



An explainable algorithm for detecting drug-induced QT-prolongation at risk of torsades de pointes (TdP) regardless of heart rate and T-wave morphology

DOI:

[10.1016/j.combiomed.2021.104281](https://doi.org/10.1016/j.combiomed.2021.104281)

Document Version

Accepted author manuscript

[Link to publication record in Manchester Research Explorer](#)

Citation for published version (APA):

Alahmadi, A., Davies, A., Royle, J., Goodwin, L., Cresswell, K., Arain, Z., Vigo, M., & Jay, C. (2021). An explainable algorithm for detecting drug-induced QT-prolongation at risk of torsades de pointes (TdP) regardless of heart rate and T-wave morphology. *Computers in Biology and Medicine*, 131, Article 104281. <https://doi.org/10.1016/j.combiomed.2021.104281>

Published in:

Computers in Biology and Medicine

Citing this paper

Please note that where the full-text provided on Manchester Research Explorer is the Author Accepted Manuscript or Proof version this may differ from the final Published version. If citing, it is advised that you check and use the publisher's definitive version.

General rights

Copyright and moral rights for the publications made accessible in the Research Explorer are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

Takedown policy

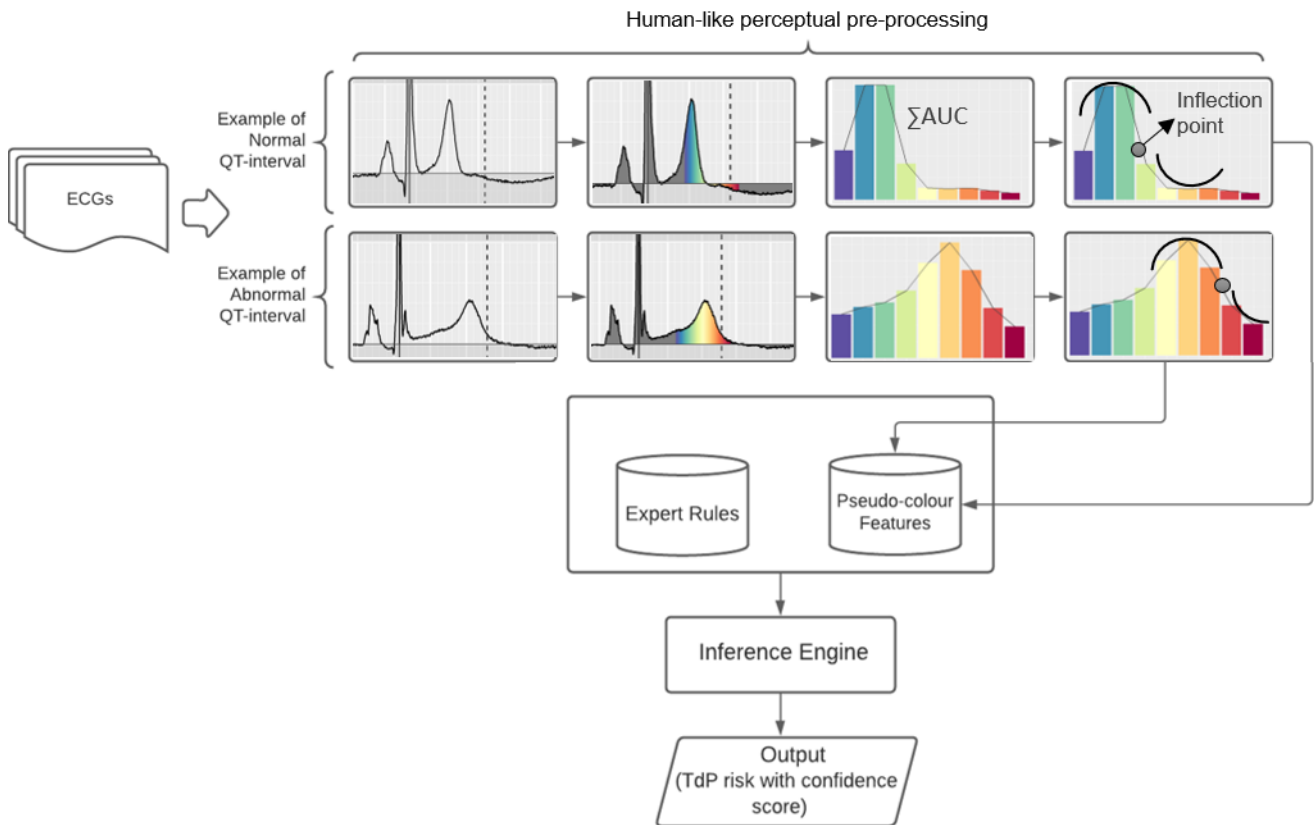
If you believe that this document breaches copyright please refer to the University of Manchester's Takedown Procedures [<http://man.ac.uk/04Y6Bo>] or contact openresearch@manchester.ac.uk providing relevant details, so we can investigate your claim.



Graphical Abstract

An explainable algorithm for detecting drug-induced QT-prolongation at risk of torsades de pointes (TdP) regardless of heart rate and T-wave morphology

Alaa Alahmadi, Alan Davies, Jennifer Royle, Leanna Goodwin, Katharine Cresswell, Zahra Arain, Markel Vigo, Caroline Jay



Highlights

An explainable algorithm for detecting drug-induced QT-prolongation at risk of torsades de pointes (TdP) regardless of heart rate and T-wave morphology

Alaa Alahmadi, Alan Davies, Jennifer Royle, Leanna Goodwin, Katharine Cresswell, Zahra Arain, Markel Vigo, Caroline Jay

- We develop an explainable algorithm to detect drug-induced QT-prolongation at risk of torsades de pointes (TdP) regardless of heart rate and robust to T-wave morphology changes using a ‘human-like’ approach, where human perception of the ECG signal is used to determine features, and diagnostic rules are determined according to the clinical literature.
- Testing the algorithm on a large number of ECGs ($n = 5050$) with variable QT-intervals at varying heart rates yields a Matthews correlation coefficient (MCC), balanced accuracy, sensitivity and specificity of 0.88, 0.97, 0.94, 0.99 respectively.
- We explore whether we can improve the algorithm by automating the rule-generation with a decision tree. We find that we can, but the rule-based decision tree algorithm overfits and generalises less well than the ‘expert’ evidence-based rule algorithm, and that whilst the decision tree rule algorithm is technically explainable, it is not clinically explainable.
- A focus group evaluation confirms the explainability of the algorithm is important, supporting an ‘expert’ rule-based approach, rather than a fully automated approach.
- Defining features according to human interpretation heuristics provides a promising route to automating ECG interpretation, and has the benefit of being explainable to patients and clinicians. Automating rule generation based on these features has potential, but here it is not sufficient to produce a reliable, trustable algorithm. A human-in-the-loop approach, where machine learning is a tool used to surface potential rules, but these are validated empirically before being implemented, may be the best approach to use in clinical practice.

An explainable algorithm for detecting drug-induced QT-prolongation at risk of torsades de pointes (TdP) regardless of heart rate and T-wave morphology

Alaa Alahmadi^{a,*}, Alan Davies^b, Jennifer Royle^c, Leanna Goodwin^c, Katharine Cresswell^d, Zahra Arain^d, Markel Vigo^a and Caroline Jay^{a,*}

^aDepartment of Computer Science, The University of Manchester, Manchester, UK

^bDivision of Informatics, Imaging and Data Sciences, School of Health Sciences, The University of Manchester, Manchester, UK

^cDigital Experimental Cancer Medicine Team, CRUK Manchester Institute, and The Christie NHS Foundation, Manchester, UK

^dCancer Precision Medicine and Cancer Prevention and Early Detection, NIHR Manchester Biomedical Research Centre (BRC), Manchester University NHS Foundation Trust, Manchester, UK

ARTICLE INFO

Keywords:

Drug-induced LQTS
QT-prolongation
Torsades de pointes
TdP
Automated ECG interpretation
Human-like algorithm
Rule-based algorithm
Visual perception
Machine perception
Explainable AI

ABSTRACT

Torsade de pointes (TdP), a life-threatening arrhythmia that can increase the risk of sudden cardiac death, is associated with drug-induced QT-interval prolongation on the electrocardiogram (ECG). While many modern ECG machines provide automated measurements of the QT-interval, these automated QT values are usually correct only for a noise-free normal sinus rhythm, in which the T-wave morphology is well defined. As QT-prolonging drugs often affect the morphology of the T-wave, automated QT measurements taken under these circumstances are easily invalidated. An additional challenge is that the QT-value at risk of TdP varies with heart rate, with the slower the heart rate, the greater the risk of TdP. This paper presents an explainable algorithm that uses an understanding of human visual perception and expert ECG interpretation to automate the detection of QT-prolongation at risk of TdP regardless of heart rate and T-wave morphology. It was tested on a large number of ECGs ($n = 5050$) with variable QT-intervals at varying heart rates, acquired from a clinical trial that assessed the effect of four known QT-prolonging drugs versus placebo on healthy subjects. The algorithm yielded a balanced accuracy of 0.97, sensitivity of 0.94, specificity of 0.99, F1-score of 0.88, ROC (AUC) of 0.98, precision-recall (AUC) of 0.88, and Matthews correlation coefficient (MCC) of 0.88. The results indicate that a prolonged ventricular repolarisation area can be a significant risk predictor of TdP, and detection of this is potentially easier and more reliable to automate than measuring the QT-interval distance directly. The proposed algorithm can be visualised using pseudo-colour on the ECG trace, thus intuitively ‘explaining’ how its decision was made, which results of a focus group show may help people to self-monitor QT-prolongation, as well as ensuring clinicians can validate its results.

1. Introduction

This paper provides a new perspective on the challenging topic of electrocardiogram (ECG) interpretation. ECGs, which represent the electrical activity of the human heart, are powerful and widely used diagnostic instruments. The data they contain is rich and extremely complex, however, and learning how to accurately interpret them can take many years. Automated systems can identify a normal sinus rhythm with reasonable accuracy, but are much poorer at reliably detecting abnormalities [22, 62, 67]. A clinically significant cardiac abnormality that automated methods have been shown to be particularly unreliable at detecting is QT-interval prolongation [22, 26, 48, 62, 73, 75]. This can be congenital, or acquired, resulting from the clinical administration of certain pharmacological drugs, and

is associated with a life-threatening arrhythmia known as Torsade de Pointes (TdP) [12, 85].

In this work, we combine knowledge of human perception with clinical expertise to develop an automated algorithm that can reliably detect patients with QT-prolongation at risk of Torsade de Pointes (TdP). We take as our starting point a visualisation technique that displays the ECG signal such that a lay person can detect QT-interval prolongation quickly and accurately. The information encoded in the visualisation is mapped to a set of features, which form the basis of two interpretation algorithms: an ‘expert’ set of rules, which are formed according to clinical practice; and a decision tree, which automates the generation of rules from the same set of features. Whilst the decision tree appears more accurate when trained and tested with cross-validation, the expert algorithm is more accurate when tested on unseen data.

2. Background and significance

2.1. Identifying patients at risk of TdP

Torsades de pointes (TdP), or ‘twisting of the points’, was a term first used in 1966 by Francois Dessertenne to de-

*Corresponding author.

✉ alaa.alahmadi@postgrad.manchester.ac.uk (A. Alahmadi); alan.davies-2@manchester.ac.uk (A. Davies); jenny.royle@digitalecmt.org (J. Royle); leanna.goodwin@digitalecmt.org (L. Goodwin); katharine.cresswell@mft.nhs.uk (K. Cresswell); zahra.arain@mft.nhs.uk (Z. Arain); markel.vigo@manchester.ac.uk (M. Vigo); caroline.jay@manchester.ac.uk (C. Jay)

scribe a form of polymorphic ventricular tachycardia, in which the continuously changing polarity and amplitude of the QRS complexes appear to twist around the isoelectric line of the electrocardiogram (ECG) [18]. It is a potentially lethal arrhythmia that can degenerate into ventricular fibrillation, the leading cause of sudden cardiac death in young individuals with structurally normal hearts [21, 58, 82, 85]. It is often precipitated by triggers such as emotional stress or exercise, especially swimming [9, 78]. Since the time of its original description by Dessertenne, it has been well-established that TdP is frequently associated with QT-interval prolongation on the ECG, and is caused by a cardiac ion channelopathy known as ‘long QT syndrome’ (LQTS)[18, 21, 82, 85].

Whilst TdP can result from both congenital and acquired long QT syndrome, acquired LQTS caused by pharmacological drugs is by far the most common cause of TdP [12, 21, 82, 85]. A steadily increasing number of medications have been reported to cause drug-induced QT-prolongation, TdP and sudden cardiac death [12, 21, 46]. This, in turn, has troubled clinicians, the pharmaceutical industry and regulatory authorities, particularly because many of these QT-prolonging drugs, including antihistamines, psychotropics, antibiotics and antiarrhythmic agents, are widely prescribed, often for self-limited diseases [82, 85].

The QT-interval represents the duration of the ventricular depolarisation and repolarisation cycle; it is measured on the ECG from the beginning of the QRS complex (reflecting ventricular depolarisation) to the end of the T-wave (representing subsequent repolarisation) [30]. In LQTS, a delay occurs in ventricular repolarisation, which increases the risk of premature ventricular contraction (PVC) occurring during the relative refractory period of repolarisation, reflected on the ECG by a unique pattern known as the R-on-T phenomenon that initiates TdP [6, 85] (see Figure 1). Dividing the ventricular repolarisation period into early and late repolarisation, shown on the ECG as the $J-T^{peak}$ and $T^{peak}-T^{end}$ intervals respectively (Figure 1), is recommended for assessing the risk of drug-induced TdP; many TdP episodes are reported following a prolonged $T^{peak}-T^{end}$ interval in particular [7, 15, 40, 42].

Measuring the QT-interval, however, is known to be a challenging task [12, 25, 56, 57], making drug-induced QT-prolongation detection difficult, even for clinicians who routinely read ECGs [70, 77]. Whilst it is relatively easy to determine the beginning of the QRS complex (or RS complex if there is no Q-wave), identifying the end of the T-wave is recognized as being the most difficult aspect of measuring the QT-interval [12, 25, 30, 56]. Numerous methods have been proposed [25, 30, 38, 45, 53, 54, 59, 60, 66], among which the tangent method is the most popular [60], but all techniques have been shown to be associated with potential inaccuracies [12, 56]. A major challenge lies in the fact that the morphology (shape) of the T-wave itself can be very variable, and QT-prolonging drugs may cause abnormal changes in it [12, 45, 76, 56]. Variability in T-

wave amplitude [52], a prolonged, flat T-wave, and fused T-U waves (also known as T-U complexes) can easily invalidate QT-interval measurement [12, 37, 49, 45]. This issue complicates both manual and automated QT-interval measurement, and the best method to use remains a subject of debate [12, 43, 67]. At present, despite numerous efforts [35, 36, 43, 53, 54, 59], no automated QT-interval measurement method is considered reliable enough to be used in the clinical assessment of the cardiac safety of a QT-prolonging drug [12, 22, 26, 48, 55, 62, 67, 73, 75]. Automated methods in the literature fall into two categories according to how the end of the T-wave is defined: threshold-based methods and slope-based methods [12, 53, 54, 59]. Threshold-based methods identify the end of the T-wave based on threshold levels, defined as a fraction ranging from 5% to 15% of the amplitude or differential of the T-wave [12, 36]. In slope-based methods, the end of the T-wave is defined as the interception of the maximum T-wave slope with the isoelectric line [12, 36]. As the steepest part of the descending portion of the T-wave is affected by T-wave morphology, different methods have been proposed to identify the maximum T-wave slope, including the tangent method, peak slope method, and least-square fitting method [12, 36, 53, 54, 79]. Generally, slope-based methods provide a better estimation of the QT-interval than threshold-based methods, as the results of the latter vary according to the threshold level and the T-wave morphology [12, 35, 83]. However, the accuracy of slope-based methods can be significantly reduced when the T-wave has a slow-moving deflection that fails to return to the ECG baseline, making the steepest slope hard to define [12]. In addition, these methods depend on an accurate identification of the isoelectric line, which is known to be influenced by noise and ECG baseline wander (a type of artefact) [12, 67]. Bizarre automated QT measurements are also sometimes reported, even for normal, noise-free ECGs [12].

Assessing the risk of drug-induced TdP also benefits from consideration of the ventricular repolarisation morphology more generally [31, 74]. Augmented U-waves, or partial/complete T-U wave fusion, where the T and U waves cannot be distinguished, are significant risk predictors of TdP [47, 49]. In most cases, this is shown to be, in fact, a prolonged biphasic, or notched T-wave [84, 76]. For certain T-U morphologies commonly reported in drug-induced long QT syndrome, measuring the QT-interval as a simple time period may lack diagnostic validity, and the presence of prolonged fused T-U waves should be viewed as a more accurate predictor of TdP [12, 31].

