Building Decision Support Systems from Large Electrocardiographic Data Sets

by

Alex Page

Submitted in Partial Fulfillment of the

Requirements for the Degree

Doctor of Philosophy

Supervised by

Professor Tolga Soyata

Department of Electrical and Computer Engineering
Arts, Sciences and Engineering
Edmund A. Hajim School of Engineering and Applied Sciences

University of Rochester
Rochester, New York

2016

Biographical Sketch

Alex Page was born in Portland, Maine, USA. He studied Computer Engineering, Physics, and Applied Mathematics at the University of Rhode Island, where he received an award for creativity in engineering, and was also recognized as the top Physics student in his graduating class. After graduating with his Bachelor of Science degree in 2011, he began graduate studies at the University of Rochester in 2012 as a Sproull Fellow, and earned his Master of Science degree in Electrical Engineering in 2015. At the University of Rochester, he served as a teaching assistant in GPU Programming and Circuits & Microcontrollers courses, while researching computer systems for medical data processing under the direction of Professor Tolga Soyata.

The following publications were a result of work conducted during doctoral study:

[1] A. Page, O. Kocabas, T. Soyata, M. Aktas, and J.-P. Couderc, "Cloud-based privacy-preserving remote ECG monitoring and surveillance," *Annals of Noninvasive Electrocardiology*, vol. 20, no. 4, pp. 328–337, 2014.

[2] A. Page, O. Kocabas, S. Ames, M. Venkitasubramaniam, and T. Soyata, "Cloud-based secure health monitoring: Optimizing fully-homomorphic encryption for streaming algorithms," in *2014 IEEE Globecom Workshops (GC Wkshps)*. IEEE, 2014, pp. 48–52.

[3] M. Hassanalieragh, A. Page, T. Soyata, G. Sharma, M. Aktas, G. Mateos, B. Kantarci, and S. Andreescu, "Health monitoring and management using internet-of-things (IoT) sensing with cloud-based processing: Opportunities and challenges," in *Services Computing (SCC), 2015 IEEE International Conference On*. IEEE, 2015, pp. 285–292.

[4] A. Page, M. Hassanalieragh, T. Soyata, M. K. Aktas, B. Kantarci, and S. Andreescu, "Conceptualizing a real-time remote cardiac health monitoring system," in *Enabling Real-Time Mobile Cloud Computing through Emerging Technologies*. IGI Global, 2015, pp. 1–34.

[5] S. Ames, M. Venkitasubramaniam, A. Page, O. Kocabas, and T. Soyata, "Secure health monitoring in the cloud using homomorphic encryption: A branching-program formulation," in *Enabling Real-Time Mobile Cloud Computing through Emerging Technologies*. IGI Global, 2015, pp. 116–152.

[6] A. Page, M. K. Aktas, T. Soyata, W. Zareba, and J.-P. Couderc, "QT clock to improve detection of QT prolongation in long QT syndrome patients," *Heart Rhythm*, vol. 13, no. 1, pp. 190–198, 2016.

[7] A. Page, T. Soyata, J.-P. Couderc, M. Aktas, B. Kantarci, and S. Andreescu, "Visualization of health monitoring data acquired from distributed sensors for multiple patients," *Global Communications Conference (IEEE GLOBECOM 2015)*, 2015.

[8] A. Page, T. Soyata, J.-P. Couderc, and M. Aktas, "An open source ECG clock generator for visualization of long-term cardiac monitoring data," *IEEE Access*, vol. 3, pp. 2704–2714, 2015.

[9] A. Page, S. Hijazi, D. Askan, B. Kantarci, and T. Soyata, "Research directions in cloud-based decision support systems for health monitoring using internet-of-things driven data acquisition," *International Journal of Services Computing*

(IJSC), vol. 4, no. 4, pp. 18–34, 2016.

[10] G. Honan, A. Page, O. Kocabas, T. Soyata, and B. Kantarci, "Internet-of-Everything Oriented Implementation of Secure Digital Health (D-Health) Systems," in *IEEE International Symposium on Computers and Communications (ISCC)*, Messina, Italy, July 2016.

[11] A. Page, J. Hellawell, P. Yue, L. Belardinelli, W. Zareba, T. Soyata, and J.-P. Couderc, "Evaluating the effect of a novel late sodium current inhibitor (eleclazine) on the QT, QTpeak, and TpTe intervals in LQT3 patients using the QT clock concept," in *2016 Computing in Cardiology Conference (CinC)*, 2016.

Acknowledgments

I have many people to thank for getting me here. So many professors and classmates have stood out in both undergraduate and graduate school. I thank the entire Electrical, Computer & Biomedical Engineering Department at the University of Rhode Island, as well as the Departments of Physics, Math, and Computer Science. At the University of Rochester, I would particularly like to thank Professors Dery, Huang, and Sobolewski for their efforts in bringing me to Rochester. I am also grateful for the guidance of Michele Foster and Professor Heinzelman, and for Professors Soyata and Bocko, who gave me everything I needed to succeed.

Thank you to my many collaborators and labmates for sharing in the work and experience, and to my rock climbing partners for their friendship and support these last two years. I am also proud to be a part of Interlock and BSidesROC, whose members continually humble me with their knowledge and creativity.

Most importantly, I would like to thank my family, whose support began long before college, and Erin O'Toole, who has shown more care and understanding in this journey than I can ever repay.

Abstract

The advent of portable medical sensors such as electrocardiograms has enabled widespread remote monitoring. The data collected from such devices presents new research opportunities and challenges to improve diagnosis and treatment for every individual patient. Condensing the sensor data into a form that is useful for doctors and researchers is the focus of this work. In particular, I investigate techniques to streamline the handling and processing of large data sets, the utility of long-term monitoring data (such as Holter ECG recordings) in decision support, and the application of machine learning in risk stratification. To support this process, I explain a system for automatically delineating key markers in ECG recordings, and displaying them in a novel way for identification of anomalies or patterns. I then show how this method can be used in drug studies, as well as to unmask potential problems in patients with certain genetic disorders. Finally, machine learning and statistical analysis techniques are used to identify patients who have a higher risk of symptoms or who require urgent care. In the process, many challenges with the use of ECG data in machine learning algorithms are addressed.

Contributors and Funding Sources

This work was supported by a dissertation committee consisting of Professor Soyata (advisor), Professor Couderc, and Professor Doyley of the Department of Electrical and Computer Engineering, and Doctor Aktas of the Department of Medicine. Graduate study was partially supported by a Sproull Fellowship from the University of Rochester. Professor Couderc and Doctor Aktas supplied the ECG databases, and outlined and explained the clinical challenges. All other work was completed independently by the student.

Table of Contents

List of Tables

List of Figures

1 Introduction

Heart disease is the leading cause of death in the United States [12]. Treatment of heart disease can require massive amounts of data collection and aggregation in order to determine the best course of action for an individual patient. Management of that data poses many challenges in terms of privacy, storage/computing constraints, analysis, and synthesis [10]. My aim is to develop solutions for data handling problems in cardiology, in order to provide better decision support.

The infrastructure for medical remote monitoring is improving every day. Several portable, inexpensive sensors including electrocardiograms and glucose monitors are now available (the Clearbridge VitalSigns CardioLeaf [13], for example). The data from these devices enables physicians and researchers to view a more complete picture of their patients' health than what would be possible in the hospital. A brief clinic checkup may not reveal important long-term patterns, for example, and even an extended hospital stay will not necessarily provide data representative of a patient's experiences in their normal daily routines. Not only is there a diagnostic incentive for long-term remote monitoring, but there is also a financial one; in-hospital tests are typically more expensive. Obviously, long-term

remote monitoring is very beneficial, but the data must be condensed in order to be manageable; a doctor cannot possibly review a table with 120,000 rows of cardiac parameters per patient per day, for example.

I will investigate efficient methods for data storage and processing ("backend"), as well as visualization and decision support ("frontend"). The goal in both ends is to reduce diagnostic time even while facing a torrent of incoming data, which may require faster processing, machine-learning-based diagnostic assistance, or more comprehensive visualization methods. Findings will be presented in the context of cardiology case studies.

1.1 Electrocardiography

Electrocardiograms allow us to measure the electrical activity of the heart via sensors placed on the surface of a subject's skin. The data from an ECG is simply the signal amplitude (voltage) vs. time, as seen in Figure 1.1. This signal may be passed into an automated annotation algorithm, which will demarcate different intervals such as RR. Physicians can then analyze the annotation statistics to check, for example, average heart rate, or maximum QTc. (The meaning and relevance of QTc will be discussed in Section 3.2.)

ECG technology dates back over 100 years, but has improved in many ways during that time. One relatively modern version of this is the *Holter monitor* [14], a portable ECG that can digitally record a patient's ECG over a period of several days. These recordings, often called *Holters*, can provide insight into the patient's daily "cardiac routine", as opposed to a clinic checkup where the doctor only views the ECG for a few seconds.

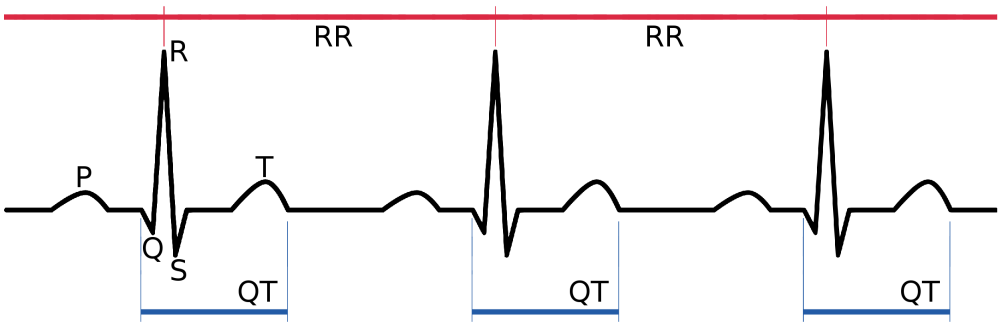


Figure 1.1: ECG tracing with some features labeled. The main features are the P, Q, R, S, and T waves/deflections. The distances (i.e. times) between some of these points are clinically important to monitor, such as RR (heart beat duration) and QT (ventricular repolarization interval).

Because 24-hour recordings contain many pieces of information for each of ∼120,000 heart beats on several different leads, the recordings are rarely looked at as a whole; instead, they are summarized in ways that are likely to omit important information. In our research, we will strive to use/present *all* of the information in a 24-hour recording.

1.2 Telemetric and Holter Warehouse (THEW)

The Telemetric and Holter Warehouse (THEW) is a collection of long-term ECG recordings maintained by a group at the University of Rochester [15]. I first used this database as a source of real ECG data for a medical data privacy project [1]. However, while performing statistical analysis on ECG recordings in the THEW, we realized that there was a great opportunity to study various cardiac features by hour, day of week, season, etc. Time and date can be extracted from Holter recordings, yet such analyses are rarely presented. With a feature like heart rate (HR), for example, we can provide much more than simply the mean or standard deviation in a given population; we can compute these statistics for *every minute*

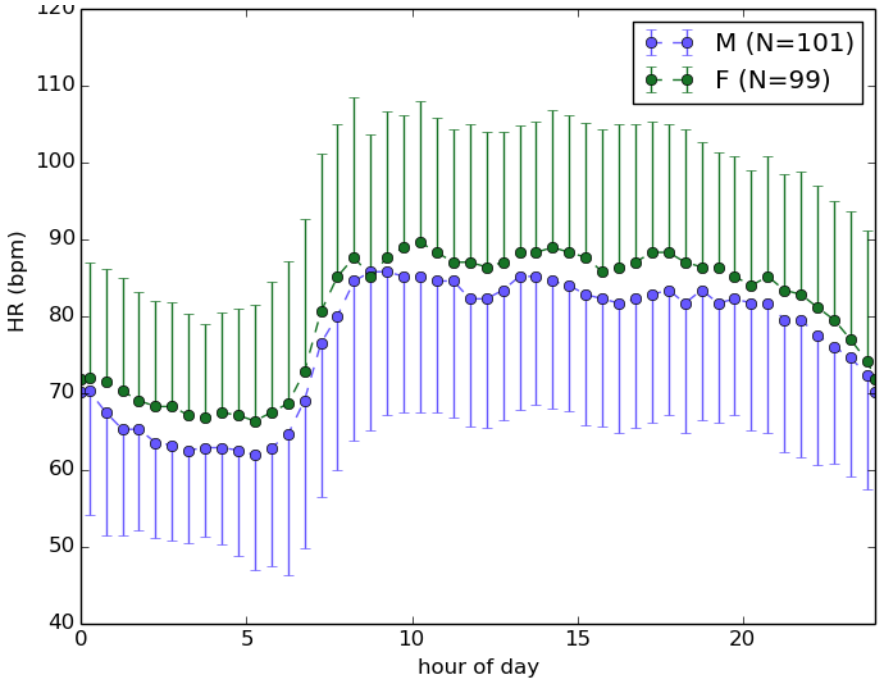


Figure 1.2: Median heart rate in healthy individuals (beats per minute). Hour 0 is midnight. Error bars indicate standard deviation, and are drawn in only one direction to avoid overlap.

of the day, yielding more accurate reference values for clinical use. Futher, by using the demographic information in the THEW database, we are also able to separate results by age, gender, the presence of certain drugs, and other factors. Our analysis of heart rate in 200 healthy subjects revealed interesting features in the daily cardiac cycle, such as the difference and transition of HR between night and day, and maxima and minima around meals [4]. One result of this work is shown in Figure 1.2. Various THEW databases will be utilized for all case studies in the future chapters.

1.3 Remote Health Monitoring

In remote health monitoring, we attempt to pass as many data crunching (i.e. automatable) tasks to machines, while leaving *reasoning* tasks to the human doctor. By drastically augmenting the computational capabilities of smartphones (or other sensory devices [16]) with the help of cloud-based processing, today's most sophisticated Machine Learning (ML) algorithms can be executed on real-time monitoring data and present results to a doctor for decision support [7, 17]. The unmatched ability of a human reasoning system can then make decisions that not only include the medical experience of the doctor, but also the computationally-intensive analytical processing results of the machine — tasks that the human brain is not good at.

A perfect example of a system that can provide such decision-supporting feedback to a doctor is a remote health monitoring system, consisting of Internet-of-Things (IoT)-based data acquisition devices, connected to a cloud-based decision support system [3]. This system could significantly improve diagnostic accuracy, healthcare quality, and the patient's quality of life. The technological components of such a system are maturing; significant research momentum is making advanced medical data acquisition devices commercially available [18], while sophisticated and powerful ML algorithms are well understood [19] and are readily available [20–22]. Once integrated remote health monitoring systems are available to incorporate these individual pieces, the Digital Health (D-Health) revolution can begin in earnest; by analyzing massive amounts of data, not only can mechanisms that cause diseases be better understood [6] — for example, by understanding the genes that contribute to them — but also personalized treatment is possible, which targets the treatment towards the individual lifestyle and genetic

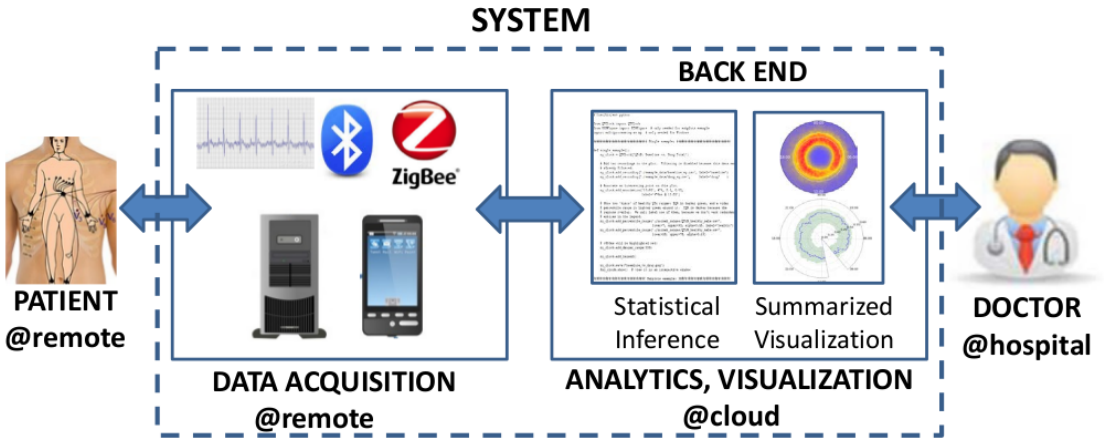


Figure 1.3: High-level overview of a remote health monitoring system.

make-up of the patient. It is also worth noting that remote sensing is actually less costly for the healthcare organization, as the patient does not need to occupy a hospital bed.

In order to understand the big picture of how medical data flows from a remote patient to the doctor, we will present an overview of such a system. This will also allow us to isolate the pieces of the system that we will be researching.

Figure 1.3 shows the layout of a medical remote monitoring system. We can break it into four main components:

1. Patient area: This includes sensors and short range wireless communication hardware.

2. Acquisition: This includes larger computers (phones, PCs) involved in aggregation/pre-processing and long-range data transmission.

3. Back End: This is where data is stored/processed, including generation of detailed reports/visualizations.

4. Doctor: This is the presentation layer or "front end" interface application which provides alerts, reports, etc.

1.3.1 Sensing

In the first parts of the system, sensory recordings of the patient will be transmitted from the patient's house (or any remote location) to the datacenter of the health care organization (HCO) in real-time in a secure fashion using well established encryption mechanisms [23]. The measured data from the sensors is sent to a data concentrator. The concentrator is typically a smart phone which is in the vicinity of the patient. Biosensors along with the concentrator form an IoT based architecture [9]. Similar to other existing work in the domain of IoT based medical sensing [24,25], access to individual sensor information is realized through an intermediate data aggregator. Combining ECG monitoring parameters with additional bio-markers improves the utility of the monitoring system to far beyond what is currently achievable with ECG-only monitoring [26] or single-biomarker monitoring (e.g., Glucose [27]).

1.3.2 Cloudlet

Because some tasks in the data acquisition block of Figure 1.3 may be overwhelming for a smartphone — especially when battery life is considered — it could be desirable to bypass the phone, replacing it with a nearby PC (connected to the patient's home power and Internet) instead. A fast, local computer in this "cloud interface/helper" configuration is sometimes referred to as a "cloudlet". [28–32]

There are many advantages to this approach; most importantly, a cloudlet is more capable than a phone in terms of hardware — both for speed, and extensi-

bility. It is also likely to have access to a faster, more reliable Internet connection, and the patient does not have to carry it or charge it. And cost-wise, a computer such as an Intel NUC [33] would be comparable to the smartphone approach. The primary disadvantage of a cloudlet compared to a phone is that it limits the mobility of the patient, requiring that they stay within range of a fixed location. Another possible disadvantage is setup time; since some patients will not know how to connect a new computer to their home network, a technician may need to set the cloudlet up for them. (This could potentially be solved by using a 4G cellular card in the cloudlet, rather than the patient's Internet connection.) Note that we are not overly concerned about the reliability of residential Internet access in this type of monitoring system; if latency and reliability were critical, the patient would be kept in the hospital.

It is also possible to use the patient's *own* computer as the cloudlet, i.e. to install an application on their home computer that performs the energy-heavy tasks, and receives sensor data via (e.g.) a USB Wi-Fi adapter. However, as there are no guarantees of speed, security, or reliability on a patient's personal computer, it is recommended instead that the hospital maintain a set of preconfigured cloudlet PCs. These PCs would be issued for remote-monitoring sessions, then wiped/reconfigured for the next patients.

1.3.3 Data Analysis and Reporting

In the later stages of the system, which are where this research lies, the data is analyzed as prescribed by the doctor. For instance, we may simply be looking for a change in status represented by a parameter crossing a threshold (e.g. "glucose too high"). Or, the doctor may want to confirm a diagnosis by comparing this

patient to others in the database, either visually or using statistics or machine learning.

While the comprehensive nature of this system substantially improves its diagnostic value, it introduces challenges in handling such a high volume of data. Visualization of such multi-dimensional data encompassing ECG parameters and multiple bio-markers is not straightforward. Well known ECG-based visualization of a patient's cardiac operation has been in use for over a century, but provides limited information for a short operational interval. In Section 3.3, we will develop a new mechanism that allows the doctor to visualize ECG measurements over $\geq$24 hours.

2 ECG Data Handling

In this chapter I will discuss the tools and workflow used to process and manage thousands of ECG recordings and their associated patient information.

2.1 Processing Individual Recordings

ECG sensors measure voltage at resolutions of about 16 bits and rates that generally range from 180–1000 Hz. A typical system will include up to 12 of these sensors (i.e. leads) in different positions on the skin. File sizes are therefore roughly 100MB per patient per lead per day [34]. A hospital monitoring 1000 patients will likely generate 100+ GB of raw data per day. (Though, this is usually compressible down to less than 40% of its original size using a utility like `zip`. More specialized compression algorithms may be used to take advantage of the specific data structure of ECG to achieve higher compression rates).

A typical ECG waveform for three heartbeats was shown in Figure 1.1. Each lead of the ECG monitor will capture between about 100 and 1000 amplitude samples during a single beat. What the doctor is interested in, though, are values

like the heart rate or the QT interval. There are roughly 20 such measurements we may want to store for every heart beat, such as the location or amplitude of certain features in the QRS complex. These annotations may very well occupy more space than the original recording. Accordingly, the hospital may need to plan for a couple of terabytes per day to cover their 1000 patients.

From the raw ECG data (in ISHNE format [34]), we must build a hierarchical database that has the original data at its lowest layer, commonly-requested features such as heart rate at the highest layer, and primitives such as "R peak locations" in between. This structure allows us to generate results more quickly than building them from the raw data on every request, and it also allows us to standardize the interface to clinically-relevant features at the highest layers when dealing with different types of sensors. We construct the database for our LQTS application (Chapter 3) in two major steps.

Stage 1: Raw Data → Primitives The primitives will be extracted from the ECG recording, resulting in a table like that of Figure 2.1. This process is performed by a DSP algorithm, which, in the case of ECG analysis, is typically based around wavelet transforms. The computational complexity of this stage is an important factor in determining where it should run; as alluded to Section 1.3.2, there is a tradeoff between running it in the datacenter vs. on a device nearer to the patient. The primitives for most interesting ECG-based values are the locations and amplitudes at the start, beginning, and end of the labeled features of Figure 1.1 (P, Q, R, S, T).

Stage 2: Primitives → Clinical Markers From the table in Figure 2.1, the distance between different features can be computed. Many of these values

time	lead	feature	amplitude
15:00:05.135	1	(P	0.15
15:00:05.150	2	(P	0.36
15:00:05.165	1	P	0.19
15:00:05.175	2	P	0.34
15:00:05.190	1	P)	0.15
15:00:05.195	2	P)	0.37
15:00:05.230	1	(Q	0.12
15:00:05.245	2	(Q	0.43
15:00:05.270	2	R	0.96
15:00:05.280	1	R	1.68
15:00:05.335	2	S)	0.32
15:00:05.340	1	S)	0.12
15:00:05.445	2	(T	0.57
15:00:05.540	2	T	0.69
15:00:05.560	1	(T	0.24
15:00:05.630	2	T)	0.55
15:00:05.640	1	T	0.42
15:00:05.715	1	T)	0.26
15:00:06.145	1	(P	0.04
15:00:06.170	1	P	0.12
15:00:06.170	2	(P	0.32
15:00:06.195	2	P	0.31
15:00:06.210	2	P)	0.31
15:00:06.215	1	P)	0.07
15:00:06.250	1	(Q	0.09
15:00:06.265	2	(Q	0.38
15:00:06.285	2	R	0.96
15:00:06.300	1	R	1.59
15:00:06.355	1	S)	0.06
15:00:06.355	2	S)	0.36
15:00:06.460	2	(T	0.55
15:00:06.555	2	T	0.66
15:00:06.575	1	(T	0.24
15:00:06.650	2	T)	0.53
15:00:06.660	1	T	0.33
15:00:06.735	1	T)	0.14

Figure 2.1: ECG processing: output of first stage. The raw recording has been turned into a list of important feature locations and amplitudes.

time	lead	interval	value
15:00:05.230	1	PR	0.095
15:00:05.245	2	PR	0.095
15:00:05.270	2	RRbpm	54.54
15:00:05.280	1	RRbpm	55.04
15:00:05.335	2	QRS	0.09
15:00:05.340	1	QRS	0.11
15:00:05.630	2	QT	0.385
15:00:05.630	2	QTcB	0.367
15:00:05.630	2	QTcF	0.372
15:00:05.715	1	QT	0.485
15:00:05.715	1	QTcB	0.464
15:00:05.715	1	QTcF	0.471
15:00:06.250	1	PR	0.105
15:00:06.265	2	PR	0.095
15:00:06.285	2	RRbpm	59.11
15:00:06.300	1	RRbpm	58.82
15:00:06.355	1	QRS	0.105
15:00:06.355	2	QRS	0.09
15:00:06.650	2	QT	0.385
15:00:06.650	2	QTcB	0.382
15:00:06.650	2	QTcF	0.383
15:00:06.735	1	QT	0.485
15:00:06.735	1	QTcB	0.480
15:00:06.735	1	QTcF	0.481

Figure 2.2: ECG processing: output of second stage. Important intervals such as QT and RR have been measured using the output of the previous stage (Fig. 2.1). We may also want to record some amplitudes here (e.g. T wave height, ST elevation) to avoid having to refer back to the lower level data.

