A Unique Digital Electrocardiographic Repository for the Development of Quantitative Electrocardiography and Cardiac Safety: The Telemetric and Holter ECG Warehouse (THEW)

Jean-Philippe Couderc, PhD[*]

Jean-Philippe Couderc: jean-philippe.couderc@heart.rochester.edu

[*] Center for Quantitative Electrocardiography and Cardiac Safety, University of Rochester Medical Center, Rochester, NY 14618 USA, Phone: 585-275-1096; Fax: 585-273-5283

Abstract

The sharing of scientific data reinforces open scientific inquiry; it encourages diversity of analysis and opinion while promoting new research and facilitating the education of next generations of scientists. In this article, we present an initiative for the development of a repository containing continuous electrocardiographic information and their associated clinical information. This information is shared with the worldwide scientific community in order to improve quantitative electrocardiology and cardiac safety.

First, we present the objectives of the initiative and its mission. Then, we describe the resources available in this initiative following three components: data, expertise and tools. The Data available in the Telemetric and Holter ECG Warehouse (THEW) includes continuous ECG signals and associated clinical information. The initiative attracted various academic and private partners whom expertise covers a large list of research arenas related to quantitative electrocardiography; their contribution to the THEW promotes cross-fertilization of scientific knowledge, resources, and ideas that will advance the field of quantitative electrocardiography. Finally, the tools of the THEW include software and servers to access and review the data available in the repository.

To conclude, the THEW is an initiative developed to benefit the scientific community and to advance the field of quantitative electrocardiography and cardiac safety. It is a new repository designed to complement the existing ones such as Physionet, the AHA-BIH Arrhythmia Database, and the CSE database. The THEW hosts unique datasets from clinical trials and drug safety studies that, so far, were not available to the worldwide scientific community.

Keywords

database; ECG/EKG; electrocardiogram; cardiac safety; QT; long QT syndrome; acute myocardial infaction; coronary syndrome; acute myocardial infarction; thorough QT studies; healthy normal; torsades de pointes

I. INTRODUCTION

Despite the effort implemented to reduce the number of sudden cardiac death in the U.S. by early advanced care, there is a clear need for the improvement of risk stratification techniques to optimize the use of prophylactic therapies such as implantable defibrillators and drug therapies. Meanwhile, cardiac safety is also one of the most challenging hurdles in the development of new molecular entities. It has been estimated that as many as 86% of all drugs tested in pharmaceutical development show specific inhibitory activity of potassium ion kinetics, which in some cases can lead to torsades de pointes and potentially to sudden cardiac death.

The University of Rochester Medical Center and the Heart Research follow-Up Program (NY) enabled the creation of the "Center for Quantitative Electrocardiography and Cardiac Safety" (CES). The CES activities are defined around 1) the development and maintenance of computer resources (data storage and computing center) for data sharing and analysis, 2) the deployment of platforms for the sharing of medical information, and 3) the formation of a scientific network to initiate collaborative research projects. The CES distributes the resources from the Telemetric and Holter ECG Warehouse (THEW) to the international scientific community, sharing unique clinical and ECG information for the design and validation of technologies to improve quantitative electrocardiography and cardiac safety.

In addition, the THEW includes a partnership with the U.S. Food and Drug Administration (FDA, see FDA private-public partnership (PPP) in appendix). Under its public health mission, FDA is interested in the development of improved technologies to evaluate drug safety and efficacy [1]. This FDA partnership was designed to leverage resources and expertise toward the implementation of collaborative projects among FDA and other public and private stakeholders. Out of this collaborative effort, specific projects were started to expand the data in the CES repository (THEW), and facilitate scientific projects toward the development, testing and validation of ECG-related technologies.

Finally, the THEW has developed collaborations with multiple academic centers in the United States and in Europe. Private corporations have submitted research proposals as well and the initiative has spawned collaborative projects between academia and industries. In this article, we provide insights into the roles, the structures and the resources we have developed that led to these achievements.

II. The THEW INIATIVE

A. Mission Statement

The objective of the Telemetric and Holter ECG Warehouse (THEW) is to provide access to electrocardiographic data to research organizations for the design and validation of analytic methods to advance the field of quantitative electrocardiography with a strong focus on cardiac safety.

B. Information Technologies Infrastructures

The overall infrastructure of the warehouse is schematized in Figure 1. The ECG signals from the warehouse are currently hosted in two servers: 1) a SFTP server located at University of Buffalo (NY) and 2) a client-based access server located at University of Rochester Medical Center (NY). Free access to the data is available to academic centers and to not-for profit organizations.

Our resources include an IBM BlueGene/P super computer as well. The Blue Gene/P system consists of 1,024 nodes, 4,096 CPU cores, 2 TB of RAM, and 180 TB of storage. Access to

this computer resource to THEW users is currently in development. The computer is hosted at the Rochester Center for Research Computing (see Appendix).