2.2. A human-like approach to automated ECG interpretation

We have shown in previous work that superimposing pseudo-colouring, a technique that represents continuously varying values using a sequence of colours [80], on the ECG significantly improves people’s ability to detect QT-prolongation at risk of TdP, regardless of heart rate [4]. Our technique draws from the field of pre-attentive

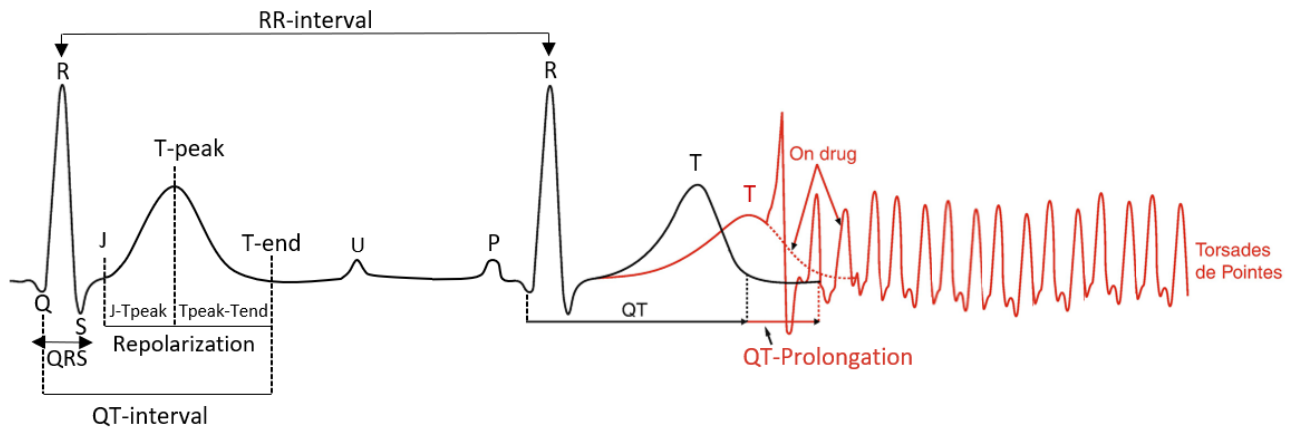


Figure 1: An illustration of drug-induced QT-interval prolongation and the R-on-T phenomenon that initiates torsades de pointes (TdP) arrhythmia on the electrocardiogram (ECG).

processing theory in human vision, which outlines a set of visual properties including colour that can be detected rapidly and accurately by the human eye [32, 33, 81]. This form of data presentation helps to overcome the problem of identifying the end of the T-wave, directing the observer instead to inspect repolarisation morphology. The technique was designed to support ‘intuitive’ visual perception of drug-induced LQTS, both for patients on a QT-prolonging drug, and clinicians with less training in ECG interpretation. The technique colours the area under the curve of the ECG signal within the ventricular repolarisation period relative to the R-peak, with a gradient of cool colours (purple to blue to green) indicating normal QT-interval ranges, and warm colours (yellow to orange to red) showing abnormal QT-interval ranges. As the ventricular repolarisation period is represented by the T-wave on the ECG [12], the pseudo-colour highlights its position in relation to the inter-heartbeat time dimension, without needing to identify either the peak or end of the T-wave. The pseudo-colouring was adjusted for heart rate using the ‘QT nomogram’, which is a risk assessment method designed specifically for identifying patients at risk of drug-induced TdP according to heart rate [13, 24]. Figure 2 shows examples of ECGs with pseudo-colouring that have different heart rates, but similar levels of TdP risk according to QT-interval; the dashed lines represent the QT nomogram line value at risk of TdP for a given heart rate. More details about the pseudo-colouring technique and its evaluation can be found in [3, 4].

Here, we exploit an understanding of how humans use pseudo-colouring to determine QT-interval length to inform a new approach to the automated detection of QT-prolongation. We term this approach *human-like*, as the knowledge representation and reasoning processes used within the algorithm are inspired by the way in which humans interpret ECGs.

Long QT syndrome refers to a prolongation of the ventricular repolarisation period, represented by the T-wave

on the ECG [12]. Although the interval of interest starts at the Q-wave, it is the prolongation of the T-wave in particular that is the ECG marker for LQTS, rather than the QRS complex that represents the ventricular depolarisation period [12]. As the T-wave generally has the largest area under the curve (AUC) of the ECG signal, the pseudo-colouring provides a way to approximate QT prolongation, alleviating the need to measure the QT-interval distance precisely.

Although using AUC to assess the risk of drug-induced LQTS has yet to appear in the automated ECG interpretation literature, it has been theoretically suggested by Bonate *et al.* (1999), who proposed a univariate summary measure that calculated the total area under the QT interval for both the ECG baseline and the post-dose ECG [10]. However, this method still required identification of the beginning of the Q-wave and the end of the T-wave to locate the QT-interval for comparison, and lacked clarity on whether both positive and negative values from the ECG isoelectric line be used in the calculation, or only positive values.

Our proposed algorithm avoids the necessity of locating the start and end of the QT-interval by considering the AUC in reference to the easily detectable R-peak, as described in Section 3.2.1. In a pilot study, we took the ECG signal with superimposed pseudo-colour as a starting point, and calculated the AUC for each colour in the sequence using the trapezoidal rule [2]. The QT-interval was considered ‘prolonged’ if the warm pseudo-colours (yellow to orange to red) occupied more than 50% of the AUC of the ECG signal, i.e. when the T-wave contained more warm than cool colours; otherwise it was considered ‘normal’. When tested on a clinical dataset of 40 ECGs, the algorithm was more accurate in detecting QT-prolongation at risk of TdP than current signal processing techniques, including threshold-based and slope-based QT-interval measurement methods.

However, this pilot version of the algorithm has two significant limitations. Firstly, as it relies on the ratio of warm to cool pseudo-colour under the T-wave, it

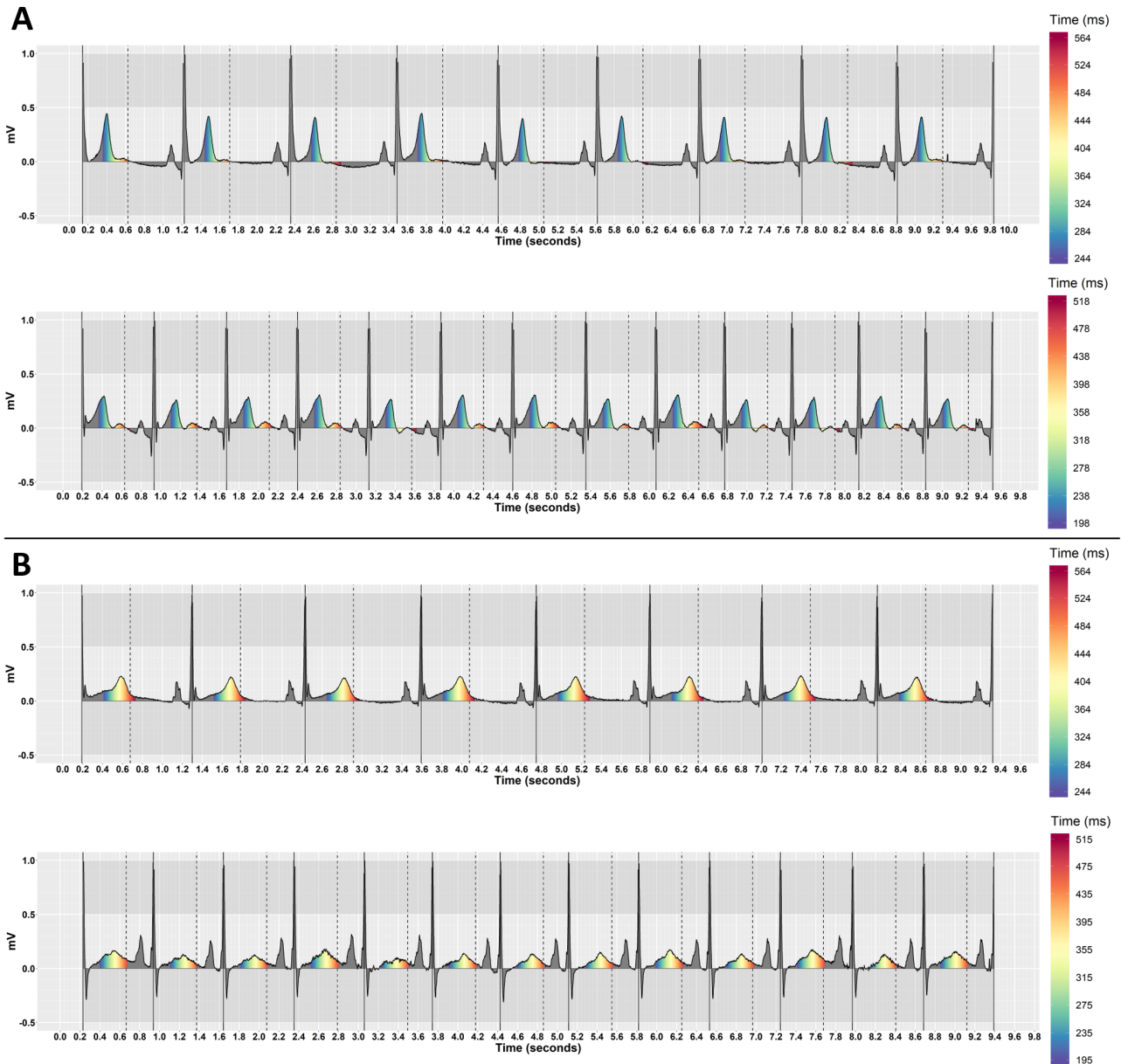


Figure 2: Pseudo-coloured ECGs that have different heart rates, but similar QT-interval TdP risk levels. The dashed lines represent the QT nomogram line value at risk of TdP. (A) QT-intervals are below the nomogram line by 120ms for both ECGs (no risk of TdP). The top stimulus has a low heart rate (HR = 55, QT = 361) and the bottom has a high heart rate (HR = 83, QT = 329). (B) QT-intervals are above the nomogram line in both ECGs (risk of TdP). The top stimulus has a low heart rate (HR = 52, QT = 579) and the bottom has a high heart rate (HR = 85, QT = 470).

may inaccurately classify the QT-interval as normal where drugs prolong both early and late ventricular repolarisation, shown on the ECG as prolonged $J-T^{peak}$ and $T^{peak}-T^{end}$ intervals, or when a patient has an electrolyte abnormality such as hypercalcemia alongside QT-prolongation. In these cases, the width of the T-wave may increase, and thus contain proportionally more cool colours than warm. Figure 3 shows an example of this issue, where the pseudo-coloured ECG has wide T-waves with clinically prolonged $J-T^{peak}$ and $T^{peak}-T^{end}$ intervals, and the percentage of warm colours area is therefore only 28.67%—considerably

below the 50% threshold required for a ‘prolonged’ result from the algorithm. Secondly, calculating the AUC alone, without inspecting the shape of the curve, makes it difficult to distinguish between T and U waves, particularly where there are significant T-wave morphology changes.

In this paper, we combine the information about QT-interval length provided by the pseudo-colour with additional knowledge about T-wave morphology to generate a set of visual features that form the basis of an automated algorithm that can determine risk of TdP across a wide range of cases. We empirically compare two approaches to build-

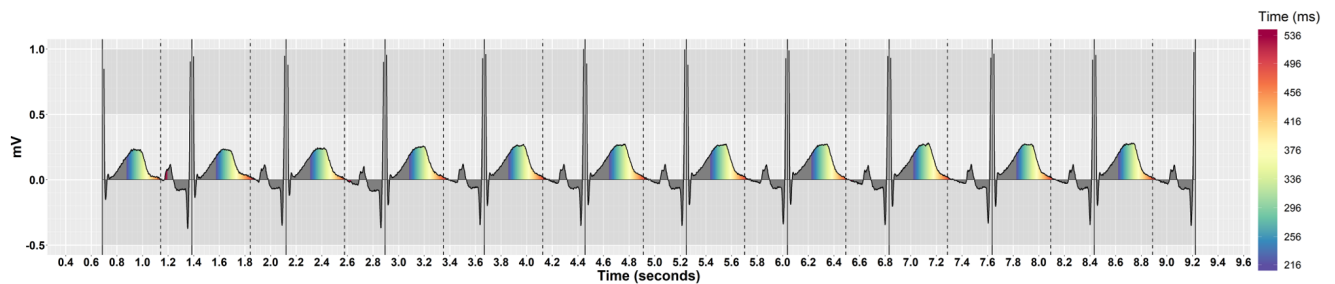


Figure 3: A pseudo-coloured ECG that has wide T-waves with prolonged $J-T^{peak}$ and $T^{peak}-T^{end}$ intervals but only 28.67% warm colours. The dashed lines represent the QT nomogram line value at risk of TdP.

ing the algorithm: a manually-curated ‘expert’ algorithm that incorporates knowledge from the clinical literature; and a decision tree, which automates the generation of the rules from the data. In addition, we report the results of two focus groups—one consisting of patients, the other of clinicians—which explored the relevance of our approach to clinical practice.

3. Materials and methods

In this section, we provide a detailed description of how the algorithm was developed in 3.2 and explain the evaluation process in 3.3. The source code for the algorithm can be found in [5].

3.1. ECG data acquisition

The ECG datasets ($n = 5050$) were acquired from a clinical trial that assessed the effect of four known QT-prolonging drugs versus placebo on healthy subjects [40]. The 10-second lead-II recording was selected from each 12-lead ECG, as this is typically used to measure the QT-interval [12]. The heart rates of the ECGs ranged from 40 to 96 beats per minute (bpm), and the QT-interval values ranged from 300 to 579 ms. The ECGs were from 22 subjects who received a single dose of a pure hERG potassium channel blocker (‘Dofetilide’), and three drugs that block hERG and either calcium or late sodium currents (‘Quinidine’, ‘Ranolazine’, and ‘Verapamil’), during a placebo-controlled cross-over trial.

As part of the clinical study methodology, QT-intervals and heart rates were calculated for all ECGs, and it is these QT/HR values that were used as ground truth for our subsequent evaluation of the algorithm. According to the QT/HR pair plots of all ECGs on the nomogram [13], 180 ECGs were on or above the nomogram line, showing risk of TdP, while the other ECGs ($n = 4870$) were below the nomogram line, as shown in Figure 4. The open ECG datasets are available online from the PhysioNet database [29].

3.2. Algorithm development

3.2.1. Pseudo-colouring application

The R-peaks in the ECG signal were detected using an automated math function that finds the greatest peaks (maxima), according to regular pattern in the signal [11]. The

average RR-interval and heart rate were calculated. The TdP risk threshold was calculated for each heart rate using the nomogram line [13].

In clinical practice, the QT-interval is measured by counting the small squares (each representing 40 ms) on the standard ECG background grid from the beginning of the Q-wave to the end of the T-wave [12]. An approximate time for the ventricular depolarisation and repolarisation cycle was calculated for each heartbeat from the R-peak minus 20 ms (which served as a proxy for the beginning of the Q-wave) to the maximum potential QT-prolongation at risk of TdP, which was estimated as the QT-nomogram line value at risk of TdP plus two small squares (80 ms). This formed an additional inter-heartbeat time dimension, to which the pseudo-colour could be mapped.

As the time period of interest is the duration of ventricular repolarisation, the pseudo-colouring sequence was applied to the area between the isoelectric line (where amplitude is zero) and the ECG signal, starting at the time of the QT-nomogram line value at risk of TdP plus two small squares (80 ms) to six small squares (240 ms) below the nomogram line. The time of the QT-nomogram line value at risk of TdP was mapped to dark orange, and values 40 ms and 80 ms above the nomogram line were mapped to red and dark red respectively, showing the higher risk of TdP. Time values below the nomogram line were mapped to progressively cooler colours, showing no risk of TdP. This resulted in nine indices on the pseudo-colouring scale, where each index was mapped to a colour code and represented a small square on the ECG. Figure 5 illustrates how the pseudo-colouring technique was applied according to the QT nomogram line and the standard ECG background grid.

3.2.2. Calculating the pseudo-coloured area under the curve (AUC) of the ECG signal

The pseudo-coloured area under the curve (AUC) of the ECG signal was calculated using the trapezoidal rule. In mathematics, and more specifically in integral calculus, the trapezoidal rule is a common method for approximating the area under the curve of a linear function, which works by dividing it into a number of intervals of equal width [86]. That is, given a linear function $f(x)$ of a real variable x and an interval $[a, b]$, the rule estimates the area under the

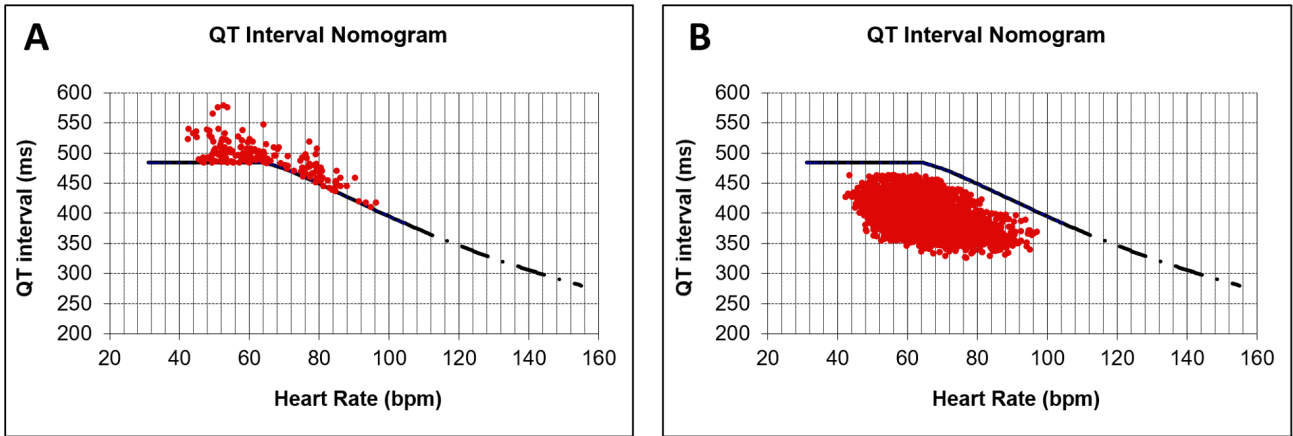


Figure 4: Assessment of the TdP risk for all ECGs ($n = 5050$) using the QT-nomogram. According to the acquired QT/HR pair values from the clinical trial study: (A) A total of 180 ECGs were on or above the nomogram line, showing risk of TdP. (B) The other ECGs ($n = 4870$) were below the nomogram line, showing no risk of TdP.

