

A scientific repository to support the research and development of technologies related to quantitative electrocardiography: The Telemetric and Holter ECG Warehouse (THEW)

Jean-Philippe Couderc

Center for Quantitative Electrocardiography and Cardiac Safety,
University of Rochester Medical Center, Rochester, New York, USA

Abstract

We implemented an ECG data-warehousing project to promote cross-fertilization of scientific knowledge, resources and ideas that will advance the understanding of quantitative ECG and cardiac safety. While other open ECG databases are available, our Telemetric and Holter ECG Warehouse (THEW) proposes unique sets of continuous digital ECGs recorded in cardiac patients, healthy individuals and individuals exposed to non-cardiac and cardiac drugs. In this paper, we present briefly the content and the current research triggered by this initiative. (Cardiol J 2010; 17, 4: 416–419)

Key words: database, electrocardiogram, cardiac safety, QT, long QT syndrome, acute myocardial infarction, coronary syndrome, thorough QT studies, healthy, torsades de pointes

Introduction

The need for accurate and fast exchange of information has been the rationale for the development of computer technologies and the driving force for the continuous automation of our surroundings. In the medical arena, digital conversion has happened at every level, from the patient/care providers' interaction to the electronic data capture systems in large clinical trials. Digital tools have streamlined the information flow from patients to scientists through physicians (and other health care providers). Yet, the information acquired in clinical studies and other medical investigations is rarely shared with the broader scientific community, even years after the study findings have been released. Legal protection of patient information, and technological challenges,

were the main barriers to better public access to medical data.

In 1995, the European Parliament and the Council of the European Union (E.U.) adopted a "Directive on the Protection of Individuals with Regard to the Processing of Personal Data and on the Free Movement of Such Data". The following year, the U.S. Congress enacted the Health Insurance Portability and Accountability Act (HIPAA) in order to improve the efficiency and effectiveness of the nation's healthcare system by encouraging the widespread use of electronic data interchange. Today, clinical studies must ensure the protection of private health information through data de-identification i.e. deletion of information that would allow a patient to be identified from any electronic dataset: 18 fields were defined as identifiable data, including name, date of

Address for correspondence: Jean-Philippe Couderc, PhD, Center for Quantitative Electrocardiography and Cardiac Safety, University of Rochester Medical Center, Rochester, NY 14618 USA, tel: 585 275 1096, fax: 585 273 5283, e-mail: jean-philippe.couderc@heart.rochester.edu

Table 1. Databases available in the THEW.

Study	Leads	ECGs	Sampling	Rec.	Pts.	Size
Healthy	3	24-h	200	201	201	22 GB
Coronary artery disease	3	24-h	200	271	271	29 GB
Acute myocardial infarction	3	24-h	200	160	93	18 GB
Thorough QT study	3	24-h	200	175	34	6 GB
TdPs	12	24-h	180	6	6	2 GB
History of TdPs	12	4 min	1,000	84	37	242 MB
Atrial fibrillation	12	15–20 min	1,000	73	73	1.7 GB
Thorough QT study	12	24-h	1,000	140	70	190 GB
Genotyped congenital LQTS	2–3	24-h	200	480	307	14 GB

TdPs — torsades de pointes; LQTS — long QT syndrome; Leads — number of leads; ECGs — average duration of the electrocardiograms; Sampling — sampling frequency expressed in Hz; Rec. — number of recordings; Pts. — number of patients; Size — size of the dataset excluding the clinical information

birth, social security number and contact information (a complete list can be found in the Appendix 1). Consequently, the legal and technological barriers no longer exist, and developing an open database depends primarily on logistic and financial resources. The storing and sharing of medical data are associated with costs for infrastructure development, maintenance, and technical support to users. For an organization to bear such costs while ensuring an altruistic scientific mission is a challenge.

In this paper, we describe how our group has designed the Telemetric and Holter ECG Warehouse (THEW) initiative to facilitate the sharing of a large database of medical information, including primarily continuous ECG data. Legal, technical and financial solutions were designed and successfully implemented, triggering collaborative projects with both federally-funded organizations and private companies. We describe the content of the THEW, as well as the research activities which have been initiated based on its data.

Mission statement

The objective of the THEW is to provide access to ECG data for the design and validation of analytic methods to advance the field of quantitative ECG, with a strong focus on cardiac safety.

Data available in the THEW (Table 1)

The THEW databases encompass ECG recordings from cardiac patients and healthy subjects. The group of cardiac patients consists of individuals with acute myocardial infarction, coronary artery disease, or patients with atrial fibrillation. Patients with congenital or acquired forms of long QT syndrome are hosted as well. The group of healthy individuals en-

compasses ECGs from ‘normal’ individuals, where normal is defined as the absence of cardiac history, no blood pressure, no chronic illness and normal physical examination.

Clinical files describing patients’ clinical information are associated with each database of ECG files. The content of each clinical information file is described in detail on the website of the initiative (www.thew-project.org).

Technical specifications of the ECG signals

The hosted ECG recordings are continuous, with a length varying from ten minutes to 24 hours, and with a sampling frequency of 180 Hz, 200 Hz and 1000 Hz. According to the database, the ECG recordings have different amplitude resolutions (coded on 10, 12 or 16 bits) and different numbers of leads (3- or 12-lead recordings). Three-lead signals are recorded using a pseudo-orthogonal configuration (X, Y, and Z). Twelve leads follow the Holter configuration, in which limb leads are reported to the torso (treadmill configuration or Mason-Likar lead placement) and the precordial leads follow the standard resting 12-lead ECG configuration. See Table 1 for details about the technical specifications of the ECG recordings in each THEW database.