("segments" or "intervals") are good markers of potential cardiac issues. We may also be interested in values *computed from* intervals; heart rate in beats per minute, for example, is $\frac{60}{\text{RR (sec)}}$. RR is measured from R to R in consecutive beats. Finally, some *amplitudes* such as T wave height or ST segment elevation may also be needed.

Raw ECG data comes in many formats (ISHNE, text, MIT, dat, etc.); in the THEW databases [15], the International Society of Holter and Noninvasive Electrocardiology (ISHNE) format is used. We use an open source annotation algorithm developed by Yuriy Chesnokov [35] to perform Stage 1. This algorithm was written for a PhysioNet QT Measurement challenge, and we therefore expect it to be well-suited to our LQTS objectives. This award-winning open-source algorithm was tested on 548 sample ECG recordings and delivered a 17.30ms RMS error (i.e. $\sim$4% error) in QT interval measurement. The algorithm also provides

fiducials for the PR interval and QRS complex for all ECG leads. We ported this code from Windows to Linux, and modified it to read ISHNE-formatted data files from the THEW database. The resulting code is much more cross-platform than the original — i.e. it is closer to being executable on a smartphone or embedded microprocessor, which will be helpful during later development. It is also very fast compared to other software such as ECG-kit [36], which is important with the amount of data we need to process.

Once all primitive features have been annotated, we can compute the values of interest such as QT and heart rate. Although these computations are relatively simple – e.g. subtracting Q from T – there are ∼100,000 heart beats per patient per day, detected on 3 separate leads. This begins to add up to a lot of computation if we wait until the doctor asks for it. Further, if we want to aggregate results, perhaps to see the average heart rate for a group of 1000 people, we are much better off having pre-computed it across each recording. So this step will save a lot of time for future queries. These results are stored in a table in a SQLite database associated with each recording. This database allows us to perform queries such as "What percentage of the time was the patient's heart rate above 90bpm?" or "What was the patient's average QTc between 2 and 3 AM?". This is an improvement on the current system, which may only provide min/average/max heart rates from an entire ECG recording. However, we can improve this even more by providing a view of the *entire* period (e.g. 24 hours); this will be explored in Section 3.3

Since we have data from multiple leads — usually 2–3, but sometimes up to 12 — there is a lot of redundant information in the annotations. To create a leaner and more accurate table of annotations, we merge this information by

simply looking for clusters of a feature (i.e. detection of the same feature on several leads at about the same time) and choosing the "best" measurement to keep (or perhaps a weighted average). With "R" detection, for example, we will keep the median time of detection across all leads. The criteria for determining the best measurement may depend on the diseases of interest, signal quality, or lead placement. With QT, we are concerned about large values, so a conservative choice may be to look at the max QT rather than median. Further, QT detection relies on strong T wave presence in the signal, so we may choose to measure QT and related features on the lead with the strongest T waves. Lastly, we will throw out any unreasonable values (such as heart rate of 25bpm or QT of 800ms) and also any values that were computed in the vicinity of ectopic heart beats. This cleanup process could be considered "Stage 3".

In our work, we maintain separate SQLite databases for each Holter, containing tables with primitive locations and clinical marker values. (In the long term, a different database system such as MySQL [37] or MariaDB [38] will likely be a better solution, but for now, SQLite simplifies portability across our test systems.) Why might we need the lower-level data later? Machine learning may benefit from things we didn't expect, we may need to look for a different disease than originally intended, we may want to hand-check data that was used for a decision, verify annotations, apply some improved function to a layer, etc. Maintaining all levels of data saves processing time in these types of "rework" by allowing reuse of higher-layer data instead of returning to the raw recording.

2.2 Separating Patients into Cohorts

One fundamental task that we must achieve is being able to easily analyze subgroups of a large database, in order to compare them to each other. In Fig. 1.2 for example, we separate men and women in the THEW E-HOL-03-0202-003 database to compare their heart rates.

To allow us to separate ECGs into different groups, we use metadata provided with each THEW database. The databases each contain a spreadsheet with demographic information about the subject of each Holter. This information may include things like weight, gender, history of hypertension, etc., and the fields are not consistent across databases. From these spreadsheets, I created a single database table that merges all of the information into one location. Using this database, I can sort/query across recordings from four different THEW databases. For example, I may want to create 4 groups of Holters — healthy females, healthy males, LQTS females, LQTS males — and can do this from a single interface.

While a primary goal of this system is to provide a picture of an individual's health status to the doctor, this assistance can be greatly improved by allowing comparison of the individual to various groups (such as other people of the same age or gender, taking the same medications). In order to perform these comparisons, we need to develop "norms" for different populations by utilizing existing databases (such as the THEW [15] or PhysioNet [39] databases). To this end, we have found that it is beneficial to store a set of pre-computed statistics for each major population. Populations are segregated by characteristics that are known to influence each feature, such as gender or age. For each population and feature, we store the percentile values of the feature at regular intervals throughout the day — one set of values per minute has been sufficient for our use. Table 2.1

Table 2.1: Heart rate range in healthy males. For every 1-minute window of the day, all applicable heart rate values in the cohort are added to a list. Then, the percentiles are computed. In this case, we look at 10% intervals, but in our actual data, we maintain these tables down to 1% intervals.

time	10%	20%	30%	...	70%	80%	90%
00:00-00:01	57	61	65		77	83	94
00:01-00:02	55	60	64		78	85	95
...							
11:23-11:24	65	71	75		94	99	107
...							
23:59-24:00	56	61	64		78	85	94

shows a subset of one such data set.

Incidentally, though typical ranges for QTc (and other ECG parameters) have been thoroughly investigated [40], reference ranges that take precise time of day into account do not exist. Because of the time-dependence of LQTS that we will observe in the next chapter, we have developed our own reference ranges from some of the THEW databases. Some (summarized) tables are provided in Appendix A. (You should be warned, however, that these ranges were generated by an annotation algorithm that has not been used in a clinical setting, operating on a relatively small database.)

3 Decision Support for Long QT Syndrome (LQTS) Patients

In this (and the following) chapter, we will discuss the specific medical conditions that we intend to provide assistance with, the methods we use to do so, and results of several case studies.

3.1 Objectives

From the methods of Chapter 2, we have obtained useful ECG measurements from the raw data. We must now present and interpret these measurements.

There are many diseases that can be detected via ECG. Our novel course of research will include the use of 24-hour ECG data, rather than the conventional 10-second snapshots, to attempt to provide more relevant information to the doctor. This restricts the scope of diseases we can investigate to those that manifest with some slow variation or circadian rhythm. In the Long QT Syndrome (LQTS), patients with one genetic defect, LQT2, tend to show symptoms much more strongly during sleep. LQTS is normally diagnosed based on two features — QT and RR (see Fig. 1.1) — and a value computed from them called QTc.

Perhaps a machine learning algorithm could improve diagnostic accuracy by including other features as input, or by computing QTc differently. Or, without involving machine learning, we can attempt to provide visual representations of QTc for a 24-hour period that will also greatly aid a physician.

From the available data for LQTS patients, we define the following objectives:

1. Visualize and compare long-term QTc values of individuals and groups [7].
2. Apply machine learning methods to classify patients as:
   (a) healthy, LQT1, or LQT2,
   (b) at risk, or not at risk for events.
3. Investigate ways to unmask QT prolongation concealment.
4. Visualize and quantify the effects of a new drug on QT and its subintervals.

In the simplest cases, we may only be interested in QTc crossing a threshold. More realistically, though, we will need to look for trends.

One tool we will develop, the QT clock (Section 3.3.2), may be used to monitor patients receiving known QT-prolonging drugs, specifically during inpatient dofetilide initiation when the dose-dependent effect of the drug needs to be assessed over a couple of days. The QT clock may also be used during the ambulatory outpatient period to assess the long-term safety and efficacy of a QT prolonging drug, or the benefits of a QT shortening drug. This is what we will explore in relation to Objective 4.

3.2 Background on LQTS

3.2.1 The QT and QTc Intervals

The QT interval indicates how long the heart's ventricular repolarization process takes. QT varies with heart rate, so it is usually corrected based on the current heart rate. The corrected version, called QTc, is more stable and is usually around 400 ms in a healthy person, and it may go up to 500 ms or even higher with LQTS. QTc is usually computed with one of the following two equations:

$$QTcB = \frac{QT}{\sqrt{RR/s}} \tag{3.1}$$

$$QTcF = \frac{QT}{\sqrt[3]{RR/s}} \tag{3.2}$$

where the "B" and "F" indicate that these are the Bazett [41] and Fridericia [42] corrections, and the division by one second is to preserve the units of QT. Prolonged QTc is an important marker for potentially fatal events [43], and subjects with prolonged QTc are said to have *Long QT Syndrome* (LQTS). The normal range of QTcB is roughly 356–449 ms in men, and 369–460 ms in women [40]. QTc may be prolonged by drugs or because of genetic factors. LQTS is responsible for an estimated 3000–4000 sudden deaths in children and young adults in the US every year [44]. At least thirteen genes have been identified that contribute to Long QT; LQT1 and LQT2 are the most common [45]. In both the congenital and drug-induced cases, QT prolongation increases the risk of torsades de pointes (TdP), an arrhythmia which can lead to serious issues including syncope, fibrillation, and death [46]. QTc is therefore an important value for cardiologists to monitor, particularly on patients with one of the LQTS mutations, or patients

who have been prescribed drugs that are known to prolong the QTc interval.

3.2.2 Drug-Induced LQTS

As of today, there are hundreds of drugs available on the US market that can slightly prolong the QTc interval. While these drugs are generally safe, the accumulation of their small QT effects when patients are prescribed with multiple drugs can become a health concern. [47] The acquired LQTS is modulated by more than just drug interaction; diet, genetic predisposition, and circadian variation of heart regulations are amongst a set of factors that can play a crucial role in the patient response to a drug with potential QT effect. Therefore, a solution is to continuously monitor patients and enable QT surveillance. This concept implies that QTc intervals are continuously assessed from an ambulatory ECG signal — a Holter — and corrected for heart rate.

3.2.3 Congenital LQTS

The congenital Long QT syndrome (LQTS) is an inherited channelopathy that is associated with a prolonged duration of ventricular repolarization, predisposing such patients to the occurrence of life-threatening ventricular tachyarrhythmias such as torsades de pointes and sudden cardiac death [48]. Hundreds of mutations have been identified in the LQTS, however mutations in the KCNQ1 and KCNH2 genes are the most commons forms of the disease — LQT1 and LQT2, respectively. LQT1 is associated with a reduction of the slow component of the late repolarizing potassium current (IKs), while LQT2 is associated with a reduction of the rapid component (IKr) [49]. Heart rate corrected QT prolongation (QTc),

type of mutation, gender, history of syncope, and age are recognized modulating risk factors associated with the LQTS [50].

3.2.4 Concealment of QTc Prolongation

The genetic mutations that can cause LQTS are denoted LQT1, LQT2, ... LQT13 [45]. The LQT2 and LQT3 mutations are associated with more cardiac events at night [51], when the heart rate is low (i.e. when RR is long), whereas LQT1 patients are more likely to experience symptoms during exercise [50]. Additionally, certain prescription drugs can prolong QT in ways that may not be fully characterized during clinical tests, resulting in more prolongation when the patient goes home than the doctor was able to predict from in-hospital monitoring. This means that daily summaries of QTc value reviewed by a doctor are unlikely to show the full scope of a patient's LQTS. When a subject has periods of prolonged QT that are not always present, we say that they have concealed LQTS. Goldenberg et al. [52] reported that about 25% of genotyped-confirmed patients have a concealed form of the disease, wherein genotype positive patients have an ECG tracing exhibiting normal range QTc intervals (QTc $\leq$440 ms) [52].

The severity of the functional defects and the phenotypic penetrance in the LQTS is modulated by numerous factors amongst which age, gender, type and location of LQT mutation play crucial roles [53]. But this also means that genetic tests do not fully capture individual risk. Further, while the standard 12-lead ECG is an important diagnostic tool used in the investigation, evaluation, and monitoring of patients with LQTS, its utility in predicting risk may be limited by the fact that QT concealment may give the false impression of low risk in certain subsets of patients. The presence of QTc concealment in LQTS has been

studied previously using standard 10-second ECG tracings. The limitation of such an approach is that these ECGs are usually acquired during clinical hours, and they may not provide the most accurate assessment of risk but rather should be considered a snapshot measurement of risk.

The cardiac events in LQTS have been shown to be strongly associated with triggers linked to inappropriate QT adaptation to changes in heart rate. Therefore, challenging the ventricular repolarization process in these patients using protocols to exacerbate the QTc prolongation have been proposed such as brisk standing, [54] exercise, or epinephrine challenge [54–56]. Such protocols exploit the fact that β-adrenergically enhanced potassium currents (IKr and IKs) are required to counterbalance the larger inward depolarizing current of the L-type Ca2+ current (also enhanced by β-adrenergic stimulation). [57] Abnormal regulation of these potassium currents under condition of an adrenergic tone put the LQT1 and LQT2 patients at risk for life-threatening ventricular arrhythmias. The cellular and environmental factors that affect repolarization are of course dynamic, and therefore a static assessment of QTc may be limited in conveying the true extent of abnormal repolarization for any given patient.

We envision a long-term remote-monitoring system that can upload ECG signals to the healthcare provider for automated analysis of QTc. Ideally, this system will provide a 24-hour picture to the doctor in a simple form containing all key information; i.e. we want to summarize, while avoiding under-sampling or over-averaging of the data.

3.3 Visualization Methods

One of the most useful types of decision support is not for a computer to generate specific recommendations, but to simply present the data in a manner that allows the doctor to fully understand the situation. Based on this presentation, the doctor can make his or her own decision. The challenge here is to condense many sensor measurements spanning a long period of time into a very concise summary.

An important consideration in building visual aids for decision support is knowing which features are relevant to the condition being investigated. In the Long QT Syndrome (LQTS), for example, many ECG measurements such as QT, RR, or TpTe (T wave peak to T wave end) may all carry some information about the disease, not just QT. Additionally, there are several ECG leads (sensor locations) to choose from, and certain leads may be better for QT measurement. We also know that LQTS manifests differently throughout the day based on patient genotype, so perhaps there are a few key times of day that should be checked (as opposed to looking at an overall average of the available data).

We are building a sizable array of factors that are relevant to this disease, and circling back around to the original problem: displaying it all to the doctor in a form that can be digested very quickly. Remember that in addition to each of these factors — ECG marker and lead, time of day, etc. — the doctor may also have 20–30 patients. Further, the advent of long-term remote monitoring means that each patient will be generating more data than ever before. So we will investigate new infographic concepts to summarize a patient's day.

The first techniques we will apply to LQTS monitoring involve the removal of redundant information from the ECG recording. For instance, while many ECG measurements may contain some information related to the patient's illness,

we may focus simply on QTc (which combines two measurements, QT and RR). Further, since many ECG leads are available, we will combine data from all of them using e.g. a median or average. (We could also choose to look only at a single lead, perhaps the least noisy, as discussed in Section 2.1.)

Now that we are focused on a single (computed) feature on a single ("merged") lead, our visualization problem is much more focused. We must plot or tabulate the values of QTc for ∼120,000 heart beats per day. Again, we know that certain times of day are more critical based on genotype. However, as they are based on sleep and exercise patterns, they will still vary significantly between patients. So, we would like to show the entire day if possible.

The most obvious way to present the remaining data would be to simply plot it. However, the scale of the plot must be determined to ensure that short duration events are still visible. In the case of LQTS, we are mainly interested in events that last for several minutes. It is therefore practical to plot a full 24 hours in a fairly "typical" plot size (e.g. "half page"), which allows us to see with at least 1-minute resolution.

Finally, we note that for data spanning 24 hours or more, polar axes can be beneficial. By using the angle of a polar plot to represent time of day, and the radius to represent the value of some feature, multi-day data can simply continue to circle around the plot. Even with single-day data, this representation makes it unnecessary to adjust the axes range to view different recordings. (For example, should the x axis start where the recording does, or at some other time like midnight?) We have found it best to standardize on a 24-hour polar axis. In this section, we will present our ideas to develop a scalable and intuitive visualization mechanism for long term QTc data that prevents data overload for the cardiologist

while preserving the original data as much as possible. The main method we designed we have dubbed the "QT Clock" or "ECG Clock".

Comparing an individual patient's QTc to different groups can help the doctor more confidently decide whether certain therapies (e.g., beta blockers or defibrillator implantation) should be prescribed. The 24-hour plots can also be useful as a diagnostic tool. If a doctor suspects LQTS based on a 10-second ECG during a clinical visit, (s)he can obtain a 24-hour Holter recording of the patient, and superimpose their patient's measurements onto various groups that they suspect the person may fall into.

The doctor will also be able to prescribe or recommend against certain activities; for example, they may notice an issue at 10AM on a Saturday, ask what the patient was doing at that time, find out that they were mowing their lawn, and advise against that activity. Data mining may also reveal bad habits common to almost all patients, e.g. going to bed too late on Sunday nights. Physicians would then be able to remind their patients of this hazard. We also foresee *predictive* capabilities using this technology. The effects of various drugs on markers such as QTc may be characterized based on drug trial results to predict the effects on a new patient. i.e., we can project ahead of time what the QTc (or other) plot will look like from a patient's baseline and from recordings of similar individuals, and even project what the impact will be in several years. Such predictions can change the protocols for prescribing certain drugs, saving patients and hospitals time and money while providing more accurate care for the patient.

To augment the ECG Clocks, we would like to add thresholds and statistics for comparison. However, static "warning" thresholds for cardiac features do not account for a patient's age, prescriptions, congenital disorders, for the time of

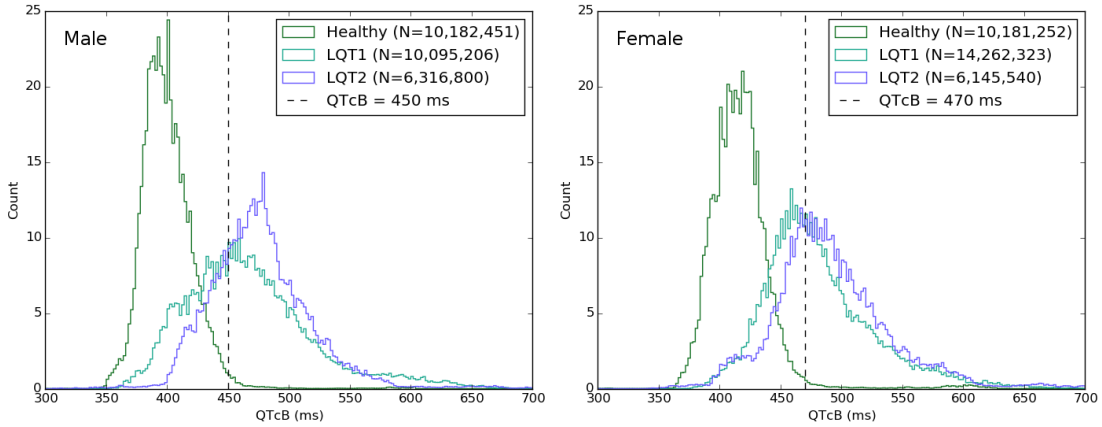


Figure 3.1: QTc distribution in healthy, LQT1, and LQT2 cohorts. We see that there is significant overlap of the LQTS patients with the healthy cohort, severely hindering the ability to diagnose the disease with a QTc threshold alone. 450ms and 470ms are typical clinical warning thresholds for males and females, respectively.

day, or many other factors. It is currently possible to build a database from drug trials and other clinical data sets (such as the THEW [15] or PhysioNet [39] databases), constructing a picture of the *norm* for features like QTc within a particular group. As more sensor data (with demographic information) becomes available due to continuous monitoring, these reference ranges will become very well defined for all populations. We have begun to generate pictures of the normal QTc ranges for several groups, using recordings from the THEW database. Some histograms from this work are given in Figure 3.1. Tables with other intervals and times are provided in Appendix A.

3.3.1 matplotlib

Much of our work involves plotting, at least in order to get an idea about what quantitative analyses to pursue. Because most of the scripts and machine learning tools we intend to use are written in Python, we look for a Python-based plotting

solution to standardize on. matplotlib is a Python library with a MATLAB-like interface that is capable of generating any plot we could conceivably need. It will also be used as a platform to develop the "ECG Clocks" we describe in the next section.

3.3.2 ECG Clocks

We developed a novel graphical concept designed to facilitate the review of a large set of QTc measurements and to provide an infographic understanding of 24-hour QTc dynamics. The QT clock is a circular plot representing a 24-hour clock (00:00 to 24:00) with midnight at the top of the clock. The radius of the clock represents the QTc interval values varying from 300ms to 600ms from the center to the perimeter. This format provides consistency when looking for problems that can depend on the time of day, and recordings that start at different times. (For monitoring intervals much longer or shorter than 24 hours, it may be advisable to "unwind" the plot into a typical Cartesian plane, though.) The clock is used to present information for two different purposes:

1. Viewing the expected QTc range for a population (i.e. their typical QTc values across the day).

2. Monitoring changes in QTc for an individual patient. QTc values extracted from a single Holter recording may provide insight about personal daily variation and period of maximum QTc prolongation.

Examples of the visualization we've just described are given in Figures 3.2 and 3.3. Because our data was still fairly noisy even after all the preprocessing steps,

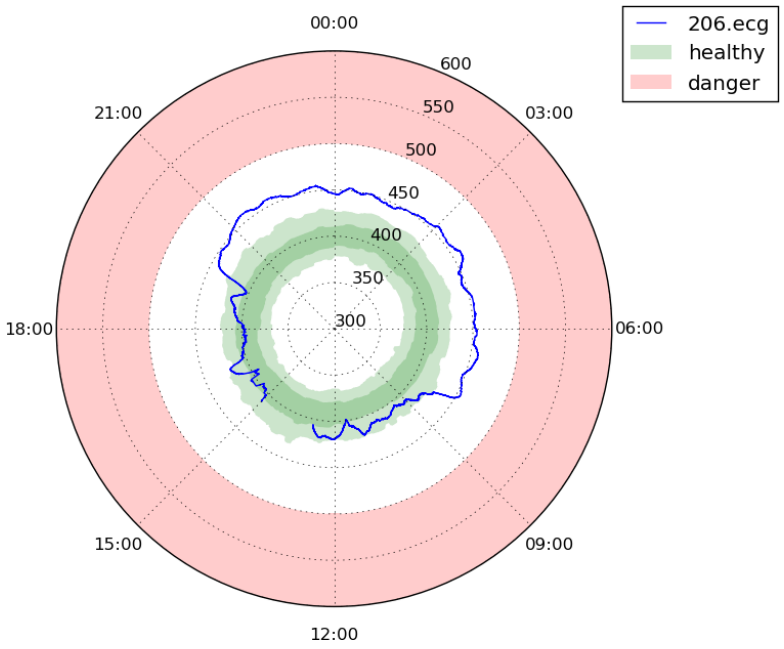


Figure 3.2: Example "QT Clock" to illustrate the ECG Clock concept. This plot is for a 1-year-old female with a genetic disorder. ECG clocks are polar plots of some feature value (e.g. QTc) vs. time. Time proceeds around the plot like a 24-hour clock. In this case, we have also highlighted some regions to indicate "good" and "bad" values.