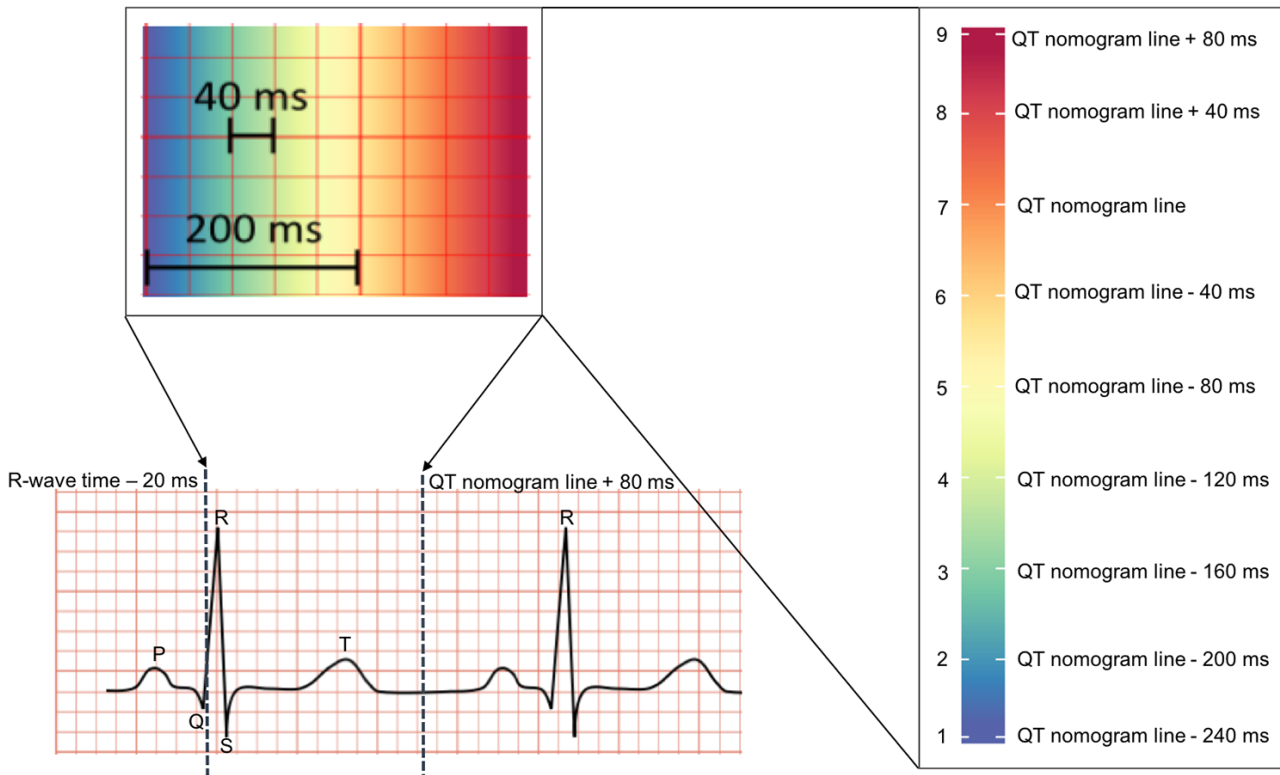


Figure 5: An illustration of how the pseudo-colouring technique was applied according to the QT nomogram line and the standard ECG background grid. A small square on the grid is equal to 40 ms.

graph of the function $f(x)$ as a trapezoid, calculating its area as follows:

$$\int_a^b f(x)dx \approx \frac{(f(a) + f(b))}{2} (b - a) \quad (1)$$

The time dimension of the raw ECG signal is represented by integer numbers ranging from 1 to 10000 mil-

liseconds for a 10-second ECG recording. To maximise the precision of the calculation, we considered $[a, b]$ to be the interval of two successive timestamps in the ECG signal, where the difference between them is equal to one millisecond. If we consider the ECG time interval of x_1 and x_2 , the trapezoidal rule was applied by taking the average amplitude of the ECG signal on the Y-axis of this interval as $f(x_1)$ and $f(x_2)$, and multiplying it by the difference in time between x_1 and x_2 , which is always equal to one. The

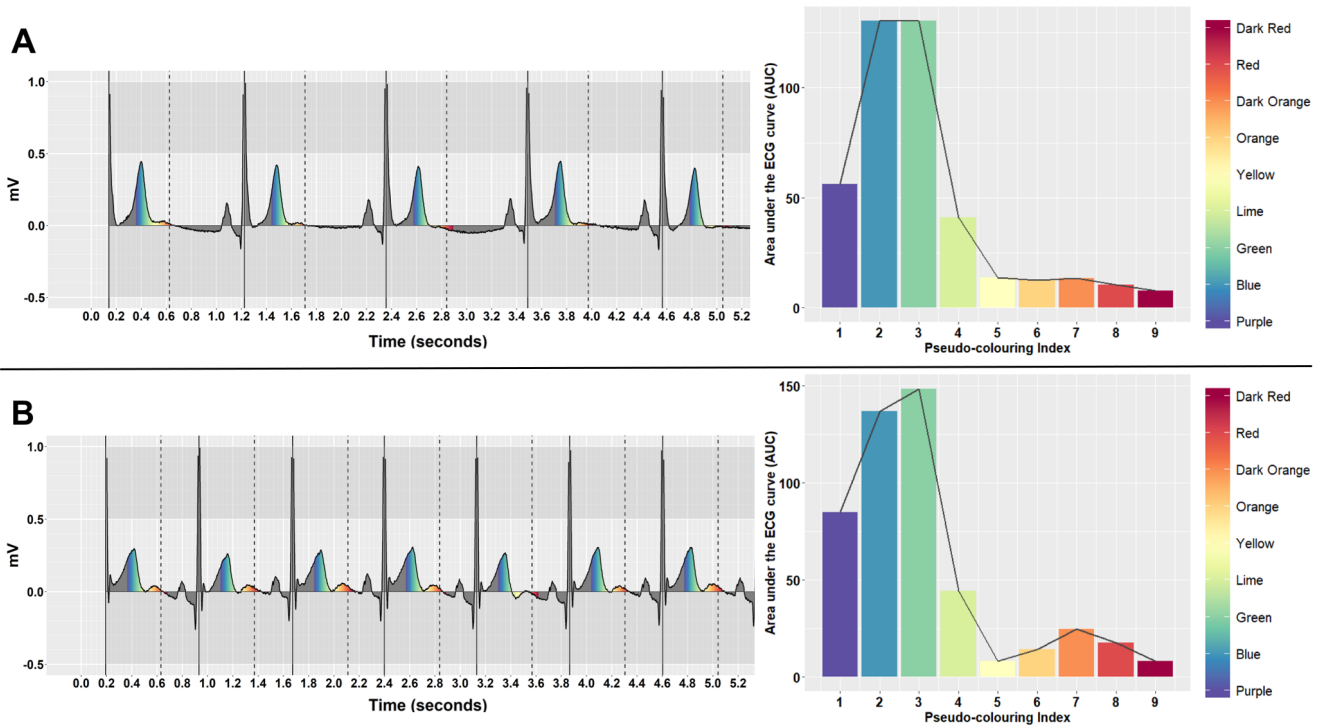


Figure 6: Pseudo-coloured ECGs with normal QT-intervals (no risk of TdP), and their global representation graph, showing the calculated $\sum AUC$ for each pseudo-colour across the whole ECG signal. (A) The ECG has a low heart rate (HR = 55, QT = 361), and displays a greater area of cool pseudo-colours. It also has a negative small red area below the ECG baseline, which was included in the AUC calculation. (B) The ECG has a high heart rate (HR = 83, QT = 329), with a cool pseudo-coloured T-wave and warm pseudo-coloured U-wave.

trapezoidal rule was thus applied to each one millisecond subinterval, and the results were then summed for every 40 ms colour index on the pseudo-colouring scale to produce a total AUC value for each small square on the standard ECG background grid. The negative area below the ECG baseline was treated in the same way as the positive area, as the baseline is not usually fixed, and in some abnormal cases the T-wave can be negative in Lead-II, a condition known as ‘T wave inversion’. As such, the absolute AUC was calculated.

As a spectrum-approximation pseudo-colouring sequence was used, the indices from 1 to 4 (purple, blue, green, lime) were considered to be cool spectral colours, while the indices from 5 to 9 (yellow, orange, dark orange, red, dark red) were considered warm spectral colours. By plotting the calculated AUC as a function of pseudo-colouring sequence, we obtain a global representation of the pseudo-coloured repolarisation morphology across the whole 10-second ECG signal. Figures 6 and 7 show examples of pseudo-coloured ECGs with normal (no risk of TdP) and prolonged (at risk of TdP) QT-intervals, and their global representation plot, showing an approximation of the total pseudo-coloured AUC.

3.2.3. Locating intervals of concavity and inflection points

As T-wave morphology can change substantially and other ECG waves may present during ventricular repolarisation, including an augmented U-wave, a simple comparison of the relative amounts of cool to warm pseudo-colour in the AUC may not be sufficient for recognising a prolonged QT-interval. We therefore also consider the T-wave morphology by inspecting the shape of the ECG signal curve.

In differential calculus, the second derivative $f''(x)$ of a linear function $f(x)$ is the derivative of the derivative of $f(x)$, i.e. the derivative of the first derivative $f'(x)$. While the first derivative measures the rate of change of a quantity, e.g. the slope or gradient of a line that is tangential to the $f(x)$ curve at a particular point, the second derivative measures how the rate of change of a quantity is itself changing. The tangent method commonly used to identify the end of the T-wave is equal to the first derivative of the ECG signal function at the T-peak point; the second derivative will measure the concavity of the T-wave at the same T-peak point, showing the rate at which the T-wave signal is increasing or decreasing over time. This provides a new approach to estimating the end of the T-wave, which builds on the widely utilised tangent method [60].

The second derivative measures the concavity of the graph of $f(x)$, and identifies inflection points at which the shape of the function changes from concave down to con-

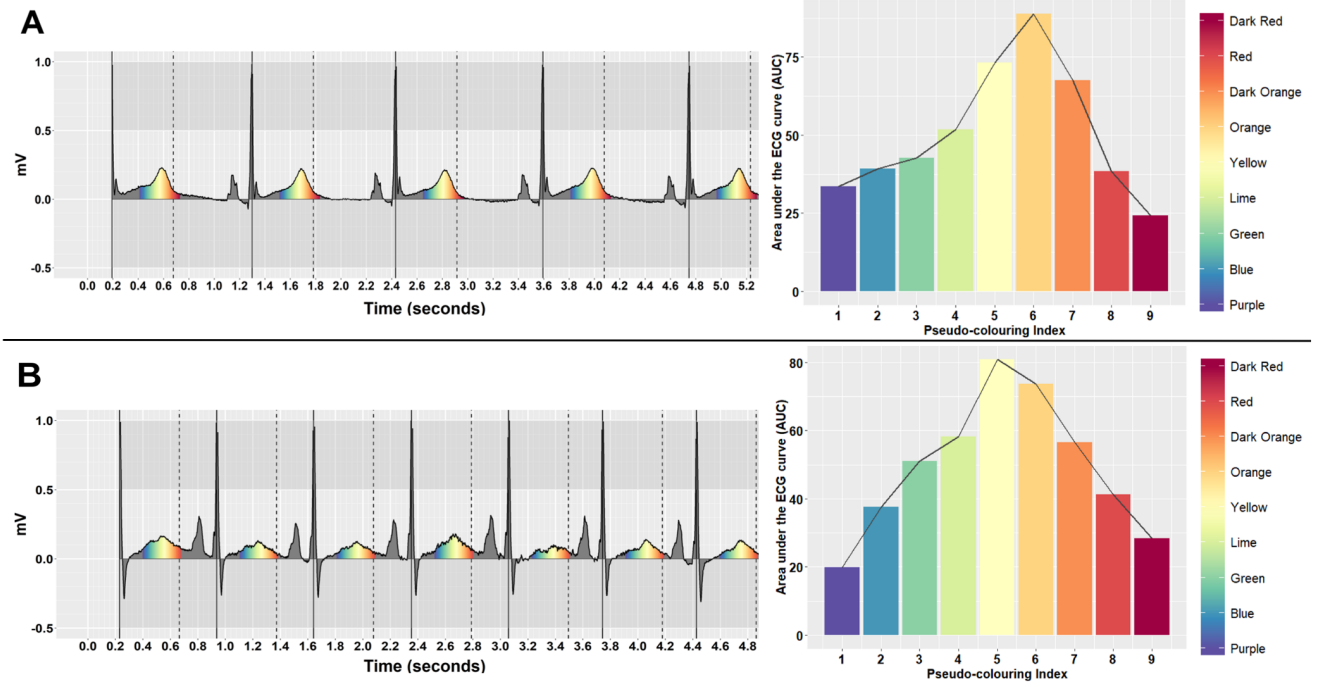


Figure 7: Pseudo-coloured ECGs with prolonged QT-intervals (at risk of TdP), and their global representation graph, showing the calculated $\sum AUC$ for each pseudo-colour across the whole ECG signal. (A) The ECG has a low heart rate (HR = 52, QT = 579), and displays a greater area of warm pseudo-colours. (B) The ECG has a high heart rate (HR = 85, QT = 470), with a notched T-wave that contains a greater area of warm pseudo-colours.

cave up (or vice versa). If the second derivative of a function is positive, then the shape of this function is concave up (convex), and if it is negative, then its shape is concave down (concave).

We thus determined the second derivative of the graph representing the global pseudo-coloured AUC, to estimate the location of the maximum concavity downward within the ECG signal and its subsequent inflection point, at which the shape of the signal changes to convex. This helps with identifying the location of the maximum rate of change of the maximum T-wave down slope on the pseudo-colouring scale, and the subsequent pseudo-colour index at which the T-wave signal starts to rise again, indicating its end.

When using Leibniz's notation for derivatives, the first and second derivatives of the graph representing the global pseudo-coloured AUC, considering the AUC as a dependent variable y with respect to an independent variable x representing the pseudo-colouring index, are written as follows:

$$f'(x) = \frac{dy}{dx} \text{ and } f''(x) = \frac{d^2y}{dx^2} \quad (2)$$

The difference between any two successive indices dx on the pseudo-colouring scale is always equal to one. That is, the first derivative is measured by calculating the difference in the AUC between each two successive indices on the pseudo-colouring scale. This produced 8 derivatives. The second derivative was calculated by taking the deriva-

tive of these 8 derivatives, which produced further 7 derivatives. Then, the maximum concave down and its subsequent inflection point (if any) were identified with respect to the pseudo-colouring index. Figure 8 shows an illustration of this process.

As the QT nomogram line represents the QT value at risk of TdP, only the first five second derivatives were used as predictors of at risk QT-prolongation, shown by the grey blocks in Figure 8. In addition, the inflection point index, defined as the index between any two pseudo-colour indices at which the maximum concave down changes to convex, was also used as a predictor. As such, six predictors were calculated for each ECG.

3.2.4. Generating the expert rules

A rule-based algorithm was developed to classify the QT-interval as 'normal' (no TdP risk) or 'abnormal' (at risk of TdP), and rate the confidence of the classification on a 6-point scale of 'very likely normal' (1), 'probably normal' (2), 'possibly normal' (3), 'possibly abnormal' (4), 'probably abnormal' (5), and 'very likely abnormal' (6).

As the inflection point is a potential proxy for the end of the T-wave, we hypothesised that the probability of QT-prolongation at risk of TdP increases as the inflection point index increases. If the inflection point index is located within the cool pseudo-colours, i.e., an index of 3.5 or less, this indicates a normal QT-interval (no TdP risk). Conversely, if the inflection point index is located within the warm pseudo-colours, i.e., an index 5.5 or more, this indi-

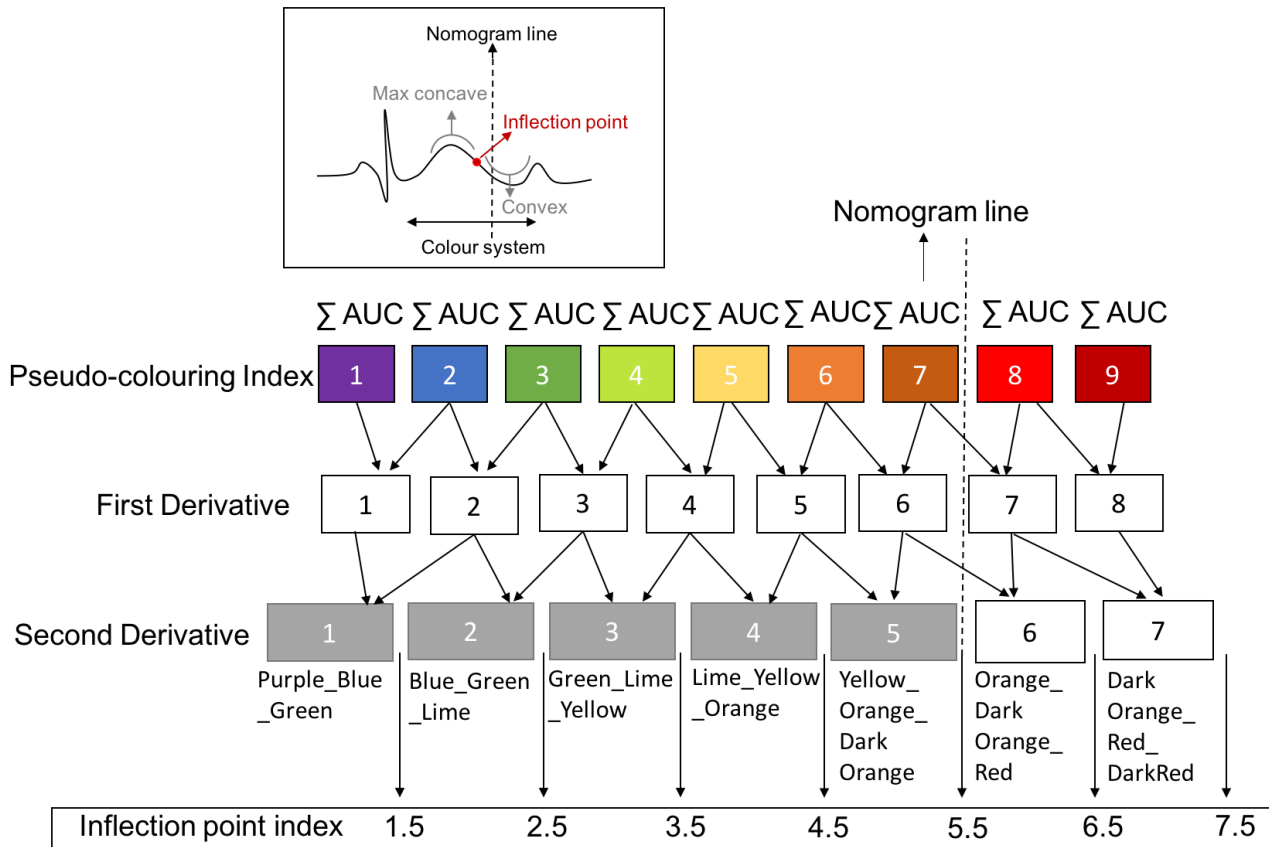


Figure 8: An illustration of how the second derivative of the pseudo-coloured AUC was measured and how the inflection point index was determined.

icates an abnormal QT-interval at risk of TdP.