Accessing data from the THEW

All databases from the THEW are available worldwide via the internet. Because 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), whose 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 so as to improve data content, data structure or organizational processes. For-profit organizations do not have to send such a form, but they are instead required to pay a membership fee to access data from the repository.

Two options are offered to access the THEW data. First, the THEW client application is a software developed in collaboration with Global Instrumentations (Syracuse, NY, USA) providing: 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 annotation. The second option, the secured file transfer protocol (SFTP), allows the downloading of ECG files in a well-documented format (the so-called ISHNE format) for both the raw ECG signal and the cardiac beat annotations [1].

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 property based on the THEW data is fully owned by the inventor and cannot be claimed by either the THEW or the organization(s) that provided the data to the THEW. Importantly, data from the THEW cannot be shared between organizations without prior consent from the THEW steering committee (regardless of their status). Such a requirement is necessary in order to have for-profit companies help us continue to develop our activities through membership data-access fees.

Current research

During this past year, 11 organizations have joined our initiative to conduct research and develop ECG-related technologies. While private companies have focused on the validation of current technologies, such as the automation of QT interval measurements for drug safety trials [2], the work from academic centers is related to new technologies including ECG markers of drug-induced torsades de pointes [3, 4], QT/RR modeling [5], and ECG filtering [6]. With an increasing number of groups joining our initiative, we have spawned many more activities encompassing other arenas of quantitative electrocardiography, they include repolarization dynamics, atrial fibrillation detection, and phenotype-genotype correlation in patients with the congenital long QT syndrome.

The THEW and other ECG-related initiatives

Helping the scientific community by developing an ECG repository is not a novel concept. There are several examples of ECG databases built over recent decades. The MIT initiative around Physionet, the AHA-BIH Arrhythmia Database, and the CSE database are examples of ECG databases that have greatly benefited scientists worldwide. The Physiobank [7] is probably one of the most successful ECG databases available today (see Appendix 1). 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); and 2) our initiative will facilitate the analysis of large sets of long-term digital Holter recordings (we host primarily 24-hour recordings). The ECG data contribution of the THEW is expected to grow quickly. Over the past two years, our initiative has received ECG recordings from for-profit organizations and academic centers encompassing 1,600 recordings from 1,092 individuals, representing close to 283 GB of continuous digital ECGs.

Conclusions

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 future generations of scientists. We have started an initiative to develop a unique repository of medical information and continuous ECG signals. After less than two years of activities, 15 publications have been released based on data from our repository, focusing primarily on designing improved or novel technologies for the quantitative analysis of ECG. A stream of organizations has already joined our initiative, and we invite more organizations and individuals to join our endeavor by using data from the repository, sharing their data and/or initiating collaborative projects.

Acknowledgements

The author does not report any conflict of interest regarding this work.

This work is supported by the National Heart, Lung, and Blood Institute of the U.S. Department of Health and Human Services grant # U24HL096556.

References

1. Badilini F. The ISHNE Holter standard output format. *Ann Non-inv Electrocardiol*, 1998; 3: 263–266.
2. Khawaja A, Petrovic R, Safer A, Baas T, Doessel O, Fischer R. Analyzing thorough QT study 1 and 2 in the Telemetric and Holter ECG Warehouse (THEW) using Hannover ECG system (HES®): A validation study. *Computing in Cardiology Proceedings*, 26–29 September 2010, Belfast (UK) [in press].
3. Jie X, Rodriguez B, Pueyo E. Torsadogenic drug-induced increased short-term variability of JT-area. *Computing in Cardiology Proceedings*, 26–29 September 2010, Belfast (UK) [in press].
4. Couderc JP, Xia X, Kaab S, Zareba W. Electrocardiographic markers associated with sotalol-induced torsades de pointes. *Computing in Cardiology Proceedings*, 26–29 September 2010, Belfast (UK) [in press].
5. Halamek J, Jurak P, Lipoldova J, Leinveber P. QT/RR coupling and gender difference. *Computing in Cardiology Proceedings*, 26–29 September 2010, Belfast (UK) [in press].
6. Johannesen L, Grove USL, Stampe Soerensen J, Schmidt ML, Couderc JP, Graff C. A wavelet-based algorithm for delineation and classification of ECG waves. *Computing in Cardiology Proceedings*, 26–29 September 2010, Belfast (UK) [in press].
7. Goldberger AL, Amaral LA, Glass L et al. PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. *Circulation*, 2000; 101: E215–E220.

Appendix 1. List of the fields to be removed from medical datasets so as to be HIPAA compliant

1. Name
2. Postal address (geographic subdivisions smaller than state)
3. All elements of dates, except year (birth date, if over 89, must be aggregated)
4. Phone number
5. Fax number
6. E-mail address
7. Social Security number
8. Medical Record number
9. Health Plan number
10. Bank account number
11. Certificate/license numbers
12. URL
13. IP address
14. Vehicle identifier
15. Device ID
16. Biometric ID
17. Full face/identifying photo
18. Any other unique identifying number, characteristic or code