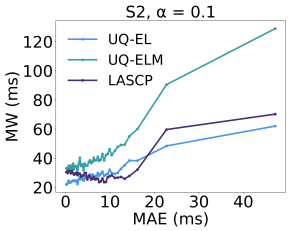

Figure 3: Error-based calibration

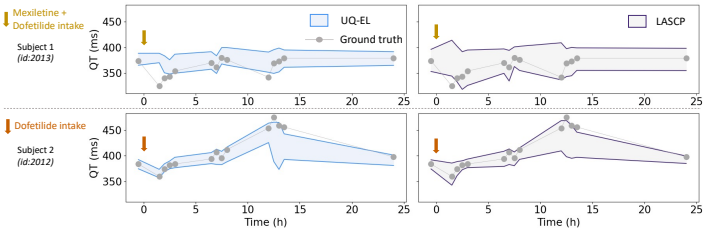

Figure 4: 24-hour QT monitoring following drug intake (90% PIs)

Results. UQ-EL and LASCP yield PIs with comparable CP, MW and MAD (cf. Table 1). As shown in Figure 2, the coverage probabilities of these 2 methods are less well-calibrated than UQ-ELM but as a trade-off lead to tighter PIs. Furthermore, the out-of-bound target QT intervals do not fall far from the PIs as reflected by the low MADs. Figure 3 shows that PI width increases more consistently with model error with the Bayesian methods than with LASCP, which makes them useful for human review of less accurate DL-based QT measurements flagged based on PI width. Figure 4 illustrates how the PIs generated with UQ-EL and LASCP provide reliable information on drug-induced QT prolongation (or lack thereof). We refer to Appendix A.4 for ECG illustrations of the effects of Dofetilide, known to significantly prolong the QT interval.

The results of UQ-ELM imply that most target QT intervals do lie somewhere within the distribution of the $k \times L$ QT samples. Better weighing of lead reliability and model confidence might provide, when necessary, single point estimates that are more accurate than mean or median predictions. To this aim, we could adapt the expectation–maximization algorithm proposed by Zhu et al. (2014) to aggregate crowd-sourced expert QT annotations. We could also replace the authors hand-crafted ECG signal quality indices with the deep learned ECG features used in LASCP.

4 Conclusion

TdP risk could be better managed with individualized, frequent and automated QT monitoring, specially amongst non-cardiologists—for instance, drug-induced QT prolongation is prevalent in psychiatry (Ali et al., 2020). In this paper, we improve the trustworthiness of DL-based QT estimators by providing high-quality PIs. In particular, approximate Bayesian inference through lead integration and implicit deep ensembling would be a simple, reliable and easy to interpret clinical decision support tool for drug safety monitoring.

REFERENCES

Zahid Ali, Mohammad Ismail, Zahid Nazar, Fahadullah Khan, Qasim Khan, and Sidra Noor. Prevalence of QTc interval prolongation and its associated risk factors among psychiatric patients: a prospective observational study. *BMC psychiatry*, 20(1):1–7, 2020.

Anastasios Angelopoulos, Stephen Bates, Jitendra Malik, and Michael I. Jordan. Uncertainty sets for image classifiers using conformal prediction. *arXiv preprint arXiv:2009.14193*, 2020.

Arsenii Ashukha, Alexander Lyzhov, Dmitry Molchanov, and Dmitry Vetrov. Pitfalls of in-domain uncertainty estimation and ensembling in deep learning. *arXiv preprint arXiv:2002.06470*, 2020.

Murat S. Ayhan and Philipp Berens. Test-time Data Augmentation for Estimation of Heteroscedastic Aleatoric Uncertainty in Deep Neural Networks. In *International conference on Medical Imaging with Deep Learning*, 2018.

James J. Clark and Alan L. Yuille. Data fusion for sensory information processing. *Kluwer Academic Publishers*, 10:978–1, 1990.

Mously D. Diaw, Stéphane Papelier, Alexandre Durand-Salmon, Jacques Felblinger, and Julien Oster. AI-Assisted QT Measurements for Highly Automated Drug Safety Studies. *IEEE Transactions on Biomedical Engineering*, 2022.

