PhysioBank, PhysioToolkit, and PhysioNet
Components of a New Research Resource for Complex Physiologic Signals

Ary L. Goldberger, MD; Luis A.N. Amaral, PhD; Leon Glass, PhD; Jeffrey M. Hausdorff, PhD; Plamen Ch. Ivanov, PhD; Roger G. Mark, MD, PhD; Joseph E. Mietus, BS; George B. Moody, BS; Chung-Kang Peng, PhD; H. Eugene Stanley, PhD

Abstract—The newly inaugurated Research Resource for Complex Physiologic Signals, which was created under the auspices of the National Center for Research Resources of the National Institutes of Health, is intended to stimulate current research and new investigations in the study of cardiovascular and other complex biomedical signals. The resource has 3 interdependent components. PhysioBank is a large and growing archive of well-characterized digital recordings of physiological signals and related data for use by the biomedical research community. It currently includes databases of multiparameter cardiopulmonary, neural, and other biomedical signals from healthy subjects and from patients with a variety of conditions with major public health implications, including life-threatening arrhythmias, congestive heart failure, sleep apnea, neurological disorders, and aging. PhysioToolkit is a library of open-source software for physiological signal processing and analysis, the detection of physiologically significant events using both classic techniques and novel methods based on statistical physics and nonlinear dynamics, the interactive display and characterization of signals, the creation of new databases, the simulation of physiological and other signals, the quantitative evaluation and comparison of analysis methods, and the analysis of nonstationary processes. PhysioNet is an on-line forum for the dissemination and exchange of recorded biomedical signals and open-source software for analyzing them. It provides facilities for the cooperative analysis of data and the evaluation of proposed new algorithms. In addition to providing free electronic access to PhysioBank data and PhysioToolkit software via the World Wide Web (http://www.physionet.org), PhysioNet offers services and training via on-line tutorials to assist users with varying levels of expertise. **(*Circulation*. 2000;101:e215-e220.)**

Key Words: aging ■ databases ■ death, sudden ■ electrophysiology ■ heart rate ■ nervous system, autonomic ■ nonlinear dynamics

The purpose of this article is to provide a brief introduction to the newly established Research Resource for Complex Physiologic Signals* and to invite participation by the biomedical community in a cooperative research enterprise.

Background and Objectives

Clinical diagnoses and basic investigations are critically dependent on the ability to record and analyze physiological signals. Examples of such signals include ECG and heart rate recordings from patients at a high risk of sudden death and healthy control subjects (Figure 1), fluctuations of hormone and other molecular biological signal messengers and transducers in neuroendocrine dynamics, and multiparameter recordings in sleep apnea (Figure 2) and epilepsy. Over the past few decades, however, the clinical and investigative analyses of these signals has not substantially changed, despite the technological advances that allow for the recording and storage of massive datasets of continuously fluctuating signals. Furthermore, although these typically complex signals represent processes that are nonlinear and nonstationary in nature (Figure 1),[1,2] the analytic tools and models used to study such data often assume linearity and stationarity. Such conventional techniques include analysis of means, standard

From the Margret and H.A. Rey Laboratory for Nonlinear Dynamics in Medicine, Beth Israel Deaconess Medical Center/Harvard Medical School, Boston, Mass (A.L.G., J.M.H., J.E.M., C.-K.P); the Center for Polymer Studies and Department of Physics, Boston University, Boston, Mass (L.A.N.A., P.Ch.I., H.E.S.); the Division of Health Sciences and Technology, Harvard University/Massachusetts Institute of Technology, Cambridge, Mass (R.G.M., G.B.M.); and the Centre for Nonlinear Dynamics in Physiology and Medicine, Department of Physiology, McGill University, Montréal, Québec, Canada (L.G.).

