PhysioNet: An NIH Research Resource for Complex Signals

Madalena Costa,*,† George B. Moody,‡ Isaac Henry,* and Ary L. Goldberger*

Abstract: The Research Resource for Complex Physiologic Signals, supported by the National Institutes of Health (NIH), is intended to promote and facilitate investigations in the study of cardiovascular and other complex biomedical signals. The resource website (www.physionet.org) has 3 interdependent components: 1) PhysioBank is an archive of well-characterized digital recordings of physiologic signals and related data, including databases of electrocardiogram and heart rate time series from patients with heart failure, coronary disease, sleep apnea syndromes, and cardiac arrhythmias; 2) PhysioToolkit is a library of open-source software for physiologic signal processing and analysis; and 3) PhysioNet, for which the resource is named, is an on-line forum for dissemination and exchange of recorded biomedical signals and open-source software for analyzing them. PhysioNet, in cooperation with the annual Computers in Cardiology conference, hosts a series of challenges inviting participants to tackle clinically interesting problems that are either unsolved or not well solved. PhysioNet invites contributions of databases and software from the biomedical community. **Key words:** algorithms, arrhythmias, database, electrocardiogram, heart failure, ischemia, PhysioNet, signal analysis, sleep apnea.

PhysioNet (http://www.physionet.org) is a web-based resource created in 1999 under the auspices of the National Center for Research Resources (NCRR) of the National Institutes of Health (NIH) to initiate, promote, and catalyze basic-to-bedside research in the study of complex physiologic data (1–5). It provides physiologic data in open, Internet-accessible archives, open-source software for the analysis of physiologic data, and an on-line forum to facilitate cooperative analysis of data and the evaluation of proposed new algorithms. The need for this unique NCRR Biotechnology Center was motivated by the importance of complex signals in basic and clinical investigations and by the fact that vitally important, hypothesis-driven research on complex physiologic signals, both basic and clinical, had been hindered by the critical lack of data, analytic and human/communications resources.

*From the *Beth Israel Deaconess Medical Center, Harvard Medical School, Boston, MA; †Faculty of Sciences, University of Lisbon, Lisbon; and ‡Harvard-MIT Division of Health Sciences and Technology, Boston, MA.*

This work was supported by the National Institutes of Health/National Center for Research Resources (P41-RR13622).

Reprint requests: Ary L. Goldberger, MD, Cardiovascular Division, Beth Israel Deaconess Medical Center, 330 Brookline Ave, Boston, MA 02215; e-mail: agoldber@bidmc.harvard.edu.

0022-0736/03/360S-0037$30.00/0
doi:10.1016/j.jelectrocard.2003.09.038

Structure of the Resource

PhysioNet integrates *core* and *collaborative research*, *service*, *dissemination*, and *training* functions related

to complex physiologic signals via an integrated structure centering around three key interrelated components: *PhysioBank*, a data resource; *PhysioToolkit*, an analytic/software resource; and the *PhysioNet* web site, a dissemination and communications resource, from which the resource takes its name.

PhysioBank is an archive of well-characterized biomedical signals for use by the research community. As we build PhysioBank, we collect, characterize, and document databases of multiparameter signals from healthy subjects and patients with pathological conditions that have major public health implications, eg, epilepsy, congestive heart failure, sleep apnea, sudden cardiac death, myocardial infarction, movement disorders, and aging (6). *PhysioToolkit* is a growing library of signal processing and analytical techniques implemented in open-source software. The PhysioToolkit library includes software for physiological signal processing and analysis; the detection of physiologically significant events using both classic methods and novel techniques from statistical physics, fractal scaling analysis, and nonlinear dynamics; the analysis of nonstationary processes; and interactive display.

The *PhysioNet* web site provides a two-way dynamic link between the resource and the research community for efficient retrieval and submission of data and software from and to PhysioBank and PhysioToolkit. PhysioNet is an on-line forum for the dissemination and exchange of recorded biomedical signals and the software for analyzing such signals; it provides facilities for the cooperative analysis of data and algorithms, where both can be submitted, discussed, evaluated, reviewed, and examined in detail by any investigator wishing to join this on-line community. PhysioNet publishes original data and software, in addition to the text of articles. Traditionally, peer-reviewed articles in biomedicine have been published without access to the original waveform or other signal-related data on which these studies are based. PhysioNet, for the first time, has provided a platform to publish such critical primary data and analytic algorithms, not just the text and summary statistics. This innovation allows validation of the findings and also provides unique databases for future studies. As an educational component, PhysioNet provides on-line tutorials to assist clinicians, students, and basic researchers in making the best use of the Resource (7). In conjunction with Computers in Cardiology, PhysioNet sponsors time-series competitions focusing on important challenges in medical diagnostics and forecasting.