Barbara J. Drew, Robert M. Califf, Marjorie Funk, Elizabeth S. Kaufman, Mitchell W. Krucoff, Michael M. Laks, Peter W. Macfarlane, Claire Sommargren, Steven Swiryn, and George F. Van Hare. Practice standards for electrocardiographic monitoring in hospital settings: an American Heart Association scientific statement from the Councils on Cardiovascular Nursing, Clinical Cardiology, and Cardiovascular Disease in the Young: endorsed by the International Society of Computerized Electrocardiology and the American Association of Critical-Care Nurses. *Circulation*, 110(17):2721–2746, 2004.

Barbara J. Drew, Michael J. Ackerman, Marjorie Funk, W. Brian Gibler, Paul Kligfield, Venu Menon, George J. Philippides, Dan M. Roden, Wojciech Zareba, American Heart Association Acute Cardiac Care Committee of the Council on Clinical Cardiology, 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 endorsed by the American Association of Critical-Care Nurses and the International Society for Computerized Electrocardiology. *Journal of the American College of Cardiology*, 55(9):934–947, 2010.

Marc O. Ernst and Martin S. Banks. Humans integrate visual and haptic information in a statistically optimal fashion. *Nature*, 415(6870):429–433, 2002.

Christopher R. Fetsch, Alexandre Pouget, Gregory C. DeAngelis, and Dora E. Angelaki. Neural correlates of reliability-based cue weighting during multisensory integration. *Nature neuroscience*, 15(1):146–154, 2012.

Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *International Conference on Machine Learning*, pp. 1050–1059. PMLR, 2016.

John R. Giudicessi, Matthew Schram, J. Martijn Bos, Conner D. Galloway, Jacqueline B. Shreibati, Patrick W. Johnson, Rickey E. Carter, Levi W. Disrud, Robert Kleiman, Zachi I. Attia, et al. Artificial intelligence–enabled assessment of the heart rate corrected QT interval using a mobile electrocardiogram device. *Circulation*, 143(13):1274–1286, 2021.

Steven A. Hicks, Jonas L. Isaksen, Vajira Thambawita, Jonas Ghouse, Gustav Ahlberg, Allan Linneberg, Niels Grarup, Inga Strümke, Christina Ellervik, Morten Salling Olesen, et al. Explaining deep neural networks for knowledge discovery in electrocardiogram analysis. *Scientific reports*, 11(1):1–11, 2021.

Robert A. Jacobs. Optimal integration of texture and motion cues to depth. *Vision research*, 39(21): 3621–3629, 1999.

Lars Johannesen, Jose Vicente, Jay W. Mason, Carlos Sanabria, Kristin Waite-Labott, Mira Hong, Ping Guo, John Lin, Jens Stampe Sørensen, Loriano Galeotti, et al. Differentiating Drug-Induced Multichannel Block on the Electrocardiogram: Randomized Study of Dofetilide, Quinidine, Ranolazine, and Verapamil. *Clinical Pharmacology & Therapeutics*, 96(5):549–558, 2014.

Lars Johannesen, Jose Vicente, Jay W. Mason, Cassandra Erato, Carlos Sanabria, Kristin Waite-Labott, Mira Hong, John Lin, Ping Guo, Abdul Mutlib, et al. Late Sodium Current Block for Drug-Induced Long QT Syndrome: Results from a Prospective Clinical Trial. *Clinical Pharmacology & Therapeutics*, 99(2):214–223, 2016.

David C. Knill and Jeffrey A. Saunders. Do humans optimally integrate stereo and texture information for judgments of surface slant? *Vision research*, 43(24):2539–2558, 2003.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30, 2017.

Jing Lei, Max G'Sell, Alessandro Rinaldo, Ryan J. Tibshirani, and Larry Wasserman. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113 (523):1094–1111, 2018.

Robert M. Lester, Sabina Paglialunga, and Ian A. Johnson. QT assessment in early drug development: the long and the short of it. *International Journal of Molecular Sciences*, 20(6):1324, 2019.