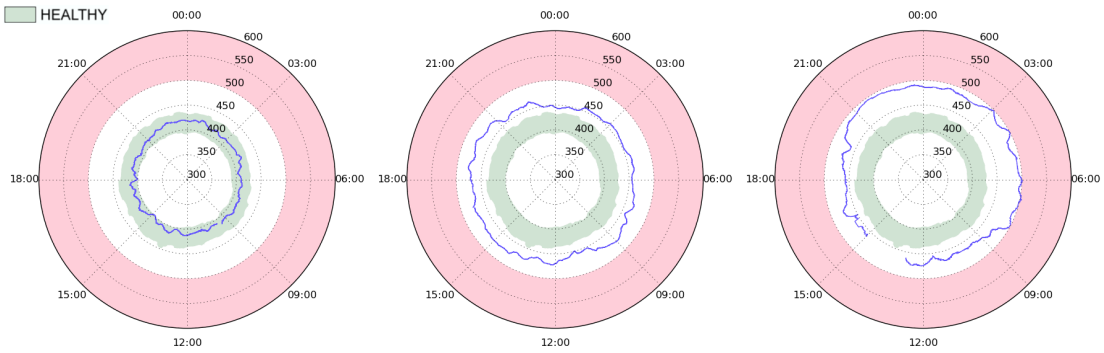


Figure 3.3: Example QTc clocks from three individuals. The first patient is healthy, and her QTc stays in the normal range all day. The second patient has the LQT1 defect, which manifests as slightly elevated QTc throughout the day. The third patient has a LQT2 defect, which is noticeably worse at night than during the day.

we used a median filter to smooth it. This type of filtering is described in the next section.

There are two particularly important features to note in the figures: (1) the blue line, representing the value of a single cardiac feature (QTc) for a specific patient, and (2) the green area, representing the range of normal values for that feature based on analysis of recordings from healthy subjects. This presentation has many uses in clinical and research areas, as we will demonstrate in the later sections.

The ECG Clock library is built on top of matplotlib, and its primary purpose is to generate plots of ECG interval values on a 24-hour axis. There is a wide range of applications for such plots, some of which will be demonstrated in Section 3.4. Users may modify the code if desired; it is released under the permissive MIT License, and is available at `https://bitbucket.org/atpage/ecgclock/`. Extending the functionality of the library to cover new cases usually only requires minor changes.

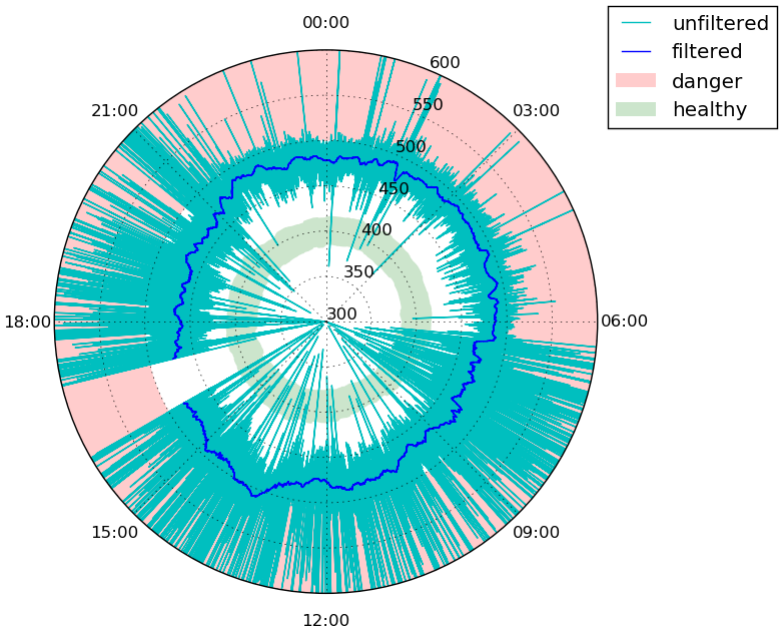


Figure 3.4: The importance of filtering beat-to-beat data, particularly with noisy annotations. Directly plotting ∼80,000 QTcF values for this 32-year-old female LQT1 patient yields the cyan line. Applying a 10-minute median filter produces the blue line.

Although an unlimited number of recordings can be added to the same plot — to view a patient's response to different prescriptions, for example — we find that the plots tend to get cluttered with more than 3–4 recordings. Incorporation of additional information (e.g. from more sensors, which are not necessarily cardiac) on the same axes is an ongoing research challenge [3]; we expect that plots of heart rate and QTc together, or QTc and TpTe, for example, will make it easier to gauge the interaction between related features.

3.3.3 Filtering

The plot in Figure 3.4 illustrates the effects of noise when we attempt to simply view QTc vs. time on one of our "clock" plots. Noise is not washed out like it may be in a histogram; a line is being drawn to every outlier, and even relatively small

error rates can produce a few thousand outliers over the course of a day (which consists of ∼100,000 heart beats). This is amplified by the fact that a single faulty detection can result in two incorrect values; with heart rate, for example, wrongly detecting an extra heart beat would make the heart rate appear to jump up for 2 beats and then return to normal. Further, QTc is somewhat dynamic; much of its variation isn't "noise". To smooth the plot, we apply a median filter to the list of QTc values, replacing each point with the median of the points around it. In other words:

$$F_i = \text{median}([O_{i-N}, \ldots, O_i, \ldots, O_{i+N}])$$

where F is the filtered signal, O is the original signal, and the width of the filter is 2N.

This approach will cause problems, though, if the doctor is interested in short-duration events; events that occur for less than ∼5 minutes, for example, are likely to be removed by the filter. The best solution for this is to collect a cleaner signal (e.g. using better sensors) and to apply more advanced annotation techniques. It is also important to eliminate errors at each stage in the database construction. Because there are so many data points to work with, it is generally safe to discard all questionable values. Relatively wide filters do not cause a problem for the QTc case study, but physicians will need to select filtering windows that make clinical sense.

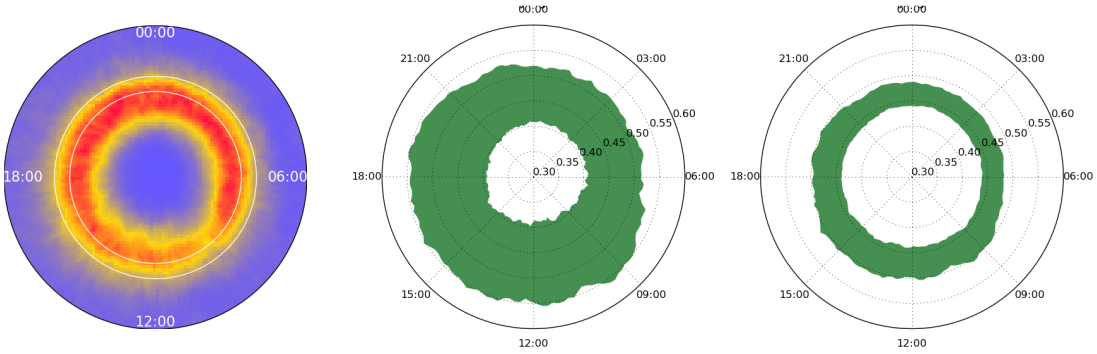


Figure 3.5: Early QT clocks in which we attempt to visualize the typical range of values for a feature (QTcB) in specific cohort (LQT1 female subjects who are not on beta blockers). We start with a 3D histogram (heat map) on the left, with white circles indicating the conventional "warning" range. In the center, we look at median +/- SD, and find that the standard deviation is so wide that the picture is not very helpful. Finally, on the right, we view median +/- MAD, which gives a clearer idea of what values are "normal".

3.4 Visualization Case Studies

3.4.1 Gender- and Genotype-Specific QTc Prolongation

In the histogram in Figure 3.5 (left side), we have plotted QTcB for every heartbeat from 94 24-hour recordings – approximately 10 million data points in total. We then produce a similar plot showing points within 1 standard deviation of the median as a solid color. Median is used rather than mean because we expect to have a non-negligible number of erroneous values in our data set due to the noisy environment and imperfect annotation algorithm, and we want to avoid giving weight to these bad values. However, these outliers still affect the standard deviation; the width of the band in the center plot is a result of this. Further, the standard deviation across multiple patients gives a false sense of how much variability is really normal for a single patient. To get a more representative view of QTcB, we produce the same plot using median absolute deviation (MAD) instead

of standard deviation. This results in the final plot in Figure 3.5.

Population and gender-specific distribution of QTc for our LQTS study cohorts are plotted on the QT clocks in Figure 3.6. The lower and higher boundaries of these patterns correspond to the 16th and 84th percentile (i.e., inner 68% of patients), with 1-minute resolution. This percentile range was chosen to highlight the equivalent of ± 1 standard deviation during each 1-minute epoch, without assuming a normal distribution. Such presentations reveal that there is a very stable, circular, "healthy" range of QTc across 24-hour variations (defined by our healthy cohort, dark green areas). We superimposed on these plots the 24-hour variation of QTc for LQT1 in the upper panels, LQT2 in the middle panels, and LQT3 in the lower panels. This figure reveals different configurations of 24-hour QTc dynamics according to LQTS types. The most striking one is the change in symmetry of the pattern in LQT2 and LQT3 groups revealing the exacerbation of the QTc interval prolongation during the night period. This phenomenon is stronger in LQT3 male patients while it is not present in LQT1 patients. Note that the "healthy" and "LQTS" regions rarely overlap, in contrast to what we see in Figure 3.1. This is mainly because the clocks only show the inner 68% of each group – the tails of their respective histograms are omitted. Additionally, Figure 3.1 removes all time dependence; overlap may be different depending on time of day.

In Figure 3.7, we expand the percentile range of Figure 3.6, which begins to introduce more noise due to the relatively small size of our database. However, some interesting features emerge, particularly the times of overlap in LQT1 and LQT2 with their healthy counterparts. For example, we see that LQT2 becomes quite distinct at $\sim$3AM, even though there is some concealment of prolongation

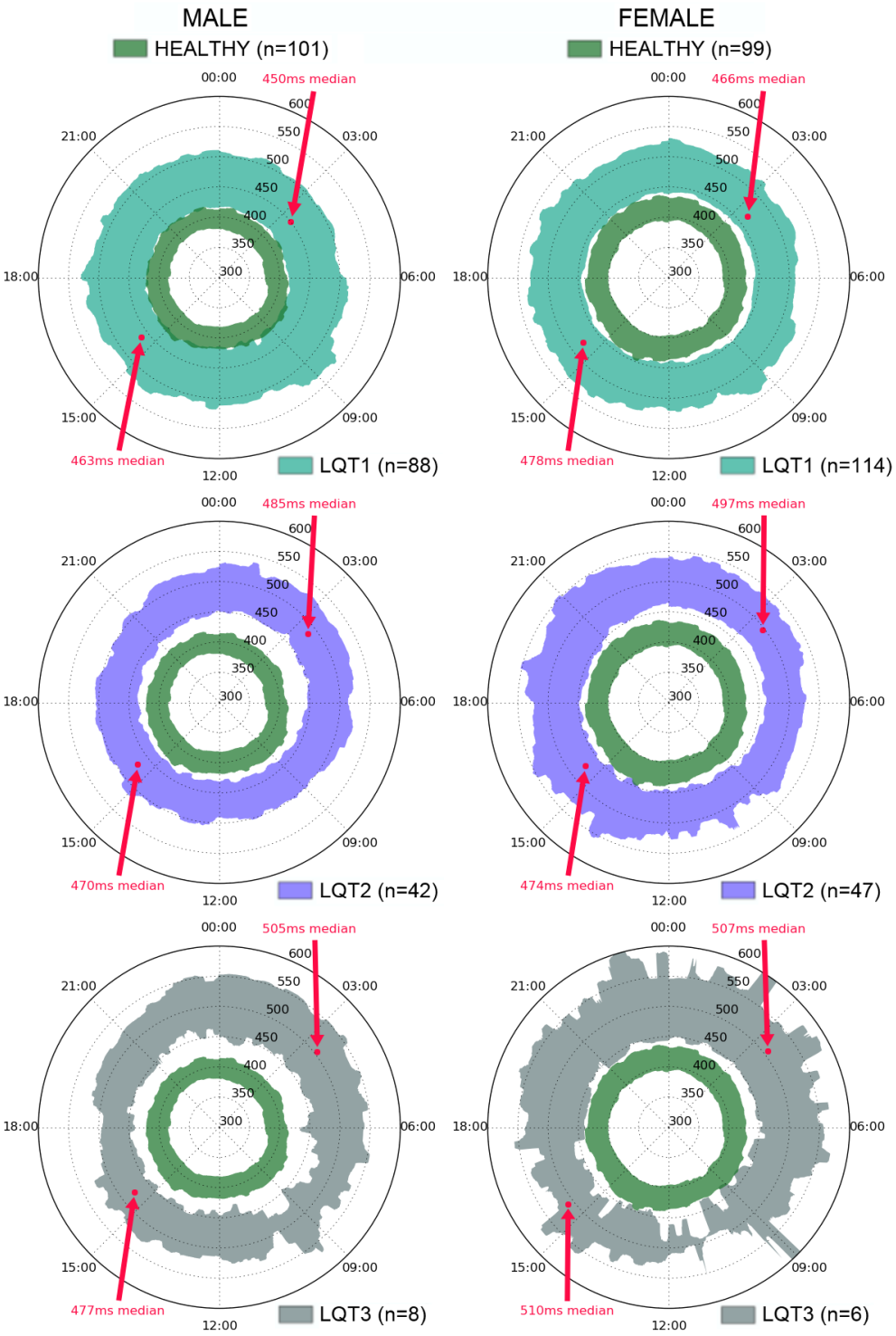


Figure 3.6: QTc clocks for LQT1, LQT2, and LQT3 cohorts compared with healthy peers. Note the asymmetry in the LQT2 and LQT3 cohorts, who experience greater QTc prolongation at night.

during the day. We also note that the healthy range remains relatively narrow, despite expanding the scope to the inner 90 percent.

3.4.2 Drug Effects

QTc clocks could also be useful for clinicians to see that a treatment is working, to adjust prescriptions, to look for harmful side effects, or to verify that a patient is taking their medication correctly. This would involve plotting and comparing QTc from a baseline Holter vs. a followup Holter. For example, dofetilide (Tikosyn) is a prescription drug known to prolong QT, so the doctor will want to see how much it prolonged QT over an entire day or week to make adjustments to the dose. Further, the effects of beta blockers are quite obvious on a heart rate clock; if a patient does not take their prescription, the doctor will be able to tell from the plot.

Beta Blockers In Figures 3.8 and 3.9, we display QTc values of patients on and off beta blockers. In Fig. 3.8, all patients are plotted, meaning that we cannot say that beta blockers *cause* the differences we are seeing; instead, it may be the case that patients with naturally higher QTc were more likely to be prescribed beta blockers. In Fig. 3.9, we fix this by only viewing data for patients before and after starting beta blocker therapy. However, we do not have data from enough subjects to draw conclusions from this figure.

Sotalol A typical use of plotting multiple Holters on the same axes is shown in Fig. 3.10, where we compare a patient's baseline QTc and heart rate to his QTc and heart rate on an antiarrhythmic drug. The drug was administered in the morning, and we can see its effect on QTc increase into the afternoon, drop

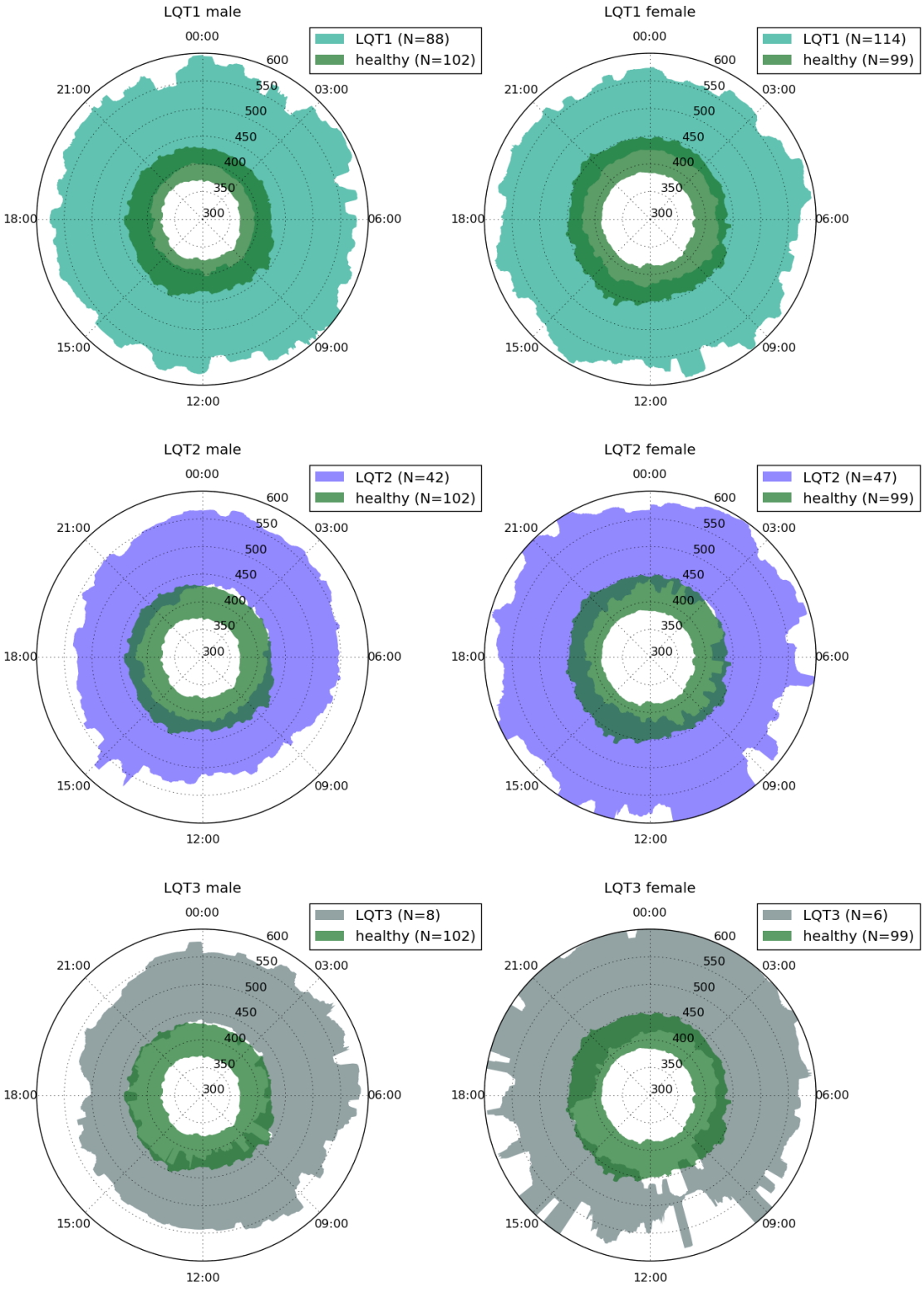


Figure 3.7: QTc in LQTS patients vs. healthy subjects. The range shown is the 5th–95th percentile. This is similar to Figure 3.6, which has a narrower percentile range that emphasizes the differences between the groups.

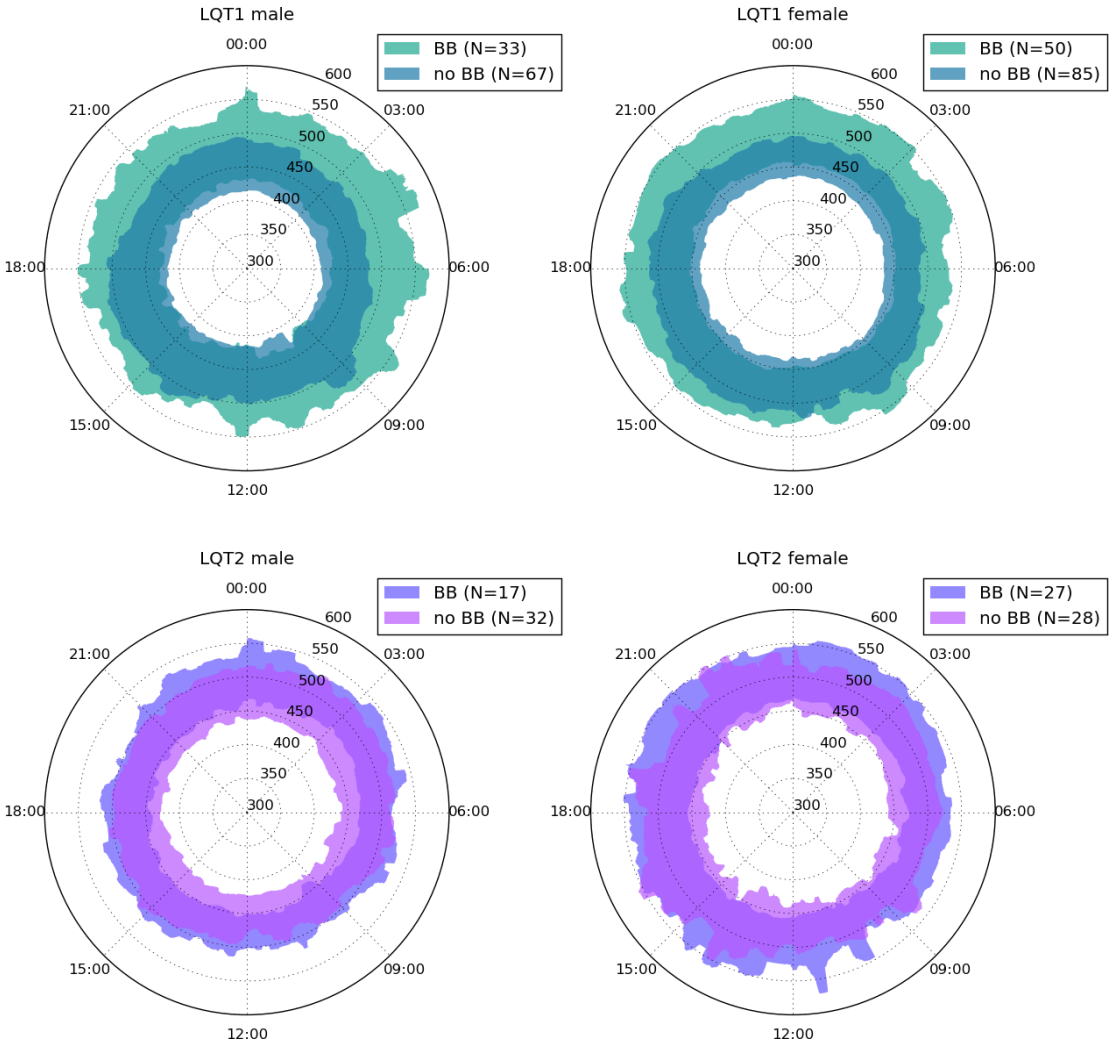


Figure 3.8: Comparing QTcB in Holters of patients on vs. off beta blockers. Patients on beta blockers have longer QTc, but this is likely the cause of the prescription rather than the effect, as we are looking at separate groups of patients. The range shown is the 16th–84th percentile.

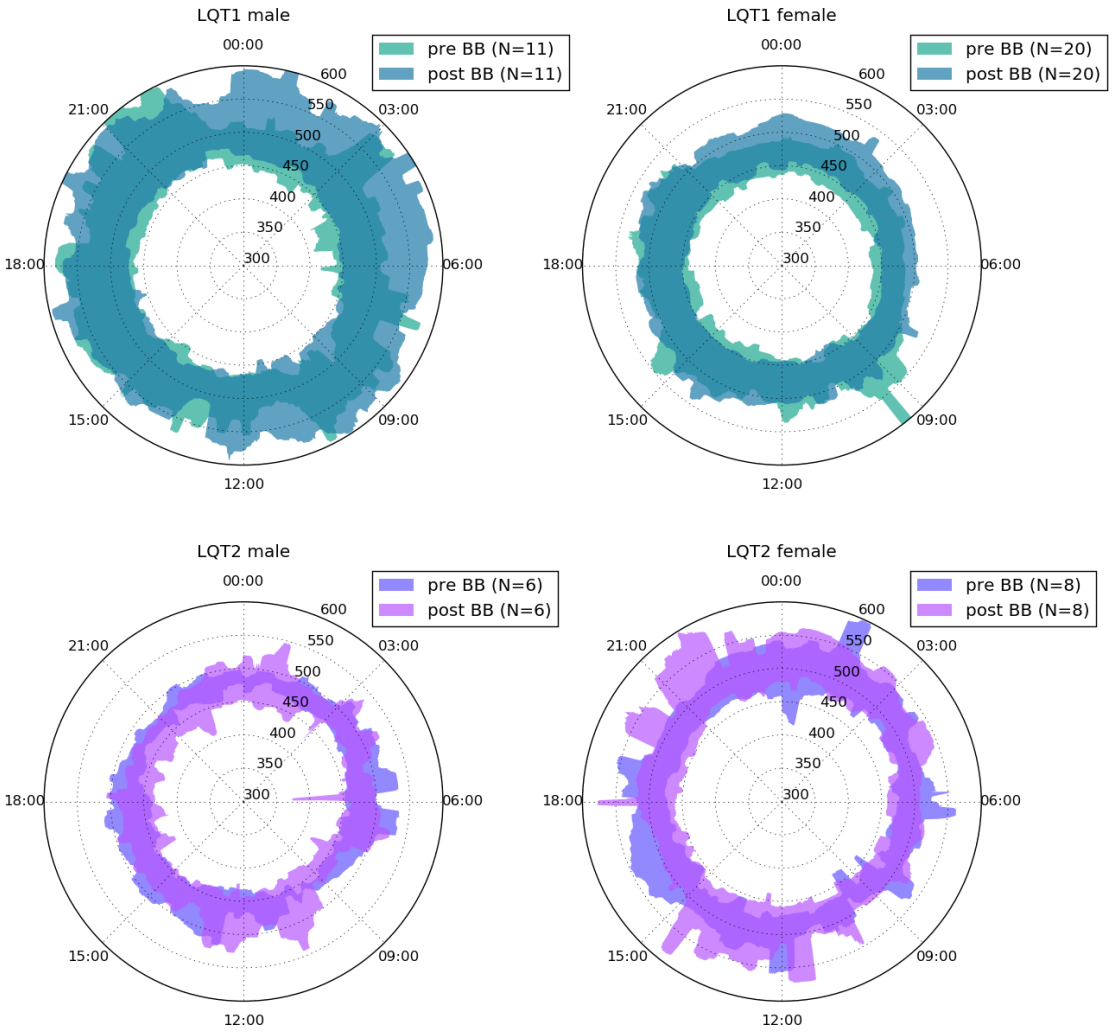


Figure 3.9: QTcB before and after beta blocker therapy. The range shown is the 16th–84th percentile. In this case, we lack information about the type of beta blockers used, and the sample size is also very small. Still, the apparent trend of increased nighttime QTc prolongation in LQT1 patients after beta blocker initiation may merit further research.

