American National Standard

ANSI/AAMI EC57: 2012

Testing and reporting performance results of cardiac rhythm and ST segment measurement algorithms

AAMI
Advancing Safety in Medical Technology

Objectives and uses of AAMI standards and recommended practices

It is most important that the objectives and potential uses of an AAMI product standard or recommended practice are clearly understood. The objectives of AAMI's technical development program derive from AAMI's overall mission: the advancement of medical instrumentation. Essential to such advancement are (1) a continued increase in the safe and effective application of current technologies to patient care, and (2) the encouragement of new technologies. It is AAMI's view that standards and recommended practices can contribute significantly to the advancement of medical instrumentation, provided that they are drafted with attention to these objectives and provided that arbitrary and restrictive uses are avoided.

A voluntary *standard* for a *medical device* recommends to the manufacturer the information that should be provided with or on the product, basic safety and performance criteria that should be considered in qualifying the device for clinical use, and the measurement techniques that can be used to determine whether the device conforms with the safety and performance criteria and/or to compare the performance characteristics of different products. Some standards emphasize the information that should be provided with the device, including performance characteristics, instructions for use, warnings and precautions, and other data considered important in ensuring the safe and effective use of the device in the clinical environment. Recommending the disclosure of performance characteristics often necessitates the development of specialized test methods to facilitate uniformity in reporting; reaching consensus on these tests can represent a considerable part of committee work. When a drafting committee determines that clinical concerns warrant the establishment of *minimum* safety and performance criteria, referee tests must be provided and the reasons for establishing the criteria must be documented in the rationale.

A *recommended practice* provides guidelines for the use, care, and/or processing of a medical device or system. A recommended practice does not address device performance *per se*, but rather procedures and practices that will help ensure that a device is used safely and effectively and that its performance will be maintained.

Although a device standard is primarily directed to the manufacturer, it may also be of value to the potential purchaser or user of the device as a frame of reference for device evaluation. Similarly, even though a recommended practice is usually oriented towards healthcare professionals, it may be useful to the manufacturer in better understanding the environment in which a medical device will be used. Also, some recommended practices, while not addressing device performance criteria, provide guidelines to industrial personnel on such subjects as sterilization processing, methods of collecting data to establish safety and efficacy, human engineering, and other processing or evaluation techniques; such guidelines may be useful to health care professionals in understanding industrial practices.

In determining whether an AAMI standard or recommended practice is relevant to the specific needs of a potential user of the document, several important concepts must be recognized:

All AAMI standards and recommended practices are *voluntary* (unless, of course, they are adopted by government regulatory or procurement authorities). The application of a standard or recommended practice is solely within the discretion and professional judgment of the user of the document.

Each AAMI standard or recommended practice reflects the collective expertise of a committee of health care professionals and industrial representatives, whose work has been reviewed nationally (and sometimes internationally). As such, the consensus recommendations embodied in a standard or recommended practice are intended to respond to clinical needs and, ultimately, to help ensure patient safety. A standard or recommended practice is limited, however, in the sense that it responds generally to perceived risks and conditions that may not always be relevant to specific situations. A standard or recommended practice is an important *reference* in responsible decision-making, but it should never *replace* responsible decision-making.

Despite periodic review and revision (at least once every five years), a standard or recommended practice is necessarily a static document applied to a dynamic technology. Therefore, a standards user must carefully review the reasons why the document was initially developed and the specific rationale for each of its provisions. This review will reveal whether the document remains relevant to the specific needs of the user.

Particular care should be taken in applying a product standard to existing devices and equipment, and in applying a recommended practice to current procedures and practices. While observed or potential risks with existing equipment typically form the basis for the safety and performance criteria defined in a standard, professional judgment must be used in applying these criteria to existing equipment. No single source of information will serve to identify a particular product as "unsafe". A voluntary standard can be used as one resource, but the ultimate decision as to product safety and efficacy must take into account the specifics of its utilization and, of course, cost-benefit considerations. Similarly, a recommended practice should be analyzed in the context of the specific needs and resources of the individual institution or firm. Again, the rationale accompanying each AAMI standard and recommended practice is an excellent guide to the reasoning and data underlying its provision.

In summary, a standard or recommended practice is truly useful only when it is used in conjunction with other sources of information and policy guidance and in the context of professional experience and judgment.

INTERPRETATIONS OF AAMI STANDARDS AND RECOMMENDED PRACTICES

Requests for interpretations of AAMI standards and recommended practices must be made in writing, to the AAMI Vice President, Standards Policy and Programs. An official interpretation must be approved by letter ballot of the originating committee and subsequently reviewed and approved by the AAMI Standards Board. The interpretation will become official and representation of the Association only upon exhaustion of any appeals and upon publication of notice of interpretation in the "Standards Monitor" section of the *AAMI News*. The Association for the Advancement of Medical Instrumentation disclaims responsibility for any characterization or explanation of a standard or recommended practice which has not been developed and communicated in accordance with this procedure and which is not published, by appropriate notice, as an *official interpretation* in the *AAMI News*.

American National Standard

ANSI/AAMI EC57:2012
(Revision of ANSI/AAMI EC57:1998/(R)2008)

Testing and reporting performance results of cardiac rhythm and ST segment measurement algorithms

Developed by
Association for the Advancement of Medical Instrumentation

Approved 18 December 2012 by
American National Standards Institute, Inc.

Abstract: This recommended practice establishes a method for testing and reporting the performance of algorithms used to detect cardiac rhythm disturbances, including the ST segment.

Keywords: arrhythmia database, arrhythmia monitoring, ST segments, heart rate variability

AAMI Recommended Practice

This Association for the Advancement of Medical Instrumentation (AAMI) recommended practice implies a consensus of those substantially concerned with its scope and provisions. The existence of an AAMI recommended practice does not in any respect preclude anyone, whether they have approved the recommended practice or not, from manufacturing, marketing, purchasing, or using products, processes, or procedures not conforming to the recommended practice. AAMI recommended practices are subject to periodic review, and users are cautioned to obtain the latest editions.

CAUTION NOTICE: This AAMI recommended practice may be revised or withdrawn at any time. AAMI procedures require that action be taken to reaffirm, revise, or withdraw this recommended practice no later than five years from the date of publication. Interested parties can obtain current information on all AAMI documents by calling or writing AAMI.

All AAMI standards, recommended practices, technical information reports, and other types of technical documents developed by AAMI are *voluntary*, and their application is solely within the discretion and professional judgment of the user of the document. Occasionally, voluntary technical documents are adopted by government regulatory agencies or procurement authorities, in which case the adopting agency is responsible for enforcement of its rules and regulations.