The data and software available via PhysioNet are assigned to one of 3 categories:

1. Class 1 databases and software are fully supported. Databases belonging to this class have been carefully scrutinized and are accompanied by comprehensive sets of annotations that have been derived by at least two experts working independently, with all discrepancies resolved either by consensus or by independent review of a third expert. Class 1 software has been extensively and rigorously tested. Any reports of errors are promptly and publicly logged and carefully reviewed.
2. Class 2 databases and software support published research. Data and software in this group are contributed by authors of the published article or by the journals in which the articles are published. We maintain copies of the original data and software together with corrections submitted by the authors. We encourage users to report errors directly to the authors.
3. Class 3 databases and software comprise other contributed material. Databases belonging to this class are works in progress, which have been considered of interest to the research community although they may have been less thoroughly studied than those in class 1. Software in this class includes code that may need further testing or development.

Data contributed to PhysioNet are placed in classes 2 and 3 on acceptance, and may be admitted to class 1 after review and a public comment period.

PhysioBank

Currently, PhysioNet includes databases of multiparameter 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. In this paper, we describe only the electrocardiogram (ECG) and RR interval databases belonging to classes 1 and 2.

1. European ST-T Database [Class 1]: This database was created to support development and evaluation of algorithms for QRS detection in the presence of ST-T abnormalities, in addition to detectors of ST segments and T-wave changes (8,9). The database comprises 90 two-hour annotated excerpts of ambulatory ECG recordings from 79 participants. Myocardial ischemia was diagnosed or suspected for each

subject. Additional selection criteria were established in order to obtain a representative selection of ECG abnormalities in the database, including baseline ST-segment displacement resulting from conditions such as hypertension, ventricular dyskinesia, and effects of medication. All recordings are annotated beat-by-beat and with respect to rhythm, signal quality, ST, and T-wave changes. Header files for each record summarize pathology, medications, electrolyte imbalance, and technical information about each recording. The creators of the database have contributed 48 complete records, together with header and annotation files for all 90 records to PhysioBank

2. Long-term ST Database [Class 1]: This database was created to be used as a reference for the development and evaluation of automated ischemia detectors as well as basic research into mechanisms and dynamics of myocardial ischemia (10–12). The records were chosen to exhibit a variety of ST-segment changes, including ischemic ST episodes, axis-related nonischemic ST episodes, episodes of slow ST level drift, and episodes containing mixtures of these phenomena. Currently, 43 of 86 records are available on PhysioNet, together with 3 papers describing the database in detail. Each record is 21 to 24 hours long and contains 2 or 3 ECG signals. All records are annotated beat-by-beat (with ST level measurements for each beat) and with respect to ST, rhythm, and signal quality changes.
3. MIT-BIH Arrhythmia Database [Class 1]: This database comprises 48 half-hour excerpts of two-channel, 24-hour, ECG recordings obtained from 47 subjects with several types of arrhythmias (13). Twenty-three records were chosen randomly from a set of over 4,000 recordings collected from a mixed population of inpatients (about 60%) and outpatients (about 40%) at Boston's Beth Israel Hospital; the remaining 25 recordings were selected from the same set to include less common but clinically significant arrhythmias. ECGs were annotated beat-by-beat and with respect to signal quality and rhythm changes. Currently, 25 complete recordings, the first 10 minutes of each of the remaining recordings, and the reference annotation and header files for all records are available online.
4. MIT-BIH Noise Stress Test Database [Class 1]: Twelve half-hour ECG recordings and 3 half-hour recordings of noise typical in ambulatory ECG recordings (14). The ECG recordings were created by adding calibrated amounts of noise (recorded using ambulatory ECG recorders) to clean ECG recordings from the MIT-BIH Arrhythmia Database. This database was created to assess the noise tolerance of arrhythmia detectors.
5. ANSI/AAMI EC13 Test Waveforms [Class 1]: 10 short recordings specified by the current American National Standard for testing various devices that measure heart rate (15).
6. BIDMC Congestive Heart Failure (CHF) Database [Class 2]: This database comprises 15 long-term ECGs (about 20 hours each) from subjects with severe CHF, annotated beat-by-beat (NYHA class 3-4) (16).
7. Post-Ictal Heart Rate Oscillations in Partial Epilepsy [Class 2]: Seven annotated single-lead ECG recordings, with times of seizures indicated (17).
8. QT Database [Class 2]: Over 100 fifteen-minute 2-lead ECG recordings (many excerpted from other databases), with onset, peak, and end markers for P, QRS, T, and (where present) U waves from 30 to 50 selected beats in each recording (18).