An inflection point index of 4.5 is considered borderline, as the maximum concave down, representing the T-wave, is located between the cool and warm pseudo-colours of green, lime, yellow and orange. As the QT-interval may be very close to or on the nomogram line at this point, it is difficult to determine the risk of TdP precisely using the inflection point alone. In this situation, we thus also used the pseudo colouring index of the maximum concave down. If most of the maximum concave down colours are cool pseudo-colours, i.e. 50% or more of the concave is located within green, lime and yellow, then the QT-interval is not considered at risk of TdP; if 50% or more of the concave is located within lime, yellow, and orange, it is considered at risk. The expert algorithm's inference process is shown in Figure 9, and Figure 10 shows a flowchart of the expert rules. The confidence rating is also determined based on the inflection point index and the maximum concave location on the pseudo-colouring index. As the location of the maximum concavity on the pseudo-colouring scale increases, the likelihood of QT-prolongation at risk of TdP also increases, as illustrated in Figure 10.

The inflection point index can also help estimate the range in which the numerical value of the QT-interval is likely to be. This can support an ECG reader measuring the QT-interval, by providing an estimated 40 ms time win-

dow (representing a small square on the ECG grid) at which the T-wave probably ended. Figure 8 shows how the second derivative was calculated by examining the shape of the ECG curve within three successive pseudo-colours, and then moving one pseudo-colour at a time (i.e. inspecting the shape of the ECG curve again within the two backward pseudo-colours and one forward pseudo-colour). The inflection point index is located between the maximum concave down (corresponding to a negative second derivative that decreases within the three successive pseudo-colours) and the subsequent convex, which has a positive second derivative that is increasing within the previous two pseudo-colours and one subsequent pseudo-colour, representing an increasing curve, at which point the T-wave probably ended. If the maximum concave down in Figure 8 was located in the first grey block of the second derivative (i.e. within the purple, blue and green pseudo-colours), and the inflection point was 1.5, the T-wave was decreasing within the blue and green pseudo-colours, but increasing within the lime pseudo-colour. As each colour represents a 40 ms time scale, then the QT range can be estimated accordingly. Table 1 illustrates how the QT range relative to the pseudo-colouring index and the QT-nomogram was estimated.

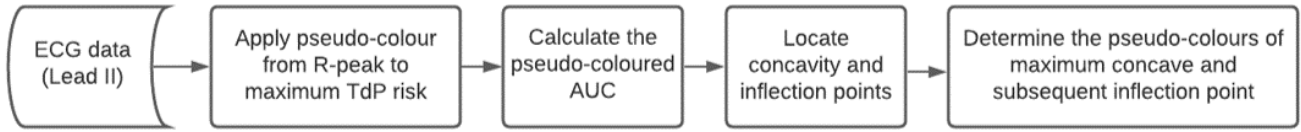


Figure 9: The expert algorithm’s inference process.

Table 1

The confidence rating and the corresponding maximum concave pseudo-colours, inflection point index and the QT range relative to the pseudo-colouring index and the QT-nomogram.

Rating	Likelihood	Maximum concave pseudo-colours	Inflection point index	QT range (Pseudo-colouring index)	QT range (Nomogram Line = NL)
1	Very likely normal	Purple,Blue,Green	1.5	3 (Green) to 4 (Lime)	NL - 160ms to NL - 120ms
2	Probably normal	Blue,Green,Lime	2.5	4 (Lime) to 5 (Yellow)	NL - 120ms to NL - 80ms
3	Possibly normal	Green,Lime,Yellow	3.5	5 (Yellow) to 6 (Orange)	NL - 80ms to NL - 40ms
4	Possibly abnormal	Lime,Yellow,Orange	4.5	6 (Orange) to 7 (Dark Orange)	NL - 40ms to NL
5	Probably abnormal	Yellow,Orange,Dark Orange	5.5	7 (Dark Orange) to 8 (Red)	NL to NL + 40ms
6	Very likely abnormal	Yellow,Orange,Dark Orange	6.5	8 (Red) to 9 (Dark Red)	NL + 40ms to NL + 80ms

3.2.5. ‘Explaining’ the algorithm

A textual interpretation explaining the output of the algorithm was provided with the classification. It described what is visualised in the pseudo-coloured ECG signal and how the algorithm reached its decision. In some cases, the ECG may have an ST-elevation, a sign of myocardial infarction (heart attack), represented by the shape of the ECG signal as a concave (representing the ST-elevation), followed by a convex (representing the beginning of the T-wave, i.e. the $J-T^{peak}$ interval), then a maximum concave down representing the T-wave. This information was included in the result text report if the ECG had this pattern. Below, we show examples of how the algorithm results were explained for four ECGs with different QT-intervals and heart rates. Figure 11 shows normal QT-intervals, and Figure 12 shows at risk QT-prolongation, where the ECG (A) also has an ST-elevation.

“ The QT-interval of this ECG is very likely normal, and the patient is not considered at risk for TdP. This decision has been made based on the assumption that the QT-interval is considered normal when the area under the T-wave contains cool pseudo-colours (purple to blue to green) and prolonged with risk of TdP when it contains warm pseudo-colours (yellow to orange to red). The maximum concave down in the pseudo-coloured area was considered to be the T-wave, and most of its colours are cool (purple to blue to green), with a greater amount of blue than green, indicating a normal QT-interval. The T-wave probably ends within the green colour region, which indicates that the QT/HR falls below the nomogram line by approximately 120 ms or more,

showing no risk of TdP. Based on the pseudo-colouring scale, the estimated value of the QT-interval ranges from 310 to 350 ms, and the HR is 71.

[Figure 11A]

“ The QT-interval of this ECG is probably normal, and the patient is not considered at risk of TdP. This decision has been made based on the assumption that the QT-interval is considered normal when the area under the T-wave contains cool pseudo-colours (purple to blue to green) and prolonged with risk of TdP when it contains warm pseudo-colours (yellow to orange to red). The maximum concave down in the pseudo-coloured area was considered to be the T-wave, and most of its colours are cool colours (purple to blue to green), with a greater amount of green than blue, indicating a normal QT-interval. The T-wave is probably ends within the green region, which indicates that the QT/HR falls below the nomogram line by approximately 80 ms or more, showing no risk of TdP. Based on the pseudo-colouring scale, the estimated value of the QT-interval ranges from 343 to 383 ms, and the HR is 75.

[Figure 11B]

“ The QT-interval of this ECG is possibly abnormal, and the patient is considered at risk of TdP. This decision has been made based on the assumption that the QT-interval is considered normal when the area under the T-

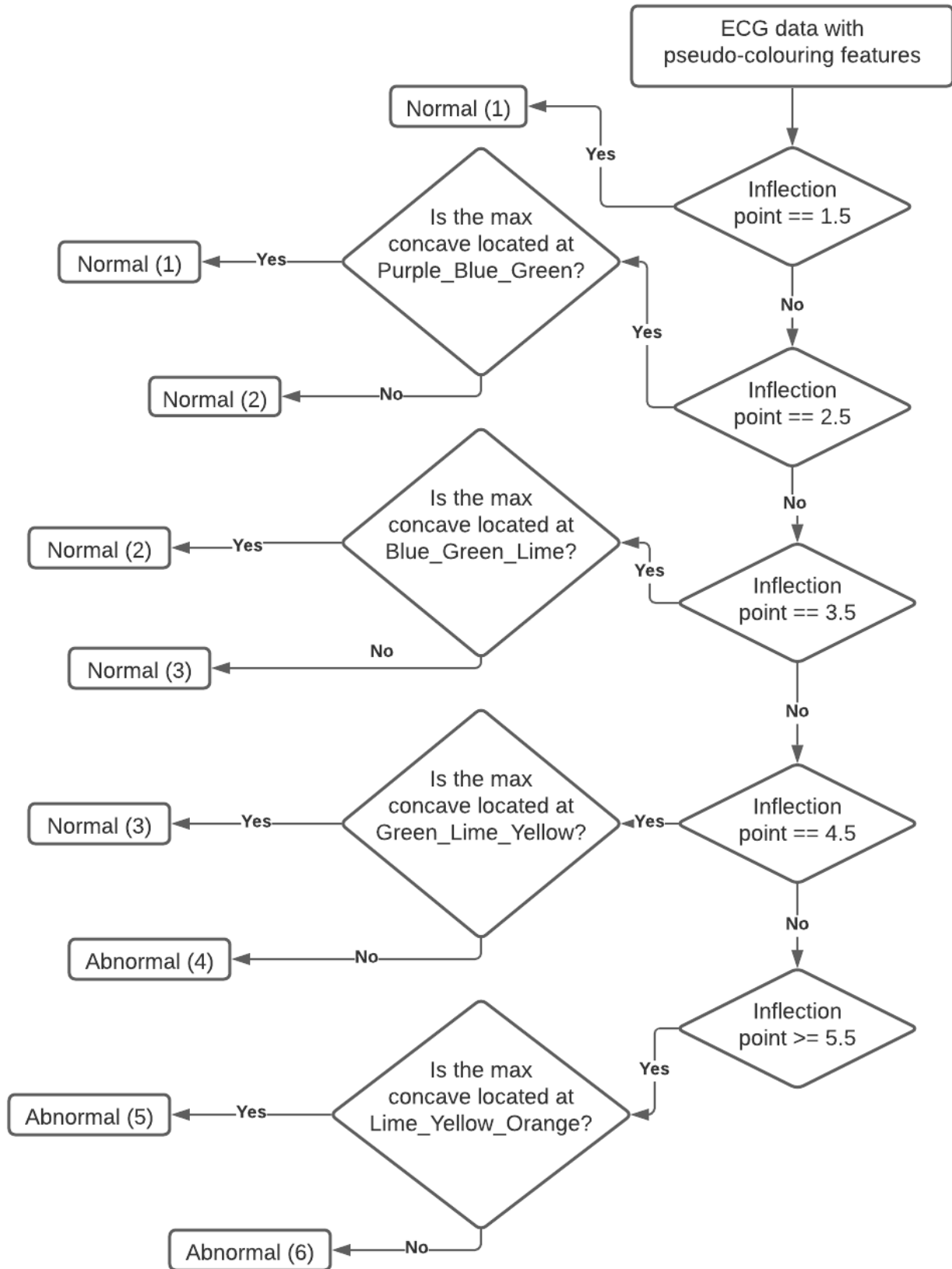


Figure 10: A flowchart of the expert IF-THEN rules. The number in round brackets represents the confidence rating of the algorithm.

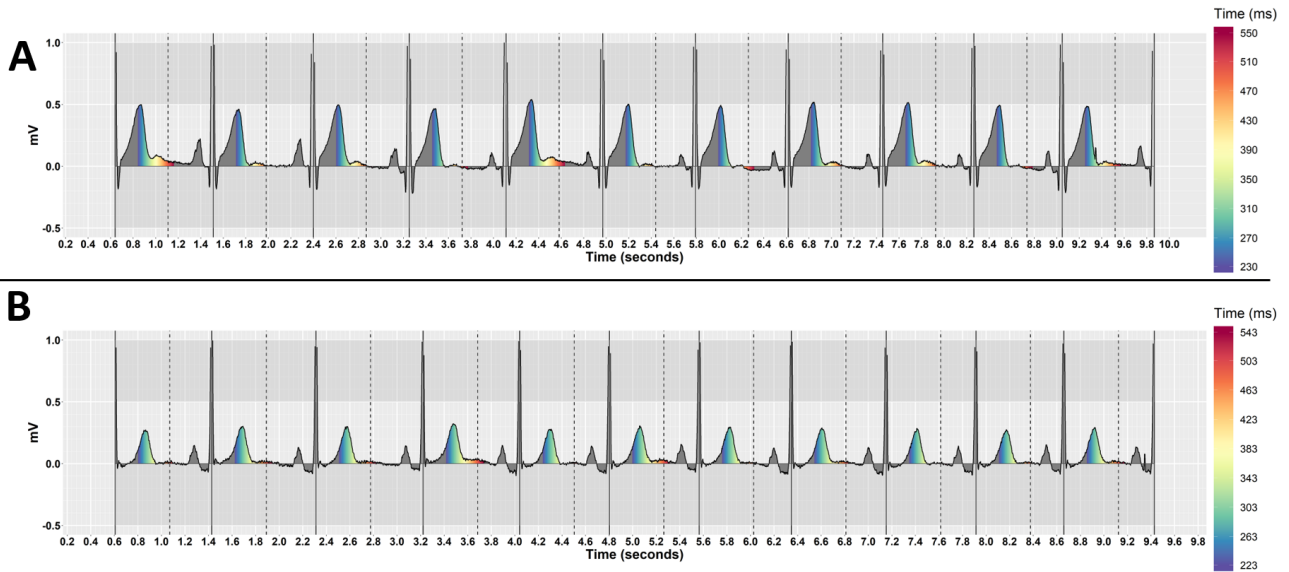


Figure 11: The pseudo-coloured ECG examples with normal QT-intervals showing no risk of TdP interpreted by the expert algorithm. Based on the actual QT/HR acquired from the clinical trial study: (A) The ECG has a normal QT-interval (QT = 345, HR = 71, Difference from the nomogram line = -125, Drug = Ranolazine). (B) The ECG has a normal QT-interval (QT = 378, HR = 75, Difference from the nomogram line = -85, Drug = Placebo).