*This multicenter Resource was established, as of September 1, 1999, under the auspices of the National Center for Research Resources, National Institutes of Health. Participating Core Centers include Beth Israel Deaconess Medical Center/Harvard Medical School, Boston University's Center for Polymer Studies, Harvard University-Massachusetts Institute of Technology's Division of Health Sciences and Technology, and McGill University's Centre for Nonlinear Dynamics in Physiology and Medicine.

Correspondence to Ary L. Goldberger, MD, Cardiovascular Division, GZ-435, Beth Israel Deaconess Medical Center, 330 Brookline Avenue, Boston, MA 02215. E-mail ary@astro.bidmc.harvard.edu

Circulation **is available at http://www.circulationaha.org**

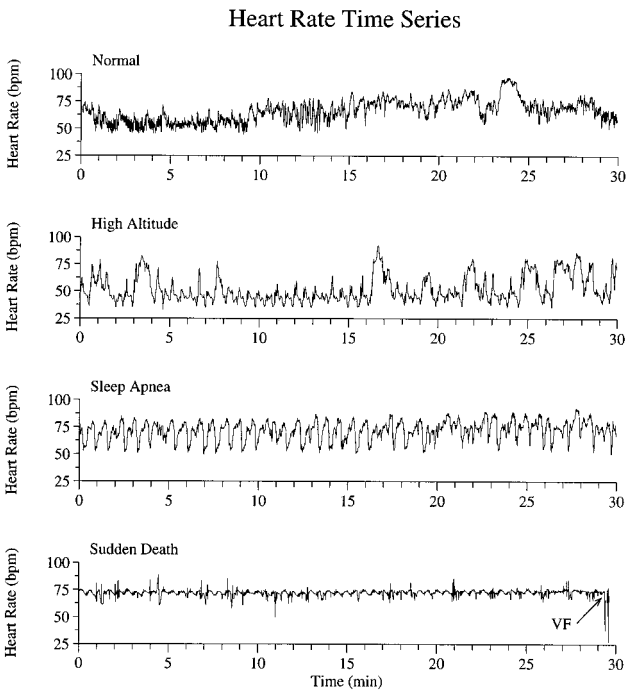

Figure 1. Representative complex physiological fluctuations. Heart rate (normal sinus rhythm) time series of 30 minutes from (top) a healthy subject at sea level, (second row) a healthy subject at a high altitude (4700 m), (third row) a subject with obstructive sleep apnea, and (bottom) a sudden cardiac death subject with ventricular fibrillation (VF). Note the highly nonstationary heart rate variations in the first and second rows, such that the statistical properties change over relatively short time periods. Nonstationarity, as well as sustained oscillations, as seen in the bottom 2 rows, suggest underlying nonlinear dynamics.[1,8,9]

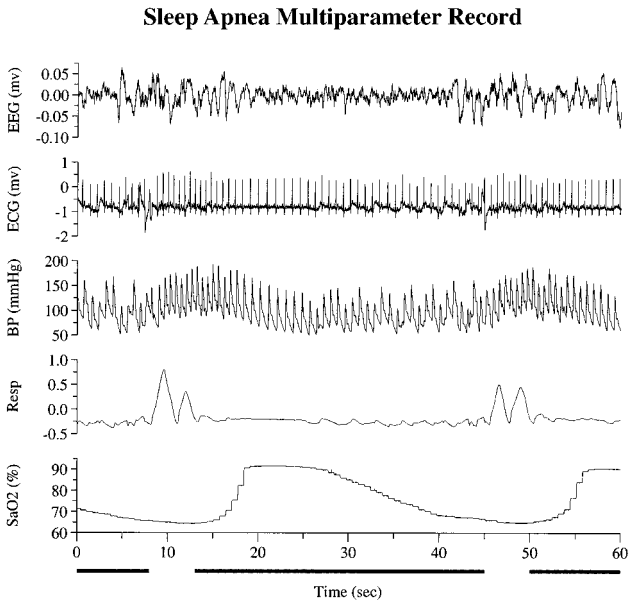

Figure 2. Multichannel record of patient with obstructive sleep apnea recorded during sleep study. Data show complex coupling of cardiopulmonary and electroencephalographic (EEG) dynamics. The shaded lines below the signals indicate prolonged sleep apnea episodes, characterized by periodic cessation of breathing. BP indicates blood pressure; Resp, respiration; and SaO2, arterial oxygen saturation.

deviations, and other features of histograms and classic power-spectrum analysis.