Published by

Association for the Advancement of Medical Instrumentation
4301 N. Fairfax Dr., Suite 301
Arlington, VA 22203-1633
www.aami.org

ISBN 1-57020-478-0

Contents

Figures

Glossary of equivalent standards

International Standards adopted in the United States may include normative references to other International Standards. AAMI maintains a current list of each International Standard that has been adopted by AAMI (and ANSI). Available on the AAMI website at the address below, this list gives the corresponding U.S. designation and level of equivalency to the International Standard.

www.aami.org/publications/standards/glossary.pdf

Committee representation

Association for the Advancement of Medical Instrumentation

Electrocardiograph (ECG) Committee

This recommended practice was developed by the ECG Committee of the Association for the Advancement of Medical Instrumentation. Committee approval of the standard does not necessarily imply that all committee members voted for its approval.

At the time this document was published, the **AAMI Electrocardiograph Committee** had the following members:

Cochairs:
Richard A. Sunderland
Ahmet Turkmen
Brian J. Young

Members:
Robert William Bain, CBET, Baltimore Medical Engineers & Technician Society
Robert E. Bruce
Scott Coggins, Covidien
Prakash C. Deedwania, MD, The VA Medical Center
Laura Dhatt, Physio-Control
Sreeram Dhurjaty, Dhurjaty Electronics Consulting LLC
Richard Diefes, ECRI Institute
Greg Downs, Spacelabs Medical Inc.
Arthur R. Eddy, Jr.
James J. Greco, Medapprove Inc
Richard Gregg, Philips Electronics North America
Janice M. Jenkins, PhD, University of Michigan College of Engineering
Carolyn Lall, Draeger Medical Systems Inc
Dongping Lin, PhD
Walter G. Lloyd, Childrens Hospital Boston
Peter W. Macfarlane, Royal Infirmary
Luis A. Melendez, Partners Healthcare
George Moody, Massachusetts Institute of Technology
Cadathur Rajagopalan, PhD SMIEEE, Mindray DS USA Inc
Linda Ricci, FDA/CDRH
Johann-Jakob Schmid, Schiller AG
Jonathan Steinberg, MD, St Lukes Roosevelt Hospital Center
Richard A. Sunderland, Welch Allyn, Inc.
Ahmet Turkmen, BS MS PhD, University of Wisconsin-Stout
Jeffrey Wiser, 3M Healthcare
Ted Yantsides, Conmed Corp
Brian J. Young, GE Healthcare

Alternates:
Mark J. Callahan, Covidien
Kejian Chen, 3M Healthcare
Yu Chen, PhD, Draeger Medical Systems Inc
Steve Duke, Physio-Control
Charles S. Ho, PhD, FDA/CDRH
Richard Richardson, GE Healthcare
Serkan Sezer, Schiller AG
Donald Stewart, Spacelabs Medical Inc.
Anna Varlese, Conmed Corp
John J. Wang, Philips Electronics North America
Yinqi Zhang, Spacelabs Medical Inc

NOTE—Participation by federal agency representatives in the development of this standard does not constitute endorsement by the federal government or any of its agencies.

Foreword

This recommended practice was developed by the Arrhythmia Monitoring Working Group of the AAMI Electrocardiograph (ECG) Committee. It reflects the conscientious efforts of health care professionals, in cooperation with manufacturers of arrhythmia monitoring devices, to develop recommendations for testing and reporting performance results of algorithms for cardiac arrhythmia detection and ST segment measurement.

The first edition of this document was issued in 1987 under the title *Recommended practice for testing and reporting performance results of ventricular arrhythmia detection algorithms* (AAMI ECAR:1987). The document was developed to assist in the comparison of ventricular arrhythmia detection algorithm performance through the promulgation of a generally accepted method for testing and reporting such performance. Major changes were incorporated into this revision, retitled *Testing and reporting performance results of cardiac rhythm and ST segment measurement algorithms* (ANSI/AAMI EC57:1998), including updated references to databases that have become available since 1987 and the addition of mechanisms for testing and reporting ST measurement and heart-rate variability performance along with supraventricular ectopic performance statistics. As with cardiac ventricular rhythm measurements, these additional parameters are intended to benefit users who are comparing algorithm performance. This current revision makes minor changes to the 1998 standard and updates the information for the databases.

It is not intended that these recommendations be construed as universally applicable to all circumstances. It is also recognized that these recommendations may not be achievable in all situations.

This recommended practice must be reviewed and updated periodically to assimilate progressive technological developments. The concepts incorporated in this recommended practice should not be considered inflexible or static.

As used within the context of this recommended practice, “shall” indicates requirements strictly to be followed to conform to the recommended practice; “should” indicates that among several possibilities one is recommended as particularly suitable, without mentioning or excluding others, or that a certain course of action is preferred but not necessarily required, or that (in the negative form) a certain possibility or course of action is discouraged but not prohibited; “may” is used to indicate that a course of action is permissible within the limits of the recommended practice; and “can” is used as a statement of possibility and capability. “Must” is used only to describe “unavoidable” situations.

Suggestions for improving this recommended practice are invited. Comments and suggested revisions should be sent to Standards Dept., AAMI, 4301 N. Fairfax Dr., Suite 301, Arlington, VA 22203-1633. Comments may also be e-mailed to: standards@aami.org.

NOTE—This foreword is not a part of the AAMI Recommended Practice, *Testing and reporting performance results of cardiac rhythm and ST segment measurement algorithms* (ANSI/AAMI EC57:2012), but it does provide important information about the development and intended use of the document.

Testing and reporting performance results of cardiac rhythm and ST segment measurement algorithms

1 Scope

1.1 General

The availability of annotated arrhythmia and ST databases has permitted different automated arrhythmia detection algorithms to be tested on the same data. This recommended practice provides a protocol for a reproducible test with realistic clinical requirements, and emphasizes record-by-record presentation of results that reflect an algorithm's ability to detect events of clinical significance. Beat-by-beat comparisons are used to measure performance in QRS (see 2.7), ventricular ectopic beat (VEB), and supraventricular ectopic beat (SVEB) detection. Run-by-run comparisons are used to measure an algorithm's ability to detect consecutive VEBs and SVEBs. Detection of ventricular flutter, atrial flutter, ventricular fibrillation, and atrial fibrillation are addressed. The evaluation of heart-rate variability measurement algorithms and ST segment measurement algorithms are also examined.

Although this document seeks to establish clinically relevant measures of performance for the comparison of algorithms, it must be recognized that certain clinical concerns cannot be addressed within the context of this recommended practice. Available databases do not yet contain a representative sample of nonventricular arrhythmias, paced patients or artifacts typical of a very significant portion of ECG signals originating in the clinical setting. In addition, these databases have a limited bandwidth and should be used with caution when testing algorithms designed for full ECG diagnostic bandwidth devices. Therefore, the clinical implications of a test are necessarily limited by the size, scope, and characteristics of the databases used for testing. Performance measures derived from such testing should be regarded as uncertain indicators of performance in clinical settings.