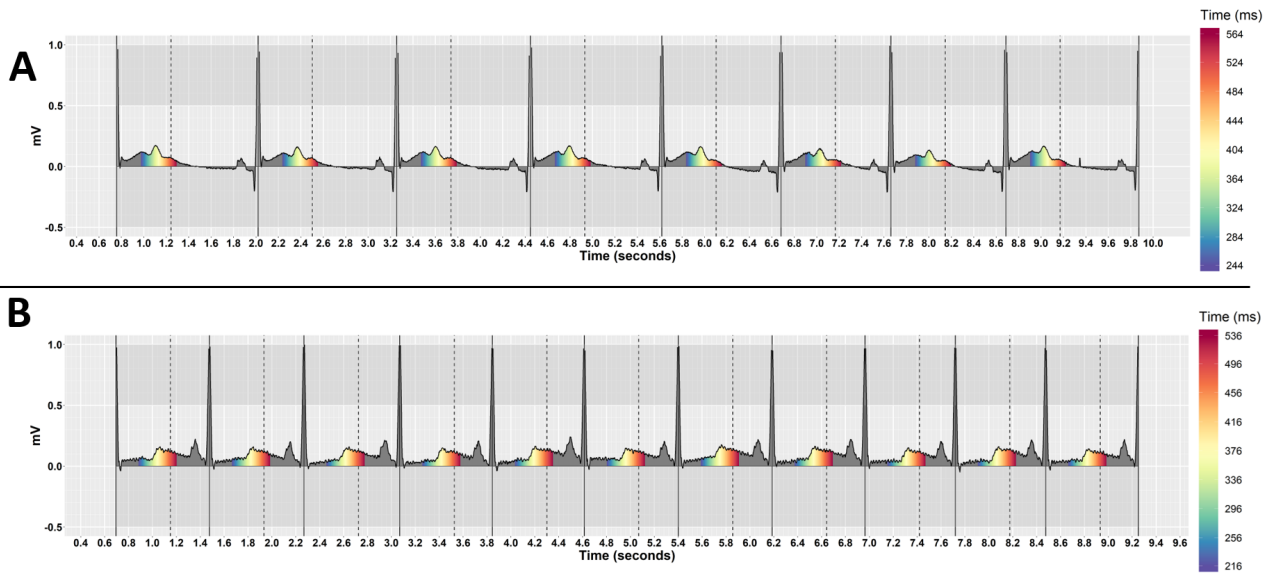


Figure 12: The pseudo-coloured ECG examples with prolonged QT-intervals at risk of TdP interpreted by the expert algorithm. Based on the actual QT/HR acquired from the clinical trial study: (A) The ECG has at risk QT-prolongation and ST-elevation (QT = 532, HR = 52, Difference from the nomogram line = + 48, Drug = Dofetilide). (B) The ECG has at risk QT-prolongation (QT = 518, HR = 77, Difference from the nomogram line = + 62, Drug = Quinidine).

wave contains cool pseudo-colours (purple to blue to green) and prolonged with risk of TdP when it contains warm pseudo-colours (yellow to orange to red). The maximum concave down in the pseudo-coloured area was considered to be the T-wave, and most of its colours are warm colours (green to yellow to orange), with a greater amount of yellow than

green, indicating an abnormal QT-interval. It looks like there is an ST-elevation as there is a purple-blue wave before the T-wave. The T-wave probably ends within the yellow-orange region, which indicates that the QT/HR falls on or very close to the nomogram line, showing risk of TdP. Based on the pseudo-colouring scale, the estimated value of the QT-interval

ranges from 444 to 484 ms, and the HR is 53. ”

[Figure 12A]

“ The QT-interval of this ECG is very likely abnormal, and the patient is considered at risk of TdP. This decision has been made based on the assumption that the QT-interval is considered normal when the area under the T-wave contains cool pseudo-colours (purple to blue to green), and prolonged with risk of TdP when it contains warm pseudo-colours (yellow to orange to red). The maximum concave down in the pseudo-coloured area was considered to be the T-wave, and most of its colours are warm colours (orange to red), indicating an abnormal QT-interval. The T-wave probably ends within the orange-red region, which indicates that the QT/HR falls above the nomogram line, showing risk of TdP. Based on the pseudo-colouring scale, the estimated value of the QT-interval ranges from 496 to 536 ms, and the HR is 77. ”

[Figure 12B]

3.3. Evaluation design

The algorithm was evaluated in four ways. Firstly, we evaluated the diagnostic accuracy of the algorithm for detecting QT-prolongation at risk of TdP on all ECG datasets ($n = 5050$). Secondly, as different types of QT-prolonging drug can affect the T-wave morphology in different ways, we systematically evaluated the effect of drug type on the algorithm's sensitivity to increases in the QT-interval using psychophysical methods. In vision science, psychophysical experiments investigate the relationship between the intensity of a physical stimulus and human perception, by systematically varying the properties of the stimulus along one or more physical dimensions [72]. They are widely used in computer vision research to evaluate an algorithm's behaviour by measuring the exemplar by exemplar difficulty and modeling the algorithm's pattern of errors over different levels of object visibility and saliency, making the algorithm's classification inference more explainable (see examples in [19, 27, 28, 34, 63]). In addition, we assessed how the drug type affected the algorithm's classification confidence score.

Thirdly, we compared the human expert-generated rules with a set of rules automatically generated through statistical machine learning. In this case the same six predictors used for the expert algorithm were used as inputs to train a C4.5 decision tree classifier. The comparison was performed under two conditions: 1) generating the decision tree model with imbalanced class datasets, i.e. using all 5050 ECGs, and 2) with balanced class datasets using down-sampling, which involves randomly removing datasets from the majority class. A 5-fold cross-validation with 20% holdout procedure was used to evaluate the rule-based decision tree performance under both conditions.

That is, 80% of the data was used to train the model and this was tested using 5-fold cross-validation withholding 20% of the data, as shown in Figure 13. The source code of the decision tree classification model can be found in [5]. Finally, we compared the classification accuracy of the expert algorithm with the results of a previous study conducted with humans [4].

3.4. Statistical analysis

The diagnostic accuracy of the algorithm was measured by calculating the area under the receiver operating characteristic (ROC) curve (AUC), sensitivity, specificity, precision, accuracy, F1-score and error rate. QT-intervals below the nomogram line were classified as negative (i.e. 'normal', no risk of TdP), and QT-intervals on or above the nomogram line as positive (i.e. 'abnormal', at risk of TdP). As the ECG dataset has a class imbalance issue, i.e. the number of samples in the negative class ($n = 4870$) is much larger than the number of samples in the positive class ($n = 180$), we used the additional statistical measures of the balanced accuracy (i.e. the average of the sensitivity and specificity), area under the precision-recall curve (PR-AUC) and Matthews correlation coefficient (MCC), which have been shown to be particularly useful for evaluating imbalanced binary classification [14, 65].

The effect of drug type on the algorithm's sensitivity to increases in the QT-interval was modeled using the psychometric function and just noticeable difference (JND) threshold [50]. The psychometric function is an inferential model used in psychophysical detection and discrimination tasks, here used to model the proportion of the ECGs classified as 'QT-prolongation at risk of TdP' by the expert algorithm, as a function of QT-interval increase across the four QT-prolonging drugs and placebo. The just noticeable difference (JND) threshold is defined in psychophysics as the minimum amount of a change in a stimulus required for it to be 'just noticeable' [61]. In this study, we defined it as the minimum difference in the QT-interval from the nomogram line required for the TdP risk to be detectable. The JND threshold was determined by fitting the psychometric function using a logistic function with maximum likelihood estimation (MLE). The correlation between the QT-interval difference from the nomogram line and the algorithm's confidence score for each drug type and across all drugs was calculated using a Spearman's rank correlation.

The comparison with the rule-based decision tree model involved testing the expert algorithm on the same testing data used for each fold (then averaged across the five folds), and on the final 20% holdout testing datasets, under both conditions (imbalanced and balanced class datasets). Evaluation metrics included sensitivity, specificity, precision, balanced accuracy, F1-score, and Matthews correlation coefficient (MCC).

The comparison with human interpretation was conducted by testing the expert algorithm on ECGs ($n = 40$) used in a previous study conducted with humans to evaluate the pseudo-colouring technique [4]. The ECGs were

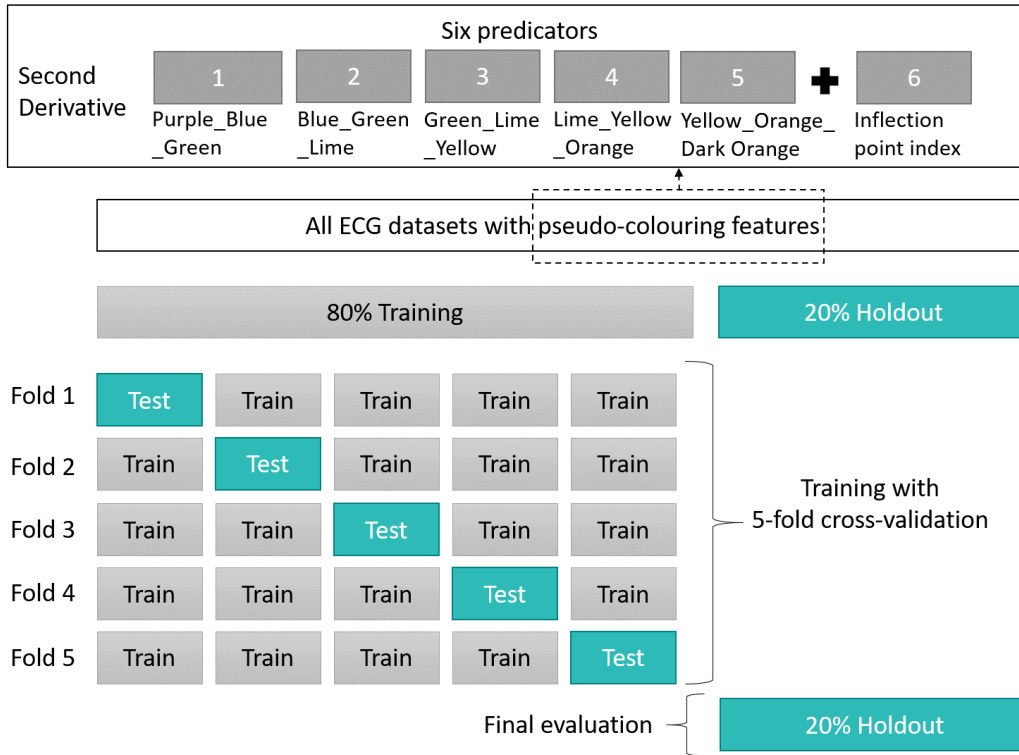


Figure 13: An illustration of the 5-fold cross-validation with 20% holdout procedure used to develop a rule-based decision tree classification model for pseudo-coloured ECGs.

acquired from the same clinical trial study used in the current paper, and were selected from multiple patients with different values of the QT-interval and heart rate; 20 ECGs were below the nomogram line (no TdP risk), and 20 ECGs were on or above the nomogram line, showing QT-prolongation at risk of TdP. Evaluation metrics included sensitivity, specificity and accuracy.

4. Results

All pseudo-coloured ECGs and related metadata underpinning the findings reported in this article can be found in [5].

4.1. Diagnostic accuracy

The results show that the expert algorithm is reliable and accurate in detecting QT-prolongation at risk of the TdP regardless of heart rate. The algorithm was tested on the 5050 ECGs (TdP risk $n = 180$, no risk $n = 4870$), with various values of QT-interval and heart rate across different QT-prolonging drugs, and a placebo. It achieved an accuracy of 0.99, balanced accuracy of 0.97, sensitivity of 0.94, specificity of 0.99, precision of 0.83, F1-score of 0.88, ROC (AUC) of 0.98, precision-recall (AUC) of 0.88, Matthews correlation coefficient (MCC) of 0.88, and error rate of 0.01. The full results are shown in Table 2. Figure 14 shows the algorithm's ROC and precision-recall curves.

The expert algorithm provides an estimated 40 ms range of the QT-interval value, and we found that 74%

Table 2

The diagnostic accuracy results of the expert algorithm.

Evaluation Metric	Proportion
Accuracy	0.99
Balanced Accuracy	0.97
Recall/sensitivity	0.94
Specificity	0.99
Precision/positive predictive value (PPV)	0.83
F1-score	0.88
ROC (AUC)	0.98
Precision-recall (AUC)	0.88
Matthews correlation coefficient (MCC)	0.88
Error rate	0.01

of the ECGs were within the algorithm's estimated range. In addition, the expert algorithm showed superior performance to the pilot version, which considered a QT-interval to be prolonged with risk of TdP if the warm pseudo-colours (yellow to orange to red) occupied more than 50% of the area under the ECG signal. Figure 15 shows a comparison of sensitivity, specificity and balanced accuracy, when classifying all ECGs ($n = 5050$), between the pilot version and the modified version reported in this paper. This indicates that locating concavity and inflection points with respect to the pseudo-colouring scale significantly increased sensitivity to detecting QT-prolongation.

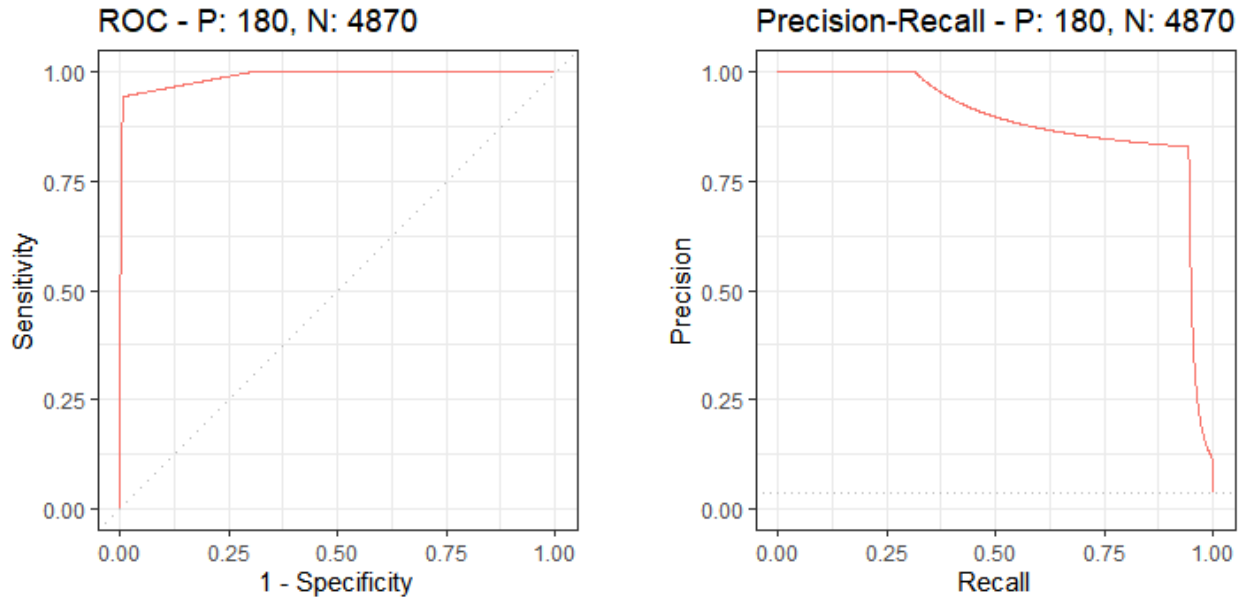


Figure 14: The algorithm's ROC and precision-recall curves when tested on all ECGs (Positive cases (P) $n = 180$, Negative cases (N) $n = 4870$). ROC (AUC) = 0.98 and precision-recall (AUC) = 0.88.



Figure 15: A comparison of the sensitivity, specificity and balanced accuracy, when classifying all ECGs ($n = 5050$), between the pilot version and the modified version of the expert algorithm.

4.2. Modeling the effect of drug type on QT-prolongation detection

The clinical trial study from which the ECGs were acquired assessed the effect of QT-prolonging drug type on T-wave morphology [76]. The results showed that patients on a pure hERG blocker (Dofetilide) developed flat, asymmetric, and notched T-waves, whilst patients on multi-channel blocking drugs (Quinidine and Ranolazine) had equal or greater T-wave morphology changes including distorted and bizarre T-waves. In particular, 'Dofetilide' and 'Quinidine' were shown to cause significant T-wave morphology changes, while substantially prolonging the QT-interval. No significant T-wave changes were observed for patients on 'Verapamil', or the placebo.

In this study, we examined the robustness of the expert algorithm to T-wave morphology changes by evaluating its sensitivity in detecting at risk QT-prolongation across drug types. The psychometric function was plotted as

the proportion of the ECGs classified as 'QT-prolongation at risk of TdP' by the expert algorithm, as a function of the QT-interval difference from the nomogram line. The results show that the sensitivity of the expert algorithm in detecting at risk QT-prolongation increased as the QT-interval difference from the nomogram decreased, regardless of T-wave morphology changes (Figure 16). We estimated the 75% just noticeable difference (JND) threshold as the value of the QT-interval with respect to the nomogram line at which the proportion of the ECGs classified as 'QT-prolongation at risk of TdP' by the expert algorithm is equal to 0.75. It was estimated for the drugs that caused a QT-prolongation at risk of TdP, which were 'Dofetilide' and 'Quinidine'. Figure 17 shows the JND threshold results, which show that the expert algorithm was sensitive in detecting TdP risk for 75% of the ECGs at which the QT-interval value is below the nomogram line by approximately -9 and -3 milliseconds for 'Dofetilide' and 'Quinidine' respectively.

A Spearman's rank correlation was used to determine the relationship between the algorithm's confidence rating and the QT-interval difference from the nomogram line. The results show that there was a strong, positive correlation between the algorithm's confidence rating and the QT difference from the nomogram, which was statistically significant ($R = 0.828$, $p < 0.001$) across all drug types, and per drug (Table 3). Figure 18 shows how the algorithm's confidence rating score of TdP risk increased as the QT-interval difference from the nomogram line decreased (i.e. as the QT-value gets closer to the nomogram line). The results demonstrate the reliability of the algorithm as high risk scores (5 and 6) were given only for 'Dofetilide' and 'Quinidine' drugs, where the QT-values were above the

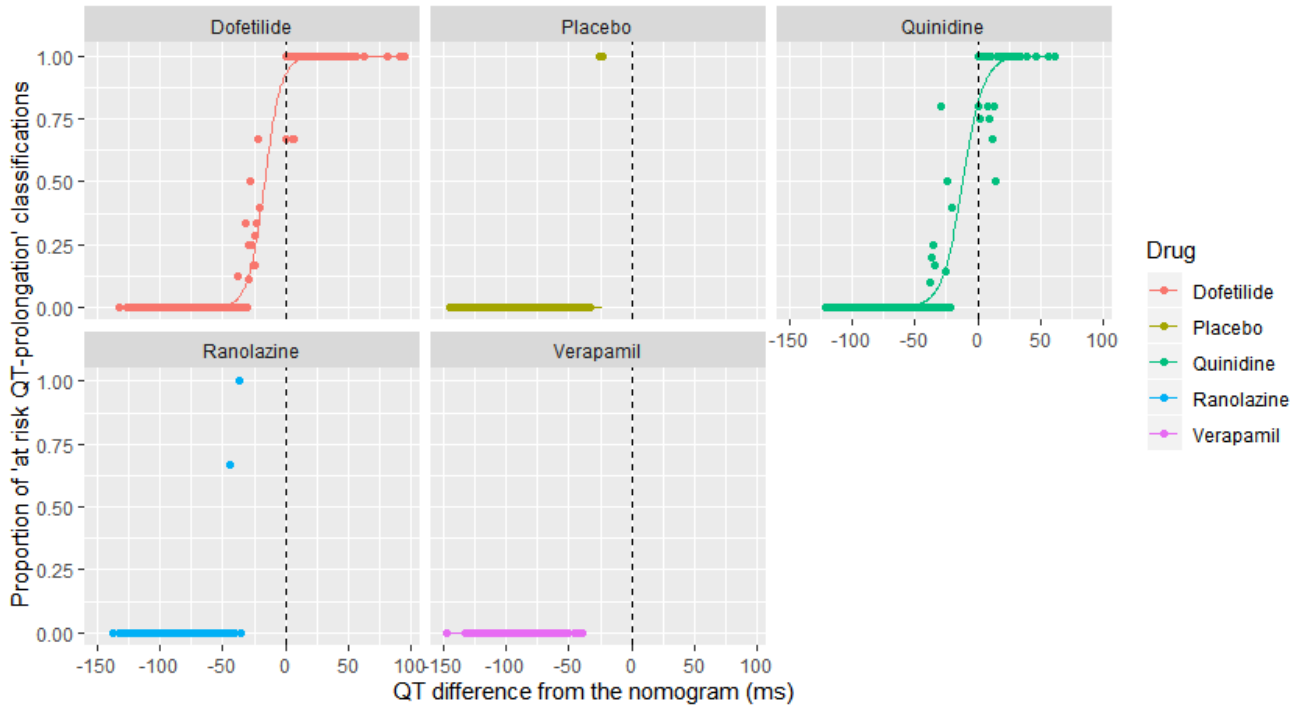


Figure 16: The psychometric function plot shows the proportion of the ECGs classified as ‘QT-prolongation at risk of TdP’ by the expert algorithm, as a function of the QT-interval difference from the nomogram line for each drug type. The QT value of the nomogram line is equal to 0 on the X-axis.