Recent findings[3–8] indicate that such complex datasets may contain "hidden information," which is defined here as information that is neither visually apparent nor extractable with conventional methods of analysis. Such information promises to be of clinical value (forecasting sudden cardiac death in ambulatory patients or cardiopulmonary catastrophes during surgical procedures). It may relate to basic mechanisms in molecular biology and physiology.[9] With the advent of sophisticated computational tools and powerful methods for storing and disseminating vast quantities of information, the biomedical research community seems to be on the cusp of a major breakthrough at both the clinical and basic levels of investigation.[10]

The Challenges

Unfortunately, vitally important, hypothesis-driven research on complex biomedical signals, both basic and clinical, has been hindered by the lack of the following 3 types of resources.

Data Resources

Researchers need, but generally lack access to, high-quality, rigorously validated, and standardized databases of biomedical signals obtained in a variety of healthy and pathological conditions. In many cases, both experimental and clinical data are collected at considerable expense to the public, analyzed once by their collectors, and filed away indefinitely. As a result, federal agencies and other research sponsors may fund repetitive and redundant projects.

Analytic Resources

Significant effort is required to develop software for signal processing, time-series analysis, and related functions needed by researchers working with these datasets. Commercial software is unavailable for many of these functions, and what little is available is generally unsuitable for use with multigigabyte datasets. Researchers frequently develop such software at considerable expense for use within a single project. Furthermore, the validation of signal processing and analysis algorithms (and of their software implementations) is rarely performed in a way that permits rigorous peer review. Frequently, researchers self-evaluate their software using the same private dataset used for its development and then report its behavior using ad hoc measures of performance.

Human and Communications Resources

Advances in the field of complex biomedical signal analysis have also been limited by the lack of concentrated and concerted research efforts. Furthermore, advanced analytic techniques developed by experts in the field are often not readily accessible to end users, who may lack the background and technical skills needed for the successful use of these new tools. Even among experts, the processes of the evaluation of new algorithms and the comparison of research results are complicated by subtle variations in software implementations of algorithms that themselves may not be thoroughly speci-

fied in research reports and by a lack of standardized test data and testing methods.

This set of problems and challenges is reminiscent of the status of research in genetics and molecular biology before the advent of GenBank, which is arguably the first successful large-scale example of a medium for the exchange and dissemination of raw research data. By allowing researchers to begin their work with instant access to all of the ever-increasing store of knowledge of DNA sequences, GenBank encourages innovative rather than redundant research, leverages research expenditures to promote the most efficient use of limited funds, and makes serendipitous discoveries more likely.

In much the same way, biomedical research in general, and cardiovascular investigations in particular, rely on large quantities of physiological data. Intersubject variability is a major focus of research interest in biomedical research (as it is in genetic research); hence, information gathered from a variety of subjects has an added value quite beyond the importance of verifying an initial set of findings. In contrast to the relatively simple alphabet of DNA sequences, however, biomedical signals are characterized by complex time-varying features and interrelationships that require nontrivial computational techniques for quantification and analysis (Figures 1 and 2).

The new resource offers researchers a medium for the exchange and dissemination of such biomedical signals and algorithms. It aims to bring to biomedical research the diverse and compelling benefits offered to molecular biology by GenBank. The central mission of the resource is to accelerate current research progress and to stimulate and bootstrap new investigations in the study of complex biomedical signals with an integrated approach.