This recommended practice has been developed for testing algorithms, not entire systems. It is not a performance standard, but rather a set of recommendations for testing cardiac rhythm and ST measurement and reporting the results of those tests. The intent of this recommended practice is that automated testing methods be reproducible.

1.2 Inclusions

This recommended practice applies to algorithms implemented in devices or systems that use automated methods to analyze the ECG.

This document applies both to human-operated, stand-alone devices that use automated methods to analyze the recorded ECG, and to so-called real-time event recorders that use automated methods to select abnormal events for recording.

1.3 Exclusions

Testing methodologies other than beat-by beat techniques, specified rhythm analysis, and ST segment analysis are outside the scope of this document. The evaluation of systems that rely on intensive interaction by a skilled user is also outside the scope of this document. However, if beat-by-beat evaluations are performed, the results of such testing should conform to this recommended practice.

2 Definitions of abbreviations

NOTE—Definitions for beat labels (N, V, F, S, Q, U, X, O) are provided in 4.2.

For the purposes of this standard, the following abbreviations apply.

2.1 **AF:** Atrial fibrillation or atrial flutter.

2.2 **BW:** Data record identified from the NST (Noise Stress Test) database.

2.3 **DB:** Database.

2.4 **EM:** Data record identified from the NST (Noise Stress Test) database.

2.5 HRV: Heart rate variability.

2.6 MA: Data record identified from the NST (Noise Stress Test) database.

2.7 QRS: The waveform presented in an ECG during ventricular depolarization.

2.8 RMS: Root-mean squared.

2.9 RRV: R-to-R variability.

2.10 SVEB: Supraventricular ectopic beat.

2.11 SVTA: Supraventricular tachycardia.

2.12 ST: Segment of the ECG between the end of the QRS complex and the start of the T-wave.

2.13 VEB: Ventricular ectopic beat.

2.14 VF: Ventricular fibrillation or ventricular flutter.

3 Algorithm testing

This section describes what constitutes a complete test of an algorithm. The term "report" refers to the evaluation procedure described in this section and not to the clinical report that the physician receives.

3.1 Databases

3.1.1 General description of available databases

At the time this document was developed, five databases were available for evaluation of cardiac arrhythmia ST algorithms:

AHA: The American Heart Association Database for Evaluation of Ventricular Arrhythmia Detectors (80 records of 35 minutes each), ECRI, 5200 Butler Pike, Plymouth Meeting, PA 19462, USA;

MIT–BIH: The Massachusetts Institute of Technology–Beth Israel Hospital Arrhythmia Database (48 records of 30 minutes each), http://physionet.org/physiobank/database/mitdb/;

ESC: The European Society of Cardiology ST-T Database (90 records of 2 hours each), http://physionet.org/physiobank/database/edb/;

NST: The Noise Stress Test Database (12 ECG records of 30 minutes each plus 3 records of noise only—supplied with the MIT–BIH database), http://physionet.org/physiobank/database/nstdb/; and

CU: The Creighton University Sustained Ventricular Arrhythmia Database (35 records of 8 minutes each—supplied with the MIT–BIH database with incomplete annotations), http://physionet.org/physiobank/database/cudb/.

Sources for these databases include the following:

ECRI, 5200 Butler Pike, Plymouth Meeting, PA 19462, USA (AHA database);

MIT–BIH Database Distribution, MIT Room E25-505, Cambridge, MA 02139, USA (MIT–BIH, NST, CU databases and the ESC database inside North America—Internet site: http://ecg.mit.edu); and

CNR Institute of Clinical Physiology, Computer Laboratory, via Trieste, 41 56100 Pisa, Italy (ESC database outside North America).

The first four of these databases (AHA, MIT–BIH, ESC, and NST) consist of digitized excerpts of two-channel Holter-type recordings, with each beat labeled. This set of annotation files, in which each beat has been identified by expert cardiologist-annotators, are referred to as "reference" annotations. The CU database contains digitized single-channel ECG recordings with rhythm changes labeled.

Database elements have been referred to as tapes and records. For the purpose of this document, the term "tapes" refers only to physical taped recordings of ECGs. Database elements are referred to as "records."

This list of standard databases is not intended to exclude others that may become available in the future. It is, however, a list of those that were both adequate and available at the time of this document's publication.

Databases should be:

— available to the public;

— fully described (standard digital format);

— clearly identifiable by name, version, date, etc.; and

— annexed with utilities and instructions for use.

If any records from a given database are used to fulfill the requirements of 3.5, device performance shall be tested and reported on a record-by-record basis for all records from that database except as excluded by 3.1.2. The first 5 minutes of each record are designated as a learning period. The remainder of each record is the test period. Device performance is measured only during the test period of each record; the entire test period shall be used for this purpose, except as noted in 3.1.2.

3.1.2 Records to be excluded during testing

Of the 80 available records in the AHA database, two are recorded from patients with pacemakers. Of the 48 records in the MIT–BIH database, four are from patients with pacemakers. In these databases, records with paced beats do not retain sufficient signal quality for reliable processing by systems that use special analog circuits for pace artifact detection or enhancement. Such systems shall exclude these six records containing paced beats from the reporting requirements. Performance on these records shall be reported for devices that are intended to analyze paced analog ECG recordings made without pacer artifact detection or enhancement, but aggregate performance statistics shall exclude these records in all cases. This exclusion of records with paced beats applies to arrhythmia algorithms as well as to ST-segment measurement algorithms.

The NST database contains three records (BW, EM, and MA) that are noise recordings only and are not intended for use in standard tests. The remaining 12 records are those on which device performance shall be tested and reported.

Segments of data in which ventricular flutter or fibrillation (VF) is present are excluded from beat-by-beat comparisons (for QRS and VEB detection) only. Well-defined QRS complexes necessary for a beat-by-beat comparison are not present during these segments, which are marked by rhythm labels in the database annotation files. These segments are included, however, in the tests of consecutive VEB detection and VF detection. Other segments of these records (i.e., those that do contain labeled beats) shall be included in the beat-by-beat comparisons.

3.2 Testing requirements

3.2.1 The accuracy of QRS detection shall be tested using the AHA DB, the MIT–BIH DB, and the NST DB at a minimum.

3.2.2 The accuracy of heart rate measurements shall be tested using the AHA DB, the MIT–BIH DB, and the NST DB. If the algorithm is claimed to measure heart rate variability (HRV) or RR interval variability (RRV), its ability to do so shall be demonstrated using the MIT–BIH databases.

3.2.3 The accuracy of VEB detection shall be tested using the AHA DB, the MIT–BIH DB, and the NST DB at a minimum.

3.2.4 If the device is claimed to detect ventricular flutter or fibrillation (VF), its ability to do so shall be tested using the CU DB, the AHA DB, and the MIT–BIH DB at a minimum.