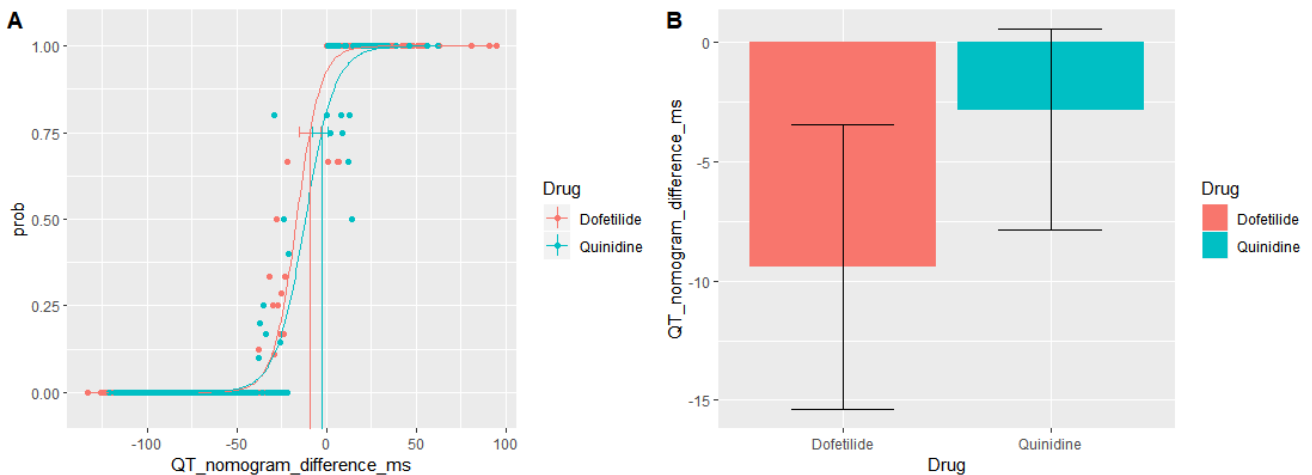


Figure 17: Psychophysical detection measures of the expert algorithm’s sensitivity. (A) The fitted psychometric function plot shows the proportion of the ECGs classified as ‘QT-prolongation at risk of TdP’ by the expert algorithm, as a function of the QT-interval difference from the nomogram line for ‘Dofetilide’ and ‘Quinidine’. The QT value of the nomogram line is equal to 0 on the X-axis. (B) The just noticeable difference (JND) thresholds plot. The error bars represent bootstrap confidence intervals.

nomogram line showing risk of TdP (Figure 18).

4.3. Comparison with rule-based decision tree classification model

Imbalanced classes are a common issue in machine learning, particularly with medical data, where rare conditions have a limited number of representative cases. Imbalance may introduce bias into the training data, but there is no one-size-fits-all solution, as in some cases using more

data, even if not balanced, may improve the model performance [1, 44, 51]. As such, we examined whether class balance affected the decision tree’s rule generation and performance. We used the down-sampling technique as we have a very large number of inputs in the negative class. The re-sampled datasets have various values of QT-interval and heart rate, and were for multiple patients on multiple QT-prolonging drugs/placebo.

Under both the imbalanced and balanced class condi-

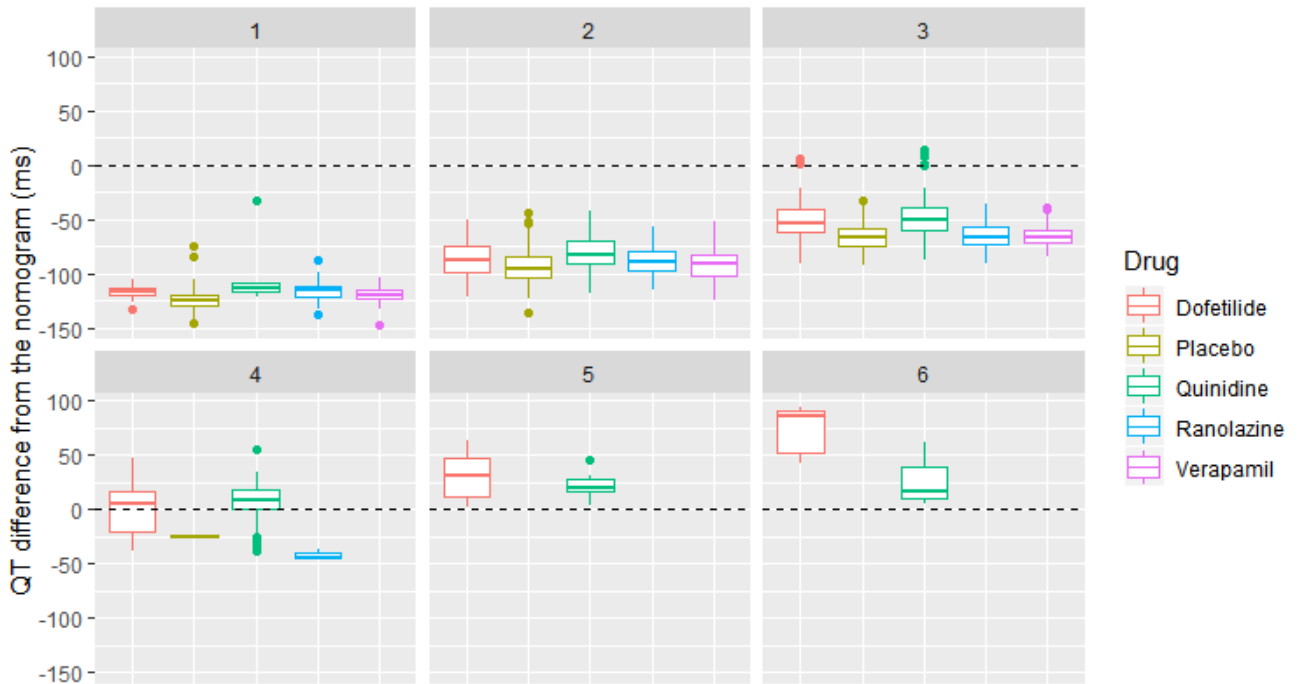


Figure 18: A boxplot shows the distribution of the ECG cases grouped by their confidence rating (1-6) as classified by the expert algorithm. The QT value of the nomogram line is equal to 0 on the Y-axis. The confidence ratings are ‘very likely normal’ (1), ‘probably normal’ (2), ‘possibly normal’ (3), ‘possibly abnormal’ (4), ‘probably abnormal’ (5), and ‘very likely abnormal’ (6).

Table 3

The results of the Spearman’s rank correlation between the algorithm’s confidence rating and the QT-interval difference from the nomogram line for each drug type and all.

Drug type	Spearman’s rho (R)	P-value
Dofetilide	0.867	<0.001
Quinidine	0.824	<0.001
Ranolazine	0.768	<0.001
Verapamil	0.750	<0.001
Placebo	0.810	<0.001
All	0.828	<0.001

tions, the results show that the expert algorithm was more accurate and generalisable to new ECG data than the decision tree model (Figure 19). The decision tree model was over-fitted to the training data under both conditions, showing higher average accuracy during cross-validation than the expert algorithm, but lower accuracy when tested on new data. Tables 4 and 5 show the evaluation results for the imbalanced and balanced class conditions respectively.

More rules were generated under the balanced class condition, which was expected given that the training dataset is smaller and thus the classification rules may be more difficult to infer. Figure 20 illustrates a flowchart of the decision tree IF-THEN classification rules for the imbalanced and balanced class conditions.

4.4. Comparison with human interpretation

A comparison with human interpretation shows that the expert algorithm was more accurate at classifying the 40 ECGs than the averaged human participants ($n = 43$). Figure 21 and Table 6 show the performance results of the expert algorithm and human interpretation in terms of sensitivity, specificity, balanced accuracy, and area under the ROC curve.

5. Focus group evaluation

We conducted two focus group discussions, one with patients and one with clinicians, to gather feedback about the potential usage of the pseudo-coloured ECG and automated algorithm within clinical practice. The focus groups were conducted as patient and public involvement/stakeholder engagement events, which are less formal than an interview study, and designed to gather feedback to direct an approach, rather than provide empirical evidence for it.

The patient focus group involved seven cancer patients (4 male and 3 female), who were recruited via the Patient and Public Involvement and Engagement (PPIE) coordinators of the Cancer Precision Medicine themes, within the NIHR Manchester Biomedical Research Centre (BRC), UK. Table 7 shows the patients’ demographic information and Table 8 summarises the focus group discussion with patients using the short form of GRIPP2 (Guidance for Reporting Involvement of Patients and Public)[71].

All patients had been regularly monitored for drug-induced ECG changes. Some patients ($n = 3$) had at-

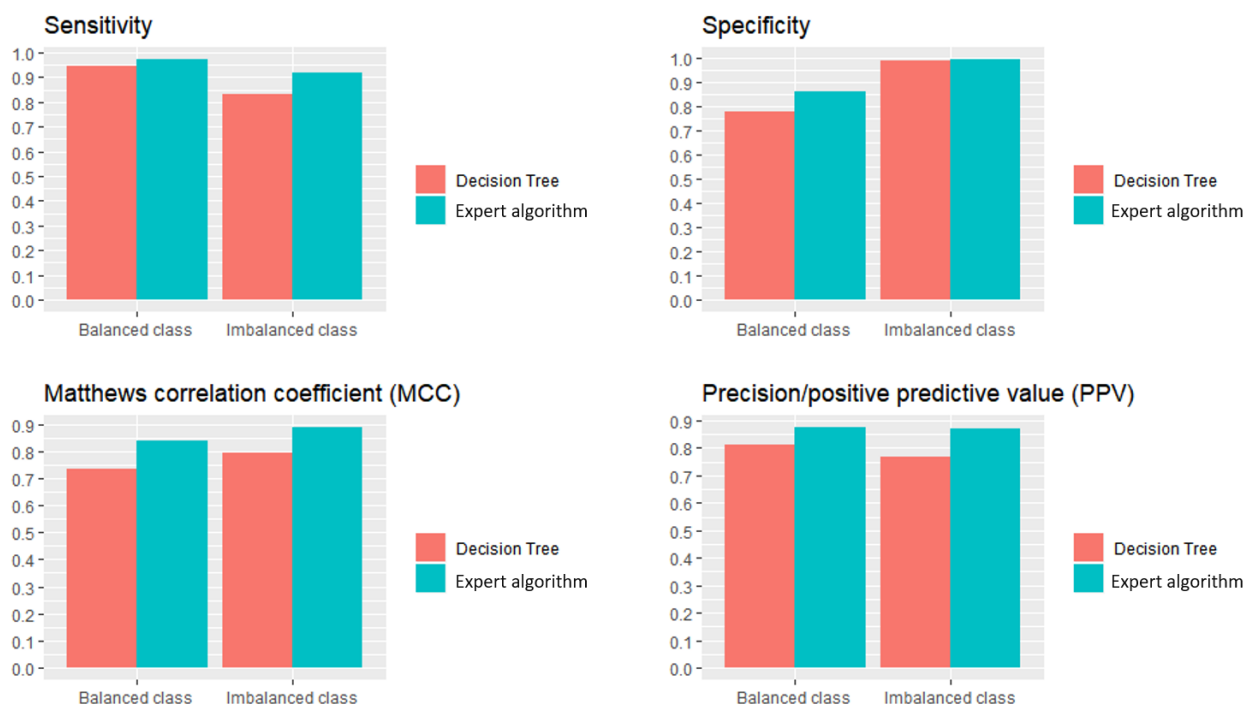


Figure 19: A comparison of the expert algorithm and the rule-based decision tree model, under imbalanced and balanced class conditions, when tested on 20% holdout datasets. Evaluation metrics include sensitivity, specificity, Matthews correlation coefficient (MCC) and precision, also known as positive predictive value (PPV).

Table 4

A comparison of the rule-based decision tree model and the expert algorithm, under the imbalanced class condition, showing the evaluation metrics on each fold (K), averaged across the 5 folds, and on 20% holdout datasets. MCC = Matthews correlation coefficient.

		MCC	Precision	Recall/Sensitivity	Specificity	Balanced Accuracy	F1-score
K1	Expert algorithm	0.882	0.837	0.939	0.992	0.965	0.885
	Decision tree model	0.919	0.937	0.909	0.997	0.953	0.923
K2	Expert algorithm	0.961	0.925	0.967	0.997	0.998	0.961
	Decision tree model	0.979	0.961	1	0.998	0.999	0.980
K3	Expert algorithm	0.828	0.730	0.950	0.991	0.970	0.826
	Decision tree model	0.903	0.863	0.950	0.996	0.973	0.904
K4	Expert algorithm	0.876	0.822	0.948	0.989	0.969	0.880
	Decision tree model	0.859	0.914	0.820	0.996	0.908	0.864
K5	Expert algorithm	0.839	0.774	0.923	0.991	0.957	0.842
	Decision tree model	0.943	0.928	0.962	0.997	0.980	0.945
Averaged across 5 folds	Expert algorithm	0.877	0.818	0.952	0.992	0.972	0.879
	Decision tree model	0.921	0.921	0.928	0.997	0.962	0.923
20% holdout datasets	Expert algorithm	0.888	0.868	0.917	0.995	0.956	0.892
	Decision tree model	0.793	0.769	0.833	0.991	0.912	0.800

tempted to interpret their ECG signal results before, but they found them difficult to understand. The investigator gave a short introduction explaining drug-induced LQTS and how to detect it on both standard and pseudo-coloured ECGs.

All patients found the pseudo-colouring technique to be effective in distinguishing between normal and prolonged

QT-intervals, and preferred the coloured ECG over the non-coloured one.

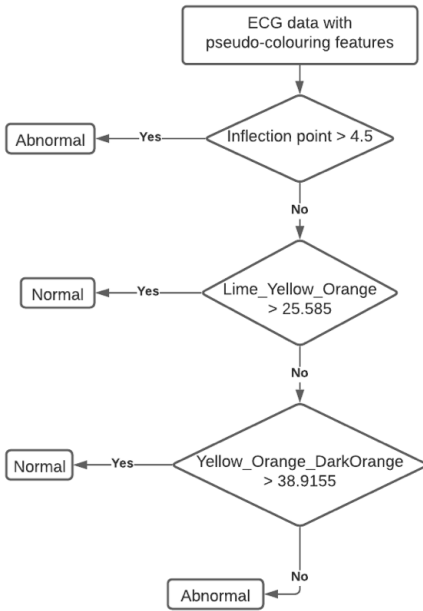
Most patients ($n = 5$) agreed that empowering people to self-monitor for drug-induced LQTS had the potential to save lives, provide a cost-effective healthcare solution, and support more informed shared decision-making between patients and clinicians. Three patients agreed a

Table 5

A comparison of the rule-based decision tree model and the expert algorithm, under the balanced class condition, showing the evaluation metrics on each fold (K), averaged across the 5 folds, and on 20% holdout datasets. MCC = Matthews correlation coefficient.