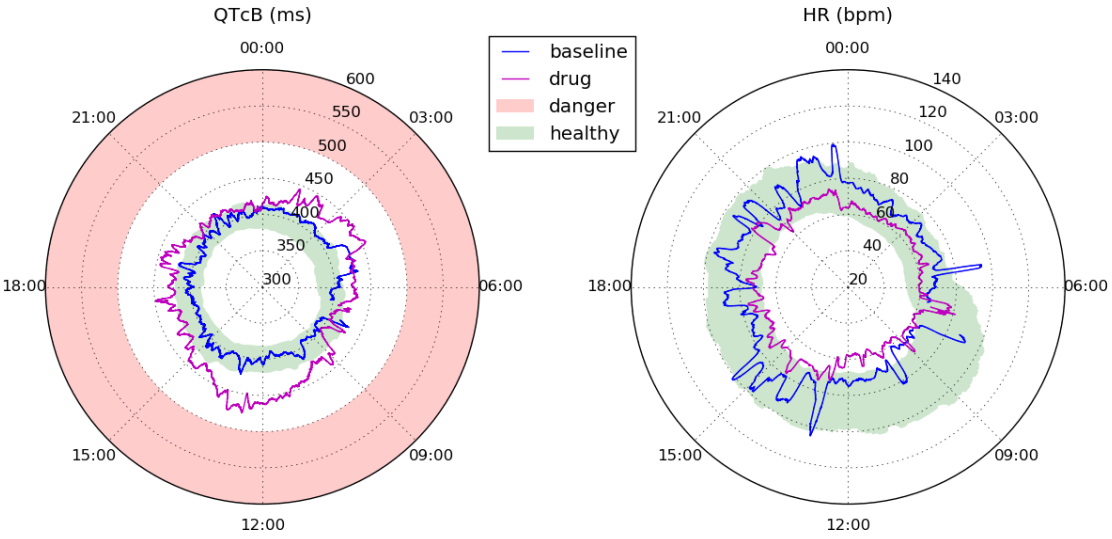


Figure 3.10: ECG clocks of QTc and HR during a drug trial. One individual's baseline was recorded (blue line), and then he was recorded again on sotalol (purple line). We see that the drug increased QTc, and decreased heart rate and heart rate variability (HRV).

off until roughly midnight, and then re-emerge during sleep. The effect on heart rate is more immediate and consistent throughout the day. This presentation may supplement existing measurements during drug trials. A cardiologist would be able to use similar plots for their patients to determine if prescriptions were working as expected, and also to monitor medication adherence.

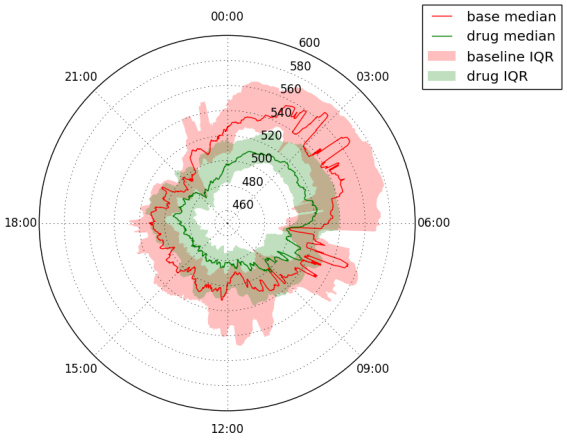
Some patients may have drastically different sleep schedules than the average person, e.g. due to working third shift. In the figure, we notice that this patient's heart rate pattern appears to be misaligned by a couple of hours compared to the "normal" range. In cases like this, it may be desirable to rotate the "expected" range to match the patient's schedule. We will be adding an `offset` parameter to the ECG Clock Library to allow this. (This will also be useful to adjust annotation data containing incorrect timing information.)

Eleclazine One of the most important applications of ECG Clocks is to analyze the impact of drugs on a patient or group of patients. The University of Rochester Medical Center's Heart Research Follow-up Program has provided ECG annotations for several cohorts in a study for a new drug, "eleclazine", which was developed by Gilead Sciences, Inc. Eleclazine is a late sodium current inhibitor which may shorten QTc in LQT3 subjects. In a recent study, we compared a group of baseline Holters to a group of "on drug" Holters [11]. This type of plot tells a much more complete story than a list of basic statistics from the recordings. Figure 3.11 displays various repolarization intervals before and after drug induction to verify and quantify its effects throughout the day.

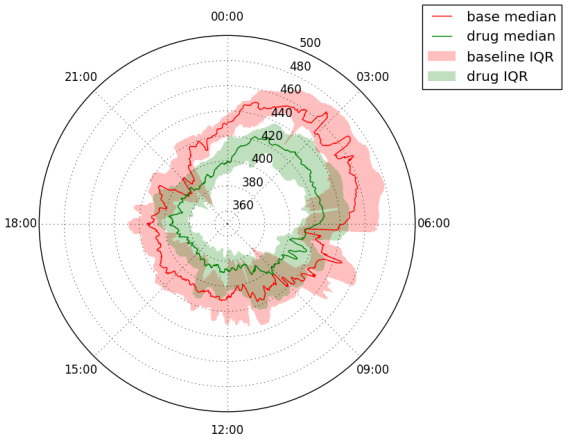
3.4.3 Annual Checkups

In Fig. 3.12, we show the same patient's QTc recorded at five different ages. On the left, we see that QTc is very stable when comparing ages 3, 4, and 6. At ages 5 and 7, though, QTc is prolonged; this is shown on the right, along with the age 6 plot for reference. This patient's QTc indicated relatively low risk in the recordings on the left, yet it indicates high risk during the two "anomalous" recordings [58]. Unfortunately, we do not have information about prescriptions or other possible causes for the prolongation at ages 5 and 7, but his physician would immediately investigate the cause when presented with the plot. Finally, note the distinct "LQT2-like" shape of the plots, where QTc increases at night. This asymmetry could aid in diagnosis, preempting genetic testing in some cases.

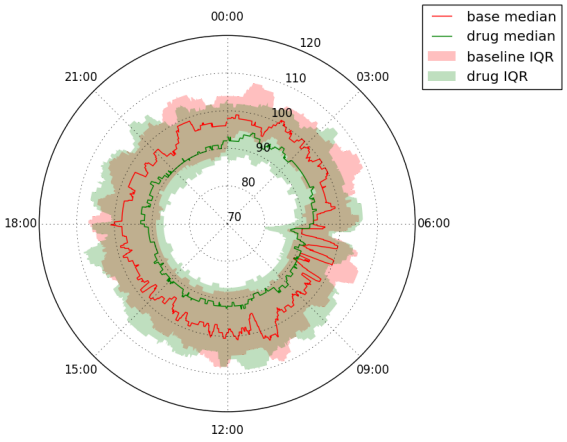
The same representation helps to easily pinpoint periods of the day associated with QTc prolongation, and therefore counsel patients about their daily periods/activities associated with risk.



(a) QTcF



(b) QTpcF



(c) TpTe

Figure 3.11: Baseline and on-drug QTcF, QTpcF, and TpTe interval measurements in eleclazine trial (ms). There are only 4 subjects in this cohort, but results are fairly clear; the drug mainly impacts QTp (not TpTe), as intended.

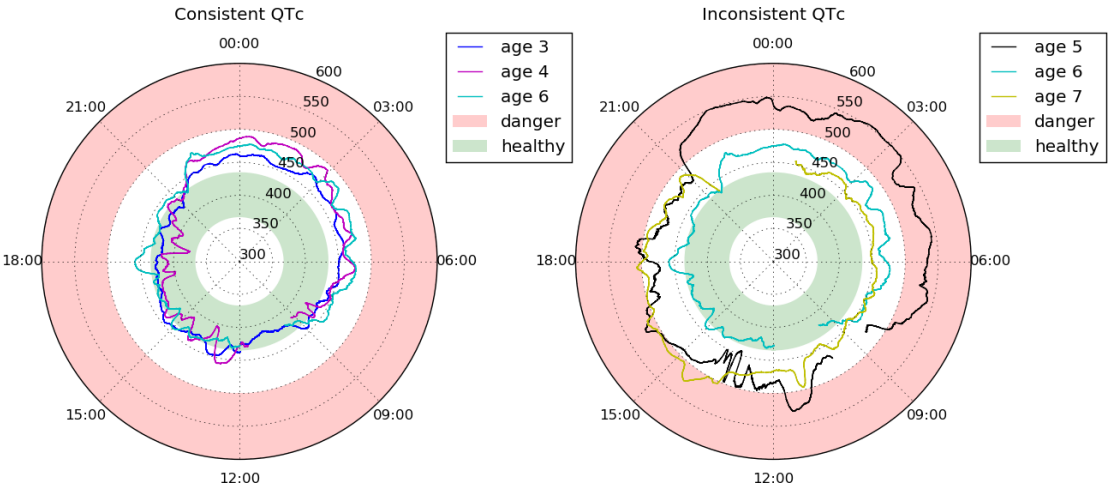


Figure 3.12: Monitoring QTcF changes at different ages in a male LQT2 patient. We note that the recordings at ages 5 and 7 show higher QTc than the others, which could flag the physician to investigate drug interactions or other possible causes.

3.4.4 Event Prediction

Figure 3.13 shows the QTc differences between LQTS subjects with and without symptoms (i.e. "events" like syncope). Unfortunately, the plots are not particularly useful from a clinical perspective, because we do not know if the Holters were performed before or after the events. Further, it is not surprising that patients with symptoms have longer QTc. However, these plots can provide an idea of QTc thresholds where symptoms become more likely. We will attempt a more quantitative analysis of these groups in Section 3.5.6.

3.4.5 QTc Prolongation Concealment

We measured the beat-to-beat RR and QT intervals in 24-hour Holter recordings for healthy patients (n=200) and genotype positive LQTS patients (n=305). The number of QTc intervals measured was around 10 million in each healthy group (male and female), 12 million in each LQT1 group, and 6 million in each LQT2

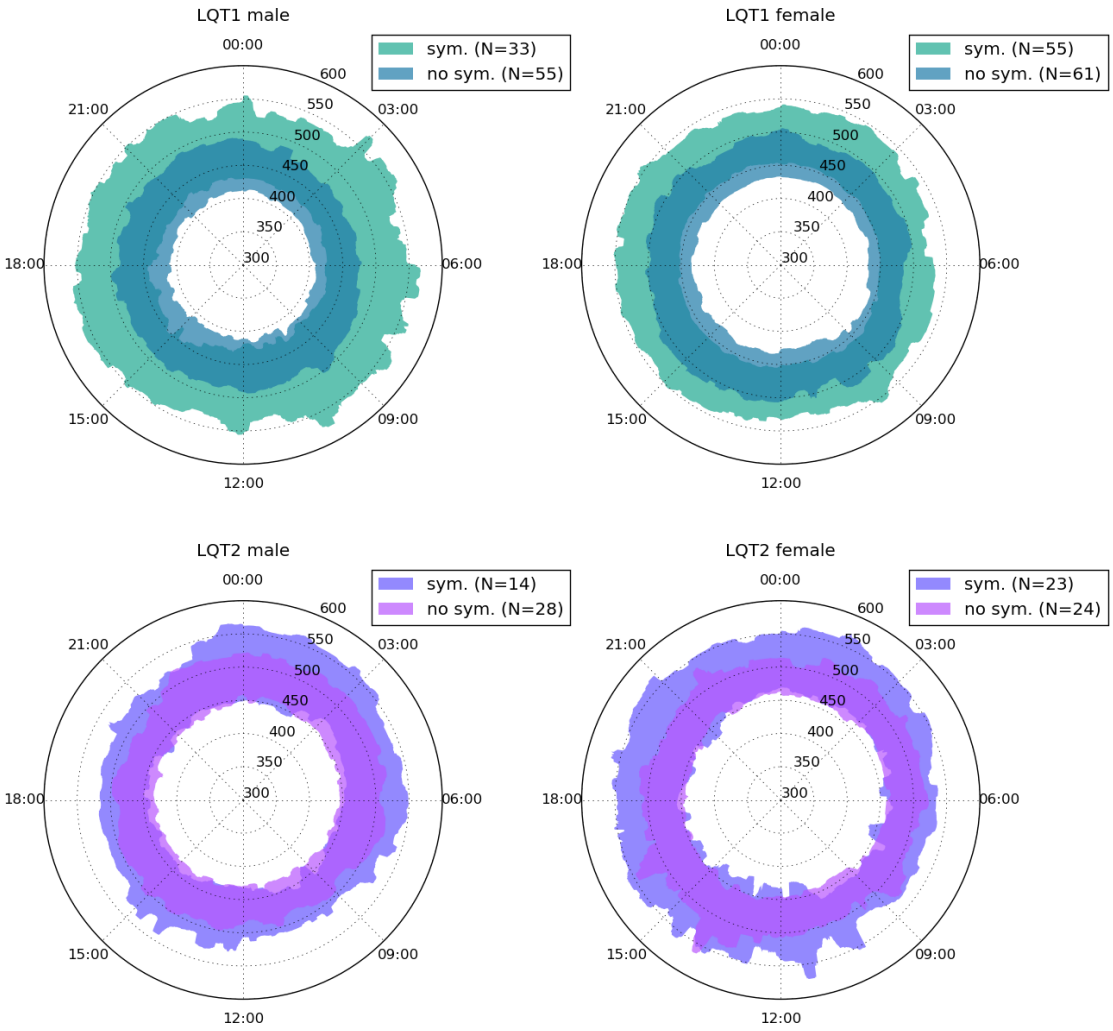
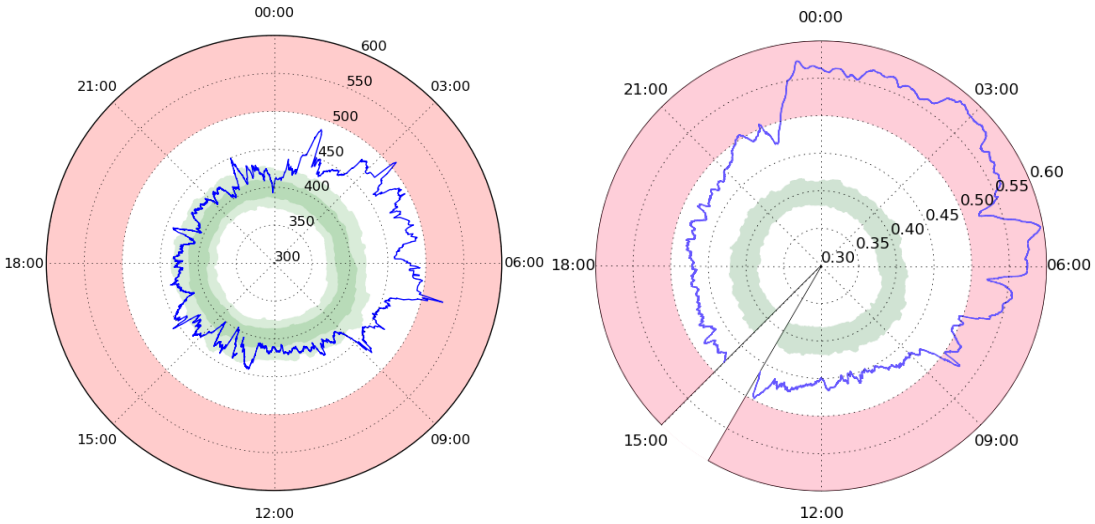


Figure 3.13: QTcB in patients with and without symptoms. The range shown is the 16th–84th percentile.



(a) 35yo male (b) 38yo male

Figure 3.14: Extreme cases of QT prolongation concealment. Both patients have LQT2 mutations. The plots confirm that some ECG biomarkers (such as QTc) need to be evaluated throughout a 24-hour period.

group. Figure 3.1 highlighted the gender-specific QTc distribution for healthy patients compared with LQT1 and LQT2. There is a difference in the distribution of QTc intervals in men versus women. In men, LQT2 is associated with longer median QTc intervals when compared to those with LQT1 (471ms vs. 455ms, p=0.03) while this difference is weaker in females (479ms vs. 470ms, p=0.04). The QTc distribution plot in LQT1 males shows the largest common area with the distribution of QTc in healthy males, indicating that this group of LQTS patients may have the least QTc prolongation, i.e. the most concealment.

To illustrate the potential consequences of concealment, in Figure 3.14 we plot QTc for two LQT2 subjects with extreme differences in daytime and nighttime QTc values. In particular, the second patient in the figure clearly has serious problems throughout the night, a fact that would be lost during a daytime clinical visit or even a long-term average. If a cardiologist had only looked at these patients' *average* QTc, it would have obscured the the fact that their QTc is

dangerously elevated for several hours. The plot takes minimal time to process, and is infinitely more useful than a simple average. Note that with a quick glance, the doctor can assess *if* and *when in the day* a person has dangerously prolonged QTc.

While we've focused on QTc and 24-hour observation periods, the process and framework will be similar to monitor other medical markers such as O_2 saturation, glucose levels, or body temperature, and over different intervals.

3.4.6 Mutation Location and QT Clock Profile

Figure 3.15 presents the gender-specific QT clocks for the groups of LQT1 patients with C-loop versus non-C-loop mutations, and LQT2 patients with pore versus non-pore mutations. In male LQT1 patients, we observed no clear differences in the QTc clock profiles between patients with mutations inside or outside the C-loop regions, though the C-loop region appears slightly worse overall (i.e. more QTc prolongation). In women with LQT1, QTc values remained >450 ms throughout the day and nearly overlapping QTc clock profiles were observed for those with C-loop and non C-loop mutations. However, among male patients with LQT2, those with a pore region mutation consistently had more QTc prolongation when compared to those with non-pore mutations. This obervation is consistent with prior report in LQT2 patients which revealed that male LQT2 patients with non-pore mutation were associated with lower risk for events [59]. Finally, in female LQT2 patients, the trend for increased QTc prolongation in those with pore mutations also appears to be present, but has more overlap with the non-pore subjects than we see in the male cohort.

The percentage of patients on beta-blockers was not significantly different in

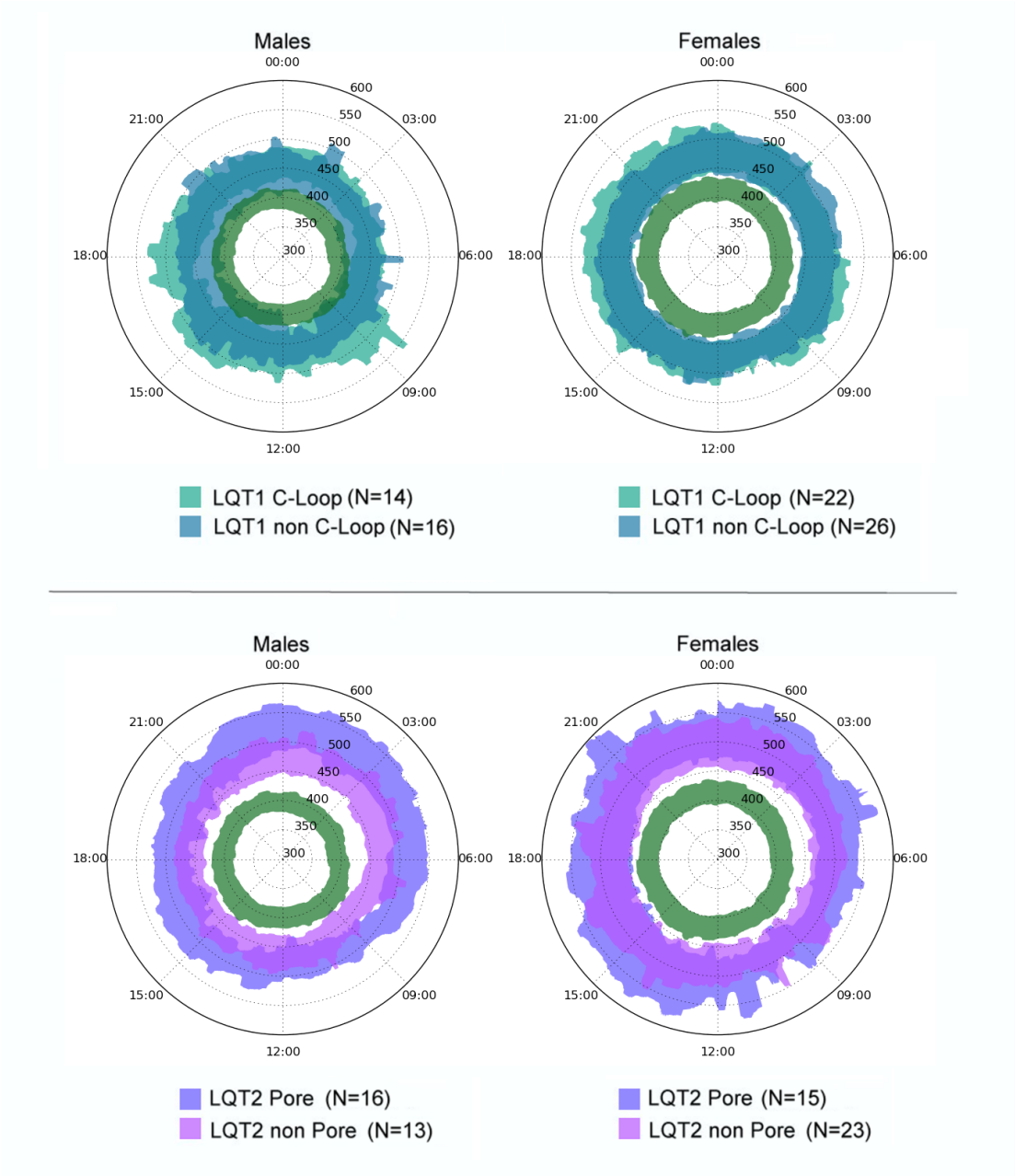


Figure 3.15: QTcB for LQT1 and LQT2 patients with missense mutations, separated by gender and mutation location. The most striking fact is that men with LQT2 have much higher QTc when the mutation is in the pore region.

LQT1 patients with C-loop mutations vs. non C-loop mutations (33% vs. 24%, p=0.15) or in LQT2 patients with mutations in pore vs. non-pore regions (47% vs. 42%, p=0.47).

3.5 Quantitative Analysis and Machine Learning

The main objective of the visualization techniques developed in Sections 3.3 and 3.4 was to present enough data for a doctor to make a decision. However, especially in the case of rare diseases with which the physician may not be experienced, it would be good for the computer to also provide some suggestions. That is, in addition to presenting plots or tables of ECG measurements, the computer should also attempt to classify/diagnose illnesses based on those measurements. In this section, we begin to investigate ways to augment the visualizations using machine learning (ML) algorithms. The goal is to utilize (a subset of) the same data used to generate plots to compute the likelihood that a patient has a particular medical condition. In this section, we will continue to focus on the LQTS.

In some cases, we will be comparing groups using statistical tests. Statistical analysis was performed using the SciPy stats module. The Mann-Whitney U test was used when assessing statistical differences between different populations, and the Wilcoxon signed-rank test was used when comparing results within the same population. A chi-square test was used in the case of binary variables. In all cases, p-values ≤ 0.05 were considered statistically significant.