3.2.5 If the device is claimed to detect supraventricular ectopic beats, or atrial flutter or fibrillation (AF), its ability to do so shall be tested using the MIT–BIH DB and the NST DB at a minimum.

If the device is claimed to measure ST segment deviations or to detect ST segment changes, its ability to do so shall be tested using the ESC DB at a minimum, unless the characteristics of the database conflict with the algorithm under test.

3.3 Test environment

Algorithm testing using standardized digital databases occurs, by definition, outside the context of the complete monitoring device's clinical setting. Yet, a correlation between algorithm performance and the device's actual clinical performance must be ensured for the results to be meaningful.

To conduct an evaluation that accurately reflects the capabilities of the algorithm as implemented in a monitoring device, it is preferable to perform the test using hardware comparable to the monitoring device although it is

recognized that the nature of the algorithm testing process might require modifications of the hardware or software. Additionally, signals should be presented to the algorithm in a method comparable to the method employed in clinical settings. The computational environment used to perform algorithm testing shall be disclosed.

When algorithm evaluations are conducted under conditions or constraints grossly different from those encountered by the monitoring device in an actual clinical setting, the algorithm results might not represent the true performance of the device. Actual devices can have limited processor speed, computational precision, filtering, and so on. Testing or analysis shall be performed indicating that the algorithm performance in an actual monitoring device can reasonably be expected to correlate with performance in the simulated test environment. This validation shall be disclosed.

Of special concern are monitoring devices intended to monitor more than one patient simultaneously. The algorithm for each patient may be identical and may be tested in isolation to determine the capabilities of the algorithm. In the actual monitoring device, the computing resource provided to each patient is dependent on the computing resources required by all the others. Therefore, validation of algorithm performance in the presence of other patient inputs shall be disclosed.

One method of multipatient monitor performance validation is to provide all patient inputs of the device with the same test waveform. Algorithm performance for all patient inputs shall be reported; the tester is not allowed to choose the best-performing patient input. In the event that a system cannot simultaneously process the same data on all patient inputs, this fact shall be reported, and the number of patient inputs that can be simultaneously processed shall be disclosed.

3.4 Multiple-lead analysis

Any algorithm that analyzes a multiple number of leads simultaneously shall be permitted to report the results as a single test. For any database that has more leads available than can be simultaneously analyzed, the actual combination of channels used shall be disclosed. For any system that can analyze more channels than are available in the database, the disclosure shall state how the data were entered. At no time during the processing of the entire database is the operator allowed to change the combination of leads used. Results shall be reported on a record-by-record basis.

3.5 Requirements for the evaluation report

3.5.1 Required statistics

For each record, the statistics below shall be reported as required in 3.5.2 and 3.5.3. Aggregate statistics based on the record-by-record reports summarizing the performance of the algorithm under test for each of the databases employed shall be reported as required. Formal definitions of the statistics are provided in the annex as noted.

The following symbols and abbreviations are used in the following tables:

R = required reporting of this statistic from this database

O = optional reporting of this statistic from this database

\- = no reporting of this statistic required from this database

✓ = aggregate statistic required

3.5.2 Requirements for all arrhythmia algorithms

The requirements for all algorithms are given in Table 1.

Table 1—Requirements for all arrhythmia algorithms

Record-by-record statistics required for each record	Formal definition	Gross statistic	Average statistic	AHA DB	MIT–BIH DB	NST DB	CU DB	ESC DB
QRS sensitivity	A.3.5.2	✓	✓	R	R	R	-	O
QRS positive predictivity	A.3.5.2	✓	✓	R	R	R	-	O
VEB sensitivity	A.3.5.2	✓	✓	R	R	R	-	O
VEB positive predictivity	A.3.5.2	✓	✓	R	R	R	-	O
VEB false positive rate	A.3.5.2	✓	✓	R	R	R	-	O
RMS heart rate error	A.3.5.3	✓	✓	R	R	R	-	O
Ventricular couplet sensitivity	A.3.5.3	✓	✓	R	R	-	-	-
Ventricular couplet positive predictivity	A.3.5.3	✓	✓	R	R	-	-	-
Ventricular short run sensitivity	A.3.5.3	✓	✓	R	R	-	-	-
Ventricular short run positive predictivity	A.3.5.3	✓	✓	R	R	-	-	-
Ventricular long run sensitivity	A.3.5.3	✓	✓	R	R	-	-	-
Ventricular long run positive predictivity	A.3.5.3	✓	✓	R	R	-	-	-
% Beats missed during shutdown	A.3.5.2	✓	✓	R	R	R	-	O
% N missed during shutdown	A.3.5.2	✓	✓	R	R	R	-	O
% V missed during shutdown	A.3.5.2	✓	✓	R	R	R	-	O
% F missed during shutdown	A.3.5.2	✓	✓	R	R	R	-	O
Total shutdown time	A.3.5.2	✓	✓	R	R	R	-	O

3.5.3. Requirements for algorithms with optional capabilities

Requirements for algorithms with optional capabilities are given in Table 2.

Table 2—Requirements for algorithms with optional capabilities

Record-by-record statistics required for each record IF such capability claimed	Formal definition	Gross statistic	Average statistic	AHA DB	MIT–BIH DB	NST DB	CU DB	ESC DB
HRV or RRV result	A.3.5.3	-	-	-	R	-	-	-
VF episode sensitivity	A.3.5.3	✓	O	R	R	-	R	-
VF episode positive predictivity	A.3.5.3	✓	O	R	R	-	R	-
VF duration sensitivity	A.3.5.3	✓	O	R	R	-	R	-
VF duration positive predictivity	A.3.5.3	✓	O	R	R	-	R	-
VF false positive report	A.3.5.3	-	-	R	R	-	R	-
VF time to detection	A.3.5.3	-	✓	R	R	-	R	-
SVEB sensitivity	A.3.5.2	✓	✓	-	R	-	-	-
SVEB positive predictivity	A.3.5.2	✓	✓	-	R	-	-	-
SVEB false positive rate	A.3.5.2	✓	✓	-	R	-	-	-
Supraventricular couplet sensitivity	A.3.5.3	✓	✓	-	R	-	-	-
Supraventricular couplet positive predictivity	A.3.5.3	✓	✓	-	R	-	-	-
Supraventricular short run sensitivity	A.3.5.3	✓	✓	-	R	-	-	-
Supraventricular short run positive predictivity	A.3.5.3	✓	✓	-	R	-	-	-
Supraventricular long run sensitivity	A.3.5.3	✓	✓	-	R	-	-	-
Supraventricular long run positive predictivity	A.3.5.3	✓	✓	-	R	-	-	-
AF episode sensitivity	A.3.5.3	✓	-	-	R	R	-	-
AF episode positive predictivity	A.3.5.3	✓	-	-	R	R	-	-
AF duration sensitivity	A.3.5.3	✓	-	-	R	R	-	-
AF duration positive predictivity	A.3.5.3	✓	-	-	R	R	-	-
AF false positive report	A.3.5.3	-	-	-	O	O	-	-
AF time to detection	A.3.5.3	-	✓	-	O	O	-	-
ST mean error; all measurements	A.3.5.3	✓	✓	-	-	-	-	R
ST standard deviation; all measurements	A.3.5.3	✓	✓	-	-	-	-	R
ST mean error; - 200 µV to + 200 µV	A.3.5.3	✓	✓	-	-	-	-	R
ST standard deviation; - 200 µV to + 200 µV	A.3.5.3	✓	✓	-	-	-	-	R
ST slope mean error; - 2 mV/Sec to + 2 mV/Sec	A.3.5.3	✓	✓	-	-	-	-	R