Structure of the Resource: Data, Software, Interchange

The Research Resource for Complex Physiologic Signals has the following 3 key, interrelated components: PhysioBank, a data resource; PhysioToolkit, an analytic/software resource; and PhysioNet, a dissemination/communications resource (Figure 3).

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). This component will also include other databases that will contain signals obtained from selected in vitro and in vivo experiments, as well as from physiologically-motivated algorithms that generate complex time series.[11] A large and growing collection of these databases is now available to the scientific community via the PhysioNet website and on CD-ROM.

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; interactive display and characterization of signals; the creation of new databases to support further development of PhysioBank; the simulation of physiological and other signals, when such signals may be useful for the study of algorithm behavior; and the quantitative evaluation and comparison of analysis algorithms.

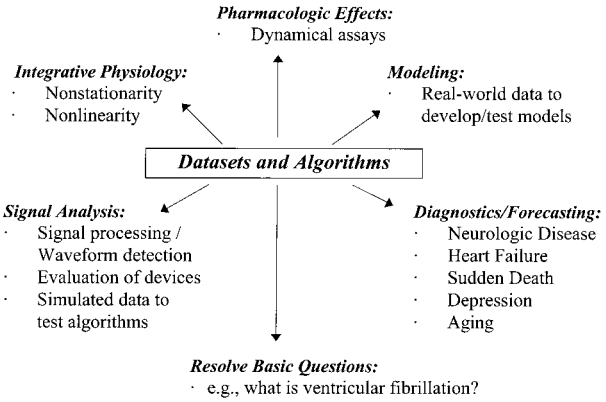

Figure 3. The datasets and algorithms provided by this resource are intended to enable a wide range of basic and clinical investigations.

PhysioNet provides a 2-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 via the World Wide Web (http://www.physionet.org). 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 the evaluation of proposed new algorithms. It provides a meeting place for physiological data and algorithms, where both can be submitted, discussed, evaluated, reviewed, and examined in detail by any investigator willing to join this on-line community. PhysioNet also provides a means to resolve differences in results that may result from errors in the interpretation of algorithm descriptions, errors in algorithm implementation, or fundamental errors in algorithm design. As an educational component, PhysioNet provides on-line tutorials to assist clinicians, students, and basic researchers in making the best use of the resource (see Appendix). In conjunction with *Computers in Cardiology 2000*, PhysioNet is supporting a time-series competition focusing on the challenge of detecting obstructive sleep apnea from the ECG (http://www.physionet.org/cinc-challenge-2000.shtml).

Data and software that are available via PhysioNet fall into the following 3 categories:

1. Fully supported. PhysioBank data and PhysioToolkit software belong to this category, which consists of well-characterized, carefully and multiply reviewed data and rigorously tested software. We attempt to correct any errors in fully supported data and software before including them in PhysioBank and PhysioTool-

Potential Benefits of the Research Resource for Complex Physiologic Signals

Stimulate new investigations on the dynamics of physiological systems
Facilitate ongoing basic and clinical studies
Provide larger, more comprehensive databases than any single center can collect
Provide a model for the future collection and analysis of cardiovascular and other biomedical data based on time series rather than averaged quantities
Provide a permanent repository for time-series data from large multicenter studies, other publicly-funded studies, and publications
Protect integrity and reliability of raw data and analyses
Allow systematic testing of traditional and novel diagnostic/prognostic algorithms on standard databases
Provide analytic tools and guidance in their use to experimentalists
Facilitate data-leveraging and data-mining: obtain maximal information from databases, including unanticipated findings
Provide data to theoreticians and nonspecialists to encourage new biomedical applications of novel methods
Foster technology transfer
Support interdisciplinary studies among basic and clinical investigators, bioengineers, applied mathematicians, physiologists, computer scientists, and biophysicists

kit. Any reports of errors are promptly and publicly logged and carefully reviewed.

