32nd Annual International Conference of the IEEE EMBS
Buenos Aires, Argentina, August 31 - September 4, 2010

The Telemetric and Holter ECG Warehouse Initiative (THEW): a Data Repository for the Design, Implementation and Validation of ECG-related Technologies

Jean-Philippe Couderc

Abstract—**we present an initiative supported by the National Heart Lung, and Blood Institute and the Food and Drug Administration for the development of a repository containing continuous electrocardiographic information to be shared with the worldwide scientific community. We believe that sharing data reinforces open scientific inquiry. It encourages diversity of analysis and opinion while promoting new research and facilitating the education of new researchers.**

In this paper, we present the resources available in this initiative for the scientific community. We describe the set of ECG signals currently hosted and we briefly discuss the associated clinical information (medical history. Disease and study-specific endpoints) and software tools we propose. Currently, the repository contains more than 250GB of data from eight clinical studies including healthy individuals and cardiac patients. This data is available for the development, implementation and validation of technologies related to body-surface ECGs.

To conclude, the Telemetric and Holter ECG Warehouse (THEW) is an initiative developed to benefit the scientific community and to advance the field of quantitative electrocardiography and cardiac safety.

I. Introduction

Cardiac safety remains a major public health concern. Sudden cardiac death (SCD) is responsible for half of all heart disease deaths and is the largest cause of natural death in the U.S. (representing about 325,000 adults each year). Despite the effort implemented to reduce this number 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 subsequently to sudden cardiac death.

Manuscript received April 1st, 2010. This work was supported by the National Heart, Lung, and Blood Institute of the U.S. Department of Healthy and Human Services grant R24HL096556.

J. P. Couderc is with the Center for Quantitative Electrocardiography and Cardiac Safety, University of Rochester Medical Center, Rochester, NY 14618 USA (585-275-1096; fax: 585-273-5283; e-mail: jean-philippe.couderc@thew-project.org).

The National Heart, Lung and Blood Institute (NHLBI) provided the resources to University of Rochester (NY) to enable the creation of the "Center for Quantitative Electrocardiography and Cardiac Safety" (CES). These resources support the inceptive CES activities around 1) the development and maintenance of computer resources (data storage and computing center), 2) the deployment of medical information, and 3) the formation of a scientific network. The CES distributes these resources to the international scientific community by sharing unique clinical and ECG information for the design and validation of technologies to improve quantitative electrocardiography and cardiac safety.

In addition, the CES developed a partnership with the U.S. Food and Drug Administration (FDA, see FDA private-public partnership (PPP)[1]). Under its public health mission, FDA is interested in the development of improved technologies to evaluate drug safety and efficacy [2]. This FDA partnership was designed to leverage resources and expertise toward the implementation of collaborative projects among FDA, University of Rochester 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 CES has developed collaborations with academia, U.S. and international corporations to submit research proposals to federal agencies, to share scientific data and technologies, and to spawn collaborative projects between academia and industries.

II. The THEW iniative

A. Mission Statement

The objective of the Center for Quantitative Electrocardiography and Cardiac Safety and its 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 described 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. This system consists of one rack of the 13.9 TFLOPS IBM Blue Gene/P massively parallel processing (MPP) supercomputer, one IBM System p front-end node, one IBM System p service node, and 8 IBM System x I/O nodes connected to 180 TB IBM System Storage. The Blue Gene/P system consists of 1,024 nodes, 4,096 CPU cores, 2 TB of RAM, and 180 TB of storage. Each node consists of a quad-core PPC450 with 8 MB of cached and 2 GB of RAM. Access to this computer resource to THEW users is currently in development. The computer is hosted at the Rochester Center for Research Computing [3].