Table 2—Requirements for algorithms with optional capabilities (continued)

Record-by-record statistics required for each record IF such capability claimed	Formal definition	Gross statistic	Average statistic	AHA DB	MIT–BIH DB	NST DB	CU DB	ESC DB
ST Slope standard deviation; -2 mV/Sec to + 2 mV/Sec	A.3.5.3	✓	✓	-	-	-	-	R
ST slope mean error; all measurements	A.3.5.3	✓	✓	-	-	-	-	R
ST slope standard deviation; all measurements	A.3.5.3	✓	✓	-	-	-	-	R
ST episode sensitivity	A.3.5.3	✓	-	-	-	-	-	R
ST episode positive predictivity	A.3.5.3	✓	-	-	-	-	-	R
ST duration sensitivity	A.3.5.3	✓		-	-	-	-	R
ST duration positive predictivity	A.3.5.3	✓		-	-	-	-	R

NOTES

1. RMS measurement errors and mean reference measurements shall be reported separately for each type of heart rate measurement made by the device under test.

2. Results shall be reported separately for each type of HRV and/or RRV measurement made by the device under test. The definitions of each index and alternative units (i.e. ms or ms^2 or μV) shall be disclosed.

3. For devices claiming ST measurement capabilities, the time and voltage resolution of ST segment amplitude and/or slope measurements, the number of leads analyzed, the filtering employed, and the treatment of ectopic and noisy beats by the ST analysis algorithm shall be disclosed.

3.6 Simulated test patterns

Some aspects of algorithm performance are best evaluated with simple deterministic test patterns. For these patterns, the proper algorithm result can be predicted. This was recommended by the ESC/NASPE special report.*

If the device is claimed to measure heart-rate variability (HRV) or RR interval variability (RRV), its ability to do so shall be tested using special simulated ECG patterns with predictable variability. One pattern (test pattern 1; see 4.3.3.3) establishes a noise floor measurement and gives guidance on how sensitive the system can be for very low variability patients. Other patterns (test patterns 2–5; see 4.3.3.3) establish accuracy of calculation and a minimum upper range for high variability patients.

4 Automated analysis

The requirement that evaluations be reproducible implies that evaluations must be performed without human intervention.

4.1 Use of standard databases

Each record shall be supplied to the algorithm continuously from the beginning to the end (i.e., without rewinding or "fast forwarding"). This requirement applies only to the manner in which the evaluator presents ECG samples to the device under test and in no way is to be construed as a restriction on the manner in which the device performs its analysis.

If the digitized ECG signals from the database records are preprocessed in any way before they are presented as input to the device under test, the preprocessing shall be disclosed in sufficient detail to permit a third party to reproduce the test. Preprocessing includes, but is not limited to the following:

— resampling (i.e., conversion to a sampling rate different from that used in the standard database files);

— reformatting (i.e., conversion of byte order, sample precision, or numeric coding);

*_Heart Rate Variability, Standards of Measurement, Physiological Interpretation, and Clinical Use_, by the European Society of Cardiology and the North American Society of Pacing and Electrophysiology, _Circulation_, 1996; 93:1043-1065. See especially page 1061.

— rescaling (altering the signal amplitude, i.e., changing the gain);

— filtering performed by software or hardware not employed in the normal operating mode of the device under test; and

— conversion from digital to analog signals.

If the evaluation of the device under test is performed using signals converted into analog form and supplied to the normal analog inputs of the device, the device's automatic gain control (AGC) will be allowed to adjust the gain automatically. If the evaluation is performed using digital data and the AGC is not digital but part of the analog front end of the device, the device may simulate its AGC capabilities by an alternative method. This alternative method allows the "test mode" that generates the "test annotations" to emit an announcement that a "gain adjustment" would be required prior to proceeding with analyzing the ECG for each patient record. This announcement should instruct the evaluator to adjust the gain of the ECG for one or all of the ECG channels. The evaluator shall then run the "xform"** (or equivalent) program to adjust the ECG's gain based on the instructions provided by the program. (If another program is used, then this shall be disclosed and made available.) This process shall be repeated until "no gain change" is announced; the device under test shall then automatically proceed with the ECG analysis.

Beat-by-beat comparisons, following the protocol described in 4.3, shall be used to derive QRS Sensitivity (QRS Se), QRS positive predictivity (QRS +P), VEB Sensitivity (VEB Se), VEB positive predictivity (VEB +P), VEB false positive rate (VEB FPR), supraventricular ectopic beat false positive rate (SVEB FPR), and, where applicable, supraventricular ectopic beat sensitivity (SVEB Se) and supraventricular ectopic beat positive predictivity (SVEB +P). Run-by-run comparisons, following the protocol described in 4.4, shall be used to derive VE couplet Se and +P, VE short run Se and +P, VE long run Se and +P, and, where applicable, SVE couplet Se and +P, SVE short run Se and +P, and SVE long run Se and +P. The protocol described in 4.5 shall be used to derive VF and AF episode Se and +P, and VF and AF duration Se and +P, where applicable. ST comparisons, following the protocol described in 4.6, shall be used, where applicable, to derive the data necessary to satisfy the reporting requirements of 3.5.3.

4.2 Use of annotation files

The test protocols described in 4.3 through 4.6 require that, for each record, the output of the device has been recorded in an annotation file (the "test annotation file"), in the same format as the reference annotation file for that record. The device need not produce this file directly. Any automated procedure for doing so is acceptable as long as it is disclosed. The programs "bxb," "rxr," "epic," and "mxm"* (either the versions supplied on the MIT–BIH Arrhythmia Database CD-ROM or any later versions released by MIT) or equivalent should be used to perform the comparisons between the test annotation files and the reference annotation files as described in 4.3 through 4.6. The reference annotation files distributed with the databases and used as input to these programs may not be altered in any way, with the exception that (where applicable) corrected reference annotation files obtained from the database suppliers may be substituted for those originally distributed with the databases. An exception to this is that location data will be altered by the "xform" program when resampling. The source of the annotation shall be disclosed.