Marek Malik. Errors and misconceptions in ECG measurement used for the detection of drug induced QT interval prolongation. *Journal of Electrocardiology*, 37:25–33, 2004.

Eugene Ndiaye. Stable conformal prediction sets. In *International Conference on Machine Learning*, pp. 16462–16479. PMLR, 2022.

Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, David Sculley, Sebastian Nowozin, Joshua Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. *Advances in neural information processing systems*, 32, 2019.

Harris Papadopoulos, Kostas Proedrou, Volodya Vovk, and Alex Gammerman. Inductive confidence machines for regression. In *European Conference on Machine Learning*, pp. 345–356. Springer, 2002.

Harris Papadopoulos, Alex Gammerman, and Volodya Vovk. Normalized nonconformity measures for regression conformal prediction. In *Proceedings of the IASTED International Conference on Artificial Intelligence and Applications (AIA 2008)*, pp. 64–69, 2008.

Harris Papadopoulos, Vladimir Vovk, and Alexander Gammerman. Regression conformal prediction with nearest neighbours. *Journal of Artificial Intelligence Research*, 40:815–840, 2011.

Tim Pearce. *Uncertainty in neural networks; bayesian ensembles, priors & prediction intervals*. PhD thesis, University of Cambridge, 2020.

Marijana Putnikovic, Zoe Jordan, Zachary Munn, Corey Borg, and Michael Ward. Use of electrocardiogram monitoring in adult patients taking high-risk QT interval prolonging medicines in clinical practice: systematic review and meta-analysis. *Drug Safety*, 45(10):1037–1048, 2022.

Mervyn Stone. The opinion pool. *The Annals of Mathematical Statistics*, pp. 1339–1342, 1961.

James E. Tisdale, Mina K. Chung, Kristen B. Campbell, Muhammad Hammadah, Jose A. Joglar, Jacinthe Leclerc, Bharath Rajagopalan, American Heart Association Clinical Pharmacology Committee of the Council on Clinical Cardiology, Council on Cardiovascular, and Stroke Nursing. Drug-induced arrhythmias: a scientific statement from the American Heart Association. *Circulation*, 142(15):e214–e233, 2020.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic learning in a random world*. Springer Science & Business Media, 2005.

Guotai Wang, Wenqi Li, Michael Aertsen, Jan Deprest, Sébastien Ourselin, and Tom Vercauteren. Aleatoric uncertainty estimation with test-time augmentation for medical image segmentation with convolutional neural networks. *Neurocomputing*, 338:34–45, 2019.

Andrew G. Wilson and Pavel Izmailov. Bayesian deep learning and a probabilistic perspective of generalization. *Advances in neural information processing systems*, 33:4697–4708, 2020.

Tingting Zhu, Alistair E. W. Johnson, Joachim Behar, and Gari D. Clifford. Crowd-sourced annotation of ECG signals using contextual information. *Annals of biomedical engineering*, 42: 871–884, 2014.

A Appendix

A.1 Probabilistic inference for multisensory integration

The brain must reason probabilistically for optimal performance as uncertainty is intrinsic to most tasks. To this aim, the human visual system performs sensory cue (or multisensory) integration over space and time by first making estimates based on each available cue or sensory signal alone before linearly combining all estimates, weighing each cue in proportion to their reliability. This phenomenon, termed weak fusion (Clark & Yuille, 1990), has been modelled using Bayesian probability but the notion of cue reliability is fairly intuitive. Take the experiment conducted by Ernst & Banks (2002) on height estimation based on visual and tactile cues. We understand that these two input signals are not equally reliable. For instance, in complete darkness, any estimate based on vision is nothing but noise. Similar ideas are expressed in the field of probabilistic opinion pooling when individual opinions are aggregated to form a consensus (Stone, 1961).