The recordings belonging to the Interbeat (RR) Interval Databases contain beat annotations only; the original ECG signals are unavailable.

1. Congestive Heart Failure RR Interval Database [Class 2]: Beat annotation files (about 24 hours each) from 29 subjects with congestive heart failure (NYHA classes 1, 2, and 3).
2. Exaggerated heart rate oscillations during 2 meditation techniques [Class 2]: Additional data from spontaneously and metronomically breathing controls, and from highly trained athletes are available. A paper describing these data is also available (19).
3. Normal Sinus Rhythm RR Interval Database [Class 2]: Beat annotation files (about 24 hours each) from 54 subjects in normal sinus rhythm.

PhysioToolkit: PhysioNet's open-source software archive, PhysioToolkit, is based on the WFDB (WaveForm DataBase) software package written in C and portable between Linux, Unix, MacOS (Apple Computer, Inc, Cupertino, CA), and MS-Windows (Microsoft Corporation, Redmond, WA). The package includes about 60 applications for digital filtering, power spectrum density estimation, QRS detection, graphical annotation editing, printing high-quality "chart recordings," just to mention some examples. PhysioToolkit also includes peer-

reviewed implementations of algorithms for simulation and modeling, physiologic signal processing and linear and non-linear time series analysis. Several of these algorithms are accompanied by publications and/or tutorials. We mention a few examples:

1. The detrended fluctuation analysis (DFA) [Class 1] software provides a new way of quantifying the extent of long-range correlations in a time series (20).
2. The *heartprints* [Class 2] software package provides a novel way to visualize the dynamics of ventricular ectopic activity (21,22).
3. The EDR [Class 2] software package can derive a respiration signal from 1 or 2 ECG signals (23).
4. The *apdet* [Class 2] software package provides automated detection of obstructive sleep apnea by analysis of interbeat intervals in the ECG (24).
5. *Ecgpuwave* [Class 2] is software for locating waveform boundaries in the ECG (ie, the beginnings, peaks, and ends of P, QRS, and T waves) (25).
6. The *pNNx* [Class 2] software package offers a novel method for characterizing HRV with a family of time-domain statistics of which pNN50 is the best known (26).

Challenges

In cooperation with the annual Computers in Cardiology conference, PhysioNet hosts a series of challenges, inviting participants to tackle clinically interesting problems that are either unsolved or not well solved. The first of these challenges began in February 2000.

We select challenge topics based on intrinsic scientific interest, seeking problems for which a variety of approaches appear possible to implement and evaluate in a relatively short time, and for which PhysioNet can provide a "virtual laboratory" in which participants can get to work with all of the prerequisite experimental data. A new challenge topic is announced in February each year.

Challenge 2000: Detecting Sleep Apnea From the Electrocardiogram ECG

Obstructive sleep apnea is a very common and clinically significant breathing disorder. Diagnosis typically requires a polysomnogram, an expensive and often uncomfortable in-clinic test. The results of the challenge demonstrate that a variety of methods for analyzing an ambulatory ECG recording, which can be obtained for a tenth of the cost at home with little or no discomfort, are just as reliable for diagnosis of sleep apnea (27,28).

Challenge 2001: Predicting Paroxysmal Atrial Fibrillation

New technology offers the possibility of averting the onset of atrial fibrillation, the most common serious abnormal cardiac rhythm; but this technology will be most beneficial if coupled with reliable methods for forecasting this arrhythmia, a problem that had received little attention prior to this challenge.

Challenge 2002: RR Interval Time Series Modeling

Variations in instantaneous heart rate (as measured using RR intervals, the time intervals between heart beats) have been recognized for many years as sensitive, noninvasive indicators of a variety of cardiac and neurologic disorders. Researchers, however, have lacked access to realistic models of heart rate variability, which are invaluable tools for the development of more accurate diagnostic tests and measures of therapeutic efficacy.