		MCC	Precision	Recall/Sensitivity	Specificity	Balanced Accuracy	F1-score
K1	Expert algorithm	0.889	0.972	0.945	0.952	0.949	0.958
	Decision tree model	0.887	0.947	0.972	0.904	0.938	0.960
K2	Expert algorithm	0.734	0.800	0.965	0.750	0.857	0.875
	Decision tree model	0.806	0.828	1	0.785	0.892	0.906
K3	Expert algorithm	0.893	0.956	0.916	0.970	0.943	0.936
	Decision tree model	0.928	0.958	0.958	0.970	0.964	0.958
K4	Expert algorithm	0.894	0.960	0.923	0.967	0.945	0.941
	Decision tree model	0.894	0.960	0.923	0.967	0.945	0.941
K5	Expert algorithm	0.861	0.928	0.928	0.933	0.930	0.928
	Decision tree model	0.896	0.962	0.928	0.966	0.947	0.945
Averaged across 5 folds	Expert algorithm	0.854	0.923	0.935	0.914	0.925	0.927
	Decision tree model	0.882	0.931	0.956	0.919	0.937	0.942
20% holdout datasets	Expert algorithm	0.839	0.875	0.972	0.861	0.917	0.921
	Decision tree model	0.732	0.810	0.944	0.778	0.861	0.872

Imbalanced Class



Balanced Class

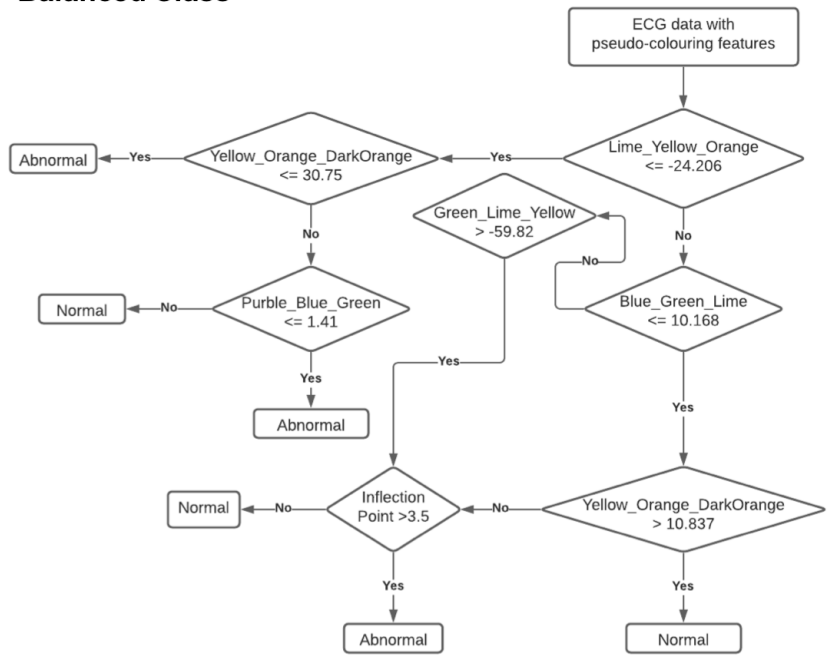


Figure 20: A flowchart of the decision tree IF-THEN rules when trained on imbalanced and balanced class data.

verbal/text-based explanation feature would support them in interpreting the pseudo-coloured ECG, and felt that this would be particularly helpful for elderly patients. Two patients believed that having a risk assessment feature that showed the severity of warm colours relative to the QT-prolongation level could help people with anxiety, so they know when not to worry, and when to seek medical advice.

While two patients raised concerns about trusting automated ECG interpretation, the other patients welcomed

the idea of the expert rule-based ECG interpretation algorithm as a supportive tool that may help to overcome the potential challenges associated with human interpretation. Most patients ($n = 5$) wanted to understand why the algorithm made a certain decision, particularly if it suggested the QT-interval to be abnormal.

The focus group with clinicians involved two female nurses and one male doctor, who are working on early phase clinical trials of new experimental cancer drugs,

Table 6

The sensitivity, specificity, balanced accuracy, and area under the ROC curve of the expert algorithm and human participants (mean values) when classifying the 40 ECGs (TdP risk $n = 20$, no risk $n = 20$).

	Expert algorithm	Human	Difference
Sensitivity	1	0.83	0.17
Specificity	0.95	0.90	0.05
Balanced accuracy	0.98	0.87	0.11
ROC (AUC)	0.98	0.93	0.05

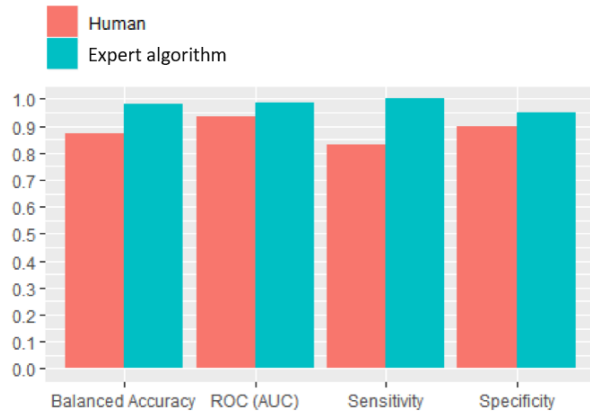


Figure 21: The sensitivity, specificity, balanced accuracy, and area under the ROC curve of the expert algorithm and human participants (mean values) when classifying the 40 ECGs (TdP risk $n = 20$, no risk $n = 20$).

Table 7

Patients' demographic information.

ID	Sex	Age	Education	Occupation
1	Female	50-59 years	Bachelors	Disabled, not working
2	Male	< 30 years	Diploma	Working full time
3	Male	< 30 years	Diploma	Working full time
4	Female	30-39 years	Masters	Unemployed
5	Male	50-59 years	Bachelors	Working part time
6	Female	50-59 years	Masters	Working full time
7	Male	50-59 years	Masters	Working full time

within the Experimental Cancer Medicine Team (ECMT) at The Christie NHS Foundation Trust, Manchester, UK. They were recruited via our research collaborators within the digital ECMT. Table 9 shows the clinicians' demographic information.

As a part of their routine clinical practice, the nurses record patients' ECGs, but they do not interpret them, while the doctor interprets the ECGs and monitors QT-interval changes. The doctor stated that the standard approach to assessing the QT-interval in clinical trials is by using a rigid manual QT-interval measurement, and they do not rely on any kind of automated QT measurements at

present. The clinicians identified the potential benefits of using the pseudo-colouring technique within clinical practice as speeding up the process of interpreting ECGs, and supporting QT-interval monitoring out of hours. All clinicians perceived the pseudo-colouring technique to be particularly useful for clinicians who have less training in ECG interpretation, as it can assist them in visually detecting drug-induced QT-prolongation.

The doctor, who frequently interprets ECGs, recommended having an automated expert algorithm that provided an alert about QT-prolongation, while showing the visualised pseudo-coloured ECG at the same time. He commented this would be particularly useful in reducing common errors associated with manual QT-interval measurement, and in resolving issues with inter-observer variability, particularly in clinical trial settings testing a new drug where accurate QT-monitoring is crucial. He said that well-known difficulties with measuring the QT-interval in clinical practice included slow and fast heart rates and T-wave morphology changes (e.g. the fused T-U complex). He felt that the expert algorithm would be very useful in clinical practice if it could detect QT-prolongation regardless of these issues. All clinicians believed that an expert, explainable automated algorithm would have a greater potential of being trusted and adopted in clinical practice than a fully automated algorithm where the basis for decisions was less clear.

6. Discussion

Many pharmacological drugs have been shown to prolong the QT-interval on the ECG and reported to cause drug-induced TdP and/or sudden cardiac death [12, 21, 46]. Despite its clinical importance, predicting the TdP risk for most of these drugs is difficult, even for clinicians who routinely read ECGs [70, 77]. Major challenges include measuring the QT-interval and determining the TdP risk at varying heart rates [12, 85]. While many modern ECG machines provide automated measurements of the QT-interval, these are usually correct only in noise-free normal sinus rhythm, in which the T-wave morphology is well defined [12, 67]. As QT-prolonging drugs often affect the morphology of the T-wave, this can easily invalidate automated QT measurement [22, 26, 48, 62, 73, 75].

This study demonstrates that an automated ECG algorithm, developed with a *human-like* approach—using human perceptual heuristics to determine features, and expertise from the clinical literature to determine rules—is highly effective in detecting drug-induced QT-prolongation at risk of TdP regardless of heart rate (Table 2; Figure 14). The psychophysical detection measures show that the sensitivity of the expert algorithm in detecting TdP risk increased as the QT-interval approached the nomogram risk line, regardless of drug type (Figures 16 and 17). The JND threshold results indicate that although multi-channel blocking drugs are known to affect T-wave morphology to a greater extent than pure hERG blocking drugs [76], the algorithm was in fact more sensitive with 'Quinidine' than 'Dofetilide'

Table 8

The short form of GRIPP2 (Guidance for Reporting Involvement of Patients and Public) for the focus group discussion with patients.

Section and topic	Extra detail
Aim	To gather feedback and opinions about the potential usage of the pseudo-coloured ECG, and the automated algorithm that interprets it, within clinical practice.
Methods	An online focus group meeting with patients was held in May 2020. Seven patients who were under frequent ECG monitoring were approached by the PPIE coordinators for the Cancer Precision Medicine themes. A semi-structured focus group guide was followed, seeking views on 1) self-monitoring drug-induced LQTS using the pseudo-colouring technique, and 2) automated ECG interpretation. The meeting was led by one of the research team, facilitated by two members of the Manchester BRC's staff.
Study results	All patients perceived pseudo-colouring as an effective way of distinguishing between normal and prolonged QT-intervals, and preferred the coloured ECG over the non-coloured one. They said that empowering people to self-monitor ECGs is going to save lives, be cost-effective, and improve patient-clinician communication, supporting more informed shared decision-making. While two patients raised concerns about trusting automated ECG interpretation, other patients welcomed the idea of an automated algorithm that used the same process of reading the pseudo-colours as expert humans.
Discussion and conclusions	All patients had a positive attitude towards using the pseudo-colouring technique, and perceived the expert automated algorithm as a supportive tool that may help to overcome the potential challenges associated with human interpretation of the pseudo-colours.
Reflections and critical perspective	Involving patients demonstrated the potential of the expert algorithm as a risk assessment tool that could help to quantify the severity of QT-interval prolongation through the area of warm colours displayed, and highlighted the advantage of producing explainable results in a way that could support and guide the lay interpreter of the pseudo-coloured ECG.

Table 9

Clinicians' demographic information.

ID	Sex	Age	Education	Job title
1	Female	< 30 years	Diploma	Clinical Research Nurse
2	Female	30-39 years	BNurs	Clinical Research Nurse
3	Male	30-39 years	MRes	Clinical Fellow

(Figure 17). Research has estimated that 2.0% to 8.8% of patients treated with 'Quinidine' will develop TdP, which is a high number [8, 12, 16, 64, 68]. Although the difference in the JND thresholds between the two drugs is very small ($\approx 6ms$), one possible reason behind it could be that 'Dofetilide' can cause greater T-U fusion/complexes, which are known to precede the development of TdP [17, 39, 87]. In fact, in this case measurement of the QU-interval is recommended [20, 30, 69]. This, in turn, will prolong the maximum concave and inflection point in the ECG signal, which potentially increases the sensitivity of the algorithm to TdP risk, whilst the QT-value is below, but very close, to the nomogram line. In addition, a Spearman's rank correlation shows a strong, positive relationship between the algorithm's confidence rating and the QT difference relative to the nomogram line regardless of drug type, demonstrating the reliability of the expert algorithm (Table 3; Figure 18).

The comparison with the rule-based decision tree model showed the value of using expert knowledge in the development of the algorithm. Both the expert and decision tree algorithms were highly effective, demonstrat-

ing the promise of using perceptual heuristics as features; whilst the decision tree was more effective under cross-validation, however, it was poorer at generalising than the expert algorithm. The decision tree inference mechanism depends on the 'value' of certain features to separate the two classes, which cannot be standardised across all ECGs. For example, under the imbalanced class condition in Figure 20, the second decision point separated the two classes based on the inflection point index and the value of the 'Lime_Yellow_Orange' feature. If the value of this feature is convex greater than 25.585, and the inflection point index is less than 4.5, then the QT-interval is classified as normal. However, the rate of change of a convex/concave depends on the amplitude of the ECG wave, which is known to differ substantially across individuals, and is affected by factors including age, race, sex and health status, body mass and electrode position [12, 30]. The decision of the expert algorithm is not based on inferred feature values, but rather on the relationships between features acquired from the medical knowledge used to detect risk of TdP. For instance, at the fourth decision point in Figure 10 the expert algorithm separates a normal from an abnormal QT-interval, despite the fact that they have the same inflection point index, based on the location of the maximum concave. As shown in Figure 8, if the inflection point index is equal to 4.5 and the maximum concave is located at 'Green_Lime_Yellow' and not 'Lime_Yellow_Orange', then this means the T-wave is farther from the nomogram line, and there is thus no TdP risk. Furthermore, we encountered the issue of cognitive and technical bias in selecting good, representative training datasets. Research has shown that this issue is rel-

atively common when selecting training medical datasets [1, 44, 51]. Despite numerous statistical re-sampling methods including up/downsampling and k-fold cross validation techniques, acquiring representative datasets that support the generalisability of machine learning algorithms, remains a challenging problem.

An additional benefit of the human-like approach is its inherent explainability. The expert algorithm is able to both precisely specify the reasons for a decision, and—due to the algorithm and human interpreter sharing the same, pseudo-coloured representation of the data—the reasons for the decision are straightforward for someone to perceive. The shared representation of the ECG data, and the shared model of how to interpret it, are also important when considering the application of this approach within clinical practice. All new regulations for adopting artificial intelligence (AI)- and machine learning (ML)-based technologies in healthcare emphasise the importance of explainability [41, 23]. Results from the focus groups with clinicians and patients also provided evidence for the value of the expert algorithm. The explanations it provides (Figures 11 and 12) may be helpful in training lay people to read pseudo-coloured ECGs to self-monitor QT-prolongation, as well as supporting clinicians in supervising and validating its automated results in clinical practice, using a human-in-the-loop approach. Clinicians commented that our approach would be particularly useful in reducing common errors associated with manual QT-interval measurement, and in resolving issues with inter-observer variability, particularly in clinical trial settings testing a new drug where accurate, frequent QT-monitoring is crucial. Whilst the decision tree algorithm is also technically explainable, its decisions cannot be mapped directly to the medical evidence, and are therefore harder to understand, and potentially spurious, if inferred from a biased dataset. Nevertheless, machine learning has been shown to have promise in many areas of medical decision making, and we highlight in particular its potential for efficiently inferring new knowledge, which, once validated, could be used in future expert algorithms.

6.1. Limitations and future work

In this study we only examined the detection of QT-prolongation at risk of TdP, and it is not clear whether the pseudo-colour heuristics would support ECG interpretation of other abnormalities that may increase the risk of TdP, including electrolyte imbalance (e.g. hypokalemia and hypocalcemia) and changes in ST-segment elevation.

This study showed the preliminary results of using human-like perceptual pre-processing of the ECG signal data to facilitate a single decision-tree machine learning algorithm, and future work should explore whether this can aid other machine learning techniques, including computer vision, where pseudo-colouring could be used to improve ECG information segmentation in a pre-processing step. Only a small number of people participated in the focus groups, and further usability evaluations are necessary to

determine the utility of the technique in clinical practice and with more diverse clinical populations.

7. Conclusion

This study demonstrates that combining a data representation based on human perceptual heuristics with expert clinical knowledge results in accurate, reliable and explainable automated detection of drug-induced QT-prolongation at risk of TdP regardless of heart rate, and robust to T-wave morphology changes. The results indicate that a prolonged ventricular repolarisation area can be a significant risk predictor of TdP, and it is potentially easier and more reliable to automate detection of this than automating the measurement of the QT-interval distance *per se*.

8. Author contributions

AA and CJ conceptualised and devised the idea for the work. AA developed and evaluated the algorithm, carried out the ECG data acquisition, designed and analysed the study and wrote the paper, with CJ, AD, and MV contributing significant edits. CJ assisted with study design and drew out the theoretical contribution. AD acted as the electrocardiogram domain expert throughout. JR, LG, KC and ZA coordinated and organised the focus group events.

9. Conflicts of interest

The authors have declared that no competing interests exist.

10. Acknowledgments

This work was supported by the first author's (Alaa Alahmadi) sponsor (Taibah University, College of Computer Science and Engineering, Kingdom of Saudi Arabia) and was funded for her PhD research by Saudi Arabian Cultural Bureau in London (grant number TAU388). We would like to thank the Digital Experimental Cancer Medicine Team (digitalECMT), and the Patient and Public Involvement and Engagement (PPIE) coordinators within the NIHR Manchester Biomedical Research Centre (BRC), for their feedback and support throughout the research project.

References

- [1] Adamson, A.S., Smith, A., 2018. Machine learning and health care disparities in dermatology. *JAMA dermatology* 154, 1247–1248.
- [2] Alahmadi, A., Davies, A., Dempsey, K., Vigo, M., Jay, C., 2021a. Human-Machine Perception of Complex Signal Data. Human-Like Machine Intelligence, Muggleton and Chater, Eds. In preparation, Oxford University Press.
- [3] Alahmadi, A., Davies, A., Royle, J., Vigo, M., Jay, C., 2019. Evaluating the impact of pseudo-colour and coordinate system on the detection of medication-induced ecg changes, in: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, pp. 1–13.
- [4] Alahmadi, A., Davies, A., Vigo, M., Jay, C., 2020. Pseudo-colouring an ecg enables lay people to detect qt-interval prolongation regardless of heart rate. *PLoS one* 15, e0237854.