As a sanity check of our annotations during ventricular repolarization intervals, we computed 24-hour averages of HR, QT, QRS, and QTc in our study

populations. These averages are presented in Table 3.1.

Table 3.1: Gender-based description of the baseline ECG measurements for healthy and LQTS cohorts using 24-hour averages. Values were computed using the process described in Section 2.1. *: P<0.05 vs. men of same genotype.

	Healthy		LQT1		LQT2		LQT3	
	Men (N=101)	Women (N=99)	Men (N=88)	Women (N=114)	Men (N=42)	Women (N=47)	Men (N=8)	Women (N=6)
HR (bpm)	77±9	81±10*	72±11	77±14*	77±17	76±15	72±15	77±7
QT (ms)	358±21	365±20*	430±48	435±48	433±59	450±54	463±58	453±65
QRS (ms)	93±5	90±4*	95±10	92±8*	94±9	92±8	97±11	94±6
QTc (ms)	400±16	416±16*	466±53	482±36*	476±42	491±40*	490±35	497±57

Fig. 3.16 depicts the conceptualized work flow of a healthcare system that is well-suited to our purposes. This system stores long term patient records in electronic format (denoted as e-Health records). The data acquired from a patient is pre-processed and filtered to reduce data dimensionality. The results are incorporated into the e-Health records (EHR), thereby gradually enriching the database quality to improve the accuracy of future ML results. This assumes, however, that we are storing *every possible piece* of (relevant) clinical information; a database with many patients may not be as useful as a smaller database with more information on each patient.

As shown in the figure, an effective ML-based healthcare system should capitalize on the vast computational capability of the *machine* and the unprecedented reasoning ability of the *human* (the doctor). Both the machine and the doctor are looking for patterns, but the doctor cannot analyze every heartbeat of every patient, nor can he be expected to be familiar with the nuances of every disease. The machine *can* do this, and can present its "suspicions" to the doctor for confirmation. We envision such a system to provide decision support to doctors in three different ways:

- **Visualization** of the long term monitoring data in a concise and intuitive format [8] could significantly reduce the data burden for the doctor, allowing him/her to make decisions quickly and accurately. This aspect was thoroughly covered in the previous sections.

- **Alerts** are simple alarms which may be triggered by a value crossing a threshold, or by a very reliable ML algorithm. For instance, the doctor may wish to be notified immediately if a patient's QTc exceeds 500 ms.

- **Classification** is the process of predicting what group a patient falls under.

> There may be several correct answers: people with a specific genotype, people at risk for certain cardiac events, etc. Prediction of short-term outcomes is a primary goal here, in which the machine may predict, for example, a high risk for myocardial infarction (MI, a.k.a. heart attack) in the next 12 hours.

We reiterate that the physician is still at the head of this process, ordering tests, analyzing records, adjusting prescriptions, etc. The visualizations and recommendations from the machine are simply additional tools to support the process — i.e., a second opinion. In time, not only will the machine make more accurate classifications due to expanding database sizes, but also the doctor will develop an intuition for how/when a machine makes accurate classifications.

It is also important to consider the privacy implications inherent to this system, where many patient's records are being aggregated and analyzed. Most Protected Health Information (PHI) can be safely removed from records in accordance with HIPAA [60], such as the patients' names and birthdays. However, as we have found in our own research, the inability to trace a record back to a specific person, in order to obtain more detailed information from their physician, can limit the effectiveness of data mining tools. Further, the criteria for PHI includes information that could be combined to statistically reveal the identity of a patient. This information — such as age, gender, race, and genetic disorders — is critical to developing good decision support systems, but including too much information on the wrong computer system risks violating HIPAA. Finally, some applications may require explicitly-protected information such as a patient's voice print [61] or city. Researchers must keep these restrictions in mind during all stages of a study.

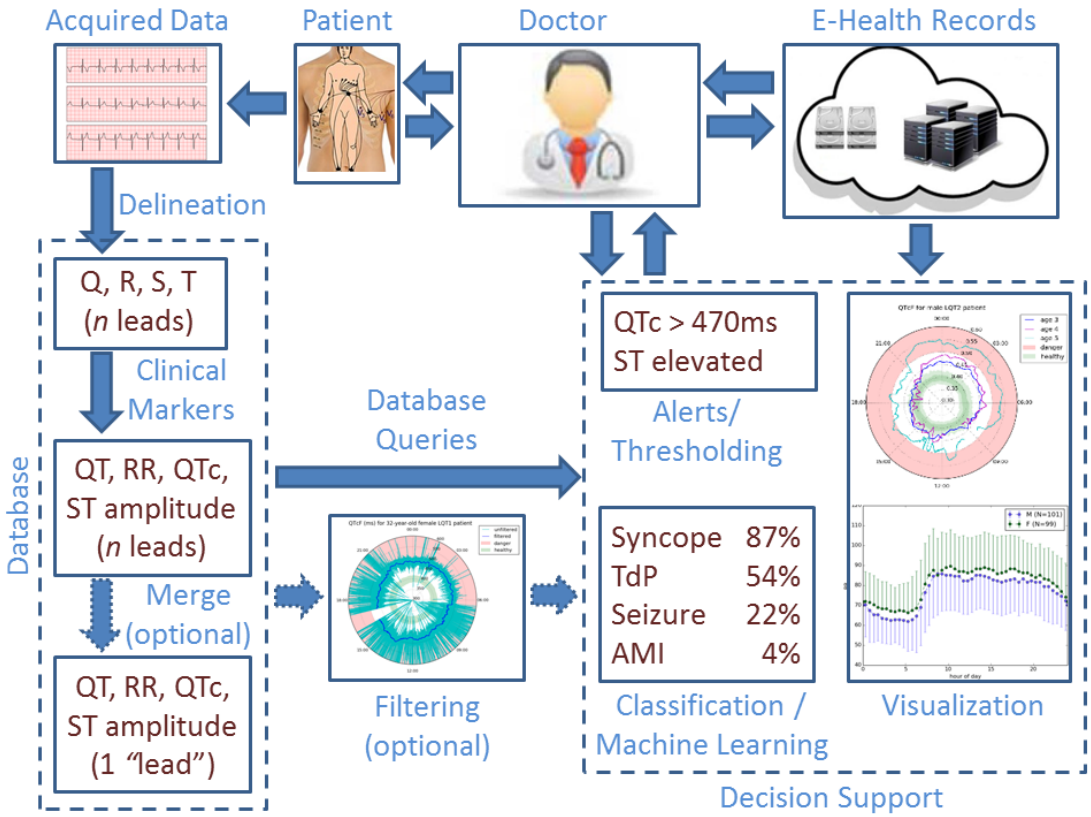


Figure 3.16: Complete ECG analysis workflow. Features for visualization and machine learning are extracted from the raw data and stored in a database. To analyze the data, queries are performed which select some subset/function of the full feature set. These results then become inputs for the decision support functions. There is feedback from the doctor to tune parameters of those functions.

3.5.1 Background on Machine Learning Techniques

We consider machine learning algorithms from three general categories:

1. "Conventional" supervised learning methods, such as SVM, decision tree, and nearest neighbors.
2. Clustering techniques such as GMM, K-means, and DBSCAN.
3. Artificial neural networks (ANN).

We will mainly discuss the first category, but will present some formulation and results from the third. We will also consider "ensemble" techniques such as AdaBoost and Random Forest, which attempt to use the results from several classifiers to improve accuracy. Clustering methods will not be discussed, as we have not yet found satisfactory parameters to achieve good results with these. Table 3.2 offers an overview of the characteristics of the main four ML algorithms that we considered in our evaluation.

To start with, we will test one of the simplest machine learning algorithms: **nearest neighbors**. This method simply selects the "closest" training sample to the presented sample (i.e. shortest Minkowski distance). An extension of this takes a weighted average of the N closest samples. While nearest neighbors is simple to implement, it faces a few drawbacks; outliers have just as much weight as other data points, and we need to store (and search) the entire training set to perform classification.

Support vector machines (SVM) are also very common, and simple to train and interpret. Depending on the nature of the data, they can be highly accurate. They operate by defining hyperplanes which separate the data into

Table 3.2: Pros and cons of some frequently-used machine learning classifiers. This overview helps us choose the best method(s) based on the input data characteristics and computational resources available.

Classifier	**Advantages**	**Disadvantages**
Nearest Neighbors	Simple to implement. Easy to understand/interpret.	Sensitive to noisy data and anomalies. Computationally expensive for large data sets.
SVM	Flexibility with non-linear data. Scales up with large sets of data. Relatively resistant to curse of dimensionality.	Difficult to interpret feature importance. Confidence estimates may be unreliable.
Random Forest	Alleviates the over-fitting problem. Easy to extract feature importance. Scales up with large sets of data.	Increased bias compared to single decision tree. Retraining on same data can produce different results.
AdaBoost	Automatically reduces dimensionality. Relatively fast.	Sensitive to noisy data and anomalies.

different groups. These planes are created from a subset of the training points, known as support vectors, in a way that maximizes the distance from the plane to the nearest data point of any class. Additionally, the feature space may be transformed using different kernels to allow nonlinear classification boundaries. Regardless of the kernel, SVM offers several advantages including memory efficiency — only a few training points (support vectors) need to be stored after training — and effective classification in high dimensional spaces. We will attempt to train SVMs using both linear and radial basis kernel functions. While this (and some other) algorithm(s) are designed for data from only two categories, the scikit-learn implementation will internally split our three-category data into two-category stages to bypass this limitation. One weakness of SVM is in the ability to judge how certain we are about a prediction; you can compute the distance from a point to the nearest separator plane, but this doesn't necessarily translate well to a "confidence percentage".

A very different method that we expect to perform well is the **random forest** algorithm. This method uses training samples to construct multiple **decision trees** — a forest — using random subsets of the given data and/or features to build each tree. It then classifies a new testing point by averaging the results of the individual trees. (A single decision tree operates by splitting the data multiple times until "leaves" are created of a single class. Then, to classify a new sample, we simply traverse the tree based on the splitting criteria until we arrive at a leaf.) By taking the mean, a random forest gets rid of the over-fitting problem that is often encountered with a single decision tree. Random forests offer other valuable features such as processing large amounts of data efficiently. They are also good at"filling in" missing data, and providing a reasonably accurate evaluation of what

variables are important in the classification. However, it does have a disadvantage in terms of "adding data later"; that is, if we want to grow our forest at some point after the initial training, it is not trivial to add more samples without fully re-training the classifier. The process of adding one (or several) training examples to a model without complete retraining is called *online* machine learning, and is not a feature of *any* of our chosen classifiers (with the exception of nearest neighbors, which doesn't really have a training stage).

The random forest is a type of "ensemble" algorithm, basing its output on the output of several other classifiers. We will test two other ensemble techniques as well: **AdaBoost**, and **voting**. The voting classifier simply takes the output of several other classifiers and does a majority vote if they disagree. In a more advanced version of this, the results of the individual classifiers will be weighted based on their confidence in it (and/or our confidence in that classifier). AdaBoost is somewhat different; it is a multi-stage classifier where each stage is trained on the failures of the previous stage. (In our case, each stage is a Decision Tree, but this can be changed.) AdaBoost is generally quite accurate, and can automatically mitigate overfitting by selecting only relevant features. However, it may lock on to noise/outliers in trying to correct its mistakes.

Finally, we will use the NVIDIA Deep Learning GPU Training System (DIGITS) [62] and Caffe [63] for **ANN**-based classification of LQTS. Artificial Neural Networks (ANNs), which are inspired by the design of our biological neurons, may be used for either supervised or unsupervised learning. "Deep" ANNs use many layers of these artificial neurons to form abstractions of the input data, leading to a final classification layer. In each layer, weights are applied to features of the previous layer to optimize performance. We have seen in Section 3.4 that proper

visual arrangement/presentation of ECG sensor data can greatly aid the doctor's decision in diagnosis and prescription. As there are many ANNs designed for visual recognition tasks, we decided to adapt our visual output (i.e. the ECG Clock) to a form that could be directly used as input for a pre-tuned ANN. One common vision task for ANNs is to classify handwritten digits from the MNIST handwriting database [64]. These are binary images, and are 28x28 px each. We simply shrink our ECG clocks (the plotted lines only) down to this size, and attempt to train an ANN to classify "healthy", "LQT1", and "LQT2", from plotted QTc values. This format essentially restricts us to 784 data points (28x28), and most of the image is blank (i.e. it is sparse); we may only plot ∼70 points. It will be interesting to see if we are providing enough data to the ANN, and if this unconventional format/structure — where the inputs are a graphical representation of the values, rather than the values themselves — performs well. The technique is shown in Fig. 3.17. Based on the examples in this figure, we expect that "healthy vs. sick" will be fairly simple to determine, but "LQT1 vs. LQT2" may be difficult.

Which ML algorithm is best to detect and classify LQTS? This really depends on properties of our data and our long term goals for how it will be used. For instance, some algorithms may be lighter in terms of storage and/or computation if we intend to continuously update the classifier (i.e. "online" machine learning). Additionally, we will want to keep the dimensionality of the data as low as possible in order to improve the accuracy of many methods. For now, we will make some assumptions — e.g. that hourly data will be sufficient, as opposed to beat-to-beat data, and that the database is small — and test the performance of a variety of conventional ML algorithms on our data. Incidentally, for machine learning

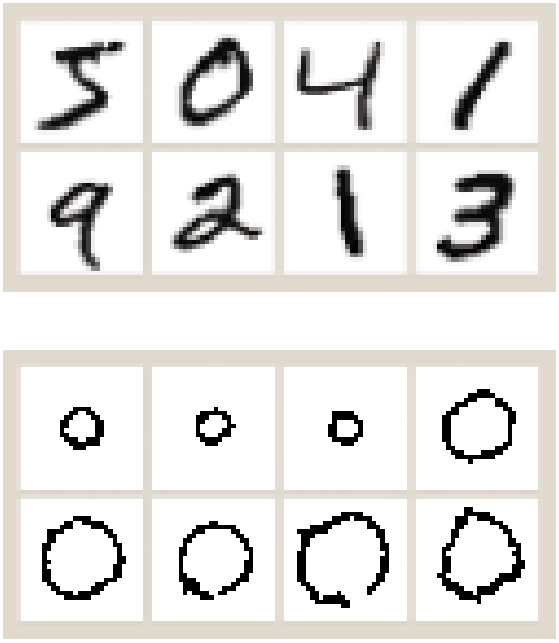


Figure 3.17: Applying artificial neural networks (ANNs) to QT clocks. We use a well-known handwriting recognition ANN and train it on reduced-resolution QT clocks (bottom) instead of the digits it was designed for (top).

purposes, our database is indeed somewhat small; we have access to 639 24-hour Holter recordings of healthy, LQT1, and LQT2 patients. LQT2 is the smallest cohort, with 145 recordings. LQT1 has 294 recordings, and healthy has 200. The scikit-learn [19] Python library will be used to perform the tests. Seventy percent of the samples will be used for training, and the remaining 30% will be used for testing. The data will initially consist of hourly QTc values for each patient, plus their gender (25 "dimensions"), and a corresponding classification (0, 1, or 2, for "healthy", "LQT1", or "LQT2", respectively).

3.5.2 Objectives

Concealment Analysis: We will attempt to identify the best way to unmask QT Prolongation concealment for each cohort (LQT1, LQT2, LQT3) using 24-

hour recordings. In this case, we are not using machine learning, but statistical analysis of the times of day at which QTc prolongation occurs.

Genotype Classification from ECG: One interesting application of machine learning will be to perform a "genetic test" from a Holter. Using hourly QTc values, we will try to make various algorithms learn the patterns that differentiate "healthy" from "LQT1" and "LQT2".

Event Prediction: We restricted our study to 434 THEW recordings of patients with the most common LQT genotypes (LQT1 and LQT2) and the most complete demographic information (such as age and gender). Our goal was to identify which of the patients had cardiac events caused by LQTS (such as seizures or syncope). In other words, we are trying to identify patterns in the ECG that can distinguish which genotype-positive patients will also suffer from related events. Given some measurements from an ECG, a classifier should simply tell us "events expected" or "no events expected," perhaps with a confidence value.

The algorithms we used to perform classifications were all implemented using scikit-learn, an open-source Python library that is built on SciPy and NumPy [19]. To assess a classifier's accuracy, we set aside 30% of the samples for testing, and trained only on the remaining 70%. Because some algorithms include inherent randomness in their operation, and because the division between training and testing data is also done at random, we repeated the cycle of selecting training data, training, and testing 50 times for each classifier. This is known as Monte Carlo cross validation. The average result from these trials tells us how well a classifier is likely to perform, whereas the worst result shows how poorly a classifier may perform due to an unlucky selection of training data.

3.5.3 Feature Selection

Quantitatively, predicting Long QT (LQT) symptoms requires training ML algorithms based on input variables that are extracted from raw ECG data. The machine learning methods described in Section 3.5.1 all have some inherent strengths and weaknesses regardless of the data. However, many of their limitations are imposed by the data we choose to provide; we know that QTc, and therefore QT and RR, are the measurements most commonly used by cardiologists to determine if a LQTS patient is in danger. We also know that people with different LQT genotypes tend to show more QTc prolongation at different times of day [6]. We therefore decided to provide hourly QT and RR measurements as input to the ML classifiers. That is, a "sample" for training or classification will consist of 48 values. We chose to split QTc into QT and RR in case the correction equations were removing information from the underlying values. Increasing the number of features risks inflicting the "curse of dimensionality," in which there are so many dimensions to work with that it is "too easy" to separate the training data into the correct groups. The learned model becomes specific to the training set, rather than generalizing to other data. This is known as *overfitting*. The apparent dimensionality can be reduced through methods like principal component analysis (PCA), or by simply removing hours that didn't seem to have much impact on classifier accuracy. In general, the fewer dimensions our input has, the less training samples we will need to achieve good performance, and the faster our classifiers will run.

Table 3.3: Percentage of cardiac beats with QTc prolongation (median values) in LQTS subjects between two periods of the day: 3–4AM and 3–4PM. Prolongation percentage is defined as the percentage of heart beats with QTc above the gender-specific threshold during the given hours. Periods starting at 3AM and 3PM were chosen as representative times to illustrate changes in QTc prolongation between sleeping and waking hours. The median QTc values at these times are identified in Figure 3.6, and prolongation percentage for the full 24 hours is shown in Figure 3.18. N is reported as the number of patients with Holter recordings containing data during the required hours.

	LQT1		LQT2		LQT3	
	Men	Women	Men	Women	Men	Women
N	82	108	40	47	8	5
Period : 3AM–4AM	48%	30%	100%	100%	100%	81%
Period : 3PM–4PM	97%	68%	87%	62%	100%	95%
p-value	0.0002	0.00001	0.0002	0.002	0.2	1.0

3.5.4 LQTS Concealment Results

Rather than comparing QTc value distributions at different times of day, it can be instructive to compare QTc prolongation percentages – i.e., the chance of QTc exceeding the clinical prolongation thresholds. In Table 3.3, we report the percentage of heart beats that showed QTc prolongation in the three LQTS cohorts for the same time periods: one nocturnal (3AM–4AM), and one diurnal (3PM–4PM). In the LQT1 and LQT2 cohorts, we observed a significant variation in median QTc prolongation (13–49%) between these two periods, $p<0.01$ (or $p<10^{-5}$ if gender is ignored). Specifically, in the LQT1 cohort, prolongation was significantly higher (regardless of gender) during the afternoon than late at night, while in LQT2 the afternoon hours showed significantly lower prolongation than at night. In LQT3, there were no statistically significant differences in prolongation when comparing these two time ranges, most likely due to the small size of our cohort.

Hourly median prolongation percentages are plotted in Figure 3.18. LQT1 and LQT2 patients appear to have very similar patterns of QTc prolongation during

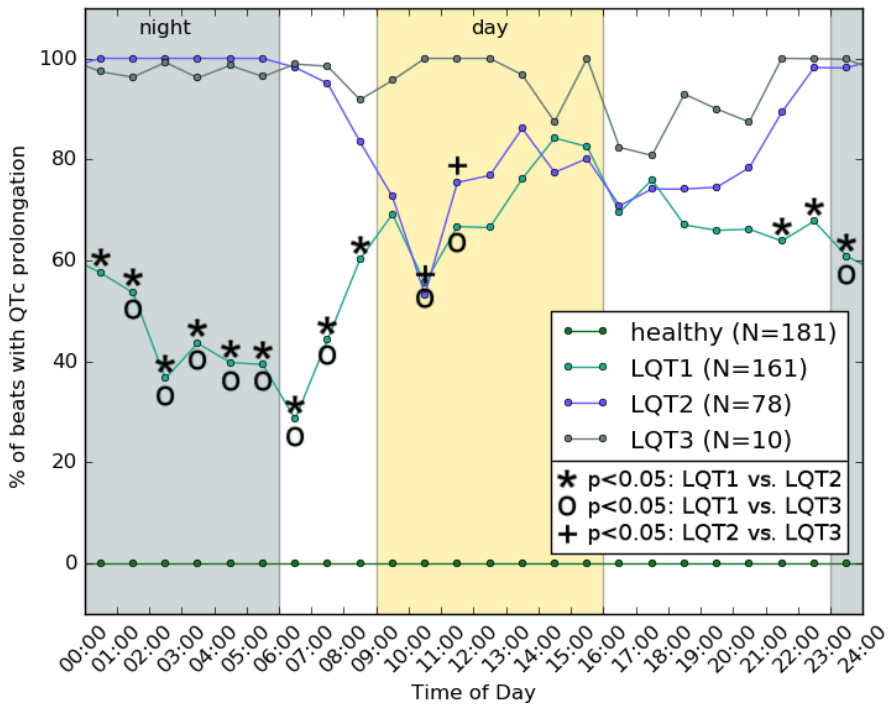


Figure 3.18: Median QTc prolongation in healthy, LQT1, LQT2, and LQT3 cohorts. Higher values mean higher chance to detect prolongation if QTc is measured at this time; 100% means that *most* beats show prolongation in a given cohort at a particular time.

clinic hours ("day"), but are quite distinct during sleeping hours ("night"). At night, QTc prolongation becomes much more prevalent in LQT2 subjects, and less common in LQT1. For readability, we did not plot separate lines for each gender in this figure. However, when the plot is separated by gender, it reveals a slightly higher QTc prolongation percentage in men than women — about 12% higher, on average. This agrees with our expectations from inspecting Figure 3.1, in which the threshold for males (450ms) appears to offer higher sensitivity than the threshold for females (470ms). Finally, we identified the two periods of the day with the most significant differences in median percentage of QTc prolongation within each LQTS family. In LQT1, the most significant difference in percentage of beats with QTc prolongation is found when comparing 05:30 (early morning) to 15:30 (afternoon) (39% vs. 83%, $p \sim 10^{-9}$), while in LQT2 the most significantly different periods of the day are around 00:30 (night) and 18:30 (evening) (100% vs. 74%, $p \sim 10^{-6}$).

3.5.5 LQT1 vs. LQT2 Results

In Figure 3.19, we present the results of several estimators on the "LQT1 vs. LQT2" problem. Classification of "healthy" vs. "long QT" was relatively accurate — about 90% — as we expected. Further differentiating between LQT1 and LQT2 was more difficult, lowering the score of each classifier to an accuracy of about 70–75% with the SVM and Random Forest methods. We should note right now that several of the recordings in our database were noisy or incomplete, which likely degraded our results. Missing data was replaced with average values, but very short recordings should probably have simply been thrown out. However, we wanted to present a realistic starting point for further research, so all data was

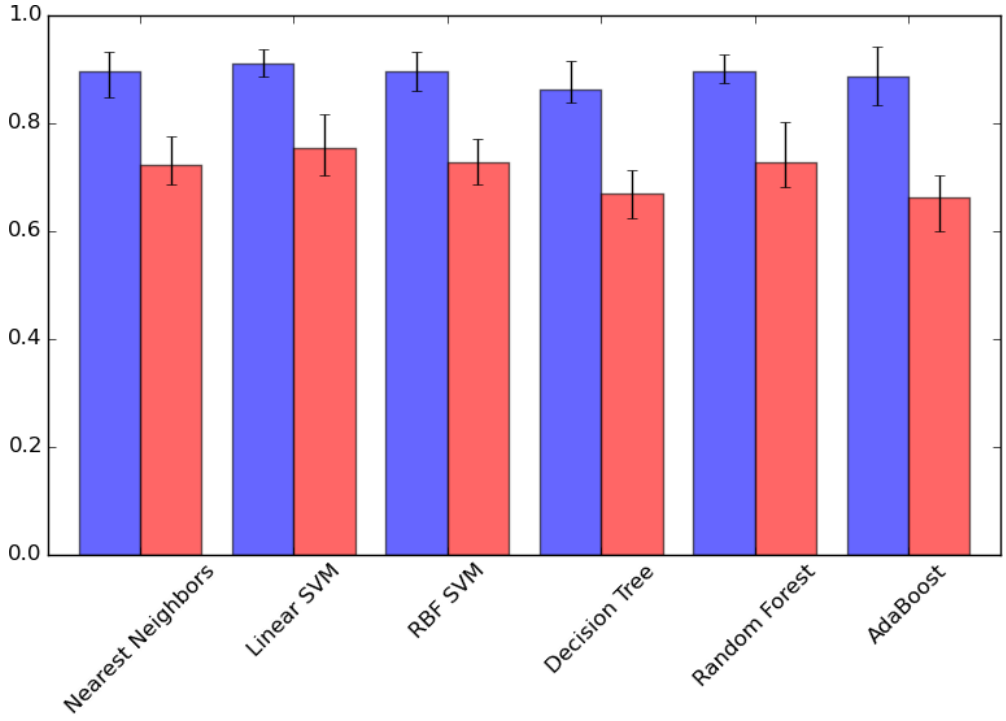


Figure 3.19: Average estimator performance in classifying LQT1 vs. LQT2 vs. healthy (red) or healthy vs. LQTS (blue). The classifiers were given 24 QTc values per patient — one per hour — as input.

retained. Another important consideration is that while our data is segregated by LQTS genotype, not all LQTS subjects show the corresponding phenotype. In other words, a handful of the LQTS patients truly do look healthy, so even a cardiologist would be likely to "misclassify" them.