Challenge 2003: Detection and Forecasting of Cardiac Ischemia

At the time of writing, the 2003 challenge is underway. This year's challenge asks if it is possible to tell the difference between transient ST changes in the ECG that are due to myocardial ischemia, and those that are not. ST changes can result from a wide variety of other causes, including changes in the heart rate, conduction pattern, position of the subject, and noise in the ECG. Using all available evidence, expert annotators have classified each significant ST change in the Long-Term ST Database as ischemic or nonischemic, and they have contributed half of that database to PhysioNet as a learning set for development of algorithms for classifying ST changes. The other half is being used to score algorithms entered into the challenge and will be posted on PhysioNet after the conclusion of the challenge in September.

Future Directions

In keeping with its original aims, PhysioNet is planning several major initiatives that will benefit the biomedical research community by developing new relevant collections of well-characterized physiologic signals and time series, rigorously tested software for study of these and similar data, and tutorial and reference materials to support their use and improvement. We briefly describe some of our ongoing projects below.

The Cardiac Arrhythmia Suppression Trial (CAST) was designed to test the hypothesis that the suppression of symptomatic or mildly symptomatic premature ventricular complexes (PVCs) in survivors of myocardial infraction (MI) would decrease the number of deaths from ventricular arrhythmias and improve survival. The CAST RR Interval Sub-Study Database was created from data gathered in the CAST and the follow-up CAST II studies, in order to investigate the relationship between survival and changes in heart rate variability in response to antiarrhythmic treatment; it has been contributed to PhysioBank and will be posted shortly (29).

We have begun to create a sudden cardiac death database, with 19 recordings from the Beth Israel Deaconess Medical Center (BIDMC)/MIT collection, including 16 patients with underlying sinus rhythm and 3 with atrial fibrillation. Each of these recordings includes a sustained ventricular tachyarrhythmia, and most include an actual cardiac arrest. We welcome contributions of similar recordings from other investigators to be included in this database, which will be annotated and reviewed in an open project hosted by PhysioNet.

The MIMIC II project is collecting data from patients admitted to the adult intensive care units of the BIDMC. Each record consists of at least four continuously digitized waveforms (typically, 2 ECG signals, arterial blood pressure, and pulmonary artery pressure), together with relevant clinical information; recordings are continuous from admission to discharge and typically last for multiple days or weeks. Over a period of 2 years, the project expects to collect approximately 7,300 patient records, from which a database will be created and contributed to PhysioBank.

In order to expand PhysioToolkit and further develop PhysioNet infrastructure, several major software projects are planned or under way; these include:

1. Extending the capabilities of the WaveForm Database (WFDB) Software Package, the Resource's core software for creating, processing, viewing, analyzing, annotating, and archiving physiologic time series, with particular emphasis on further development of cross-platform visualization tools
2. Developing software for real-time physiologic signal handling, including digitization and streaming over the Internet
3. Developing novel techniques, including trend compression, for indexing and searching physiologic signal databases
4. Creating algorithms for real-time data quality assessment, including identification of unreliable signals and false alarms in intensive care and similar settings
5. Designing forecasting algorithms for predicting outcome and clinical events based on information from physiologic signals, time series, and other sources
6. Documenting and disseminating these algorithms via PhysioNet.

Building communities of researchers working to solve common problems in the contexts of diverse studies can be a powerful stimulus to progress. PhysioNet seeks to create the conditions to support multidisciplinary collaborations that make extensive use of its freely available data and software resources, and to provide investigators with the means to bootstrap new studies.

References

1. Goldberger AL, Amaral LAN, Glass L, et al: PhysioBank, PhysioToolkit, and PhysioNet: Components of a new Research Resource for Complex Physiologic Signals. Circulation 101:e215, 2000
2. Moody GB, Mark RG, Goldberger AL: PhysioNet: A research resource for studies of complex physiologic and biomedical signals. Comput Cardiol 27:179, 2000
3. Henry IC, Moody GB, Mark RG, et al: PhysioNet: An NIH Research Resource for physiologic datasets and open-source software. Proceedings from the 14th IEEE CBMS Symposium. 14:245, 2001
4. Moody GB, Mark RG, Goldberger AL: PhysioNet: A web-based resource for the study of physiologic signals. IEEE Eng Med Biol 20:70, 2001
5. Goldberger AL, Amaral LAN, Hausdorff JM, et al: Fractal dynamics in physiology: Alterations with disease and aging. Proc Natl Acad Sci USA 99:2466, 2002 (suppl 1)
6. Korhonen I, van Gils M, Gade J: The challenges in creating critical-care databases. IEEE Eng Med Biol 20:58, 2001
7. Mukkamala R, Moody GB, Mark RG: Introduction of

