Standards and Applicable Databases for Long-term ECG Monitoring

George B. Moody,* Charles L. Feldman, SCD,† and James J. Bailey, MD‡

Continuous monitoring of the electrocardiogram (ECG) in both patients and ambulatory subjects has become a very common procedure during the past 30 years, with diverse applications that include screening for cardiac arrhythmias or transient ischemia, evaluation of the efficacy of antiarrhythmic drug therapy, and surgical and critical care monitoring. Since the first intensive care units were established in the 1960s, the need for automated data reduction and analysis of the ECG has been apparent, motivated by the very large amount of data that must be analyzed (on the order of 100,000 cardiac cycles per patient per day).

As clinical experience has led to the identification of more and more prognostic indicators in the ECG, clinicians have demanded and received increasingly sophisticated automated electrocardiographic analyzers. The early heart rate monitors rapidly evolved into devices designed to detect ventricular fibrillation, and other devices for tracking "premonitory" ventricular arrhythmias. Many newer devices attempt to detect supraventricular arrhythmias and transient ischemic ST changes.

Visual analysis of the long-term ECG is exceedingly tedious and far from simple. Accurate diagnosis of electrocardiographic abnormalities requires attention to subtle features of the signals, features that may appear only rarely and are often obscured by or mimicked by noise. Diagnostic criteria are complicated by inter- and intrapatient variability of both normal and abnormal electrocardiographic features. Given these considerations, it is not surprising that developers are faced with a difficult task in the design of algorithms for automated electrocardiographic analysis, and that the results of their efforts are less than perfect.

*From *Harvard-MIT Division of Health Sciences and Technology, Cambridge, †Brigham and Women's Hospital, Boston, Massachusetts, and ‡National Institutes of Health, Bethesda, Maryland.*

Reprint requests: George B. Moody, MIT Room 20A-113, Cambridge, MA 02139.

Since automated electrocardiographic analyzers vary in performance, and since their performance is dependent on the characteristics of their input, quantitative evaluations of these devices are essential in order to assess the usefulness of their outputs. For these reasons, several documented databases of long-term ECGs have been developed over the past 20 years with the goal of providing material for standard performance testing of automated long-term electrocardiographic monitoring algorithms and devices. The purpose of this study is to review these databases with respect to their advantages, limitations, and appropriate use in evaluation protocols.

History

In September 1972, many of those working on the development of automated arrhythmia detectors met at the second biennial Congress on the Use of Computers in Intensive Care Units and agreed that there was a need for a large, well-documented ventricular arrhythmia database for system development and evaluation. Many of the attendees subsequently participated in the Evaluation Group for arrhythmia detectors that developed recommendations for the characteristics of such a database, and a methodology for gathering data and characterizing the database.[1]

After a hiatus of several years, marked by unsuccessful attempts by the Evaluation Group for Arrhythmia Detectors to secure funding to support the database development, the National Heart, Lung and Blood Institute agreed, in 1977, to fund such an effort, to be performed under contract by the American Heart Association (AHA). Development of the database was coordinated by a group at Washington University (St. Louis, MO). The first portions of the AHA Database[2] were released in 1982; the database was completed in 1985. Between 1976 and 1980, a group

at the Massachusetts Institute of Technology and Boston's Beth Israel Hospital produced the Massachusetts Institute of Technology-Beth Israel Hospital (MIT-BIH) Database.[3] These two databases were the first standard sets of test material for evaluating long-term electrocardiographic monitors.

Although the Evaluation Group for Arrhythmia Detectors and the AHA arrhythmia subcommittee addressed the issues of the database, they did not address the details of how best to use an electrocardiographic database for evaluation purposes. An evaluation methodology for arrhythmia detectors was first proposed by the developers of the MIT-BIH Database.[4] Their proposal was refined for application to in-hospital arrhythmia detectors by the Arrhythmia Monitoring Subcommittee of the Association for the Advancement of Medical Instrumentation (AAMI) ECG committee, and was issued as an AAMI recommended practice in 1987.[5]

The most recent contribution to this 20-year effort has been made by the AAMI Ambulatory Monitoring Subcommittee, which will, in the near future, propose a standard for ambulatory monitors that will include recommended procedures for using existing and future electrocardiographic databases to evaluate computer-assisted Holter analysis systems.