C. Warehouse Content

C.1 ECG recordings—The data available in the warehouse were provided by research academic centers and major pharmaceutical companies. The current list of database is reported in Table 1. The hosted ECG recordings are continuous with a length varying from 10 minutes to 24 hours. The ECG signals in the warehouse have different sampling frequency (180Hz, 200Hz and 1000 Hz) and different amplitude resolution (coded from 10, 12 or 16 bits) depending on the database. The lead configuration depends also on the recording equipment as well. Currently, the datasets contain either 3 or 12 lead recordings. Three leads recordings are recorded using a pseudo-orthogonal configuration (X, Y, and Z). Twelve leads followed Holter configuration in which limb leads are referenced to the torso (treadmill configuration or Mason-Likar lead placement) and the precordial leads follow the standard resting 12-lead ECG configuration. The section C.3 describes the common format for the continuous ECG file and their annotation information across the databases.

C.2 Study Populations in the Repository—The THEW databases encompass ECG recordings from cardiac patients and healthy individuals. As described in Table 1, 24-hour continuous Holter ECGs from patients with acute myocardial infarction (AMI), patients with coronary artery disease (CAD), 10-minutes continuous ECGs from patients before after cardioversion for atrial fibrillation (Afib) and finally patients with the congenital or the acquired long QT syndrome are included as well. In the list of patients with history of TdPs, it is worth noting that several of these patients had a diagnostic test based on dl-sotalol IV used to unmask latent repolarization abnormalities[2]. ECGs were recorded prior, during and after sotalol infusion[3]. Several of these recordings contain documented life-threatening ventricular arrhythmias (torsades de pointes).

In addition, 24-hour continuous Holter ECGs from healthy individuals are available [4]. A first set contains individuals in normal ambulatory condition while the second set are individuals exposed to drug such as moxifloxacin (a drug used as positive control substance in drug safety trial to evaluate drug-induced QT interval prolongation, the so-called thorough QT studies [5]). Finally, the repository contains data from a full thorough QT study involving individuals exposed to a drug under development that did not reach the market because of its QT/QTc prolonging effect (study unique identifier: E-HOL-03-175-005). ECG data from all arms of the study are available in the warehouse.

Detailed information regarding these populations is provided on the initiative website (see Appendix).

C.3 ECG and annotation file formats—The technical specifications of the data are documented by the ECG file format implemented in the warehouse namely the ISHNE Holter ECG format. This format is extensively describe in [6]. Because the ISHNE format does not include cardiac beat annotation, a hybrid version of this format was developed, and shared by AMPS LLC (New-York, USA), to host the information related to cardiac beat annotation. This format is defined as follow:

1/ISHNE header as described in [6],

2/Binary annotation consisting of a 4-bytes binary data structure organized as follows:

- Label 1 [char]: beat annotation

- Label 2 [char]: for further beat description

- Location (Δ Sample) [unsigned int]: digital samples from last beat annotation

The definitions of the beat annotation for "Label 1" are as follow:

N: Normal beat

V: Premature ventricular contraction

S: Supraventricular premature or ectopic beat

C: Calibration Pulse

B: Bundle branch block beat

P: Pace

X: Artifact

Clinical information: As noted above, the THEW consists of ECG recordings from cardiac patients, healthy individuals, individual exposed to cardiac and to non-cardiac drugs. The clinical information associated to these populations is heterogeneous. Consequently, we opted to release dataset-specific files to document the clinical information rather than deploying a global unified database. These clinical files are provided in either M*icrosoft* Excel or SAS format. The list and the number of clinical variables in the clinical files vary between databases, a description of the database-specific set of information is provided in the THEW website. Importantly, none of the clinical information available in the THEW contains health private information and all available information is fully compliant with the Health Insurance Portability and Accountability Act (HIPAA).

III. Accessing Data from the THEW

All databases from the THEW are available worldwide through the internet. The accesses to the SFTP sever or through the THEW client application require the same registration process. Since the THEW is supported by the NHLBI, not-for-profit organizations can freely access the databases. However, we require these organizations to provide a single-page form describing their scientific objectives to the THEW Data Use Committee (DUC) which role is to provide feedback about potential collaborators (as an option) and scientific counseling to the submitter (s). Also, the THEW DUC receives users' feedback to improve data content, data structure or organizational processes within the THEW. The submission form is available on the download area of our website (see Appendix). For-profit organizations do not have to send such forms but they are required to pay a membership fee to access the data from the repository.

A. The THEW Client Application (CA)

The THEW CA is a Microsoft dotNET (framework 2.0) application developed in collaboration with Global Instrumentations (Syracuse, NY). This application is designed to provide: 1) easy secured access to ECG and clinical data, 2) an ECG viewer tool, 3) an ECG tool for interval (epochs) extraction from Holter recordings, 4) a system development kit based on a simple application program interface (API), and 5) a tool to download ECG signal and beat annotations. To obtain the latest version of the software, the users can send a request using the download area of our website (see Appendix).

