PhysioNet: Physiologic signals, time series and related open source software for basic, clinical, and applied research

Citation: Moody, G. B., R. G. Mark, and A. L. Goldberger. "PhysioNet: Physiologic Signals, Time Series and Related Open Source Software for Basic, Clinical, and Applied Research." in Proceedings of the 33rd Annual International Conference of the IEEE EMBS Boston, Massachusetts USA, August 30 - September 3, 2011, IEEE, 2011. 8327–8330. Web.

As Published: http://dx.doi.org/10.1109/IEMBS.2011.6092053

Publisher: Institute of Electrical and Electronics Engineers

Persistent URL: http://hdl.handle.net/1721.1/76352

Version: Author's final manuscript: final author's manuscript post peer review, without publisher's formatting or copy editing

PhysioNet: Physiologic Signals, Time Series and Related Open Source Software for Basic, Clinical, and Applied Research

George B. Moody, Roger G. Mark, and Ary L. Goldberger

Abstract— **PhysioNet provides free web access to over 50 collections of recorded physiologic signals and time series, and related open-source software, in support of basic, clinical, and applied research in medicine, physiology, public health, biomedical engineering and computing, and medical instrument design and evaluation. Its three components (PhysioBank, the archive of signals; PhysioToolkit, the software library; and PhysioNetWorks, the virtual laboratory for collaborative development of future PhysioBank data collections and PhysioToolkit software components) connect researchers and students who need physiologic signals and relevant software with researchers who have data and software to share. PhysioNet's annual open engineering challenges stimulate rapid progress on unsolved or poorly solved questions of basic or clinical interest, by focusing attention on achievable solutions that can be evaluated and compared objectively using freely available reference data.**

I. INTRODUCTION

PhysioNet (`http://physionet.org/`) is a research resource intended to stimulate current research and new investigations in the study of complex biomedical and physiologic signals. It has three major components:

PhysioBank is a large and growing archive of well-characterized digital recordings of physiologic signals, time series, and related data for use by the biomedical research community. PhysioBank currently includes more than 50 collections of cardiopulmonary, neural, and other biomedical signals from healthy subjects and patients with a variety of conditions with major public health implications, including sudden cardiac death, congestive heart failure, epilepsy, gait disorders, sleep apnea, and aging (see Tables I and II). These collections include data from a wide range of studies, as developed and contributed by members of the research community.

PhysioToolkit is a large and growing library of software for physiologic signal processing and analysis, detection of physiologically significant events using both classical techniques and novel methods based on statistical physics and nonlinear dynamics, interactive display and characterization of signals, creation of new databases, simulation of physiologic and other signals, quantitative evaluation and comparison of analysis methods, and analysis of nonequilibrium and nonstationary processes. A unifying theme of many of the research projects that contribute software to PhysioToolkit is the extraction of "hidden" information from biomedical signals, information that may have diagnostic or prognostic value in medicine, or explanatory or predictive power in basic research. All PhysioToolkit software is available in source form under the GNU General Public License (GPL).

PhysioNetWorks is a virtual laboratory for development of data and software resources that will eventually become components of PhysioBank and PhysioToolkit. By providing large, secure workspaces with redundant backup to active researchers who can easily share them with colleagues anywhere, PhysioNetWorks encourages investigators to create well-organized and documented, *usable* data and software repositories during the conduct of their research. When the research is complete and the major results have been published (or at any time the researcher wishes) the repository can be shared with a colleague, a group of colleagues, or the research community at large.

II. BACKGROUND

PhysioNet was established in 1999 as the outreach component of the Research Resource for Complex Physiologic Signals[1], [2], a cooperative project initiated by the authors at Boston's Beth Israel Deaconess Medical Center, Harvard Medical School, and MIT, together with colleagues at Boston University, McGill University, and later at numerous other institutions. Beginning in the mid-1970s, members of the PhysioNet team who were then working on some of the first microcomputer-based instruments for cardiac arrhythmia monitoring foresaw the usefulness of establishing shared databases of well-characterized ECG recordings, as a basis for evaluation, iterative improvement, and objective comparison of algorithms for automated arrhythmia analysis. A five-year effort culminated in the publication of the MIT-BIH Arrhythmia Database in 1980, which soon became the standard reference collection of its type, used by over 500 academic, hospital, and industry researchers and developers worldwide during the 1980s and 1990s. Other databases of ECGs and eventually other physiologic signals followed. By 1999, the MIT group distributed CD-ROMs containing 11 such collections, and had participated in the development of several others.