Currently Available Electrocardiographic Databases

Several databases of electrocardiographic recordings are available for evaluating electrocardiographic analyzers. They possess several important features:

1. They contain representative signals. Wide variations in electrocardiographic characteristics among subjects severely limit the value of synthesized waveforms for testing purposes. Realistic tests of electrocardiographic analyzers require large sets of real-world signals.
2. They contain rarely observed but clinically significant signals. Although it is not particularly difficult to obtain recordings of common electrocardiographic abnormalities, often those that are most significant are rarely recorded. Both developers and evaluators of electrocardiographic analyzers need examples of such recordings.
3. They contain standard signals. System comparisons are meaningless unless performance is measured using the same test data in each case, since performance is so strongly data dependent.
4. They contain annotated signals. Typically, each QRS complex has been manually annotated by two or more cardiologists working independently. The reference annotations produced as a result serve as a gold standard against which a device's analysis can be compared quantitatively.
5. They contain digitized, computer-readable signals. It is, therefore, possible to perform a fully automated, strictly reproducible test in the digital domain if desired, allowing one to establish with certainty the effects of algorithm modifications on performance.

At present, the following long-term electrocardiographic databases are available (see the Appendix for sources): the AHA Database for Evaluation of Ventricular Arrhythmia Detectors (development set: 80 records, 35 minutes each),[2] the MIT-BIH Arrhythmia Database (48 records, 30 minutes each),[3] the European Society of Cardiology (ESC) ST-T Database (90 records, 2 hours each),[6] the Noise Stress Test (NST) Database (12 records, 30 minutes each),[7] and the Creighton University Sustained Ventricular Arrhythmia Database (35 records, 8 minutes each).[8]

Each of these databases represent a very substantial effort by many workers; in particular, the AHA, MIT-BIH, and ESC databases each required more than 5 years of sustained effort by large teams of researchers and clinicians from many institutions. Nevertheless, it should be recognized that even these databases cannot represent the entire variety of real-world ECGs observed in clinical practice.

Evaluation Protocols

Between 1984 and 1987, the AAMI sponsored the development of a protocol for the use of the AHA and MIT-BIH databases, which was published as an AAMI recommended practice.[5]

More recently, the Ambulatory ECG Subcommittee of the AAMI ECG Committee has been charged with drafting a standard for ambulatory electrocardiographic monitors. Significant portions of this standard address the issue of accuracy of the automated analysis performed by some of these devices. This standard will build upon the previously adopted evaluation protocol,[5] incorporating provisions for the use of all of the databases listed above, with extensions for assessing detection of supraventricular arrhythmias and transient ischemic ST changes. The standard will break new ground in establishing standard tests for the performance of automated electrocardiographic analyzers using these databases.

A significant constraint to be imposed on evaluators by the standard is that they must obtain annota-

tion files containing the analysis results of the device under test. Although the device itself need not produce these files, the standard will require that they be produced by an automated procedure, which must be fully disclosed. The intent of this requirement is to permit reproducible independent evaluations in which neither the proprietary data of the developers (the analysis algorithms) nor that of the evaluators (the test signals and reference annotations) need to be disclosed. By defining the interface between the developer and evaluator to be the annotation file, the responsibilities of each party are clearly defined: the developer must make certain that the device's outputs are recorded in the annotation file in the manner intended by the developer, but in the language of the standard; and the evaluator must make certain that the algorithms used to compare the device's annotation files with the reference annotation files conform to the specification of the standard.

For many existing devices, it may be difficult or impossible to obtain such annotation files without the cooperation of the manufacturers. Newly designed devices should incorporate the necessary "hooks" for producing annotation files.

Accuracy of QRS detection, which is fundamental to any automated analysis, can be tested using the AHA, MIT-BIH, and NST databases, which may also be used to test the accuracy of heart rate and heart rate variability measurement, and that of ventricular ectopic beat detection. Detection of ventricular flutter or fibrillation can be tested using the Creighton University, AHA, and MIT-BIH databases. Detection of supraventricular ectopic beats, atrial flutter, or atrial fibrillation can be tested using the MIT-BIH Database. The ESC Database is appropriate for testing the measurement of ST deviation and the detection of episodes of abnormal ST deviation. The new standard for ambulatory electrocardiographic monitors will describe evaluation protocols for each item listed above.