Within annotation files, beat labels (N, S, V, F, and Q), rhythm labels (], [), and other labels (U, X, and O) are defined as follows:

N = any beat that does not fall into the S, V, F, or Q categories described below (a normal beat or a bundle branch block beat)

S = a supraventricular ectopic beat (SVEB): an atrial or nodal (junctional) premature or escape beat, or an aberrated atrial premature beat

V = a ventricular ectopic beat (VEB): a ventricular premature beat, an R-on-T ventricular premature beat, or a ventricular escape beat

F = a fusion of a ventricular and a normal beat

Q = a paced beat, a fusion of a paced and a normal beat, or a beat that cannot be classified

Other labels are needed to facilitate the beat-by-beat comparison process defined in 4.3:

U = a label that marks a segment of unreadable data

** "xform" is a utility program that can be used to transform the sample rate and amplitude of a database record. This program may be downloaded freely from http://physionet.org).

* The programs "bxb," "rxr," "epic" and "mxm" and their use are described in the WFDB Applications Guide (http://physionet.org/physiotools/wag/). These programs may be downloaded freely from PhysioNet (http://physionet.org).

U labels appear in the databases where beats cannot be located because of excessive noise or loss of signal. In the MIT–BIH and ESC databases, a pair of U labels mark the beginning and end of each unreadable segment. In the AHA database, a single U label marks the (approximate) center of each unreadable segment, which is assumed for testing purposes to begin 150 milliseconds (ms) after the previous beat label and to end 150 ms before the following beat label. Devices may also generate U labels to mark segments during which that device's analysis is suspended (shut down) for any reason (e.g., excessive noise, signal loss). Beat labels are never paired with U labels during beat-by-beat comparisons.

Extra beats are sometimes detected (false positive QRSs), and reference beats are sometimes missed (false negative QRSs). In order to perform beat-by-beat comparisons, pseudo-beat labels are added to those in the reference and test annotation files to preserve a one-to-one correspondence between beat labels. They represent the absence of a beat label. There are two types:

X = a pseudobeat label generated during a segment marked as unreadable

O = a pseudobeat label generated at any other time

In beat-by-beat comparisons, all beat labels are paired up. If either the reference or the test annotation file contains an extra beat label that has no match in the other file, the appropriate O or X label is paired with the extra label. This corresponds to a QRS detection error—either a false detection (if the extra label is in the test annotation file) or a missed beat (if it is in the reference annotation file). All such beat label pairs are counted, including those that involve O or X labels. O and X labels are not used in run-by-run comparisons (see 4.4), or for VF, AF, or ST comparisons (see 4.5 and 4.6), as it is not necessary in these instances to pair individual beat labels.

Rhythm labels mark segments of ventricular flutter or fibrillation (VF) in the AHA and MIT–BIH databases:

[= beginning of VF

] = end of VF

Beat labeling is discontinued between "[" and "]" labels. VF segments are excluded from beat-by-beat comparisons. Additional rhythm labels mark changes in rhythm in the MIT–BIH and ESC databases. Those that mark segments of atrial flutter or fibrillation (AF; see the documentation that accompanies each database) are used for evaluation of AF detection; others are ignored. Beat labels are never paired with rhythm labels.

4.3 Beat-by-beat comparison

4.3.1 General description

During a beat-by-beat comparison, reference beat labels and device beat labels are matched by pairs. To be considered a match, the absolute value of the difference between the device's estimate of the time of occurrence of a beat and the time as recorded in the reference annotation file shall not exceed 150 ms. If matching does not occur within this window, the candidate beat is considered to have been missed or to be an extra detection. The end product of a beat-by-beat comparison is a matrix in which each element is a correct count of the number of beat label pairs of the appropriate type.

Table 3—Beat label classifications

		Algorithm label						
		n	**s**	**v**	**f**	**q**	**o**	**x**
Reference	N	Nn	Ns	Nv	Nf	Nq	No	Nx
Label	S	Sn	Ss	Sv	Sf	Sq	So	Sx
	V	Vn	Vs	Vv	Vf	Vq	Vo	Vx
	F	Fn	Fs	Fv	Ff	Fq	Fo	Fx
	Q	Qn	Qs	Qv	Qf	Qq	Qo	Qx
	O	On	Os	Ov	Of	Oq		
	X	Xn	Xs	Xv	Xf	Xq		

4.3.2 Method for beat-by-beat comparison

To perform the beat-by-beat comparison, follow the steps given below.

a) Set the variable T to the time of the first reference beat label after the end of the learning period, and set the variable t to the time of the first test beat label after the end of the learning period. Set all elements of the matrix to zero.

   If T is within 150 ms of the beginning of the test period, it is possible that a matching test beat label may be placed before the beginning of the test period. If this occurs, it is counted as a match (t is set to the time of the matching test beat label before going on to step b). On the other hand, if t is within 150 ms of the beginning of the test period and there is no matching reference beat label after the beginning of the test period, the test annotation at t is not counted (t is set to the time of the next test beat label before going on to step b).

b) One of the following cases must apply:

   1) If t precedes T, set t' to the time of the next test beat label (or to a time beyond the end of the record if there are no more test beat labels). There are now two possibilities:

      i) If T is closer to t than to t' and t is within 150 ms (the match window) of T, the beat labels at T and t are paired. The variable T is reset to the time of the next reference beat label.

      ii) Otherwise, the test beat label at t is an extra detection. The extra label is paired with an O or X "pseudobeat" label. The variable t is reset to the value of t'.

   2) If t does not precede T, set T' to the time of the next reference beat label (or to a time beyond the end of the record if there are no more reference beat labels). There are again two possibilities:

      i) If t is closer to T than to T' and t is within 150 ms of T, the beat labels at T and t are paired. The variable t is reset to the time of the next test beat label.

      ii) Otherwise, the device has missed the beat at T. The extra reference beat label is paired with an O or X "pseudobeat" label. The variable T is reset to the value of T'.

c) The matrix element corresponding to the beat label pair that was generated in step b is incremented.

d) Steps b and c are repeated until both t and T are set to times beyond the end of the record.

During the derivation of the matrix, the procedure shall keep track of segments that have been marked as unreadable or as VF in either the reference or the test annotation file. During unreadable segments, pseudobeat labels are X; at all other times, pseudobeat labels are O. Test beat labels generated during reference VF segments are not counted for these purposes. Reference beat labels present during device-marked VF segments are paired with O pseudobeat labels and counted like all other missed beats. In principle, an unreadable segment or a VF segment may begin during the learning period; this possibility shall be taken into account by software designed to perform beat-by-beat comparisons.

NOTE—The reference definition of a beat appears in upper case and the algorithm annotation in lower case (e.g., REFERENCE/algorithm).