The MIT group contributed its 11 databases, and the software it had developed for exploring and analyzing them,

PhysioNet is supported by the US National Institutes of Health (NIH), its National Institute of Biomedical Imaging and Bioengineering (NIBIB), and its National Institute of General Medical Sciences (NIGMS), under NIH cooperative agreement U01-EB-008577.

GBM and RGM are co-PIs of PhysioNet, and ALG is its program director. GBM is the architect and technical director of PhysioNet, and is with the Laboratory for Computational Physiology (LCP) in the Harvard-MIT Division of Health Sciences and Technology (HST). RGM directs the LCP, and is a Distinguished Professor in HST and EECS at MIT. ALG directs the Rey Institute for Nonlinear Dynamics in Medicine at Boston's Beth Israel Deaconess Medical Center, and is a Professor of Medicine at Harvard Medical School.

Address for correspondence: George B. Moody, Massachusetts Institute of Technology, Room E25-505A, 77 Massachusetts Ave., Cambridge, MA 20139 USA. Email: `george@mit.edu`

to establish PhysioBank and PhysioToolkit. Free availability of these resources via the Internet catalyzed an even greater explosion of interest in them, as researchers and students worldwide who had no previous access to such data or software began new programs of research, and specialists began comparing their methods. These initial contributions were quickly supplemented by additional collections of data and software from their collaborators, and soon after, from many researchers worldwide. PhysioBank and PhysioToolkit have grown to many times their original sizes, and most of the growth has been thanks to the hard work and generosity of an international community of researchers.

III. ACTIVITIES

Shortly after PhysioNet was established, we initiated an annual series of open engineering challenges, in cooperation with the annual IEEE-EMBS-sponsored conference, *Computers in Cardiology* (now *Computing in Cardiology*, or CinC). We hoped to introduce PhysioNet to our international colleagues who would be attending CinC, by encouraging participation in an activity that made effective use of the facilities provided by PhysioNet to stimulate rapid progress on an unsolved problem of practical clinical significance. A timely contribution of data from Thomas Penzel made it possible to create the first PhysioNet/CinC Challenge, which attracted the attention of more than a dozen teams to the subject of detecting sleep apnea from the ECG[3]. Their efforts were broadly successful, they discussed their findings at CinC 2000, and an annual tradition was born.

In complementary ways, PhysioNet and CinC catalyze and support scientific communication and collaboration between basic and clinical scientists. The annual meetings of CinC are gatherings of researchers from many nations and disciplines, bridging the geographic and specialty chasms that separate understanding from practice, while PhysioNet provides on-line data and software resources that support collaborations of basic and clinical researchers throughout the year. The annual PhysioNet/CinC Challenges seek to provide stimulating yet friendly competitions, while at the same time offering both specialists and non-specialists alike opportunities to make progress on significant open problems whose solutions may be of profound clinical value.

The use of shared data provided via PhysioNet makes it possible for participants to work independently toward a common objective. At CinC, participants can make meaningful results-based comparisons of their methods; lively and well-informed discussions are the norm at scientific sessions dedicated to these challenges. Discovery of the complementary strengths of diverse approaches to a problem when coupled with deep understanding of that problem frequently sparks new collaborations and opportunities for further study, as occurred when participants in the first Challenge combined their efforts to obtain an even better solution to the Challenge problem[4].

Recent challenge topics have included predicting acute hypotensive episodes in intensive care unit patients[5]; developing robust methods for filling gaps in multiparameter physiologic data (including ECG signals, continuous blood pressure waveforms, and respiration), with applications in detection of clinically important events and in reduction of false alarms in the ICU[6]; and (in progress) improving the quality of ECGs collected using mobile phones.

In a prototype implementation, PhysioNetWorks supported the 2010 Challenge, collecting and scoring entries from the participants. During its first five months (to mid-June 2011) the reimplented PhysioNetWorks has attracted more than 400 members, who are using it to support 12 collaborative projects and 15 others in preparation, in addition to the challenge in progress. Current PhysioNetWorks projects include data collection and annotation development efforts, as well as efforts aimed at improving or evaluating physiologic models and other software projects focused on creation or improvement of tools for research. New contributions of data and software are channeled through PhysioNetWorks, allowing their creators to participate in all aspects of curation of their contributions, and allowing the community to provide early feedback to influence decisions that may affect usability and value to researchers.