Software to Support Evaluations

The CD-ROM that supplies the MIT-BIH, NST, and Creighton University databases also includes a suite of programs[9] that support evaluations of automated electrocardiographic analyzers in accordance with the methods described in the new standard, as well as those in the earlier AAMI recommended practice.[5] These programs are written in C language and run under MS-DOS or UNIX. By making reference implementations of the evaluation algorithms available, much needless duplication of effort may be avoided. Circulating these programs in source form permits public inspection of the accuracy of the implementations and rapid discovery and correction of bugs, with the eventual result that evaluators of devices should not have to bear the burden of evaluating the evaluation technique itself. By using these programs for evaluation, any ambiguities in the specification of the evaluation algorithms are resolved in a consistent manner for each device tested.

The major tasks facing an evaluator are presenting the reference signals to the device under test and collecting annotation files from the device. The details of these tasks vary for each device. The remaining work required—that of comparing the device's analysis against the gold standard—can be performed automatically.

Discussion

Although the databases listed above permit standardized, quantitative, automated, and fully reproducible evaluations of analyzer performance, it is risky to extrapolate from the results of these evaluations expectations of real-world performance. Such extrapolations can be particularly error prone if the evaluation data were also used to refine an analysis algorithm, since an algorithm (perhaps unintentionally) "tuned" to its training set may not perform as well in the field.

The issue of tuning was a major consideration in the design of the AHA Database, for which separate development and test sets of equal size were produced using the same selection criteria. The intention was that independent evaluators would use the test set, which has never been released, to obtain performance measurements untainted by any possibility of tuning. Long-term electrocardiographic analyzers have never been required to produce annotation files, however, making independent evaluations tedious, expensive, and error prone. As a result, the AHA Database test set, which required 8 years to complete, has never been used.

It should also be noted that the first four of the above-mentioned databases were obtained from Holter electrocardiographic recordings, with the limited frequency response common to all such recordings. Although this recording technique is not a limiting factor in the performance of many electrocardiographic analyzers, its use in these databases may tend to favor devices that are designed to

analyze Holter recordings over those that have been designed to analyze higher-fidelity input signals.

In the context of many of the tasks performed by an electrocardiographic analyzer, dealing with noise is the major problem faced by system designers. Although measurements such as ST deviation may be obtained reliably in clean signals, the presence of noise may render them inaccurate. In some instances, it is sufficient to recognize the presence of noise and either to mark the measurements as unreliable or avoid making measurements altogether. In other cases, excluding noisy data is inappropriate (eg, given the multiple correlations among physical activity, noise, and transient ischemia, excluding noisy signals is likely to introduce sampling bias in an ischemia detector).

It is difficult to measure the effects of noise on an electrocardiographic analyzer using ordinary recordings. Even if existing databases include an adequate variety of both electrocardiographic signals and noise, the sample size is certainly too small to include all combinations of noise and electrocardiographic signals that may be encountered in clinical use. In ordinary recordings, it is difficult or impossible to separate the effects of noise from the intrinsic problems of analyzing clean signals of the same type.

The NST Database[7] circumvents these problems. By adding noise in calibrated amounts to clean signals, any combination of noise and signal types is possible. Since both the noise-corrupted signal and the clean signal can be analyzed (in separate experiments) by the same analyzer, the effects of noise on the analysis are readily separable from any other problems that may arise while analyzing the clean signals. Finally, since the test can be repeated using different amounts of noise, it is possible to characterize analyzer performance as a function of the signal-to-noise ratio.

The NST Database includes a small set of electrocardiographic records with calibrated amounts of added noise. The new AAMI ambulatory electrocardiographic standard will require that performance on these records must be reported, although no specific performance levels are required. Software provided with the NST Database can be used to generate additional records for noise stress testing.