4.3.3 Heart rate, and heart rate or RR interval variability

4.3.3.1 Heart rate measurement

Many definitions of heart rate are in common use, and none is accepted universally. To evaluate the accuracy of heart rate measurement, the evaluator shall implement and disclose a method for obtaining heart rate measurements using the reference annotation files (the 'reference heart rate'). This method need not be identical to the method used by the device under test, but in general it will be advantageous if it matches that method as closely as possible. If the method is not identical, the reason for using an alternate method shall be disclosed. If the device produces a continuous heart rate signal (rather than a set of discrete measurements), this signal shall be sampled, either periodically at no less than 2 Hz, or for each beat, in order to obtain a set of discrete measurements for evaluation purposes. Each calculation of the reference HR shall be compared to the corresponding (in time) measurement of HR by the device under test. The comparison of each measurement results in a measured error expressed as a percentage of the mean of the reference heart rate measurements. If the device under test provides more than one type of heart rate measurement as an output, the provisions of this paragraph apply separately to each such type of measurement.

4.3.3.2 Heart rate variability or RR interval variability measurement from databases

The reference annotations of the MIT–BIH databases (2nd edition, published in August 1992) provide a convenient standard set of realistic heart beat sequences that can be used to compare the results of HRV algorithms from various developers as well as test the behavior of an algorithm. Because the emphasis here is on the HRV calculations and because QRS detection and classification performance are tested elsewhere, the ECG waveforms are *not* used. Only the QRS times and labels are used to assure that each developer can submit the same inputs to the HRV calculations. Although there is no widely recognized list of the expected HRV calculation results for each record of these databases, this practice will highlight any differences in HRV calculations, and over time a consensus list of expected results is likely to emerge. The following issues must be addressed to allow a comparison of HRV calculations.

In order to qualify algorithm performance, database reference labels are used as input to the HRV algorithm under test. This results in HRV performance statistics that can be compared with other algorithms. This comparison is performed with no optimization settings enabled in the HRV algorithm.

a) *Labels*: All beats understood to have a sinus node origin (those with an “n” label as defined in 4.2, including normal and bundle branch block beats) should be treated as normal. All other beats should be considered ectopic.

b) *Interrupting labels*: Certain events indicate an interruption of the heart rhythm, either physiologically (e.g., ventricular fibrillation) or artificially (e.g., unreadable signal). Any intervals that include such interruptions shall be identified and should not be used by the algorithm.

c) *Noninterrupting labels*: Some labels are informative and do not suggest an interruption of the sinus rhythm. These labels can be ignored, and the intervals that include these labels may be used.

d) *Extra intervals*: Some HRV algorithms provide for exclusion of more than one interval before and after an ectopic beat. The program should be configured to exclude only one interval before and only one interval after each ectopic beat for this test.

e) *Interval relationships*: For the purpose of this test, no intervals shall be excluded based on interval relationships (e.g., maximum and minimum allowable intervals or ratios of intervals). If a maximum limit is required (such as to avoid arithmetic overflow), that limit shall be disclosed.

f) *Quantization*: The intervals given in the database annotation files shall be requantized to the appropriate step size for the HRV algorithm to be tested. The quantization shall be applied to the absolute time (summation of full precision intervals) so that the resulting intervals do not suffer from an accumulation of round-off errors. See 4.3.3.3 f) for an elaboration.

g) *Duration*: Some indices of HRV require more than 30 min of data to be of practical use, such as SDNN (standard deviation of 24 h of intervals) or day-night difference. Still, for purposes of comparison, SDNN can be computed from just 30 min. For those algorithms that can appropriately configure for a day-night difference, this difference shall be defined as the difference between the last 15 min and the 20 min immediately prior to that of each 30-min record (in the case of longer records from the AHA DB, for example).

h) *NN50, pNN50*: These standard indices of HRV shall be defined by consecutive intervals different by more than 50 ms. The sign of the change may be in either direction, but the magnitude of the change shall be greater (and not equal) to 50 ms. This becomes important when intervals are quantized. It shall be disclosed whether NN50 is normalized to 24 h or not.

The testing outlined above is repeated with settings provided to the algorithm to reflect use of the algorithm in the clinical environment. Labels provided by QRS detection and classification are used to replace the reference labels from the database. Algorithm settings used by the manufacturer shall be disclosed. One final test run is completed with these disclosed settings, with the reference label annotations as input to the HRV algorithm.

Table 4—AHA and MIT–BIH database labels distributed for use by HRV algorithms

Use	Interrupting	Noninterrupting
N normal	~ change in signal quality	s ST segment change
L left bundle	U unreadable region	T T-wave change
R right bundle	I isolated QRS-like artifact	* Systole
B unspecified bundle	[start ventricular fibrillation	D diastole
	] end ventricular fibrillation	" Comment annotation
		= measurement annotation
	a aberrated atrial premature	p P-wave peak
	V premature ventricular	^ pacemaker artifact
	F ventricular/normal fusion	t T-wave peak
	J nodal premature	+ rhythm change
	A atrial premature	u U-wave peak
	S supraventricular premature	(waveform onset
	E ventricular escape	) waveform end
	j nodal escape	: index mark
	/ paced	< start analysis
	Q unclassifiable	> end analysis
	? beat not classified	
	e atrial escape	
	n supraventricular escape	
	x nonconducted P-wave	
	f pace/normal fusion	
	r R-on-T premature	

4.3.3.3 Heart rate variability or RR interval variability measurement of test patterns

In addition to HRV measurements made in section 4.3.3.2, it is important to evaluate the accuracy of an algorithm based on a data set that has a deterministic and known measure. This is accomplished by using an artificially created analog waveform and a set of annotation test patterns that can be presented to an algorithm and for which an expected output can be specified.

Analog test pattern: Test pattern 1 is intended to be applied through the complete signal path of the instrument. In other words, test pattern 1 is produced as an analog ECG waveform, recorded, digitized, and processed by the QRS detector. The noise floor measurement thus reveals the contributions due to sampling effects, phase lock loops, arithmetic precision, and perhaps other effects.

a) To measure HRV noise floor, connect a signal generator to the appropriate ECG inputs of the device. Adjust the signal generator to obtain a 1 mV triangular pulse with a width at the baseline of 100 ms. The repetition rate shall be between 55 and 75 pulses per min. The repetition rate shall be stable within 0.01 percent over 24 h.

b) Acquire enough signal duration to complete each HRV calculation three times. For example, if one HRV calculation is the standard deviation of all intervals in a 5-min period, then more than 15 min of data shall be acquired so three separate calculations of that index can be made. Some HRV calculations are defined only for a 24-h period. Three separate 1-day acquisitions shall be used to get the three calculations.

c) Perform three analyses of each HRV index by the device under test. Be sure each analysis is of a different segment of acquired simulated ECG data.

d) For each HRV index, record the worst case measurement (maximum variability) of the three trials. This worst case measure is the noise floor.