The epoch selection tool allows for identifying intervals of interest from the continuous ECG recordings (figure 2). Once the epoch is defined, the CA provides an interface to download the period of interest so the users do not have to download large amount of signals (when it is not needed). The download tools of the CA provide several extraction formats such as ISHNE (as

described in section C), HL7 XML (see Appendix) and ASCII files with self-explanatory data format.

To simplify the users' access to relevant epochs of recordings, we predefined sets of epochs in each ECG recordings. For example in our set of ECGs including drug-induced torsades de pointes, we created epochs covering a period preceding the occurrence of the arrhythmias. This is illustrated in Figure 2. This information speeds up the review of the data or/and the creation of personal epochs.

In addition, the user can generate her/his own list of epochs that are stored on the client side and that can be retrieved as needed between sessions.

B. The secured FTP server

The users of the THEW data have the option to access the data from the warehouse using a secured FTP server hosted at University of Buffalo with the support from the NYSTAR program (New-York State Foundation for Science, Technology and Innovation). The server is hosted at the Center for Computational Research New York State Center of Excellence in Bioinformatics & Life Sciences (see appendix). The data in the server are in ISHNE format including both the raw ECG signals and the annotation information.

Legal use of the data from the THEW—The data from the THEW can be used for research, development and educational activities. No restriction exists related to publications, inventions and patents i.e. intellectual properties based on the THEW data are fully owned by the inventor and cannot be claimed by either the THEW organization or the organization(s) that provided the data to the THEW.

Importantly, the data from the THEW cannot be shared between organizations without prior consent from the THEW organization (regardless of their status). Such requirement is necessary in order to have for-profit companies helping us continuing to develop our activities through membership data-access fees. Thus, we ask any THEW users to sign a Data Use Agreement (DUA) stipulating that the data they obtain from our repository cannot be shared outside of their organization. As today, the DUA was signed by more than 20 organizations worldwide.

IV. Discussion

Helping the scientific community by developing an ECG repository is not a novel concept. There are several examples of ECG databases built over the past decades. The MIT initiative around Physionet and the AHA-BIH Arrhythmia Database [7], the CSE database are examples of such ECG databases which benefited greatly scientists worldwide. The Physiobank [7] is probably one of the most successful initiative of ECG databases available today (see Appendix). It has significantly contributed to the development of multiple ECG technologies. We believe our initiative will complement the Physiobank ECG database for several reasons: 1) the CES will contain unique sets of ECGs and clinical data from regulatory clinical trials (not available in the Physiobank); 2) our initiave will facilitate analysis of large sets of long-term digital Holter recordings, we host primarily 24-hour recordings. The ECG data contribution of the CES/THEW is expected to grow at a fast pace. Over this past two years, our initiave has received ECG recordings from for-profit organizations and academic centers encompassing 1,100 recordings from 785 individuals representing close to 270GB of continuous digital ECGs. We expect to quadruple the size of the repository before the end of the year. Recently, we have developed further collaborations with centers from the U.S. and from Europe to add two large sets of data including more than 2,000 recordings from: 1) chest

pain patients from the emergency department and 2) genotyped congenital long QT syndrome patients. These new datasets will be available before the end of the year 2010.

V. Conclusion

The THEW initiative has rapidly grown. A stream of organizations is joining our initiative since his inceptive. We believe our repository will benefit numerous scientists and researchers by providing unique set of continuous digital ECG recordings and their associated clinical information.

This initiative provides services to researcher's worldwide by fostering and distributing resources (data and tools) needed to conduct ECG-related activities (technology development and ECG metrics). We expect this initiative to spawn various collaborative research projects and to facilitate the development of improved ECG technologies for cardiac safety.

More importantly, we believe our effort will promote cross-fertilization of scientific knowledge, resources and ideas that will advance the field of quantitative electrocardiography and cardiac safety.

Acknowledgments

References

1. FDA. Innovation, Stagnation. challenge and Opportunity on the Critical Path to New Medical Products. US Department of Health and Human Services Food and Drug Administration; 2004.