We just saw that Random Forest and Support Vector Machine (SVM) generally proved superior to other algorithms. Now, we would like to see what information they are using to arrive at their decisions. For example, based on our findings in [6], we expect that data from ∼3AM will be a very good differentiator between the classes. We also expect that afternoon QTc measurements will not help distinguish between LQT1 and LQT2. Fortunately, we can examine the internals of these trained classifiers quite easily. In Fig. 3.20, we extract the "importance" of each

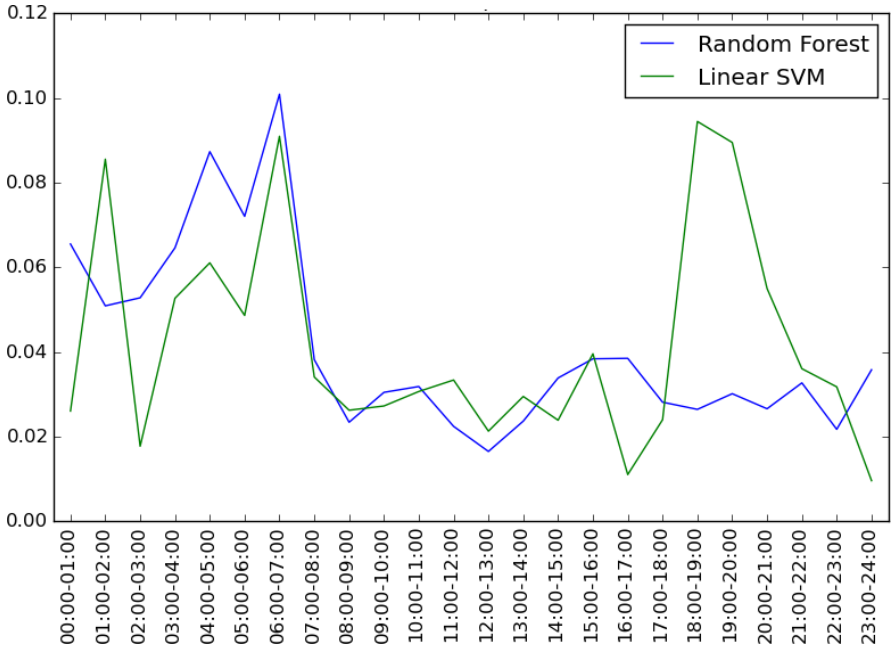


Figure 3.20: Feature importance in SVM and Random Forest when classifying LQT1 vs. LQT2 (vs. healthy). This tells us which measurements are most useful in making a prediction; the QTc value at around 6:30AM seems to be a good differentiator.

feature (hour) from the Random Forest and Linear SVM classifiers. The results are basically what we expected — late-night data is most useful to both classifiers — but SVM also used information from earlier in the evening. Because of the random nature of the training/testing data split, and the random selection of features in Random Forest, these results will not be exactly the same on every trial. However, we observed the same general trend over several trials.

In an effort to further improve the accuracy, we ran a feature selection algorithm to determine the most useful measurements in classifying LQT1 vs. LQT2. We found that T wave duration and TpTe were more useful than QT/QTc. The ST segment duration around 8PM was also somewhat important. Using hourly T wave duration and TpTe measurements improved our accuracy to 80%. We then devised custom input features based on Figs. 3.6 and 3.7; we take ratios of QTc

at one time of day vs. another. For example, one feature would be QTcF at 3AM divided by QTcF at 4PM. We expect that values less than 1 would indicate LQT1, and values greater than 1 would indicate LQT2. Using these and additional ECG measurements improves accuracy to 84-86%. Future work could likely improve this by experimenting with different feature ratios.

Finally, we attempted a very basic Artificial Neural Network analysis of the QTc data. In this case, we did not provide hourly data points, but (28x28 px) QTc clocks as shown in Fig. 3.17 (bottom). These clocks were used to train a LeNet network [65], which is known to perform well on the MNIST handwriting data set. Missing data was not "filled" in any way; we simply passed incomplete plots to the ANN. This implementation achieved ∼70% accuracy with absolutely no tuning (and ∼90% accuracy when only classifying "healthy vs. sick"). i.e., it was comparable to the classifiers we've already discussed when only given QTc. Providing QTc clocks based on a QTcF rather than QTcB improved three-way classification accuracy to ∼80%. Using this alternate QTc equation did not improve the accuracy of any of the other classifiers, only the ANN.

3.5.6 Event Prediction Results

We used the classifiers from Section 3.5.1 to "predict" if the genotype-positive LQTS patients in our database would suffer from any symptoms. "Predict" is in quotes because in this database, we are told which recordings are from patients who had symptoms, but we are *not* told if the symptoms came before or after the ECG recording. So what we are really deciding is *not* whether someone will have symptoms in the future, but whether a recording appears to be from a patient who *has had* or *will have* symptoms. Our goal is to identify these patients to the

doctor, who may then adjust medications and monitoring based on the risk level of the specific patient.

Fig. 3.21 illustrates how the performance of each classifier changes as we provide more training samples. The nearest neighbors classifier was configured to weight samples by distance (rather than uniformly), and random forest was composed of 100 trees rather than the default of 10. Random forest and both SVMs were set to use balanced class weights. All other parameters were the scikit-learn defaults. We have given the classifiers 48 data points — hourly QT and RR — as input.

In Fig. 3.21a, we plot the accuracies that can be expected from each classifier, based on 50 random selections of training data. In addition to *average* accuracy, we also looked at the *minimum* accuracy over the 50 trials. In other words, assuming that we chose the training data very poorly, how well could each algorithm do? These results are shown in Fig. 3.21b. We found that in this worst-case scenario, over 100 training samples may be required simply to break even — i.e., to break the 52% line. The highest scores, both minimum and average, came from random forest and the RBF SVM, which achieved 60–65% accuracy even with an unlucky selection of training data and as low as 50–100 training samples. The best classifier in our tests, the RBF SVM, averaged about 70% accuracy.

Obviously, we do not want the computer to make bad classifications. However, we may be able to deal with relatively low accuracy if we know when the computer "wasn't sure" of a result. We therefore test the machine's average *confidence* in its responses. When the computer is *incorrect*, we find its average confidence to be around 64–69% with the best classifiers, SVM and RF. When it is *correct*, though, its confidence is around 68–74%. It should be possible to set a threshold

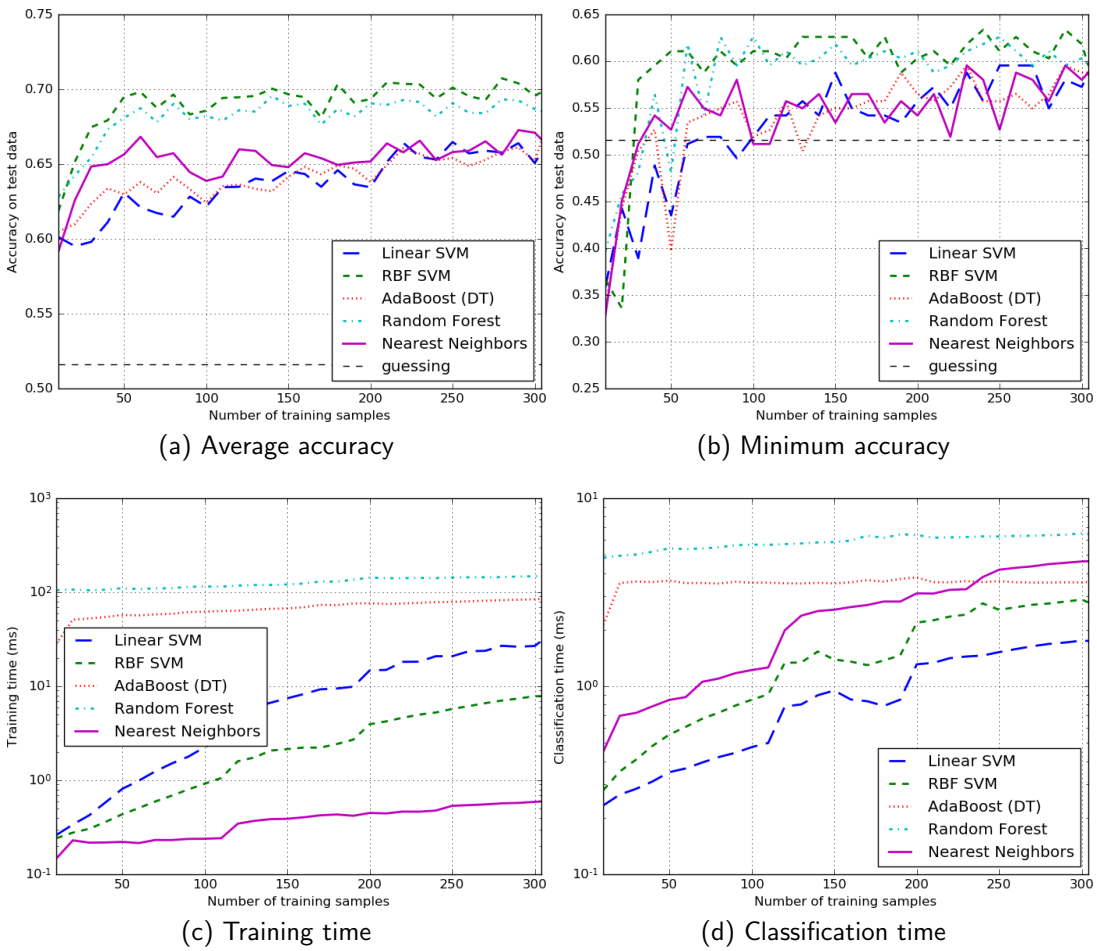


Figure 3.21: Performance of several classifiers. The input features were hourly QT and RR measurements for one day. Each data point is an average of 50 trials for which different training/testing data was randomly selected. 30% of the full dataset — 130 samples — was always used for testing, but training was conducted over increasingly larger subsets of the remaining 70% to determine how many samples were needed for optimal results. Approximately 52% of the patients did not have symptoms, so the thin dashed line in (a) and (b) represents the performance we could achieve by simply guessing "no symptoms" every time. Classification time (d) is actually a measure of how long validation took for the entire test set of 130 samples. The runtime results were computed on an Intel i7-5930K CPU.

below which the decision support system could report "inconclusive" rather than marking a patient as high or low risk.

In Fig. 3.21c and 3.21d, we measured the runtime of the training and classification stages to estimate the scalability of each classifier. Although the ensemble classifiers (AdaBoost and random forest) took longer than the others in both stages, their runtimes were barely affected by adding more training samples. Nearest neighbors, as expected, had essentially zero training time, but classification time increased with the number of samples, as it has to compute distances to every point in the training set. In fact, classification using nearest neighbors actually became slower than AdaBoost at around 240 training samples. Since the two ensemble methods have very flat runtimes, it's apparent that SVM will also become slower than even random forest, once it's given enough training samples.

Next, we investigated the impact of providing less features to the classifiers. Instead of hourly QT and RR (48 features), we provided only hourly QTc (24 features). Because QTc is designed to contain the LQTS-relevant information from QT and RR, we expected that this change would improve some of our runtimes without sacrificing any accuracy. However, we found that accuracy was decreased for random forest and AdaBoost (which are both based on decision trees), and that the learning rate appeared to decrease as well — i.e., more samples were required to reach peak performance. The peak performance of the best classifier was not hurt by removing features, though; SVM with a radial basis function kernel seems to consistently yield about 70% accuracy once it has seen enough training samples. In terms of runtimes, AdaBoost and random forest saw no improvement from the reduction in features. As expected, nearest neighbor classifications were faster, and all SVM runtimes improved.

Our use of QT, RR, and QTc was based on the knowledge of what physicians measure in practice. However, it is a good idea to make sure that we are not overlooking any other cardiac features that may be of use. We therefore decided to measure 23 different features (including QT and RR) at every hour of the day, for a total of 552 measurements. The features used include QRS, ST segment duration and elevation, QTp, JT, JTp, TpTe, and T wave duration and amplitude. Further, we use 6 features from each patient's medical record: gender, age, LQT type (1 or 2), mutation type (e.g. missense, frameshift, etc.), and mutation location.

Because we only have 434 training samples, we expect that we will need to reduce the dimensionality of this new feature set in order to mitigate overfitting. Several methods exist to reduce the dimensionality of inputs while minimizing the loss of information. We look to the PCA and χ^2 methods. PCA projects the data to a lower dimensional space, and χ^2 selects the statistically-"best" features. We measured classifier accuracy with the number of preserved features varying from 1–512. Surprisingly, we found that with both methods, only one feature/attribute was needed to achieve the $\sim$70% accuracy we'd demonstrated in the previous section. The most important features seemed to be evening QT-like measurements, where "QT-like" means QT, JT, QTp, JTp, or heart-rate corrected versions of these. Using the top 20 features with the random forest classifier yielded 72% accuracy (69% sensitivity, 75% specificity).

We noticed a pattern where the top-ranked features tended to come from approximately the same time of day. To illustrate this, in Fig. 3.22 we plot a histogram of what times the top 25 and top 100 features came from. The top 25 features all come from around 5–6PM, indicating that perhaps fatigue at the end of the workday is unmasking cardiac issues. Expanding the search to the top

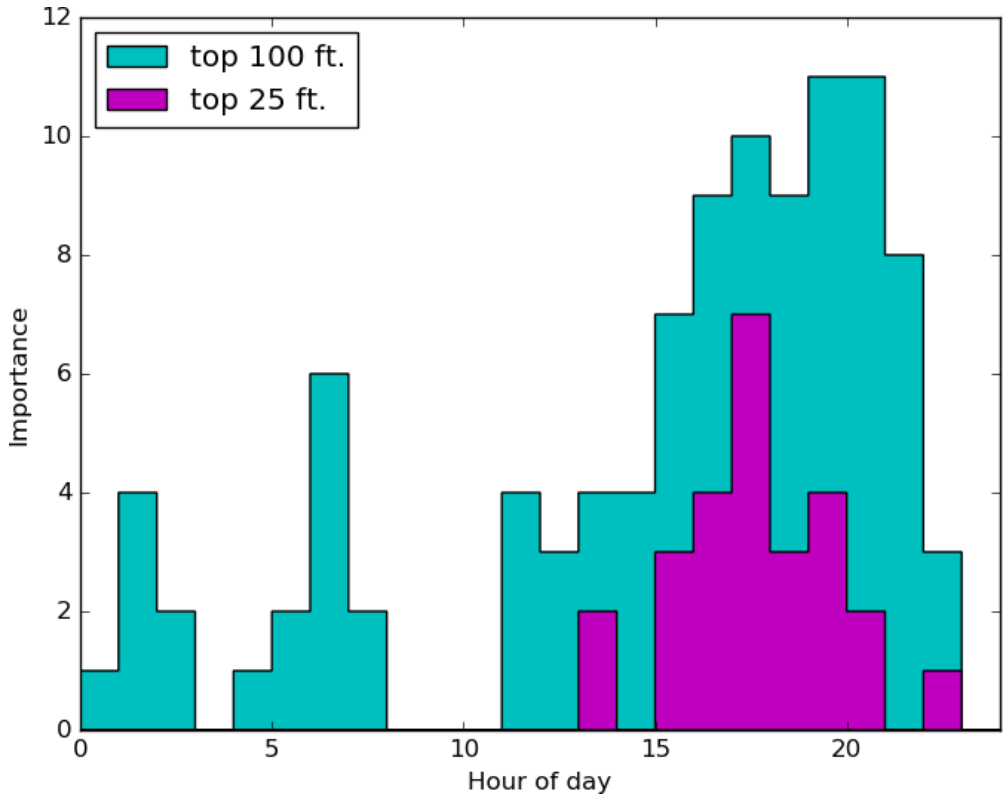


Figure 3.22: Feature importance vs. time of day (zero = midnight) when identifying events in LQTS subjects. The peaks indicate that late night, first awakening, and end of workday are the best times to detect cardiac issues in this cohort.

100 features begins to reveal other important times: first thing in the morning (6–7AM), another stressful time of day [66], and late night (1–2AM), which makes sense for the LQT2 subset of patients who tend to show more QTc prolongation during sleep [6]. Also important is the *lack* of important features around 8–11AM, which implies that a clinic checkup in the morning may not be enough to allow the physician to make an accurate assessment of a patient's risk.

Gender was expected to be a feature of significant importance in ML classification, as gender differences are known to be present in many coronary heart diseases [67], and males and females have distinct average QTc values and clinical prolongation thresholds. However, we repeated our experiments with gender as an

input, and found no significant difference in results. Upon extracting the "importance" measure from from the SVM and random forest classifiers, we found that gender was not being used in any significant way in their decisions. This process of discovering which features are significant creates new research opportunities, as the machine may identify surprising patterns or correlations that were previously unnoticed. For example, it may find that patients with a very specific genetic mutation are more at risk than others with seemingly-similar mutations. In the case of LQTS, we already know that mutations in certain regions are associated with increased risk, and it would be interesting to see if the computer would identify the same pattern.

From Fig. 3.21, we can decide which classifier works best under different constraints such as available training data or computational power. Obviously, for accuracy, we prefer to use either the RBF SVM or random forest. SVM will probably be our choice if the amount of training data is small, but we may begin to prefer random forest for faster runtimes once we are working with thousands of training samples. As runtime becomes more of a concern, we may also work to narrow down the number of features used as input. We have not discussed automated feature selection processes here, but several methods exist to reduce the dimensionality of inputs while minimizing the loss of information.

Finally, we note that all classifiers in Fig. 3.21a seem to be close to reaching a horizontal asymptote, meaning we will probably not improve their performance simply by adding more training samples. Instead, we will have to change the inputs or the parameters of the classifiers. In previous work, we found that time of day was important in classifying patients as LQT1 vs. LQT2. Both types show QT prolongation, but during different activities/times of day. That is why we

structured our inputs here as hourly data points. The dimensions also reduce well in this structure, as usually only a few hours (e.g. during sleep) are enough to differentiate LQTS types. However, perhaps a completely different format would be best for the classification experiment performed in this paper. It may not make sense to provide hourly average heart rates, as heart rate can vary greatly during a one hour period. Perhaps a different measure such as heart rate variability (HRV) would be better. Further, many high-risk LQTS patients are on beta blockers, which are drugs prescribed to slow the heart rate. So it is possible that the classifiers in this particular experiment were really noticing the presence of beta blockers (in the form of increased RR) more than any novel pattern we had hoped to find. In particular, this could explain why the decision tree based classifiers did a better job when heart rate was available. We could instead use something like the longest 10 QTc measurements throughout the day. This is a difficult measurement to make, though, in a noisy recording. We may also choose to try higher-resolution data, e.g. 5-minute data points, but this would likely require a larger database due to the increased dimensionality. We will experiment with variations of these ideas. We will also attempt to use other measurements of the T wave, such as symmetry, amplitude, and peak–end interval. Another possible approach is the use of a *voting* classifier, which attempts to aggregate the predictions of several other classifiers to reach a better result. However, our previous experience suggests that this will only be slightly more accurate than the best individual classifier. The selection of appropriate input features is an area where a physician's knowledge and intuition remains critical. As we narrow down the best features to use, we will also begin to try other classification methods such as clustering (e.g. GMM, K-means, DBSCAN) and custom artificial neural networks.

4 Decision Support for Suspected Acute Myocardial Infarction (AMI) Patients

4.1 Background and Objectives

In the previous chapter, we worked primarily with the THEW E-HOL-03-0480-013 database. This database contains 480 Holters of genotyped LQTS patients. In this chapter, we turn our focus to a different database: E-HOL-12-1172-012. This database contains 1172 Holters from patients who were hospitalized due to chest pain. In many cases, the pain was not caused by cardiac events, or was due to a non-acute cardiac condition. In other cases, the patient was indeed suffering from more immediately life threatening events such as unstable angina or heart attack (AMI). Cardiologists would like to identify which category a patient falls into as quickly as possible. Therefore, our goal with this database is to analyze the patient's record (age, family history, etc.) and test results (e.g. blood work, ECG) automatically to classify them as high or low risk. For instance, we may want to distinguish between patients with *acute cardiac* issues and patients with

Table 4.1: Example ML input from chest pain database. The first column is what the ML algorithm will learn to predict; perhaps Cohort 1 is "chest pain from non cardiac causes" and Cohort 2 is "chest pain from cardiac causes". The other columns are the features that will be used to predict the cohort.

cohort	age	smoker	diabetic	HR	STe	QTp
1	65	0	0	57	4.5e-05	338
1	90	0	0	81	-4.6e-05	316
1	33	1	0	66	5.6e-05	294
1	46	0	0	79	7.4e-05	271
2	42	0	0	73	3.1e-05	299
2	70	1	0	61	-4.5e-06	336
2	75	0	0	67	4.9e-05	277
2	74	0	1	63	5.1e-05	330
2	47	1	0	84	5.5e-05	254
2	73	0	1	68	2.7e-05	320

non-acute or *non-cardiac* issues.

4.2 Methods

Classification uses the same scikit-learn library used in Section 3.5. To streamline the initial data preparation, I have implemented a Flask [68] web frontend that allows classification cohorts to be defined (such as "patients with MI vs. patients without MI"), "static" features to be selected (such as gender or family history), and ECG parameters to be computed (such as "average QTc and STe from first 10 minutes of recording"). The output from this web frontend is a CSV file that is ready to be imported into sklearn. An example CSV is given in Table 4.1. A real spreadsheet will consist of 1172 rows (one per patient) and additional columns for other important features.

The CSV of features vs. classes is read by a python script which, as in the LQTS chapter, performs the following steps:

1. Convert inputs to floating-point values (or NaN).
2. Replace NaNs with the mean value from the column where they occurred.
3. Scale values so each column is centered on zero with unit variance.
4. Train and test a set of classifiers N times, with different testing vs. training data on each run (to determine the average performance of each classifier).

Categorical features (such as race, which may have 5–6 options) will be converted to a set of binary columns before this process begins.

In addition to the classifiers used in Chapter 3, we test three others:

1. **Linear Discriminant Analysis (LDA)**, which uses class conditional densities to determine a linear boundary between classes.
2. Linear SVM with **Stochastic Gradient Descent (SGD)** training.
3. **Naive Bayes (NB)** assuming a Gaussian likelihood for each feature.

4.3 Results

4.3.1 Initial Results

In this section, initial results are presented with untuned classifiers. Optimizations in terms of estimator parameters and input feature selection will be explored in the following section.

Table 4.2 presents the best results from all classifiers using default parameters. The following features were provided for the "record only" tests: gender, age, family history, history of coronary artery disease, history of hypertension, smoking,

Table 4.2: Predicting chest pain cause. Accuracy is the best average accuracy achieved by any classifier. In the "Best classifiers" column, we list all classifiers that performed similarly to the best (i.e., within 1–2%).