Significantly, many PhysioNetWorks projects are being established early in the grant cycles of the associated research projects. Use of PhysioNetWorks throughout the active phase of research encourages investigators to organize their work in a way that makes it easy to use community-developed exploratory and analytic tools by the research team and collaborators in the short run, thus making their final contribution more valuable to the research community at large, and avoiding the pitfalls of attempting to fulfill a mandated data-sharing requirement after funding for the project has ended and those who understand the data have begun work on other projects.

IV. CONCLUSIONS

In its first 12 years, PhysioNet has made a wide variety and large quantity of well-characterized data and related open-source software collected and created for biomedical research, often at great expense, available for re-use and further study at no cost by a worldwide community of over 40,000 researchers, clinicians, educators and students, and medical instrument and software developers. Through its open engineering challenges, it has stimulated development of inexpensive and minimally disruptive technology for detection of sleep apnea and sleep quality, prediction of adverse events and reduction in false alarms in the intensive care setting, and telemedicine. A Google Scholar search for PhysioNet and related terms finds over 5000 publications and citations as of June 2011. Finally, PhysioNetWorks has allowed us to scale up our capacity to bring new data collections and software packages on-line without requiring a proportional increase in the size of our team.

TABLE I

PHYSIOBANK COLLECTIONS OF MULTIPARAMETER AND ECG SIGNALS AND TIME SERIES

(AS OF JUNE 2011)

Collection	*Subjects*	*Duration (typical)*	*Signals and time series*	*Other*
MGH/MF Waveform Database	250	90-120 min	ECG (3 leads), ABP, PAP, CVP, respiration, airway CO_2, ...	beat annotations
Stress Recognition in Automobile Drivers	17	60-90 min	ECG, EMG, GSR, respiration	
Apnea-ECG Database	70	8 hours	ECG (subset includes respiration)	apnea annotations
Fantasia Database	40	2 hours	ECG (subset includes uncalibrated NIBP)	beat annotations
MIMIC Database	121	20-40 hours	ECG, BP, respiration, SpO_2, ...	beat labels, ICU monitor alarms
MIMIC II Waveform Database	20935	3-10 days	ECG, BP, respiration, SpO_2, ...	
MIMIC II Clinical Database	32536	3-10 days	hourly vital signs, medications, ...	ICD9 codes, lab tests, discharge summaries, ...
MIT-BIH Polysomnographic Database	16	8 hours	ECG, ABP, EEG, respiration, ...	apnea and sleep stage annotations
Sleep-EDF Database	8	8 hours	EEG (2), EOG, ...	hypnograms
SVUH/UCD Sleep Apnea Database	25	8 hours	ECG (3 leads), EEG (2), EOG (2), EMG, oronasal airflow, ribcage and abdomen movements, SpO_2, snoring, body position	apnea and sleep stage annotations
ANSI/AAMI EC13 Test Waveforms	10	1 minute	ECG	
European ST-T Database	90	2 hours	ECG (2 leads)	beat, rhythm, ST and T change annotations
Long-Term ST Database	86	24 hours	ECG (2 or 3 leads)	beat, rhythm, ST and signal quality annotations
MIT-BIH Arrhythmia Database	48	30 min	ECG (2 leads)	beat, rhythm, and signal quality annotations
MIT-BIH Noise Stress Test Database	15	30 min	ECG (2 leads) with calibrated noise	beat annotations
BIDMC Congestive Heart Failure Database	15	20 hours	ECG (2 leads)	beat annotations
Post-Ictal Heart Rate Oscillations in Partial Epilepsy	7	90-220 min	ECG	beat and seizure annotations
QT Database	100	15 min	ECG (2 leads)	annotations of onsets, peaks, ends of P, QRS, and T waves
AF Termination Challenge Database	30	1 min	ECG (2 leads)	beat annotations
Creighton University Ventricular Tachyarrhythmia Database	35	8 min	ECG (2 leads)	beat and VF annotations
Intracardiac Atrial Fibrillation Database	8	3-5 min	ECG (3 surface and 5 intracardiac leads)	beat annotations
Long-Term AF Database	84	24 hours	ECG (2 leads)	beat annotations
MIT-BIH Atrial Fibrillation Database	25	10 hours	ECG (2 leads)	beat and rhythm annotations
MIT-BIH ECG Compression Test Database	168	20 sec	ECG (2 leads)	
MIT-BIH Long-Term Database	6	24 hours	ECG (2 leads)	beat annotations
MIT-BIH Malignant Ventricular Arrhythmia Database	22	30 min	ECG (2 leads)	rhythm and signal quality annotations
MIT-BIH Normal Sinus Rhythm Database	18	24 hours	ECG (2 leads)	beat annotations
MIT-BIH ST Change Database	28	20-40 min	ECG (2 leads)	beat annotations
MIT-BIH Supraventricular Arrhythmia Database	78	30 min	ECG (2 leads)	beat annotations
Non-Invasive Fetal Electrocardiogram Database	1	5-20 min	ECG (maternal and fetal; 55 recordings over a 20 week period)	maternal beat annotations
PAF Prediction Challenge Database	100	30 min x2	ECG (2 leads)	beat annotations
PTB Diagnostic ECG Database	549	2 min	ECG (15 leads)	clinical summaries
St Petersburg INCART 12-lead Arrhythmia Database	75	30 min	ECG (12 leads)	beat annotations
Sudden Cardiac Death Holter Database	23	8-24 hours	ECG (2 leads)	beat annotations
T-Wave Alternans Challenge Database	100	2 min	ECG (12 leads, some 2 or 3)	beat annotations