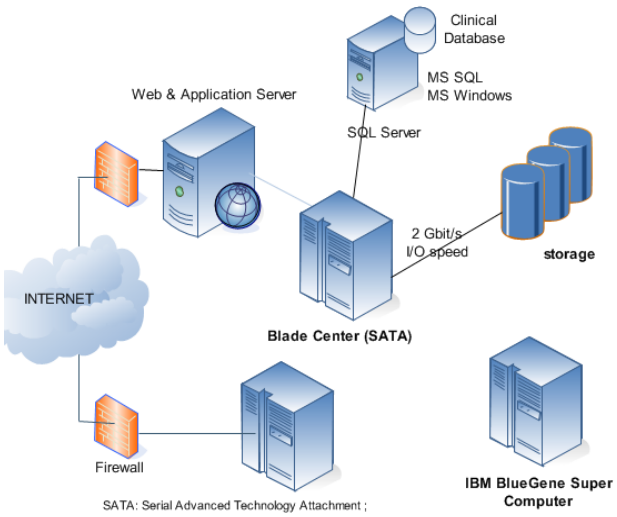

Fig. 1. Information technology structure for the THEW. Two options are proposed to access the data from the THEW: 1) Server hosted at University of Buffalo based using secured file transfer protocol (SFTP) , 2) THEW client application which is a dotNET client server application providing the access to both ECG signals and the associated clinical information

C. *Warehouse Content*

ECG recordings

The data available in the warehouse were provided by research academic centers and major pharmaceutical companies. The list of database is provided in Table 1. The hosted ECG recordings are continuous with a length varying from 10 minutes to 24 hours. The recording equipments differ between databases. The ECG signals have sampling frequency of 180Hz, 200Hz and 1000 Hz and an amplitude resolution coded from 10 to 16 bits depending on the database (see Table 1). The lead configuration depends 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 reported to the torso

TABLE I
DATABASE AVAILABLE IN THE THEW

ECG type (leads)	Label	#ECGs (#ind.)	# ECGs (SF)
24h-holter (3)	Drug study	175 (34)	18 GB (200)
24h-holter (8)	Drug study	140 (70)	190 GB (1000)
24h-holter (3)	CAD	271 (271)	29 GB (200)
24h-Holter (3)	AMI	93 (93)	18 GB (200)
24h-Holter (3)	Healthy	201 (201)	22 GB (200)
12h-Holter (12)	TdPs	6 (6)	2 GB (180)
20 minutes (12)	TdP history	7 (7)	242 MB (1000)
10 minutes (12)	Afib	74 (37)	1.7GB (1000)

CAD: coronary artery disease; AMI: acute myocardial infarction TdPs: torsades de pointes; Afib: atrial fibrillation; SF: sampling frequency expressed in Hz. The so-called "Drug studies" are thorough QT studies. Ind.: individuals

(Mason-Likar lead placement) and the precordial leads follow the standard resting 12-lead ECG configuration.

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. Several of them contain documented life-threatening ventricular arrhythmias (torsades de pointes).

Long term continuous Holter ECG from healthy individuals are available including 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 [4]).

ECG and annotation file formats

The database-specific technical specifications of the data are supported by the ECG file format used in the warehouse namely the ISHNE Holter ECG format [5], a hybrid version of this format was developed by AMPS LLC (New-York, USA) to host the information related to cardiac beat annotation. This format is described as follow:

1/ ISHNE header as described in [5]

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

· toc : digital samples from last beat annotation [unsigned int]

The definition of the label is as follow for label 1:

N: Normal beat

V: Premature ventricular contraction

S: Supraventricular premature or ectopic beat

C: Calibration Pulse

B: Bundle branch block beat

P: Pace

X: Artefact

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 are heterogeneous. Consequently, we opted to release dataset-specific files for describing clinical information (including medical history, study endpoints, etc.) rather than a global database. These clinical files are provided in either MS Excel or SAS format. The list and the number of clinical variables vary between databases, a description of the database-specific set of information is provided in our website. Importantly, none of the clinical information available in the THEW contains health private information and all available information are fully compliant with HIPAA regulation.