- [5] Alahmadi, A., Davies, A., Vigo, M., Jay, C., 2021b. An explainable algorithm for detecting drug-induced qt-prolongation at risk of torsades de pointes (tdp) regardless of heart rate and t-wave morphology. version 1.0 [computer software]. <https://doi.org/10.5281/zenodo.4415791>.
- [6] Anderson, M.E., Al-Khatib, S.M., Roden, D.M., Califf, R.M., Institute, D.C.R., et al., 2002. Cardiac repolarization: current knowledge, critical gaps, and new approaches to drug development and patient management. *American heart journal* 144, 769–781.
- [7] Bachmann, T.N., Skov, M.W., Rasmussen, P.V., Graff, C., Pietersen, A., Lind, B., Struijk, J.J., Olesen, M.S., Haunsø, S., Køber, L., et al., 2016. Electrocardiographic tpeak–tend interval and risk of cardiovascular morbidity and mortality: Results from the copenhagen ecg study. *Heart Rhythm* 13, 915–924.
- [8] Bauman, J.L., Bauernfeind, R.A., Hoff, J.V., Strasberg, B., Swiryn, S., Rosen, K.M., 1984. Torsade de pointes due to quinidine: observations in 31 patients. *American heart journal* 107, 425–430.
- [9] Behere, S.P., Shubkin, C.D., Weindling, S.N., 2014. Recent advances in the understanding and management of long qt syndrome. *Current opinion in pediatrics* 26, 727–733.
- [10] Bonate, P.L., Russell, T., 1999. Assessment of qtc prolongation for non-cardiac-related drugs from a drug development perspective. *The Journal of Clinical Pharmacology* 39, 349–358.
- [11] Borchers, H.W., 2015. *Pracma: practical numerical math functions*. R package version 1.
- [12] Camm, A.J., Malik, M., Yap, Y.G., 2008. *Acquired long QT syndrome*. John Wiley & Sons.
- [13] Chan, A., Isbister, G., Kirkpatrick, C., Dufful, S., 2007. Drug-induced qt prolongation and torsades de pointes: evaluation of a qt nomogram. *QJM: An International Journal of Medicine* 100, 609–615.
- [14] Chicco, D., Jurman, G., 2020. The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation. *BMC genomics* 21, 6.
- [15] Couderc, J.P., Zhou, M., Sarapa, N., Zareba, W., 2008. Investigating the effect of sotalol on the repolarization intervals in healthy young individuals. *Journal of Electrocardiology* 41, 595–602.
- [16] Darpo, B., 2001. Spectrum of drugs prolonging qt interval and the incidence of torsades de pointes. *European Heart Journal Supplements* 3, K70–K80.
- [17] De´ molis, J.L., Funck-Brentano, C., Ropers, J., Ghadanfar, M., Nichols, D.J., Jaillon, P., 1996. Influence of dofetilide on qt-interval duration and dispersion at various heart rates during exercise in humans. *Circulation* 94, 1592–1599.
- [18] Dessertenne, F., 1966. La tachycardie ventriculaire a deux foyers opposes variables. *Arch Mal Coeur* 59, 263–272.
- [19] Eberhardt, S., Cader, J.G., Serre, T., 2016. How deep is the feature analysis underlying rapid visual categorization?, in: *Advances in neural information processing systems*, pp. 1100–1108.
- [20] El-sherif, N., Bekheit, S.S., Henkin, R., 1989. Quinidine-induced long qt interval and torsade de pointes: role of bradycardia-dependent early afterdepolarizations. *Journal of the American College of Cardiology* 14, 252–257.
- [21] El-Sherif, N., Turitto, G., Boutjdir, M., 2020. Acquired long qt syndrome and electrophysiology of torsade de pointes, in: *Cardiac Repolarization*. Springer, pp. 201–216.
- [22] Estes III, N.M., 2013. Computerized interpretation of ecgs: supplement not a substitute.
- [23] Food, Administration, D., et al., 2019. Proposed regulatory framework for modifications to artificial intelligence/machine learning (ai/ml)-based software as a medical device (samd)-discussion paper.
- [24] Fossa, A.A., Wisialowski, T., Magnano, A., Wolfgang, E., Winslow, R., Gorczyca, W., Crimin, K., Raunig, D.L., 2005. Dynamic beat-to-beat modeling of the qt-rr interval relationship: analysis of qt prolongation during alterations of autonomic state versus human ether a-go-go-related gene inhibition. *Journal of Pharmacology and Experimental Therapeutics* 312, 1–11.
- [25] G Postema, P., AM Wilde, A., 2014. The measurement of the qt interval. *Current cardiology reviews* 10, 287–294.
- [26] Garg, A., Lehmann, M.H., 2013. Prolonged qt interval diagnosis suppression by a widely used computerized ecg analysis system. *Circulation: Arrhythmia and Electrophysiology* 6, 76–83.
- [27] Geirhos, R., Janssen, D.H., Schütt, H.H., Rauber, J., Bethge, M., Wichmann, F.A., 2017. Comparing deep neural networks against humans: object recognition when the signal gets weaker. *arXiv preprint arXiv:1706.06969*.
- [28] Gerhard, H.E., Wichmann, F.A., Bethge, M., 2013. How sensitive is the human visual system to the local statistics of natural images? *PLoS computational biology* 9.
- [29] Goldberger, A.L., Amaral, L.A., Glass, L., Hausdorff, J.M., Ivanov, P.C., Mark, R.G., Mietus, J.E., Moody, G.B., Peng, C.K., Stanley, H.E., 2000. Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals. *circulation* 101, e215–e220.
- [30] Goldenberg, I., Moss, A.J., Zareba, W., 2006. Qt interval: how to measure it and what is “normal”. *Journal of cardiovascular electrophysiology* 17, 333–336.
- [31] HABBAB, M.A., EL-SHERIF, N., 1992. Tu alternans, long qt, and torsade de pointes: clinical and experimental observations. *Pacing and Clinical Electrophysiology* 15, 916–931.
- [32] Healey, C.G., Booth, K.S., Enns, J.T., 1995. Visualizing real-time multivariate data using preattentive processing. *ACM Transactions on Modeling and Computer Simulation (TOMACS)* 5, 190–221.
- [33] Healey, C.G., Booth, K.S., Enns, J.T., 1996. High-speed visual estimation using preattentive processing. *ACM Transactions on Computer-Human Interaction (TOCHI)* 3, 107–135.
- [34] Heath, M., Sarkar, S., Sanocki, T., Bowyer, K., 1998. Comparison of edge detectors: a methodology and initial study. *Computer vision and image understanding* 69, 38–54.
- [35] Hermans, B.J., Vink, A.S., Bennis, F.C., Filippini, L.H., Meijborg, V.M., Wilde, A.A., Pison, L., Postema, P.G., Delhaas, T., 2017. The development and validation of an easy to use automatic qt-interval algorithm. *PloS one* 12, e0184352.
- [36] Hunt, A.C., 2005. Accuracy of popular automatic qt interval algorithms assessed by a ‘gold standard’ and comparison with a novel method: computer simulation study. *BMC cardiovascular disorders* 5, 29.
- [37] Ireland, R., Robinson, R., Heller, S., Harris, N., 1998. Qt measurement for tu fused ecg morphology as exhibited during hypoglycaemia, in: *Proceedings of the 20th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. Vol. 20 Biomedical Engineering Towards the Year 2000 and Beyond (Cat. No. 98CH36286), IEEE. pp. 240–242.
- [38] Isbister, G.K., Page, C.B., 2013. Drug induced qt prolongation: the measurement and assessment of the qt interval in clinical practice. *British journal of clinical pharmacology* 76, 48–57.
- [39] Jaiswal, A., Goldbar, S., 2014. Dofetilide induced torsade de pointes: mechanism, risk factors and management strategies. *Indian heart journal* 66, 640–648.
- [40] Johannesen, L., Vicente, J., Mason, J., Sanabria, C., Waite-Labott, K., Hong, M., Guo, P., Lin, J., Sørensen, J.S., Galeotti, L., et al., 2014. Differentiating drug-induced multichannel block on the electrocardiogram: randomized study of dofetilide, quinidine, ranolazine, and verapamil. *Clinical Pharmacology & Therapeutics* 96, 549–558.
- [41] Joshi, I., Morley, J., 2019. Artificial intelligence: How to get it right. putting policy into practice for safe data-driven innovation in health and care.
- [42] Kallergis, E.M., Goudis, C.A., Simantirakis, E.N., Kochiadakis, G.E., Vardas, P.E., 2012. Mechanisms, risk factors, and management of acquired long qt syndrome: a comprehensive review. *The Scientific World Journal* 2012.
- [43] Kasamaki, Y., Ozawa, Y., Ohta, M., Sezai, A., Yamaki, T., Kaneko, M., Watanabe, I., Hirayama, A., Nakayama, T., 2011. Automated versus manual measurement of the qt interval and corrected qt interval. *Annals of Noninvasive Electrocardiology* 16, 156–164.
- [44] Kaushal, A., Altman, R., Langlotz, C., 2020. Geographic distribution of us cohorts used to train deep learning algorithms. *JAMA* 324,

- 1212–1213.
- [45] Kautzner, J., Gang, Y., Kishore, R., Copie, X., Janota, T., Nagayoshi, H., Camm, A.J., Malik, M., 1996. Interobserver reproducibility of qt interval measurement and qt dispersion in patients after acute myocardial infarction. *Annals of Noninvasive Electrocardiology* 1, 363–374.
- [46] Khan, I.A., 2002. Clinical and therapeutic aspects of congenital and acquired long qt syndrome. *The American journal of medicine* 112, 58–66.
- [47] Kirchhof, P., Franz, M.R., Bardai, A., Wilde, A.M., 2009. Giant t-u waves precede torsades de pointes in long qt syndrome: a systematic electrocardiographic analysis in patients with acquired and congenital qt prolongation. *Journal of the American College of Cardiology* 54, 143–149.
- [48] Kligfield, P., Badilini, F., Denjoy, I., Babaeizadeh, S., Clark, E., De Bie, J., Devine, B., Extramiana, F., Generali, G., Gregg, R., et al., 2018. Comparison of automated interval measurements by widely used algorithms in digital electrocardiographs. *American heart journal* 200, 1–10.
- [49] Kurokawa, S., Niwano, S., Kiryu, M., Murakami, M., Ishikawa, S., Yumoto, Y., Moriguchi, M., Niwano, H., Kosukegawa, T., Izumi, T., 2010. Importance of morphological changes in tu waves during bepridil therapy as a predictor of ventricular arrhythmic event. *Circulation Journal* 74, 876–884.
- [50] Linares, D., López-Moliner, J., 2016. quickpsy: An r package to fit psychometric functions for multiple groups. *The R Journal*, 2016, vol. 8, num. 1, p. 122–131.
- [51] Liu, Y., Chen, P.H.C., Krause, J., Peng, L., 2019. How to read articles that use machine learning: users' guides to the medical literature. *Jama* 322, 1806–1816.
- [52] McLaughlin, N., Campbell, R., Murray, A., 1995a. Influence of t wave amplitude on automatic qt measurement, in: *Computers in Cardiology 1995*, IEEE, pp. 777–780.
- [53] McLaughlin, N., Campbell, R., Murray, A., 1996. Accuracy of four automatic qt measurement techniques in cardiac patients and healthy subjects. *Heart* 76, 422–426.
- [54] McLaughlin, N.B., Campbell, R., Murray, A., 1995b. Comparison of automatic qt measurement techniques in the normal 12 lead electrocardiogram. *Heart* 74, 84–89.
- [55] Miller, M.D., Co-burn, J.P., Ackerman, M.J., 2001. Diagnostic accuracy of screening electrocardiograms in long qt syndrome i. *Pediatrics* 108, 8–12.
- [56] Morganroth, J., 2001. Focus on issues in measuring and interpreting changes in the qtc interval duration. *European heart journal supplements* 3, K105–K111.
- [57] Morganroth, J., Silber, S.S., 1999. How to obtain and analyze electrocardiograms in clinical trials: focus on issues in measuring and interpreting changes in the qtc interval duration. *Annals of Noninvasive Electrocardiology* 4, 425–433.
- [58] Moss, A.J., Kass, R.S., et al., 2005. Long qt syndrome: from channels to cardiac arrhythmias. *The Journal of clinical investigation* 115, 2018–2024.
- [59] Panicker, G.K., Karnad, D.R., Natekar, M., Kothari, S., Narula, D., Lokhandwala, Y., 2009. Intra-and interreader variability in qt interval measurement by tangent and threshold methods in a central electrocardiogram laboratory. *Journal of electrocardiology* 42, 348–352.
- [60] Postema, P.G., De Jong, J.S., Van der Bilt, I.A., Wilde, A.A., 2008. Accurate electrocardiographic assessment of the qt interval: teach the tangent. *Heart Rhythm* 5, 1015–1018.
- [61] Prins, N., et al., 2016. *Psychophysics: a practical introduction*. Academic Press.
- [62] Rautaharju, P.M., Surawicz, B., Gettes, L.S., 2009. Aha/accf/hrs recommendations for the standardization and interpretation of the electrocardiogram: part iv: the st segment, t and u waves, and the qt interval: a scientific statement from the american heart association electrocardiography and arrhythmias committee, council on clinical cardiology; the american college of cardiology foundation; and the heart rhythm society: endorsed by the international society for computerized electrocardiology. *Circulation* 119, e241–e250.
- [63] RichardWebster, B., Yon Kwon, S., Clarizio, C., Anthony, S.E., Scheirer, W.J., 2018. Visual psychophysics for making face recognition algorithms more explainable, in: *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 252–270.
- [64] Roden, D.M., Woosley, R.L., Primm, R.K., 1986. Incidence and clinical features of the quinidine-associated long qt syndrome: implications for patient care. *American heart journal* 111, 1088–1093.
- [65] Saito, T., Rehmsmeier, M., 2015. The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets. *PLoS one* 10, e0118432.
- [66] Salvi, V., Karnad, D.R., Panicker, G.K., Natekar, M., Hingorani, P., Kerkar, V., Ramasamy, A., de Vries, M., Zumbrunnen, T., Kothari, S., et al., 2011. Comparison of 5 methods of qt interval measurements on electrocardiograms from a thorough qt/qtc study: effect on assay sensitivity and categorical outliers. *Journal of electrocardiology* 44, 96–104.
- [67] Schläpfer, J., Wellens, H.J., 2017. Computer-interpreted electrocardiograms: benefits and limitations. *Journal of the American College of Cardiology* 70, 1183–1192.
- [68] Selzer, A., WRAY, H.W., 1964. Quinidine syncope: paroxysmal ventricular fibrillation occurring during treatment of chronic atrial arrhythmias. *Circulation* 30, 17–26.
- [69] SHIMIZU, W., OHE, T., KURITA, T., SHIMOMURA, K., 1991. Differential response of qtu interval to exercise, isoproterenol, and atrial pacing in patients with congenital long qt syndrome. *Pacing and Clinical Electrophysiology* 14, 1966–1970.
- [70] Solomons, L., Treloar, A., Noronha, R., 2008. Competence of psychiatric clinicians in interpreting electrocardiograms and qt intervals: can they do this? does it matter? *Psychiatric Bulletin* 32, 291–294.
- [71] Staniszewska, S., Brett, J., Simera, I., Seers, K., Mockford, C., Goodlad, S., Altman, D., Moher, D., Barber, R., Denegri, S., et al., 2017. Gripp2 reporting checklists: tools to improve reporting of patient and public involvement in research. *Research involvement and engagement* 3, 13.
- [72] Stevens, S.S., 2017. *Psychophysics: Introduction to its perceptual, neural and social prospects*. Routledge.
- [73] Talebi, S., Alaleh, A., Sam, Z., Sandeep, S., Visco, F., Totouom-Tangho, H., Kalantari, H., Worku Hassen, G., 2015. Underestimated and unreported prolonged qtc by automated ecg analysis in patients on methadone: can we rely on computer reading? *Acta cardiologica* 70, 211–216.
- [74] Topilski, I., Rogowski, O., Rosso, R., Justo, D., Copperman, Y., Glikson, M., Belhassen, B., Hochenberg, M., Viskin, S., 2007. The morphology of the qt interval predicts torsade de pointes during acquired bradyarrhythmias. *Journal of the American College of Cardiology* 49, 320–328.
- [75] Tyl, B., Azzam, S., Blanco, N., Wheeler, W., 2011. Improvement and limitation of the reliability of automated qt measurement by recent algorithms. *Journal of electrocardiology* 44, 320–325.
- [76] Vicente, J., Johannesen, L., Mason, J.W., Crumb, W.J., Pueyo, E., Stockbridge, N., Strauss, D.G., 2015. Comprehensive t wave morphology assessment in a randomized clinical study of dofetilide, quinidine, ranolazine, and verapamil. *Journal of the American Heart Association* 4, e001615.
- [77] Viskin, S., Rosovski, U., Sands, A.J., Chen, E., Kistler, P.M., Kalman, J.M., Chavez, L.R., Torres, P.I., Centurión, O.A., Fujiki, A., et al., 2005. Inaccurate electrocardiographic interpretation of long qt: the majority of physicians cannot recognize a long qt when they see one. *Heart Rhythm* 2, 569–574.
- [78] Vukmirović, M., Tomašević-Vukmirović, I., Angelkov, L., Vukmirović, F., 2015. Emotional stress as a cause of syncope and torsade de pointes in patients with long qt syndrome. *Vojnosanitetski pregled* 72, 192–195.
- [79] Wan, X., Li, Y., Xia, C., Wu, M., Liang, J., Wang, N., 2016. A t-wave alternans assessment method based on least squares curve fitting technique. *Measurement* 86, 93–100.

- [80] Ware, C., 2012. Information visualization: perception for design. Elsevier.
- [81] Ware, C., 2019. Information visualization: perception for design. Morgan Kaufmann.
- [82] Woosley, R.L., Schwartz, P.J., 2020. Drug-induced long qt syndrome and torsades de pointes, in: Cardiac Repolarization. Springer, pp. 185–200.
- [83] Xue, Q., Reddy, S., 1998. Algorithms for computerized qt analysis. Journal of electrocardiology 30, 181–186.
- [84] Yan, G.X., Antzelevitch, C., 1998. Cellular basis for the normal t wave and the electrocardiographic manifestations of the long-qt syndrome. Circulation 98, 1928–1936.
- [85] Yap, Y.G., Camm, A.J., 2003. Drug induced qt prolongation and torsades de pointes. Heart 89, 1363–1372.
- [86] Yeh, S.T., et al., 2002. Using trapezoidal rule for the area under a curve calculation. Proceedings of the 27th Annual SAS® User Group International (SUGI'02).
- [87] Zipes, D.P., 1987. Proarrhythmic effects of antiarrhythmic drugs. The American journal of cardiology 59, E26–E31.