2. Kaab S, Hinterseer M, Nabauer M, Steinbeck G. Sotalol testing unmasks altered repolarization in patients with suspected acquired long-QT-syndrome--a case-control pilot study using i.v. sotalol. Eur Heart J Apr;2003 24(7):649–657. [PubMed: 12657223]

3. Couderc JP, Kaab S, Hinterseer M, McNitt S, Xia X, Fossa A, Beckmann BM, Polonsky S, Zareba W. Baseline values and sotalol-induced changes of ventricular repolarization duration, heterogeneity, and instability in patients with a history of drug-induced torsades de pointes. J Clin Pharmacol Jan;2009 49(1):6–16. [PubMed: 18957528]

4. Couderc JP, Xiaojuan X, Zareba W, Moss AJ. Assessment of the Stability of the Individual-Based Correction of QT Interval for Heart Rate. Ann Noninvasive Electrocardiol Jan;2005 10(1):25–34. [PubMed: 15649234]

5. Darpo B. The thorough QT/QTc study 4 years after the implementation of the ICH E14 guidance. Br J Pharmacol Jan;2010 159(1):49–57. [PubMed: 19922536]

6. Badilini F. The ISHNE Holter standard output format. Ann Noninvasive Electrocardiol 1998;3(3): 263–266.

7. Goldberger AL, Amaral LA, Glass L, Hausdorff JM, Ivanov PC, Mark RG, Mietus JE, Moody GB, Peng CK, Stanley HE. PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. Circulation June;2000 101(23):E215–E220. [PubMed: 10851218]

APPENDIX

A1. The set of websites used in this article are as follow

THEW website: http://www.thew-project.org

NYSTAR website: http://www.nystar.state.ny.us/

Center of Excellence in Bioinformatics & Life Sciences:
http://www.bioinformatics.buffalo.edu/

FDA Private Public Partnership website:
http://www.fda.gov/AboutFDA/PartnershipsCollaborations/
PublicPrivatePartnershipProgram/ucm166082.htm

Center for Research Computing website: http://www.rochester.edu/its/web/wiki/crc/

HL7 FDA/XML format: http://www.hl7.org/V3AnnECG/

Physionet/Physiobank: http://www.physionet.org/site-map.shtml

A2. SUI Database Naming Convention

The SUI is defined to provide basic information about the the ECG dataset. The SUI is formed
by 5 numbers defined as follow:
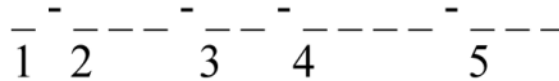
Type of signal (1 letter):

E: ECG

B: blood pressure

O: other

Type of data signal (3 letters):

HOL: Holter ambulatory ECGs

TEL: telemetric ECGs

STD: standard 12-lead resting ECGs

OTH: other types of recording

Number of leads included (2 digits).

Number of recordings included in the database.

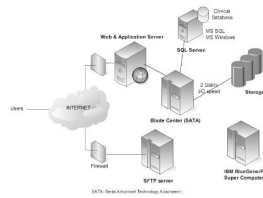Unique identification number (database entry order in the warehouse, 3 digits)

Fig. 1.
Information technology structure for the THEW: two options are proposed to access the data from the THEW: 1) Server hosted at University of Buffalo using secured file transfer protocol (SFTP); 2) THEW client application which is a dotNET client server application providing the access to both ECG signals and the associated clinical information. The THEW also include resources from a supercomputer (IBM BlueGene/P) that can be used in collaborative projects.
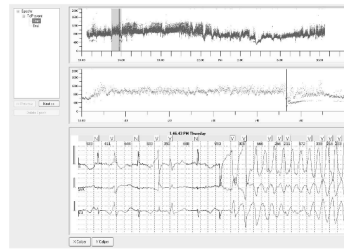
Fig. 2.
Snapshot of the THEW CA in which ECGs containing an episode of drug-induced torsades de pointes is shown. The list of epochs identify the period containing ventricular arrhythmias. The user can click on the epochs to retrieve rapidly the period of interest from the overall 24-hour period.

TABLE 1

Database available in the THEW

ECG type (leads)	SUI	Label	#ECGs (#ind.)	# ECGs (SF)
24h-holter (3)	E-HOL-03-175-005	Tqt1	175 (34)	18 GB (200)
24h-holter (12)	E-HOL-12-140-008	Tqt2	140 (70)	190 GB (1000)
24h-holter (3)	E-HOL-03-271-002	cad	271 (271)	29 GB (200)
24h-Holter (3)	E-HOL-03-160-001	Ami	160 (93)	18 GB (200)
24h-Holter (3)	E-HOL-03-201-003	normal	201 (201)	22 GB (200)
12h-Holter (12)	E-OTH-12-6-009	TDP1	6 (6)	2 GB (180)
20 minutes (12)	E-OTH-12-68-010	histTDP	68 (34)	242 MB (1000)
10 minutes (12)	E-OTH-12-73-011	Afib1	73 (73)	1.7GB (1000)

SUI: study unique identifier

Tqt: thorough QT study.

CAD: coronary artery disease

AMI: acute myocardial infarction

TdPs: torsades de pointes

histTDP: patients with an history of torsades de pointes.

Afib: atrial fibrillation;

SF: sampling frequency expressed in Hz.

Ind.: individuals