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. The THEW being supported by the NHLBI, not-for-profit organizations (NFPO) can access databases free. However, we require the NFPO to provide a single-page form describing their scientific objectives to the THEW Data Use Committee which role is: 1) to provide: feedback about potential collaborators (as an option), 2) to propose scientific counseling to the submitter(s) and 3) to receive feedback from data users in order to improve data content, data structure or organizational processes. The submission form is available on our website [6]. For-profit organizations do not have to send such form 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) ECG viewer tools, 3) ECG tools for interval (epochs) extraction from Holter recordings, and 4) a system development kit based on a simple application program interface (API). To obtain the latest version of the software, the users can send a request using the download area of our website [6].

The epoch selection tool allows for identifying intervals of interest from the continuous ECG recordings. 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 [7] and ASCII files with self-explanatory data arrangement.

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 its own list of epochs that are stored on the client side and can be retrieved as needed between sessions.

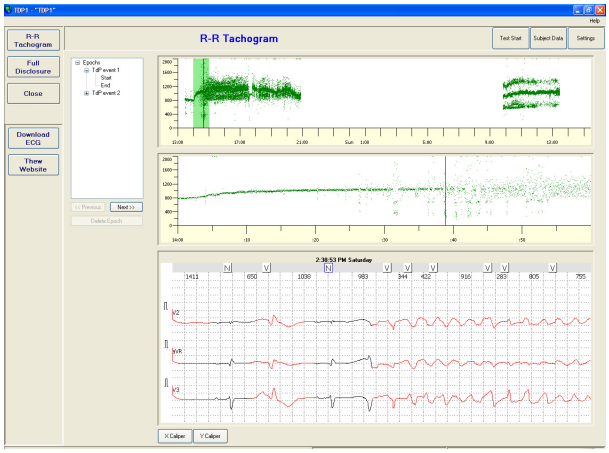

Fig. 2. Snapshot of the THEW CA in which ECGs containing an episode of drug-induced torsades de pointes is shown. The list of epochs identify the period containing ventricular arrhythmias. The user can click on the epochs to retrieve rapidly the period of interest from the overall 24-hour period.

B. The secured FTP server

The user 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 [8]. 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 is fully own 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 executed 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 [9], the CSE database are examples of such ECG databases which benefited greatly scientists worldwide. The Physiobank [10] is probably one of the most successful initiative of ECG databases available today[11]. It has significantly contributed to the development of ECG technologies. It contains ~220Gb of electrocardiographic data. 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 double up 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 2000 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 since his inceptive. The number of organizations using the data hosted in the warehouse is continuously increasing as well. 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.

Acknowledgment

The THEW has been designed with the help of many organizations: private, governmental and others. The list of these sponsors is provided in the THEW website[6].

References

[1] Food and Drug Administration, "http://www.fda.gov/AboutFDA/PartnershipsCollaborations/PublicPrivatePartnershipProgram/ucm166082.htm," 2010.

[2] 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.

[3] Center for Research Computing website, "www.rochester.edu/its/web/wiki/crc/," 2010.

[4] B. Darpo, "The thorough QT/QTc study 4 years after the implementation of the ICH E14 guidance," *Br J Pharmacol.*, vol. 159, no. 1, pp. 49-57, Jan.2010.

[5] F. Badilini, "The ISHNE Holter standard output format," *Ann. Noninvasive. Electrocardiol.*, vol. 3, no. 3, pp. 263-266, 1998.

[6] The Center for Quantitative Electrocardiography and Cardiac Safety, "www.thew-project.org,". University of Rochester Medical Center, Ed. 2010.

[7] High Level 7, "www.hl7.org/V3AnnECG/," 2010.

[8] Center of Excellence in Bioinformatics & Life Sciences, "www.bioinformatics.buffalo.edu," 2010.

[9] A. L. Goldberger, L. A. 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, p. E215-E220, June2000.

[10] A. L. Goldberger, L. A. 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, p. E215-E220, June2000.

[11] Physionet, "www.physionet.org/site-map.shtml," 2010.