Task	Input	Best classifiers	Accuracy
Acute cardiac vs. Non-acute/non-cardiac	Record only	LDA, SVM, AdaBoost, NB	75.9%
	ECG only	SVM, RF, LDA	76.4%
	All data	SVM, AdaBoost, NB, RF, LDA	79.8%
Cardiac vs. Non-cardiac	Record only	LDA, NB, SVM	76.4%
	ECG only	LDA, SVM	76.0%
	All data	SVM, LDA	76.5%
MI vs. Non-MI	Record only	LDA, SVM, NB, AdaBoost, NN	91.7%
	ECG only	SVM, RF, NN	91.4%
	All data	SVM, AdaBoost, RF	95.0%

history of high cholesterol, and diabetes. From the ECG, the following features were used (average from the first 10 minutes): HR, P wave duration, PR interval, PR segment, QRS, ST segment, ST elevation at 5 time points, QT, QTcB, QTcF, QTp, QTpcB, QTpcF, JT, JTcB, JTcF, JTp, JTpcB, JTpcF, TpTe, T wave duration, and T wave amplitude. Because so many features are provided, we expect that some overfitting will occur. In the following section, we will focus on selecting only the best features for input. Patients in this database had been tested for several cardiac markers of MI including troponin and creatine kinase-MB (CK-MB). The "All data" rows in Table 4.2 include the medical record and ECG features above, as well as the initial troponin measurement (which would be available to the cardiologist within minutes).

Unfortunately, most of these initial results are equivalent to random guessing; in the the acute vs. non-acute group, 76.2% of patients fall into "non-acute". In the cardiac vs. non-cardiac group, 76.5% of patients are "cardiac". In the MI vs. non-MI group, 91.6% of patients are "non-MI". It is promising however that the third experiment was better than guessing in two of the tests, and we believe

it can be further improved through more rigorous feature selection and classifier optimizations. The decision tree based algorithms were able to provide 100% accuracy on training data, meaning that they were almost certainly overfitting. In the final row of the table, for example, AdaBoost and Random Forest achieved 98–99% accuracy on the training set but only about 94% accuracy on the test set. Overfitting was even more pronounced in the acute vs. non-acute test where, for instance, Random Forest had 98% accuracy on training data but only 78% accuracy on the test set.

4.3.2 Optimizing

In the previous section, we found that feature selection and parameter tuning would be required to get good results for these experiments. This prompted us to run chi-squared tests to determine which variables were likely to be useful. As a result, from the "medical record" fields, we eliminated gender. From the ECG measurements, we found that heart rate, ST segment duration, corrected JT and JTp intervals, and ST elevation 40–80ms after the J point were most likely to be useful. The initial troponin measurement was also highly ranked, which is expected as it is a typical method of diagnosing MI. This analysis allows us to eliminate approximately half of the features that were used to generate Table 4.2.

Additionally, we found that there was a lot of noise in the input features due to using averages (rather than medians), failing to clean up some obviously-wrong values, and using all leads rather than the "best" lead for each measurement. We therefore re-annotated the ECGs and measured most features only on the lead with the strongest T wave amplitude. This improves the accuracy of measurements such as QT, QTp, JT, JTp, TpTe, etc. For other features like RR and QRS, we

select the lead that provides the most stable value. Stability is determined by comparing measurements to "smoothed" versions of the same measurements; in other words, the lead that would be impacted least by a median filter is the one we will use.

In the "MI vs. non-MI" experiment, using the reduced set of features with the updated annotations improves the accuracy of Random Forest to 96%. 1066 out of 1073 non-MI patients were correctly identified as such, and 58 out of 98 MI patients were correctly identified (59% sensitivity, 99% specificity).

We noticed that annotation errors *themselves* might be useful input features. For instance, finding lots of ectopic beats, or failing to detect a P wave, may be more likely in a cardiac patient than a non-cardiac one. We defined a new set of ML features, in which we compared the number of *detections* of an ECG feature to the number of HR or R peak detections. We compare to R or HR because these measurements are least prone to error. For example, if HR is successfully measured 100,000 times in a recording, but the PR interval is only measured 90,000 times, the "PR per HR" feature will be equal to 0.9. These features are of course dependent on our annotation algorithm, but we expect that similar measures of noise could be made when other annotation algorithms are used. However, we ultimately found that these features were not useful in determining chest pain cause. We may revisit them in the stenosis detection experiment (Section 5.2.6).

In the first few minutes of the ECG, the patient is likely supine. In this relaxed state, changes to some markers such as STe may not be as pronounced as they would be during a period of elevated heart rate. We may decide, then, to drop the idea of "detecting as early as possible", and look later in the recording for more useful information. Of course, the physician could obtain the same information

earlier in the recording by increasing the patient's heart rate through light exercise.

In the following chapter, we will summarize the main results from the LQTS and AMI database analyses, and outline future research directions.

5 Conclusions

In this work, we have developed several new ways to assist a physician in decisions related to cardiac disease. First, we created a visual tool, the "ECG Clock", which allows evaluation and comparison of ECG biomarkers from individual Holters or sets of Holters. This tool proved useful in many ways, including diagnoses, unmasking concealed LQTS, and risk assessment. It was also used to evaluate the effects of a new drug. Then, we applied machine learning tools to ∼400GB of long-term ECG recordings, and achieved promising results in the tasks of genotyping, cardiac event prediction, and AMI detection. In Section 5.2, we outline possible improvements to these systems.

5.1 QT Clock Discussion

The QT Clock was shown to be extremely useful in reviewing long term QTc monitoring data. The use of Holter data to view dynamic changes in repolarization during a 24-hour time period allowed us to identify trends caused by different genotypes and medications. Patients with LQT1 typically have adverse cardiac events during high sympathetic tone. When looking at the degree of QT prolonga-

tion, our results reveal that the LQT1 patients are more likely to have diagnostic QTc prolongation during the day time hours; during the night, when sympathetic withdrawal occurs, a much lower degree of QTc prolongation can be observed. On the other hand, LQT2 patients show QT prolongation during the night time hours when compared to the day time hours. The number of patients with LQT3 in our study were too few to conclude any differences in QT prolongation over time.

While modern Holter systems may provide basic reporting on the QTc interval — e.g. maximum QTc, and percentage of time that it is prolonged — the QT clock can present the time, duration, and magnitude of all prolongation events in a simple picture. This is useful in the detection of concealed LQTS. Additionally, the clocks can reveal whether a patient is taking certain prescriptions correctly, if a prescription should be adjusted, or even what dose is likely to be safe for someone being started on a new drug. Further, the database we've developed can be used for purposes other than visualization, such as ML-based decision support. The increased availability of sensor data from a wide variety of patients will yield very refined characterizations of specific groups, differentiated by genetic mutation types, drug use, age, etc., allowing software to make diagnosis recommendations and even to predict the effects a prescription would have on a certain patient. Finally, we remind the reader that long-term QTc monitoring is only one example of a medical data visualization problem. The same techniques we've presented can immediately be extended to other features (such as heart rate) and other sensors (such as glucose monitors). Without these tools, the increasing volume of sensor data will become overwhelming to the clinicians who need to process it.

5.2 Limitations and Future Work

5.2.1 Performance Optimizations

We currently manage and process hundreds of Holter recordings on a desktop computer. For simplicity, the results are stored in SQLite databases. One research objective will be to scale this to a proper cloud-based database backend capable of handling larger data sets, and sources other than ECG recordings. In parallel with this, we will create a means to automatically update the group-specific statistics when new records are added to the database.

We have implemented a few optimizations such as multithreading and database indexing to improve the speed of both the initial data processing and the later querying/visualization stages of the system. However, as the database scales, there may be more work to do here. For instance, instead of relational databases, we could investigate solutions that are more well-suited for our time series data.

5.2.2 Multi-Lead Analysis

Our analysis in the LQTS group is based on Holter recordings acquired using 2 or 3 lead configurations. These leads were not systematically reported but they were V1-, VF-, or V5-like leads. Therefore, in this analysis we do not deliver a lead-based comparison in terms of identifying prolongation but we are limited to assess the presence of prolongation using the longest QT interval measured from non-standard leads in all beats. How the lead selection may affect our analysis, and the accuracy compared to standard 12-lead ECGs remains to be elucidated. In the AMI results, there is a similar limitation; we have data from 12 labeled leads, and though we have attempted to find the "best" lead for analysis (e.g.

by using the lead with the most stable measurements, or the strongest T waves), other lead selection techniques could be investigated.

5.2.3 Multi-Modal Visualization

ECG clocks are a simple way to visualize one factor, such as QTc or HR. Research is required to determine the best ways to aggregate heterogeneous sensor data into a clinically-relevant summary for a specific patient and for specific illnesses. For instance, we may want to view information about cholesterol or troponin on the same plot. Many drugs affect more than one ECG marker. When a patient is on the drug Tikosyn [69], for example, we need to look for irregularity in HR.

Satisfying these requirements may involve multiple plots/tables, nonlinear axes, and/or methods to combine the information onto the same plot (e.g. 3D views, or interactive animations). In the case of Tikosyn, it might suffice to plot *irregularity* as a single-bit Boolean value, such as a red dot that is associated with each QTc value. From the standpoint of data compression, this has the highest information ratio, but requires a better algorithm to detect RR-irregularity.

Another visualization challenge is finding the best way to detect and highlight very short-duration events; 24-hr plots only allow us to see events that last for at least a few minutes, which has insufficient resolution in some cases. The result must take only a few seconds for the doctor to review, without excluding any key information.

5.2.4 Machine Learning

TensorFlow [20], CNTK [21], and Amazon Machine Learning [22] are relatively recent ML developments that we have yet to test. All of these solutions appear

to be relatively simple to use and to collaborate with. The Amazon product is promising as a very high level solution, that will simplify the tuning process. TensorFlow, developed at Google, is likely to be useful for researching and training more complex neural networks.

Another avenue of research will arise as data collection and collaboration increases: the analysis of trends and disease outbreaks. This analysis may be more statistical in nature than what we've presented; i.e., machine learning may not play a major role.

In Fig. 3.20, the low values during the daytime (about 9AM–5PM) indicate that perhaps we don't even need to monitor that data. We have begun to identify the most useful features in our ML case studies, and will continue to experiment with novel features in further research. If there are indeed several hours which don't require observation, it may save battery life, storage, and processing time. Many optimizations remain in terms of classifier parameters as well, but tweaking them did not affect our performance very much at this stage. We therefore believe that it is more important to find the correct features before finding the optimal classifier configurations.

At this point, our ANN results are really only a very preliminary proof of concept. We must optimize this on two fronts: 1) neural network parameters (layers, etc.) and 2) input data. The input data side will be beneficial to the "conventional" algorithms as well. This is the research we just mentioned, where we will attempt to identify other ECG parameters to include in the input, and how to best reduce the dimensionality of the feature space. Further, we will attempt to hand-select only clean, complete recordings of phenotype-positive individuals as input; from initial testing, this may reduce our error by an order of magnitude!

We may also try a different branch of research, where we look at, for example, 1 hour of ECG data, and attempt to predict if there will be a cardiac event in the following hour. This will allow us to provide real-time warnings, rather than being limited to disease classification.

5.2.5 Extending the ECG Clock Library

We have demonstrated that QT clocks could be very useful for diagnosis and monitoring of the Long QT Syndrome. They are also instructive in research involving both the congenital and drug-induced forms of this disease. Likewise, we have seen that HR clocks can be used for monitoring HRV and drug response. We expect that researchers will want to apply (and extend) these concepts to many other cardiac features, which is why we have decided to make our source code freely available.

One example of an unconventional use of the ECG Clock is given in Figure 5.1. Elevated heart rate during sleep — i.e., failure of the heart rate to "dip" to a low enough level — has been associated with cardiovascular disease and an increased risk of all-cause mortality [70, 71]. Similarly, low heart rate variability (HRV) is an indicator of risk for cardiac events [72]. We therefore take a look at heart rate, HRV, and the rate of change of HR. On the left, we see that the patient's heart rate drops from ∼80 bpm during the day to ∼65 bpm at night. From this plot, we also get an idea of heart rate variability (HRV); the heart rate appears to fluctuate across a range of ∼10 bpm throughout the day. A `HRDerivClock` class was created as an example of how the ECG Clock can be extended. In this class, we redefine the default axis range, and apply a `derivative()` function to the heart rate data as it is loaded. We can then plot either the derivative at each data point, or its

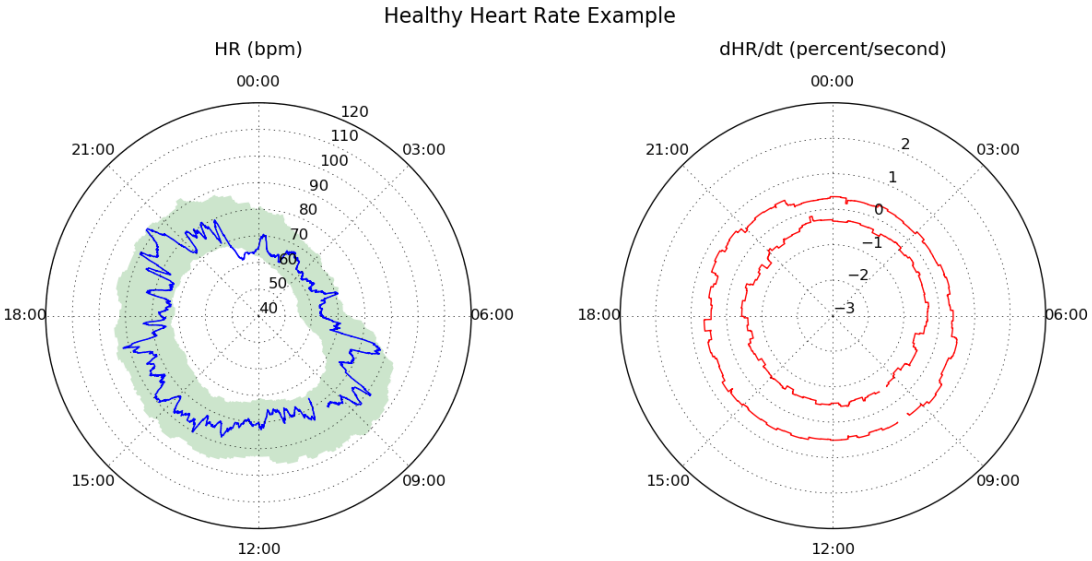


Figure 5.1: Viewing heart rate and its rate of change in a healthy individual. The plot on the left provides the doctor with a comprehensive picture of the patient's heart rate and HRV. The green region indicates the interquartile range for heart rate in healthy subjects. The red lines on the right represent the upper and lower bounds for rate of change, dHR/dt, normalized to *percentage* change rather than *bpm* change. This shows how fast the patient's heart rate is able to change, and is an example of how researchers may start to extend the ECG Clock library to display unconventional features.

upper and lower bounds within a sliding window. On the right side of Fig. 5.1, we show the upper and lower bounds. These bounds tend to stay at around ± 0.5 percent/second, meaning that a change from 80 to 65 bpm (about 20%) would take at least 40 seconds. We further note that changes take place more slowly at night. A very narrow range on this plot may indicate that the patient has trouble adapting their heart rate to different situations.

Finally, in order to make the library more accessible, we intend to develop a web interface so that non-programmers can simply upload data or annotations and generate clocks. Additionally, example IPython notebooks will simplify the process for programmers who prefer that environment.

5.2.6 Chest Pain Database Analysis

In Chapter 4, we presented only initial results from analysis of the THEW Chest Pain database. We have many goals for this database in the future:

- Improve the accuracy in the experiments already discussed, by algorithm tuning and feature selection. We may take some feature suggestions from [73] — e.g. ST slope, and maximum heart rate — and from [74] — e.g. JTp/JT, TpTe/JTp, TpTe/JT, QTp/QT, TpTe/QTp and TpTe/QT.

- Identify important features from the selection process which may previously have been unused by clinicians.

- Analyze failures of the system and see if they could be mitigated through confidence score thresholds.

- Improve our web interface that generates the ML inputs.

In addition to these goals, we would also like to explore other clinical problems, such as predicting short-term or long-term outcomes rather than chest pain cause. One interesting possibility is to determine the *location* of coronary artery stenosis — i.e., finding which artery is blocked — using features such as measurement differences between leads. This is the direction we will be pursuing in the immediate future.

Bibliography

[1] A. Page, O. Kocabas, T. Soyata, M. Aktas, and J.-P. Couderc, "Cloud-based privacy-preserving remote ECG monitoring and surveillance," *Annals of Noninvasive Electrocardiology*, vol. 20, no. 4, pp. 328–337, 2014.

[2] A. Page, O. Kocabas, S. Ames, M. Venkitasubramaniam, and T. Soyata, "Cloud-based secure health monitoring: Optimizing fully-homomorphic encryption for streaming algorithms," in *2014 IEEE Globecom Workshops (GC Wkshps)*. IEEE, 2014, pp. 48–52.

[3] M. Hassanalieragh, A. Page, T. Soyata, G. Sharma, M. Aktas, G. Mateos, B. Kantarci, and S. Andreescu, "Health monitoring and management using internet-of-things (IoT) sensing with cloud-based processing: Opportunities and challenges," in *Services Computing (SCC), 2015 IEEE International Conference On*. IEEE, 2015, pp. 285–292.

[4] A. Page, M. Hassanalieragh, T. Soyata, M. K. Aktas, B. Kantarci, and S. Andreescu, "Conceptualizing a real-time remote cardiac health monitoring system," in *Enabling Real-Time Mobile Cloud Computing through Emerging Technologies*. IGI Global, 2015, pp. 1–34.

[5] S. Ames, M. Venkitasubramaniam, A. Page, O. Kocabas, and T. Soyata, "Secure health monitoring in the cloud using homomorphic encryption: A branching-program formulation," in *Enabling Real-Time Mobile Cloud Computing through Emerging Technologies*. IGI Global, 2015, pp. 116–152.

[6] A. Page, M. K. Aktas, T. Soyata, W. Zareba, and J.-P. Couderc, "QT clock to improve detection of QT prolongation in long QT syndrome patients," *Heart Rhythm*, vol. 13, no. 1, pp. 190–198, 2016.

[7] A. Page, T. Soyata, J.-P. Couderc, M. Aktas, B. Kantarci, and S. Andreescu, "Visualization of health monitoring data acquired from distributed sensors for multiple patients," *Global Communications Conference (IEEE GLOBECOM 2015)*, 2015.

[8] A. Page, T. Soyata, J.-P. Couderc, and M. Aktas, "An open source ECG clock generator for visualization of long-term cardiac monitoring data," *IEEE Access*, vol. 3, pp. 2704–2714, 2015.

[9] A. Page, S. Hijazi, D. Askan, B. Kantarci, and T. Soyata, "Research directions in cloud-based decision support systems for health monitoring using internet-of-things driven data acquisition," *International Journal of Services Computing (IJSC)*, vol. 4, no. 4, pp. 18–34, 2016.

[10] G. Honan, A. Page, O. Kocabas, T. Soyata, and B. Kantarci, "Internet-of-Everything Oriented Implementation of Secure Digital Health (D-Health) Systems," in *IEEE International Symposium on Computers and Communications (ISCC)*, Messina, Italy, July 2016.

[11] A. Page, J. Hellawell, P. Yue, L. Belardinelli, W. Zareba, T. Soyata, and J.-P. Couderc, "Evaluating the effect of a novel late sodium current inhibitor (eleclazine) on the QT, QTpeak, and TpTe intervals in LQT3 patients using the QT clock concept," in *2016 Computing in Cardiology Conference (CinC)*, 2016.

[12] K. D. Kochanek, J. Xu, S. L. Murphy, A. M. Miniño, and H.-C. Kung, "Deaths: final data for 2009." *National vital statistics reports: from the Centers for Disease Control and Prevention, National Center for Health Statistics, National Vital Statistics System*, vol. 60, no. 3, pp. 1–116, 2011.

[13] "Clearbridge VitalSigns CardioLeaf PRO," http://www.clearbridgevitalsigns.com/pro.html, 2013.

[14] N. J. Holter, "New method for heart studies," *Science*, vol. 134, no. 3486, pp. 1214–1220, 1961.

[15] J. Couderc, "The Telemetric and Holter ECG Warehouse initiative (THEW): A data repository for the design, implementation and validation of ECG-related technologies," in *Engineering in Medicine and Biology Society (EMBC), 2010 Annual International Conference of the IEEE*. IEEE, 2010, pp. 6252–6255.

[16] T. Soyata, L. Copeland, and W. Heinzelman, "RF Energy Harvesting for Embedded Systems: A Survey of Tradeoffs and Methodology," *IEEE Circuits and Systems Magazine*, vol. 16, no. 1, pp. 22–57, Feb 2016.

[17] O. Kocabas, T. Soyata, and M. K. Aktas, "Emerging Security Mechanisms for Medical Cyber Physical Systems," *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, vol. 13, no. 3, pp. 401–416, Jun 2016.

[18] D. Son, J. Lee, S. Qiao, R. Ghaffari, J. Kim, J. E. Lee, C. Song, S. J. Kim, D. J. Lee, S. W. Jun, S. Yang, M. Park, J. Shin, K. Do, M. Lee, K. Kang, C. S. Hwang, N. Lu, T. Hyeon, , and D.-H. Kim, "Multifunctional wearable devices for diagnosis and therapy of movement disorders," *Nature Nanotechnology*, pp. 1–8, 2014.

[19] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[20] "TensorFlow," https://www.tensorflow.org/, 2016.

[21] Microsoft, "Computational Network Toolkit," https://github.com/Microsoft/CNTK, 2016.

[22] "Amazon Machine Learning," https://aws.amazon.com/machine-learning/, 2016.

[23] National Institute of Standards and Technology, "Advanced encryption standard (AES)," Nov. 2001, fIPS-197.

[24] W. Zhao, C. Wang, and Y. Nakahira, "Medical application on internet of things," in *IET Int. Conf. on Com. Tech. and Application (ICCTA 2011)*, Oct 2011, pp. 660–665.

[25] F. Hu, D. Xie, and S. Shen, "On the application of the internet of things in the field of medical and health care," in *IEEE Int. Conf. on and IEEE Cyber, Physical and Social Computing Green Computing and Communications (GreenCom),(iThings/CPSCom)*, Aug 2013, pp. 2053–2058.

[26] MegaKoto Ltd., "Mobile ECG Telemetry Solution (METS)," http://www.megakoto.fi/en/page/1747, 2013.

[27] Sensys Medical, Inc., "Near-Infrared Spectroscopy," http://www.diabetesnet.com/diabetes-technology/meters-monitors/future-meters-monitors/sensys-medical.

[28] T. Soyata, H. Ba, W. Heinzelman, M. Kwon, and J. Shi, "Accelerating mobile cloud computing: A survey," in *Communication Infrastructures for Cloud Computing*, H. T. Mouftah and B. Kantarci, Eds. IGI Global, Sep 2013, ch. 8, pp. 175–197.

[29] T. Soyata, *Enabling Real-Time Mobile Cloud Computing through Emerging Technologies*. IGI Global, Aug 2015.

[30] T. Soyata, R. Muraleedharan, C. Funai, M. Kwon, and W. Heinzelman, "Cloud-Vision: Real-Time Face Recognition Using a Mobile-Cloudlet-Cloud Acceleration Architecture," in *Proceedings of the 17th IEEE Symposium on Computers and Communications (IEEE ISCC 2012)*, Cappadocia, Turkey, Jul 2012, pp. 59–66.

[31] T. Soyata, R. Muraleedharan, S. Ames, J. H. Langdon, C. Funai, M. Kwon, and W. B. Heinzelman, "COMBAT: mobile Cloud-based cOmpute/coMmunications infrastructure for BATtlefield applications," in *Proceedings of SPIE*, vol. 8403, May 2012, pp. 84 030K–84 030K.

[32] N. Powers, A. Alling, K. Osolinsky, T. Soyata, M. Zhu, H. Wang, H. Ba, W. Heinzelman, J. Shi, and M. Kwon, "The Cloudlet Accelerator: Bringing Mobile-Cloud Face Recognition into Real-Time," in *Globecom Workshops (GC Wkshps)*, San Diego, CA, Dec 2015, pp. 1–7.

[33] "Mini PC: Intel NUC," http://www.intel.com/content/www/us/en/nuc/overview.html, 2016.

[34] F. Badilini, "The ISHNE Holter standard output file format," *Annals of noninvasive electrocardiology*, vol. 3, no. 3, pp. 263–266, 1998.

[35] Y. Chesnokov, D. Nerukh, and R. Glen, "Individually adaptable automatic QT detector," in *Computers in Cardiology, 2006*. IEEE, 2006, pp. 337–340.

[36] A. Demski and M. L. Soria, "ecg-kit: a matlab toolbox for cardiovascular signal processing," *Journal of Open Research Software*, vol. 4, no. 1, 2016.

[37] "MySQL," https://www.mysql.com/, 2016.

[38] "MariaDB," https://mariadb.org/, 2016.