computational models to PhysioNet. Comput Cardiol 28:77, 2001
8. Taddei A, Biagini A, Distante G, et al: The European ST-T database: Development, distribution and use. Comput Cardiol 17:177, 1990
9. Taddei A, Distante G, Emdin M, et al: The European ST-T database: Standard for evaluating systems for the analysis of ST-T changes in ambulatory electrocardiography. Eur Heart J 13:1164, 1992
10. Jager F, Moody GB, Taddei A, et al: Development of a long-term database for assessing the performance of transient ischemia detectors. Comput Cardiol 23: 481, 1996
11. Jager F, Taddei A, Emdin M, et al: The Long-Term ST Database: A research resource for algorithm development and physiologic studies of transient myocardial ischemia. Comput Cardiol 27:841, 2000
12. Jager F, Taddei A, Moody GB, et al: Long-Term ST database: a reference for the development and evaluation of automated ischaemia detectors and for the study of the dynamics of myocardial ischemia. Med Biol Eng Comput 41:172, 2000
13. Moody GB, Mark RG: The impact of the MIT-BIH arrhythmia database. IEEE Eng Med Biol 20:45, 2001
14. Moody GB, Muldrow WK, Mark RG: A noise stress test for arrhythmia detectors. Comput Cardiol 11: 381, 1984
15. Cardiac monitors, heart rate meters, and alarms [American National Standard (ANSI/AAMI EC13: 2002)]. Arlington, VA: Association for the Advancement of Medical Instrumentation, 2002.
16. Baim DS, Colucci WS, Monrad ES, et al: Survival of patients with severe congestive heart failure treated with oral milrinone. J Am Coll Cardiol 7:661, 1986
17. Al-Aweel IC, Krishnamurthy KB, Hausdorff JM, et al: Post-Ictal Heart Rate Oscillations in Partial Epilepsy. Neurology 53:1590, 1999
18. Laguna P, Mark RG, Goldberger AL, et al: A database for evaluation of algorithms for measurement of QT and other waveform intervals in the ECG. Comput Cardiol 24:673, 1997
19. Peng C-K, Mietus JE, Liu Y, et al: Exaggerated heart rate oscillations during two meditation techniques. Int J Cardiol 70:101, 1999
20. Peng C-K, Havlin S, Stanley HE, et al: Quantification of scaling exponents and crossover phenomena in nonstationary heartbeat time series. Chaos 5:82, 1995
21. Schulte-Frohlinde V, Ashkenazy Y, Ivanov PCh, et al: Noise effects on the complex patterns of abnormal heartbeats. Phys Rev Lett 87:068104, 2001
22. Schulte-Frohlinde V, Ashkenazy Y, Goldberger AL, et al: Complex patterns of abnormal heartbeats. Phys Rev E 66:031901, 2002
23. Moody GB, Mark RG, Zoccola A, et al: Derivation of respiratory signals from multi-lead ECGs. Comput Cardiol 12:113, 1985
24. Mietus JE, Peng C-K, Ivanov PCh, et al: Hilbert transform based sleep apnea detection using a single lead electrocardiogram. Comput Cardiol 27:753, 2000
25. Jané R, Blasi A, García J, et al: Evaluation of an automatic threshold based detector of waveform limits in Holter ECG with the QT database. Comput Cardiol 24:295, 1997
26. Mietus JE, Peng C-K, Henry I, et al: The pNNx files: Re-examining a widely used heart rate variability measure. Heart 88:378, 2002
27. Moody GB, Mark RG, Goldberger AL, et al: Stimulating rapid research advances via focused competition: The Computers in Cardiology challenge 2000. Comput Cardiol 27:207, 2000
28. Penzel T, McNames J, Chazal P, et al: Systematic comparison of different algorithms for apnoea detection based on electrocardiogram recordings. Med Biol Eng Comput 40:402, 2002
29. Stein PK, Domitrovich PP, Klieger RE, et al: Clinical and demographic determinants of heart rate variability in patients post myocardial infarction: Insights from the Cardiac Arrhythmia Suppression Trial (CAST). Clin Cardiol 23:187, 2000