One of the main results of perception modelling is that when the uncertainty associated with each cue available to the visual system is approximated by a Gaussian likelihood function, the average estimate made by the ideal observer is a weighted average of the average estimates that would be derived from each cue alone. More formally, denote s the parameter being estimated and $\{s_1, \ldots, s_n\}$ a set of n cues. Assuming that the cues are conditionally independent, we can write from Bayes' rule that

$$p(s|s_1, \ldots, s_n) \propto \left(\prod_{i=1}^{n} p(s_i|s) \right) p(s) \tag{4}$$

Then assuming that the prior $p(s)$ is uniform, i.e. all values of s are equally probable before observation, we write

$$p(s|s_1, \ldots, s_n) \propto \prod_{i=1}^{n} p(s_i|s) \tag{5}$$

Under Gaussian assumption, reasonable in light of the central limit theorem, we can define the average (or maximum likelihood) estimate $\hat{s}$, often referred to as optimal cue integration,

$$\hat{s} = \sum_{i=1}^{n} w_i \hat{s}_i \tag{6}$$

where $w_i = r_i / \sum_{i=1}^{n} r_i$ denotes the weight of cue s_i, $r_i = 1/\sigma_i^2$ its reliability and σ_i^2 its variance.

Equation 6 is at the basis of the experiments conducted in vision research to understand how the human visual system integrates multisensory information (Jacobs, 1999; Ernst & Banks, 2002; Knill & Saunders, 2003; Fetsch et al., 2012).

A.2 Data splitting

Table 2 details the number of subjects and 12-lead ECG recordings in the subsets sampled from S1 and S2.

Table 2: Number of subjects and 12-lead ECG recordings in the datasets

Dataset	Split	Subjects	12-lead ECGs
S1a	D_{train}	22	2056
S1b	I_1	14	2014
S1b	I_2	8	1149
S2	D_{val}	22	4211

A.3 CONFORMAL RESIDUAL FITTING

We trained a gradient boosting regressor using Python's Scikit-learn library (v1.1.2). The hyperparameters were optimized by cross-validated grid-search, resulting in 500 boosting stages to perform (*n_estimators* = 500) and a learning rate of 0.01. Table 3 details the resulting RMSE scores.

Table 3: Residual fitting evaluation

Split	RMSE (ms)
I_2	6.366
D_{val}	6.008

A.4 ECG ILLUSTRATIONS

Figure 5 illustrates prediction intervals on 2 ECGs from the same subject before the administration of the QT-prolonging Dofetilide, i.e. at baseline, and 12 hours afterwards. The PI is tighter at baseline where the T wave on all ECG beats are more or less prominent and end around the same time. 12 hours later, the drug affects the ECG by lengthening the T wave and lowering its amplitude, each lead in its own way, which increases uncertainty.

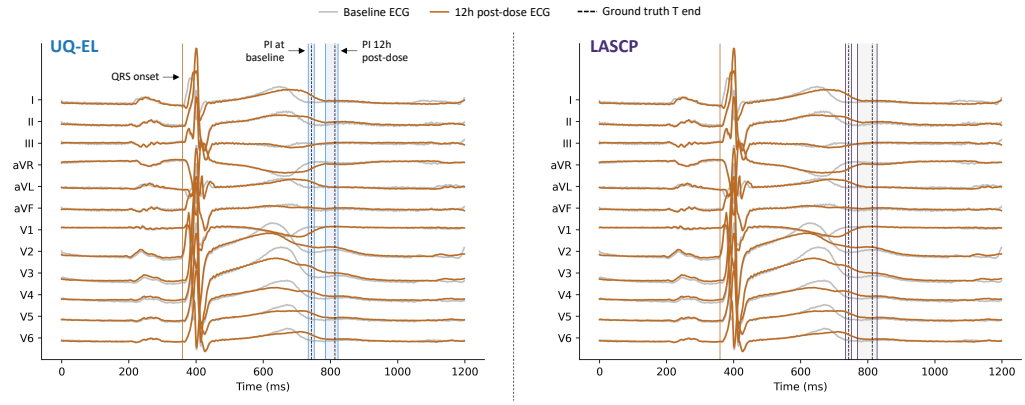

Figure 5: Examples of PIs on ECGs recorded before and 12 hours after administration of Dofetilide (drug known to prolong the QT interval) by a subject whose QT profile is illustrated in Figure 4 (id:2012). For illustration purposes, we assume that the uncertainty in the estimation of the end of the T wave is dominant compared to that of the QRS onset.