[39] A. L. Goldberger, L. A. N. Amaral, L. Glass, J. M. Hausdorff, P. C. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C.-K. Peng, and H. E. Stanley, "PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals," *Circulation*, vol. 101, no. 23, pp. e215–e220, 2000 (June 13).

[40] J. W. Mason, D. J. Ramseth, D. O. Chanter, T. E. Moon, D. B. Goodman, and B. Mendzelevski, "Electrocardiographic reference ranges derived from 79,743 ambulatory subjects," *Journal of electrocardiology*, vol. 40, no. 3, pp. 228–234, 2007.

[41] H. C. Bazett, "An analysis of time relations of the electrocardiogram," *Heart*, vol. 7, pp. 353–370, 1920.

[42] L. S. Fridericia, "Die systolendauer im elektrokardiogramm bei normalen menschen und bei herzkranken," *Acta Medica Scandinavica*, vol. 53, pp. 469–486, 1920.

[43] H. Morita, J. Wu, and D. P. Zipes, "The QT syndromes: long and short," *The Lancet*, vol. 372, no. 9640, pp. 750 – 763, 2008.

[44] G. M. Vincent, MD, "The molecular genetics of the Long QT Syndrome: genes causing fainting and sudden death," *Annual review of medicine*, vol. 49, no. 1, pp. 263–274, 1998.

[45] P. L. Hedley, P. Jørgensen, S. Schlamowitz, R. Wangari, J. Moolman-Smook, P. A. Brink, J. K. Kanters, V. A. Corfield, and M. Christiansen, "The genetic basis of Long QT and Short QT Syndromes: A mutation update," *Human Mutation*, vol. 30, no. 11, pp. 1486–1511, 2009.

[46] R. Shah, "Drug-induced qt interval prolongation: regulatory perspectives and drug development," *Annals of medicine*, vol. 36, no. S1, pp. 47–52, 2004.

[47] T. Lorberbaum, K. J. Sampson, R. L. Woosley, R. S. Kass, and N. P. Tatonetti, "An integrative data science pipeline to identify novel drug interactions that prolong the QT interval," *Drug safety*, vol. 39, no. 5, pp. 433–441, 2016.

[48] W. Zareba, A. J. Moss, P. J. Schwartz, G. M. Vincent, J. L. Robinson, S. G. Priori, J. Benhorin, E. H. Locati, J. A. Towbin, M. T. Keating *et al.*, "Influence of the genotype on the clinical course of the long-QT syndrome," *New England Journal of Medicine*, vol. 339, no. 14, pp. 960–965, 1998.

[49] W. Shimizu, A. J. Moss, A. A. Wilde, J. A. Towbin, M. J. Ackerman, C. T. January, D. J. Tester, W. Zareba, J. L. Robinson, M. Qi *et al.*, "Genotype-phenotype aspects of type 2 long qt syndrome," *Journal of the American College of Cardiology*, vol. 54, no. 22, pp. 2052–2062, 2009.

[50] P. J. Schwartz, S. G. Priori, C. Spazzolini, A. J. Moss, G. M. Vincent, C. Napolitano, I. Denjoy, P. Guicheney, G. Breithardt, M. T. Keating *et al.*, "Genotype-phenotype correlation in the long-QT syndrome gene-specific triggers for life-threatening arrhythmias," *Circulation*, vol. 103, no. 1, pp. 89–95, 2001.

[51] M. Stramba-Badiale, S. G. Priori, C. Napolitano, E. H. Locati, X. Vinolas, W. Haverkamp, E. Schulze-Bahr, K. Goulene, and P. J. Schwartz, "Gene-specific differences in the circadian variation of ventricular repolarization in the Long QT Syndrome: a key to sudden death during sleep?" *Ital Heart J*, vol. 1, no. 5, pp. 323–328, May 2000.

[52] I. Goldenberg, S. Horr, A. J. Moss, C. M. Lopes, A. Barsheshet, S. McNitt, W. Zareba, M. L. Andrews, J. L. Robinson, E. H. Locati *et al.*, "Risk for life-threatening cardiac events in patients with genotype-confirmed long-QT syndrome and normal-range corrected QT intervals," *Journal of the American College of Cardiology*, vol. 57, no. 1, pp. 51–59, 2011.

[53] A. Barsheshet, O. Dotsenko, and I. Goldenberg, "Genotype-specific risk stratification and management of patients with Long QT Syndrome," *Annals of Noninvasive Electrocardiology*, vol. 18, no. 6, pp. 499–509, 2013.

[54] S. Viskin, P. G. Postema, Z. A. Bhuiyan, R. Rosso, J. M. Kalman, J. K. Vohra, M. E. Guevara-Valdivia, M. F. Marquez, E. Kogan, B. Belhassen *et al.*, "The response of the QT interval to the brief tachycardia provoked by standing: a bedside test for diagnosing long QT syndrome," *Journal of the American College of Cardiology*, vol. 55, no. 18, pp. 1955–1961, 2010.

[55] W. Shimizu, T. Noda, H. Takaki, T. Kurita, N. Nagaya, K. Satomi, K. Suyama, N. Aihara, S. Kamakura, K. Sunagawa *et al.*, "Epinephrine unmasks latent mutation carriers with LQT1 form of congenital long-QT syndrome," *Journal of the American College of Cardiology*, vol. 41, no. 4, pp. 633–642, 2003.

[56] T. Noda, H. Takaki, T. Kurita, K. Suyama, N. Nagaya, A. Taguchi, N. Aihara, S. Kamakura, K. Sunagawa, K. Nakamura *et al.*, "Gene-specific response of dynamic ventricular repolarization to sympathetic stimulation in LQT1, LQT2 and LQT3 forms of congenital long qt syndrome," *European Heart Journal*, vol. 23, no. 12, pp. 975–983, 2002.

[57] C.-u. Choe, E. Schulze-Bahr, A. Neu, J. Xu, Z. I. Zhu, K. Sauter, R. Bähring, S. Priori, P. Guicheney, G. Mönnig *et al.*, "C-terminal HERG (LQT2) mutations disrupt IKr channel regulation through 14-3-3 epsilon," *Human molecular genetics*, vol. 15, no. 19, pp. 2888–2902, 2006.

[58] S. G. Priori, P. J. Schwartz, C. Napolitano, R. Bloise, E. Ronchetti, M. Grillo, A. Vicentini, C. Spazzolini, J. Nastoli, G. Bottelli *et al.*, "Risk stratification in the Long-QT Syndrome," *New England Journal of Medicine*, vol. 348, no. 19, pp. 1866–1874, 2003.

[59] D. Migdalovich, A. J. Moss, C. M. Lopes, J. Costa, G. Ouellet, A. Barsheshet, S. McNitt, S. Polonsky, J. L. Robinson, W. Zareba *et al.*, "Mutation and gender-specific risk in type 2 long qt syndrome: implications for risk stratification for life-threatening cardiac events in patients with long QT syndrome," *Heart Rhythm*, vol. 8, no. 10, pp. 1537–1543, 2011.

[60] U.S. Department of Health and Human Services, "Health information privacy," http://www.hhs.gov/hipaa/for-professionals/privacy/laws-regulations/, 2016.

[61] E. C. Larson, M. Goel, G. Boriello, S. Heltshe, M. Rosenfeld, and S. N. Patel, "Spirosmart: using a microphone to measure lung function on a mobile phone," in *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*. ACM, 2012, pp. 280–289.

[62] NVIDIA, "NVIDIA DIGITS — Interactive Deep Learning GPU Training System," https://developer.nvidia.com/digits, November 2015.

[63] Berkeley Vision and Learning Center, "Caffe deep learning framework," http://caffe.berkeleyvision.org/, November 2015.

[64] Y. LeCun, C. Cortes, and C. J. Burges, "The MNIST database of handwritten digits," 1998.

[65] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov 1998.

[66] W. B. White, "Cardiovascular risk and therapeutic intervention for the early morning surge in blood pressure and heart rate," *Blood pressure monitoring*, vol. 6, no. 2, pp. 63–72, 2001.

[67] A. Maas and Y. Appelman, "Gender differences in coronary heart disease," *Netherlands Heart Journal*, vol. 18, no. 12, pp. 598–603, 2010.

[68] "Flask (a python microframework)," http://flask.pocoo.org/, 2016.

[69] FDA, "Tikosyn (dofetilide), NDA 20-931, risk evaluation and mitigation strategy document," http://www.fda.gov/downloads/Drugs/DrugSafety/PostmarketDrugSafetyInformationforPatientsandProviders/UCM266277.pdf, Tech. Rep., August 2011.

[70] I. Z. Ben-Dov, J. D. Kark, D. Ben-Ishay, J. Mekler, L. Ben-Arie, and M. Bursztyn, "Blunted heart rate dip during sleep and all-cause mortality," *Archives of Internal Medicine*, vol. 167, no. 19, pp. 2116–2121, 2007.

[71] K. Eguchi, S. Hoshide, J. Ishikawa, T. G. Pickering, J. E. Schwartz, K. Shimada, and K. Kario, "Nocturnal nondipping of heart rate predicts cardiovascular events in hypertensive patients," *J. Hypertens.*, vol. 27, no. 11, pp. 2265–2270, Nov 2009.

[72] H. Tsuji, M. G. Larson, F. J. Venditti, E. S. Manders, J. C. Evans, C. L. Feldman, and D. Levy, "Impact of reduced heart rate variability on risk for cardiac events: The Framingham heart study," *Circulation*, vol. 94, no. 11, pp. 2850–2855, 1996.

[73] S. Palaniappan and R. Awang, "Intelligent heart disease prediction system using data mining techniques," in *2008 IEEE/ACS International Conference on Computer Systems and Applications*, March 2008, pp. 108–115.

[74] H. Maldonado, L. Leija, A. Vera, and C. Alvarado, "Post myocardial infarct detection with support vector machine and ECG intervals ratios JTp/JT, Tpe/JTp and Tpe/JT," in *2016 Global Medical Engineering Physics Exchanges/Pan American Health Care Exchanges (GMEPE/PAHCE)*, April 2016, pp. 1–4.

A Electrocardiographic Reference Ranges

In this appendix, we present statistics computed from beat-to-beat annotations of the THEW E-HOL-03-0202-003 (Healthy) and E-HOL-03-0480-013 (Genotyped LQTS) databases.

The age of one patient in the healthy database was not recorded; this subject is excluded from the age-dependent rows. Similarly, in the Long QT population, eight subjects have no recorded age, and three of them have no recorded gender, so they are excluded from the appropriate rows.

We note that our annotation algorithm performed poorly on PR measurements, and P wave detection in general. This is because we did not manually adjust the P wave detection parameters for each Holter, nor did we attempt to select the best lead — we simply took the median across the 2–3 available leads.

Table A.1: Heart rate statistics (in beats per minute) for healthy population.

	24-hour statistics				3-hour medians							
Group	Median	Mean	SD	N (subjects)	12AM–3AM	3AM–6AM	6AM–9AM	9AM–12PM	12PM–3PM	3PM–6PM	6PM–9PM	9PM–12AM
All	79	81	18	202	68	65	78	86	86	85	84	78
Male	77	79	18	102	66	63	77	84	84	82	82	76
Female	81	82	18	100	69	67	80	88	88	87	86	80
0-19	82	84	21	21	67	65	84	88	89	91	86	78
20-39	79	80	19	88	66	63	78	86	85	85	84	78
40-59	81	82	17	70	70	67	79	86	87	85	85	80
60-	76	79	17	22	69	66	72	85	82	82	81	75
M (0-19)	83	84	21	11	69	65	75	89	86	92	86	82
M (20-39)	75	77	18	49	63	60	72	82	82	80	80	75
M (40-59)	80	82	18	35	69	66	83	86	86	84	85	79
M (60-)	74	76	16	7	66	65	75	82	86	81	78	72
F (0-19)	81	83	21	10	65	65	86	85	92	91	86	75
F (20-39)	83	84	19	39	70	67	85	90	89	90	88	82
F (40-59)	81	82	17	35	71	67	76	87	88	85	85	81
F (60-)	77	80	18	15	70	67	71	87	81	83	81	76

Table A.2: QRS statistics (in ms) for healthy population. The resolution of the recordings was 5ms, resulting in identical median values across most times/groups. We can see that QRS is slightly longer in males, though, particularly during sleep.

	24-hour statistics				3-hour medians							
Group	Median	Mean	SD	N (subjects)	12AM–3AM	3AM–6AM	6AM–9AM	9AM–12PM	12PM–3PM	3PM–6PM	6PM–9PM	9PM–12AM
All	90	92	11	202	90	95	90	90	90	90	90	90
Male	90	94	10	102	95	95	90	90	90	90	90	90
Female	90	90	12	100	90	90	90	90	90	90	90	90
0-19	90	91	11	21	90	90	90	90	90	90	90	90
20-39	90	92	11	88	95	95	90	90	90	90	90	90
40-59	90	92	11	70	90	95	90	90	90	90	90	90
60-	90	92	13	22	90	90	90	90	90	90	90	90
M (0-19)	90	92	10	11	95	90	90	90	90	90	90	90
M (20-39)	90	95	10	49	95	95	90	90	90	90	90	90
M (40-59)	90	93	11	35	95	95	90	90	90	90	90	90
M (60-)	90	93	12	7	90	90	90	90	90	90	90	90
F (0-19)	90	90	11	10	90	90	90	90	90	90	90	90
F (20-39)	90	90	11	39	90	90	90	90	90	90	90	90
F (40-59)	90	90	12	35	90	90	90	90	90	90	90	90
F (60-)	90	91	14	15	90	90	90	90	90	90	90	90

Table A.3: PR statistics (in ms) for healthy population. These appear to be about 20ms shorter than what is found in literature. Also, we expected a fairly obvious correlation with age, which appears to be absent here.

	24-hour statistics				3-hour medians							
Group	Median	Mean	SD	N (subjects)	12AM–3AM	3AM–6AM	6AM–9AM	9AM–12PM	12PM–3PM	3PM–6PM	6PM–9PM	9PM–12AM
All	135	141	32	202	145	145	135	130	130	135	135	135
Male	140	143	30	102	145	145	140	135	135	135	135	140
Female	135	139	34	100	140	140	130	130	130	130	130	135
0-19	135	140	34	21	140	140	130	130	125	130	135	135
20-39	135	143	33	88	145	145	135	135	135	135	135	140
40-59	135	140	32	70	140	140	135	130	130	130	130	135
60-	140	141	29	22	145	145	135	135	135	135	135	140
M (0-19)	130	139	37	11	140	140	140	130	125	125	130	130
M (20-39)	140	145	30	49	145	150	140	135	135	135	135	140
M (40-59)	140	142	28	35	145	145	140	135	135	135	135	140
M (60-)	145	145	29	7	150	145	145	140	140	140	140	145
F (0-19)	135	141	30	10	150	150	130	130	130	135	135	140
F (20-39)	135	141	36	39	140	140	135	130	130	130	130	135
F (40-59)	130	137	34	35	135	140	135	130	125	125	125	130
F (60-)	135	140	29	15	145	145	130	130	135	135	135	135

Table A.4: QTcB statistics (in ms) for healthy population.

	24-hour statistics				3-hour medians							
Group	Median	Mean	SD	N (subjects)	12AM–3AM	3AM–6AM	6AM–9AM	9AM–12PM	12PM–3PM	3PM–6PM	6PM–9PM	9PM–12AM
All	407	412	39	202	404	402	408	407	409	409	408	407
Male	398	404	37	102	394	392	400	399	401	401	400	399
Female	415	421	40	100	414	411	414	414	417	416	417	415
0-19	403	411	45	21	399	397	403	402	406	408	406	405
20-39	404	410	40	88	400	398	405	406	405	406	405	405
40-59	408	413	36	70	405	404	410	407	410	409	409	408
60-	419	423	38	22	421	415	413	417	422	420	423	418
M (0-19)	397	407	46	11	397	393	394	396	398	400	398	403
M (20-39)	395	400	35	49	391	389	397	397	398	398	397	396
M (40-59)	401	406	33	35	395	394	404	401	403	403	402	401
M (60-)	412	416	42	7	406	402	410	408	420	418	416	405
F (0-19)	409	415	44	10	401	401	410	408	413	414	413	407
F (20-39)	414	420	42	39	413	409	413	414	414	416	415	415
F (40-59)	415	421	37	35	414	413	415	414	418	414	416	415
F (60-)	422	427	36	15	426	420	415	420	423	421	426	423

Table A.5: QTcF statistics (in ms) for healthy population.

	24-hour statistics				3-hour medians							
Group	Median	Mean	SD	N (subjects)	12AM–3AM	3AM–6AM	6AM–9AM	9AM–12PM	12PM–3PM	3PM–6PM	6PM–9PM	9PM–12AM
All	390	394	31	202	396	397	391	384	386	387	387	391
Male	383	387	29	102	388	389	386	378	380	381	380	384
Female	397	401	32	100	404	404	397	390	393	393	394	398
0-19	385	391	35	21	393	392	386	381	382	382	381	387
20-39	388	392	31	88	394	395	389	383	384	385	384	388
40-59	390	394	30	70	395	398	393	384	387	387	388	391
60-	404	406	32	22	411	408	401	395	402	403	405	405
M (0-19)	379	386	36	11	386	388	383	374	375	376	375	381
M (20-39)	382	385	27	49	387	390	386	378	379	379	379	382
M (40-59)	383	387	27	35	387	388	385	379	381	382	382	385
M (60-)	400	401	35	7	406	402	397	393	398	400	402	401
F (0-19)	389	396	33	10	399	396	387	385	386	387	387	391
F (20-39)	394	399	33	39	404	403	394	389	389	391	391	396
F (40-59)	398	401	30	35	404	406	401	391	395	393	395	397
F (60-)	406	408	30	15	413	410	402	397	404	404	407	406

Table A.6: Heart rate statistics (in beats per minute) for long QT population.

	24-hour statistics				3-hour medians							
Group	Median	Mean	SD	N (subjects)	12AM–3AM	3AM–6AM	6AM–9AM	9AM–12PM	12PM–3PM	3PM–6PM	6PM–9PM	9PM–12AM
All	75	78	21	480	66	64	72	81	81	82	81	73
Male	74	77	21	222	65	63	71	79	81	81	79	71
Female	75	79	22	255	67	65	73	82	81	82	81	74
0-19	79	83	24	229	69	69	76	86	86	88	87	76
20-39	70	72	17	149	63	59	67	75	75	75	75	70
40-59	72	73	16	72	65	62	69	78	77	78	77	71
60-	71	71	15	22	64	63	70	75	75	76	74	68
M (0-19)	78	81	22	127	67	67	75	83	86	87	86	74
M (20-39)	67	69	17	46	61	56	63	72	73	73	71	67
M (40-59)	71	72	18	32	62	59	67	77	78	78	75	70
M (60-)	69	70	15	15	66	64	67	71	73	75	73	69
F (0-19)	82	86	25	102	71	71	78	90	88	89	88	78
F (20-39)	71	74	17	103	63	60	69	77	77	77	77	71
F (40-59)	73	74	15	40	66	65	71	78	77	77	78	71
F (60-)	72	74	15	7	62	62	75	85	78	80	75	67

Table A.7: QRS statistics (in ms) for long QT population. Again, the 5ms resolution of the recordings makes for fairly static results, but we can still see that QRS is slightly longer in males. This group also appears to have slightly longer QRS than the healthy population.

	24-hour statistics				3-hour medians							
Group	Median	Mean	SD	N (subjects)	12AM–3AM	3AM–6AM	6AM–9AM	9AM–12PM	12PM–3PM	3PM–6PM	6PM–9PM	9PM–12AM
All	90	93	16	480	95	95	90	90	90	90	90	95
Male	95	94	17	222	95	95	95	90	90	90	90	95
Female	90	92	15	255	95	95	90	90	90	90	90	90
0-19	90	91	16	229	95	90	90	90	90	90	90	90
20-39	95	95	15	149	95	100	95	90	95	95	95	95
40-59	95	94	15	72	95	95	95	90	90	90	90	95
60-	95	96	17	22	95	100	95	95	90	95	90	95
M (0-19)	90	92	17	127	95	95	90	90	90	90	90	95
M (20-39)	95	99	15	46	100	100	95	95	95	95	95	100
M (40-59)	95	95	15	32	95	100	95	95	90	90	95	95
M (60-)	95	97	17	15	95	100	100	95	95	95	95	95
F (0-19)	90	90	15	102	90	90	90	85	85	90	90	90
F (20-39)	95	93	15	103	95	95	95	90	90	90	90	95
F (40-59)	95	93	16	40	95	95	95	90	90	90	90	95
F (60-)	95	93	15	7	100	95	95	85	90	90	90	100

Table A.8: PR statistics (in ms) for long QT population. Again, this portion of the detection algorithm did not appear to work properly.

	24-hour statistics				3-hour medians							
Group	Median	Mean	SD	N (subjects)	12AM–3AM	3AM–6AM	6AM–9AM	9AM–12PM	12PM–3PM	3PM–6PM	6PM–9PM	9PM–12AM
All	125	131	36	480	130	130	125	120	120	120	120	125
Male	120	129	37	222	125	125	120	120	120	120	120	120
Female	125	133	35	255	130	135	130	125	125	125	125	130
0-19	115	124	37	229	120	120	115	110	115	115	115	115
20-39	130	136	34	149	135	135	135	125	125	125	130	135
40-59	130	137	34	72	140	140	130	125	130	125	130	135
60-	140	147	36	22	145	145	135	135	135	140	140	140
M (0-19)	110	120	36	127	115	115	110	110	110	110	110	110
M (20-39)	130	136	33	46	135	135	130	125	130	130	130	130
M (40-59)	135	142	32	32	145	145	135	130	130	130	135	140
M (60-)	160	158	35	15	175	170	155	150	150	155	160	160
F (0-19)	125	129	37	102	125	125	120	120	120	120	120	125
F (20-39)	130	136	34	103	140	140	135	125	125	125	130	135
F (40-59)	130	133	35	40	135	135	125	120	125	125	125	130
F (60-)	125	127	28	7	135	130	125	120	115	120	125	130

Table A.9: QTcB statistics (in ms) for long QT population.

	24-hour statistics				3-hour medians							
Group	Median	Mean	SD	N (subjects)	12AM–3AM	3AM–6AM	6AM–9AM	9AM–12PM	12PM–3PM	3PM–6PM	6PM–9PM	9PM–12AM
All	460	462	49	480	457	457	461	461	461	462	461	461
Male	456	458	48	222	454	451	456	456	456	457	456	456
Female	464	467	49	255	460	461	465	464	466	466	465	464
0-19	466	468	50	229	466	466	467	465	465	467	465	467
20-39	457	460	44	149	453	451	459	459	459	460	459	459
40-59	449	450	49	72	444	442	452	451	451	451	450	447
60-	461	462	43	22	454	457	463	464	464	462	464	458
M (0-19)	465	466	49	127	468	465	467	464	463	464	463	467
M (20-39)	444	447	42	46	438	433	440	448	446	448	445	446
M (40-59)	435	436	45	32	429	424	435	438	435	442	437	435
M (60-)	456	459	40	15	454	455	453	457	458	458	459	456
F (0-19)	467	471	52	102	464	466	466	467	469	470	467	467
F (20-39)	463	466	44	103	457	457	466	463	464	464	465	463
F (40-59)	462	463	48	40	459	457	464	461	464	461	463	464
F (60-)	466	468	48	7	454	459	471	473	470	466	469	462

Table A.10: QTcF statistics (in ms) for long QT population.

	24-hour statistics				3-hour medians							
Group	Median	Mean	SD	N (subjects)	12AM–3AM	3AM–6AM	6AM–9AM	9AM–12PM	12PM–3PM	3PM–6PM	6PM–9PM	9PM–12AM
All	441	441	41	480	445	446	443	438	438	439	438	443
Male	437	438	40	222	444	442	439	435	434	435	434	440
Female	445	445	42	255	446	449	448	442	443	442	442	446
0-19	442	442	43	229	449	449	444	437	437	438	436	445
20-39	443	444	37	149	444	445	445	441	441	442	441	445
40-59	434	433	43	72	436	434	438	433	434	434	433	433
60-	450	447	39	22	447	451	453	447	451	450	452	447
M (0-19)	442	442	41	127	453	450	444	438	436	436	435	446
M (20-39)	434	436	34	46	435	435	434	434	433	436	434	436
M (40-59)	422	420	40	32	424	423	423	420	419	423	419	421
M (60-)	447	445	36	15	445	450	446	444	449	447	449	443
F (0-19)	442	443	44	102	445	447	443	437	438	440	437	443
F (20-39)	447	447	38	103	447	451	452	445	445	444	446	449
F (40-59)	447	445	42	40	450	449	451	444	447	444	445	450
F (60-)	454	450	44	7	450	455	460	453	453	452	454	454