The long history of the development of standard databases to test the performance of long-term electrocardiographic monitors and standard protocols for evaluation testifies to the importance and difficulty of the problems discussed above. Advances in analysis algorithms have driven much of the development of databases and evaluation protocols. Some of these advances have significantly reduced the effort required for database development. Inevitably, however, more ambitious analysis algorithms require new databases and testing methods. Among the subjects of current interest are late potentials (and wideband ECGs, in general), QT measurement and T wave analysis, long-term heart rate variability, paced rhythms and pacemaker function, intracardiac electrograms, and multiparameter monitoring (ECG, blood pressure, blood gases, respiration, etc.). None of these subjects is adequately addressed by existing databases or evaluation protocols.

References

1. Feldman CL: Evaluation of arrhythmia detectors. p. 21. In Ripley KL (ed): Computers in cardiology. p. 21. IEEE Computer Society Press, Los Alamitos, CA, 1974
2. Hermes RE, Geselowitz DB, Oliver GC: Development, distribution and use of the American Heart Association database for ventricular arrhythmia detector evaluation. p. 263. In Ripley KL (ed): Computers in cardiology. IEEE Computer Society Press, Los Alamitos, CA, 1980
3. Mark RG, Schluter PS, Moody GB et al: An annotated ECG database for evaluating arrhythmia detectors. p. 205. In Frontiers of engineering in health care. Proceedings of the 4th Annual Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Computer Society Press, New York, 1982
4. Schluter PS, Mark RG, Moody GB et al: Performance measures for arrhythmia detectors. p. 267. In Ripley KL (ed): Computers in cardiology. IEEE Computer Society Press, Los Alamitos, CA, 1980
5. Testing and reporting performance results of ventricular arrhythmia detection algorithms [AAMI ECAR]. Association for the Advancement of Medical Instrumentation Arlington, VA, 1987
6. Taddei A, Biagini A, Distante G et al: The European ST-T database: development, distribution, and use. p. 177. In Ripley KL (ed): Computers in cardiology. IEEE Computer Society Press, Los Alamitos, CA, 1990
7. Moody GB, Muldrow WK, Mark RG: A noise stress test for arrhythmia detectors. p. 381. In Ripley KL (ed): Computers in cardiology. IEEE Computer Society Press, Los Alamitos, CA, 1984
8. Nolle FM, Badura FK, Catlett JM et al: CREI-GARD: a new concept in computerized arrhythmia monitoring systems. p. 515. In Ripley KL (ed): Computers in cardiology. IEEE Computer Society Press, Los Alamitos, CA 1986
9. Moody GB, Mark RG: The MIT-BIH arrhythmia database on CD-ROM and software for use with it. p. 185. In Ripley KL (ed): Computers in cardiology. IEEE Computer Society Press, Los Alamitos, CA, 1990

Appendix

The following organizations are sources for the databases listed above:

1. AHA Database: ECRI, 5200 Butler Pike, Plymouth Meeting, PA 19462.
2. MIT-BIH, NST, Creighton University Databases, and software to support evaluations: MIT-BIH Database Distribution, MIT Room 20A-113, 77 Massachusetts Avenue, Cambridge, MA 02139.
3. European Society of Cardiology Database: CNR Institute of Clinical Physiology, Computer Laboratory, via Trieste, 41, 56100 Pisa, Italy.

In addition to those listed above, two other databases may be of interest. The Massachusetts General Hospital (MGH)/Marquette Foundation Waveform Database contains 250 records (375 hours in all); each record includes three electrocardiographic signals, annotated beat-by-beat, together with five simultaneously recorded hemodynamic and respiration signals. Although the MGH Database is not a long-term electrocardiographic database per se, it contains many records with significant electrocardiographic abnormalities. Finally, the Common Standards for Quantitative Electrocardiography (CSE) Database contains approximately 1,000 short records. Although the CSE database is an important standard electrocardiographic database, it is not discussed here since it was designed for evaluating diagnostic rather than long-term electrocardiographic analyzers.

4. Massachusetts General Hospital/Marquette Foundation Waveform Database: Massachusetts General Hospital, Anesthesia Bioengineering Unit, Fruit Street, Boston, MA 02114.
5. Common Standards for Quantitative Electrocardiology Database: Division of Medical Informatics, University Hospital of Leuven, Herestraat 49, 3000 Leuven, Belgium.

The evaluation software available from MIT may be used with any of these databases, except the CSE Database. Except for the AHA Database, all are available in CD-ROM format.