2. Contributed from publications. Data and software in this group are contributed by authors of published articles or by the journals in which the articles are published. These contributions offer the opportunity to gain additional insight into the experimental methods used in the associated studies, to confirm the authors' published results, and to apply other analytic methods, either for comparison with the original authors' work or to reuse their data for different purposes. We provide on-line access via PhysioNet to unmodified archival copies of these materials, and we forward reports of errors to the original contributors.
3. Other contributed material. This category includes works in progress judged to be of significant interest to the research community, such as incompletely annotated databases and implementations of novel algorithms that have not yet been rigorously tested. PhysioNet offers a forum for the peer review of such data and software and for collaborative efforts among geographically scattered researchers working to develop these materials. In some cases, we provide archival storage and on-line access via PhysioNet to these materials; in other cases, PhysioNet hosts on-line discussions only and provides links to external data and software repositories.

Potential Benefits

The combination of these 3 interrelated components—signal databases, analytic tools, and an on-line forum—is intended to make this resource useful to a wide range of researchers and clinicians (Table). Experimentalists may benefit from access to PhysioToolkit's growing collection of rigorously tested analytic software and especially from PhysioNet tutorials and on-line discussions that can help identify applicable techniques for addressing their research questions and that compare the merits of available methods across various problem domains. Theoreticians and others who lack access to well-characterized signals may benefit from access to PhysioBank's growing collections of data. Indeed, we have found, in the course of distributing a number of the PhysioBank databases on a smaller scale to the research community during the past 20 years, that the availability of such data has frequently encouraged nonspecialists with innovative methods to tackle problems in biomedical areas that they might not otherwise have attempted, often with encouraging results.

Reference databases[12–14] (http://ecg.mit.edu/ and http://reylab.bidmc.harvard.edu/) are also essential resources for developers and evaluators of algorithms that analyze biomedical signals who need to test algorithms with realistic data and to perform these tests repeatedly and reproducibly as algorithm refinements are proposed. These databases also have value in medical education by providing well-documented case studies of both common and rare but clinically significant diseases. By making well-characterized clinical data available to researchers, these databases will make it possible to formulate and answer numerous physiological questions (Figure 3), without the need to develop a new set of reference data at great cost in each case.[15] In this regard, PhysioBank can serve as a final and permanent repository for time-series data from publicly-funded studies, such as large multicenter clinical trials, or physiological studies conducted by the National Aeronautics and Space Administration (NASA). Such data are, by statute, in the public domain, yet often they cannot be readily accessed by qualified investigators, even long after the original investigators have completed their analysis. Furthermore, irreplaceable physiological data, such as electrocardiographic recordings from NASA's pre-Shuttle missions, are no longer retrievable due to a lack of mechanisms for data annotation, analysis, and archival. Such mistakes should not be repeated.[16]

Another source of concern in the biomedical community in recent years has been the problem of scientific misconduct,[17] including the publication of fraudulent data. These lapses rob not only the scientific community, which relies on published findings, but also the taxpayers who support this research. Considerable effort has been directed at designing safeguards to prevent or detect such fraudulent science. Unfortunately, even the most careful peer review may fail to discover deliberate misrepresentation or unintentional mistakes. The willingness of investigators to deposit original datasets as part of a research resource may be one of the most potent assurances of the integrity of data. The fact that these datasets can be reanalyzed by the scientific community at large permits ready double-checking of the initial findings and serves as perhaps the most efficient remedy for unintentional errors. An additional benefit is that the data can also be restudied with new techniques as they become available, allowing for "data-leveraging" or "data-mining." For federally-funded investigations, the investigators' consent to eventually bank relevant physiological signals in such a resource could become a standard part of certain research

proposals, with provisions for the absolute protection of subject anonymity.