TABLE II

PHYSIOBANK COLLECTIONS OF RR INTERVALS, GAIT, BALANCE, NEURO- AND MYOELECTRIC SIGNALS AND TIME SERIES (AS OF JUNE 2011)

Collection	*Subjects*	*Duration (typical)*	*Signals and time series*	*Other*
CAST RR Interval Sub-Study Database	809	24 hours	-	RR intervals
Congestive Heart Failure RR Interval Database	29	24 hours	-	RR intervals
Exaggerated heart rate oscillations during two meditation techniques	46	10 min - 6 hours	-	RR intervals
Normal Sinus Rhythm RR Interval Database	54	24 hours	-	RR intervals
Spontaneous Ventricular Tachyarrhythmia Database (Version 1.0 from Medtronic, Inc.)	135	5-10 min	-	RR intervals
Gait Dynamics in Neuro-Degenerative Disease	64	2 min	foot pressure	stride intervals
Gait in Aging and Disease Database	15	6-15 min	-	stride intervals
Gait Maturation Database	50	10 min	-	stride intervals
Gait in Parkinson's Disease	93	5 min	multiple foot force signals	stride intervals
Noise Enhancement of Sensorimotor Function	27	5-10 minutes	postural sway	
Unconstrained and Metronomic Walking Database	10	1 hour	-	stride intervals
CHB-MIT Scalp EEG Database	23	1-4 days	23-26 EEG signals	seizure annotations
EEG Motor Movement/Imagery Dataset	109	1-2 min	64 EEG signals	task annotations
Effect of Deep Brain Stimulation on Parkinsonian Tremor	16	1 min	rest tremor velocity	
Evoked Auditory Responses in Normals across Stimulus Level	8	5 min	evoked auditory response, oto-acoustic emissions	stimulus annotations
Term-Preterm EHG Database	300	30 min	EHG	
Examples of Electromyograms	3	10-30 sec	EMG	

REFERENCES

[1] A.L. Goldberger, L.A.N Amaral, L. Glass, J.M. Hausdorff, P.Ch. Ivanov, R.G. Mark, J.E. Mietus, G.B. Moody, C.-K. Peng, H.E. Stanley. PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals, *Circulation* 101(23):e215-e220 [Circulation Electronic Pages; http://circ.ahajournals.org/cgi/content/full/101/23/e215]; 2000 (June 13).

[2] G.B. Moody, R.G. Mark, and A.L. Goldberger. "PhysioNet: a Web-based resource for the study of physiologic signals, *IEEE Eng in Med and Biol* 20(3):70-75 (May-June 2001).

[3] G.B. Moody, R.G. Mark, A.L. Goldberger, and T. Penzel. Stimulating rapid research advances via focused competition: the Computers in Cardiology Challenge 2000, *Computers in Cardiology* 27:207-210 (2000).

[4] T. Penzel, J. McNames, P. de Chazal, B. Raymond, A. Murray and G. Moody. Systematic comparison of different algorithms for apnoea detection based on electrocardiographic recordings, *Medical & Biological Engineering & Computing* 40:402-407 (2002).

[5] G.B. Moody, L.H. Lehman. Predicting Acute Hypotensive Episodes: The 10th Annual PhysioNet/Computers in Cardiology Challenge, *Computers in Cardiology* 36:541-544 (2009). (PMID:20842209)

[6] G.B. Moody. The PhysioNet/Computing in Cardiology Challenge 2010: Mind the Gap, *Computing in Cardiology*, 37:305308 (2010).