Without common databases, such as those provided by PhysioBank, it can be impossible to resolve certain contradictory research results, ranging from understanding the dynamics of normal sinus rhythm to life-threatening cardiac arrhythmias.[9,18] A specific example of how the absence of a well-characterized database has impeded scientific progress and prevented the resolution of a major, clinically relevant problem relates to the mechanism of ventricular fibrillation (VF), the major cardiac arrhythmia associated with sudden death (Figure 1). Although multiple investigators have studied the dynamics of this electrical disturbance, there remains a remarkable lack of consensus about its underlying mechanism(s).[9,19–21] A probable source of disagreement has been that different investigators have studied different sets of waveforms obtained in diverse preparations. Furthermore, the analyses used to reach these disparate conclusions have made use of different analytic techniques or different implementations of similar algorithms. Without a standardized database of high-quality signals accepted by the community of investigators using the same algorithms, attempts to resolve this central controversy are likely to be at best incomplete and at worst reminiscent of the parable of the blind men and the elephant. To aid in the development of new approaches to defibrillation,[20,22] it would be invaluable to make available electrophysiological mapping data collected during VF in model systems. A more informed analysis of VF will lead to a deeper understanding of the mechanism of complex wave phenomena, which underlie not only sudden cardiac death[23,24] but possibly other pathophysiological dynamics, such as seizure disorders.

Finally, the peer review process itself has been shaped historically by the constraints of the publication process. It has never been feasible to publish the raw data that support research results—until now. The Internet and the near-universal availability of inexpensive, high-capacity, mass-storage media such as the CD-ROM have made it possible to consider a new paradigm for scientific publication and for peer review. Within a few years, it may not be considered acceptable for a study based on physiological signal analysis to be published in most peer-reviewed journals without making supporting raw data available for examination, and no peer review may be considered sufficiently rigorous unless it has included an examination of how the research results have been derived from these data. The resource now provides a site for authors to publish such "dynamic appendices" to accompany their articles, giving readers access to the actual time-series data on which statistical tests were performed. A precedent for the publication of such primary datasets has already been established with respect to high-resolution biomolecular structural data, which are now released at or before the time of publication of the articles describing these data.[25] "Open-source research" is a powerful idea that may sweep aside entrenched patterns of behavior in research, just as increasing awareness of the benefits of open-source software is changing the practice of software development. We hope that this new National Institutes of Health resource will now help extend these benefits and their often unanticipated rewards to those with an interest in complex physiological signals.

Appendix

Getting Started with PhysioNet

The only prerequisites for using PhysioNet are a Web browser and a connection to the Internet. To begin, please visit the PhysioNet welcome page at http://www.physionet.org. There is no charge for using PhysioNet, although free registration is required if you wish to contribute data or software. First-time visitors may wish to begin their exploration of the resource by following the links from the welcome page to PhysioBank and then to *An Introduction to the PhysioBank Archives*, which contains pointers regarding samples of the available databases and suggestions for viewing them.

PhysioToolkit software may be downloaded in source form or in precompiled versions for Linux/×86, Solaris/Sparc, or MS-DOS/MS-Windows (precompiled versions for other environments may also be available).

We invite your comments and contributions of data and software for review, discussion, and possible inclusion in PhysioBank and PhysioToolkit. Contributors are asked to review our guidelines at http://www.physionet.org/guidelines.shtml.

PhysioNet is supported by mirrored Web servers at multiple locations around the world to provide reliable access to the research community. We invite users to replicate the PhysioNet website locally and to add their sites to our list of mirrors; please visit http://www.physionet.org/mirrors/ for further information.

Acknowledgments

We wish to thank Deborah Dimond for her continued and invaluable assistance and Paul Trunfio for his many helpful suggestions. This work was supported by a grant from the National Center for Research Resources of the National Institutes of Health (P41 RR13622).

References

1. Glass L, Mackey MC. *From Clocks to Chaos: The Rhythms of Life.* Princeton, NJ: Princeton University Press; 1988.
2. Service RF. Complex systems: exploring the systems of life. *Science.* 1999;284:80–81, 83.
3. Stein KM, Karagounis LA, Anderson JL, et al. Fractal clustering of ventricular ectopy correlates with sympathetic tone preceding ectopic beats. *Circulation.* 1995;91:722–727.
4. Ho KKL, Moody GB, Peng C-K, et al. Predicting survival in heart failure case and control subjects by use of fully automated methods for deriving nonlinear and conventional indices of heart rate dynamics. *Circulation.* 1997;96:842–848.
5. Peng C-K, Hausdorff JM, Havlin S, et al. Multiple-time scale analysis of physiological time series under neural control. *Physica A.* 1998;249:491–500.
6. Vikman S, Mäkikallio TH, Yli-Mäyrey S, et al. Altered complexity and correlation properties of R-R interval dynamics before the spontaneous onset of paroxysmal atrial fibrillation. *Circulation.* 1999;100:2079–2084.
7. Kresh JY, Izrailtyan I. Evolution in functional complexity of heart rate dynamics: a measure of cardiac allograft adaptability. *Am J Physiol.* 1998;275:R720–R727.
8. Ivanov PCh, Amaral LAN, Goldberger AL, et al. Multifractality in human heartbeat dynamics. *Nature.* 1999;399:461–465.
9. Goldberger AL. Non-linear dynamics for clinicians: chaos theory, fractals, and complexity at the bedside. *Lancet.* 1996;347:1312–1314.
10. Coffey DS. Self-organization, complexity, and chaos: the new biology for medicine. *Nat Med.* 1998;4:882–885.
11. Ivanov PCh, Amaral LAN, Goldberger AL, et al. Stochastic feedback and the regulation of biological rhythms. *Europhys Lett.* 1998;43:363–368.
12. Moody GB, Mark RG. The MIT-BIH arrhythmia database on CD-ROM and software for use with it. *Computers Cardiol.* 1990;185–188.

13. Moody GB, Mark RG. A database to support development and evaluation of intelligent intensive care monitoring. *Computers Cardiol*. 1996; 657–660.
14. Moody GB, Feldman CL, Bailey JJ. Standards and applicable databases for long-term ECG monitoring. *J Electrocardiol*. 1993;26(suppl): 151–155.
15. Rockwell RC, Abeles RP. Sharing and archiving data is fundamental to scientific progress. *J Gerontol B Psychol Sci Soc Sci*. 1998;53:S5–S8. Editorial.
16. Lawler A. Space science feels budget ax in Senate. *Science*. 1999;285: 2045–2047.
17. Alberts B, Shine K. Scientists and the integrity of research. *Science*. 1994;266:1660–1661.
18. Glass L. Chaos and heart rate variability. *J Cardiovasc Electrophysiol*. 1999;10:358–360.
19. Kaplan DT, Cohen RJ. Is fibrillation chaos? *Circ Res*. 1990;67:886–892.
20. Weiss JN, Garfinkel A, Karagueuzian HS, et al. Chaos and the transition to ventricular fibrillation: a new approach to antiarrhythmic drug evaluation. *Circulation*. 1999;99:2819–2826.
21. Gilmour RF Jr, Chialvo DR. Electrical restitution, critical mass, and the riddle of fibrillation. *J Cardiovasc Electrophysiol*. 1999;10: 1087–1089.
22. Zhou X, Daubert JP, Wolf PD, et al. Epicardial mapping of ventricular defibrillation with monophasic and biphasic shocks in dogs. *Circ Res*. 1993;72:145–160.
23. Witkowski FX, Leon LJ, Penkoske PA, et al. Spatiotemporal evolution of ventricular fibrillation. *Nature*. 1998;392:78–82.
24. Mandapati R, Asano Y, Baxter WT, et al. Quantification of effects of global ischemia on dynamics of ventricular fibrillation in isolated rabbit heart. *Circulation*. 1998;98:1688–1696.
25. Time to withdraw an undesirable privilege? *Nature*. 1998;391:617. Editorial.