The following list defines the HRV index in table 5 below.

Time domain indices:

— Mean: mean of all the intervals in ms;

— SDNN: standard deviation of all intervals over the complete test duration in ms;

— SDANN: standard deviation of the 5-min means in ms;

— ASDNN: mean of the 5-min standard deviations in ms;

— NN50: count of all consecutive intervals different by more than 50 ms;

— PNN50: NN50 as a percentage of all allowed intervals;

— rMSSD: root mean square of successive differences in ms;

— TINN: triangular index interval.

Frequency domain indices:

— VLF: very low frequency power (0.00333 Hz to 0.040 Hz) in ms^2;

— LF: low frequency power (0.040 Hz to 0.150 Hz) in ms^2;

— HF: high frequency power (0.150 Hz to 0.400 Hz) in ms^2.

Table 5—Example of noise floor calculation results

HRV index	Trial 1	Trial 2	Trial 3	Noise Floor
SDNN	4.7 ms	4.8 ms	4.1 ms	4.8 ms
ASDNN	4.1 ms	3.9 ms	4.0 ms	4.1 ms
SDANN	0.2 ms	0.4 ms	0.5 ms	0.5 ms
rMSSD	5.6 ms	6.1 ms	5.7 ms	6.1 ms
pNN50	0%	0%	0%	0%
TINN	24 ms	24 ms	16 ms	24 ms
VLF	0.04 ms^2	0.04 ms^2	0.04 ms^2	0.04 ms^2
LF	0.13 ms^2	0.13 ms^2	0.13 ms^2	0.13 ms^2
HF	1.30 ms^2	1.30 ms^2	1.25 ms^2	1.30 ms^2

Digital Test Patterns: Test patterns 2 through 5 are expected to be applied in the digital domain after the QRS detector/classifier. This is to test the validity of the arithmetic in the absence of effects characterized elsewhere and to avoid the need to build an analog waveform simulator of the required complexity.

a) Define a sinusoidal test pattern as a sequence of NN interval that obeys the following rules. The values rravg, rrdev, and hrvfreq will assume different values for the different test patterns.

rravg = average rr interval in sec

rrdev = magnitude of rr variability in sec

hrvfreq = the frequency of variability in cycles per sec

T( ) = QRS times sequence

T(0) = 0.0

rr(k) = rravg + rrdev * sin(2*π*hrvfreq*T(k))

T(k+1) = T(k) + rr(k)

Specify rr() and T() in sec and use double floating point (64 bit) arithmetic in order to have sufficient precision.

Table 6—Example of HRV test results

Test Pattern	rravg	rrdev	hrvfreq	Hrvperiod
2	0.800	0.035	0.25	4 secs
3	1.000	0.070	0.10	10 secs
4	3.000	0.280	0.033333	30 secs
5	1.500	0.140	0.000278	1 hour

b) Quantitize the intervals. The QRS times sequence shall be quantitized, and the interval sequence recomputed from the quantitized times to avoid an accumulation of round-off error.

sampletime = time in seconds between allowable interval values for the algorithm under test

Tq(k) = sampletime * integer((T(k) / sampletime) + 0.5)

rrq(k+1) = Tq(k+1) - Tq(k)

c) Define all beats to be N, normal sinus initiated, and disable all rules that would exclude intervals based on relationships, such as ratios or maximum and minimum limits. If a maximum limit is required to avoid arithmetic overflow, that limit shall be disclosed. Test pattern intervals range from 0.765 sec to 3.28 sec.

d) Construct enough duration of each of the following test patterns to satisfy the requirements of each HRV index. The maximum possible computable duration shall be tested. Test pattern 5 is not required when durations as long as 60 min are not testable by the HRV index under consideration.

e) For each test pattern, predict an expected value for each HRV index (see A.3.5.3).

f) Process each list of quantitized intervals for each HRV index. Compare the measured HRV index to that expected for each test pattern (see A.3.5.3).

4.4 Run-by-run comparison

4.4.1 General description

Run-by-run comparisons are used to measure a device's ability to detect runs of consecutive ectopic beats. For each type of ectopic beat (VEB and SVEB), two run-by-run comparisons are required, one for sensitivity and another for positive predictivity. The end product of a run-by-run comparison is a pair of matrices in which each element is a count of the number of run pairs of the appropriate type.

Table 7—Run sensitivity summary matrix

		Algorithm run length						
		0	**1**	**2**	**3**	**4**	**5**	**>5**
Reference Run Length	**0**		S_{01}	S_{02}	S_{03}	S_{04}	S_{05}	S_{06}
	1	S_{10}	S_{11}	S_{12}	S_{13}	S_{14}	S_{15}	S_{16}
	2	S_{20}	S_{21}	S_{22}	S_{23}	S_{24}	S_{25}	S_{26}
	3	S_{30}	S_{31}	S_{32}	S_{33}	S_{34}	S_{35}	S_{36}
	4	S_{40}	S_{41}	S_{42}	S_{43}	S_{44}	S_{45}	S_{46}
	5	S_{50}	S_{51}	S_{52}	S_{53}	S_{54}	S_{55}	S_{56}
	>5	S_{60}	S_{61}	S_{62}	S_{63}	S_{64}	S_{65}	S_{66}

Table 8—Run positive predictivity summary matrix

		Algorithm run length						
		0	**1**	**2**	**3**	**4**	**5**	**>5**
Reference Run Length	**0**		P_{01}	P_{02}	P_{03}	P_{04}	P_{05}	P_{06}
	1	P_{10}	P_{11}	P_{12}	P_{13}	P_{14}	P_{15}	P_{16}
	2	P_{20}	P_{21}	P_{22}	P_{23}	P_{24}	P_{25}	P_{26}
	3	P_{30}	P_{31}	P_{32}	P_{33}	P_{34}	P_{35}	P_{36}
	4	P_{40}	P_{41}	P_{42}	P_{43}	P_{44}	P_{45}	P_{46}
	5	P_{50}	P_{51}	P_{52}	P_{53}	P_{54}	P_{55}	P_{56}
	>5	P_{60}	P_{61}	P_{62}	P_{63}	P_{64}	P_{65}	P_{66}

NOTE—Each entry corresponds to a combination of reference run length and algorithm run length. All run lengths greater than 5 are condensed into the last column (row). Each element is named according to the matrix to which it belongs (S or P) followed by two subscripted numerals corresponding to the reference and algorithm run lengths.

4.4.2 Terms and symbols

In the rest of this section, the general term “run” refers to a sequence of consecutive V or F labels, as defined in 4.2, (which may be mixed in any order) delineated by surrounding N, S, or Q labels (or by the beginning or end of the test period or of an unreadable segment). Recall that O and X pseudo-beat labels are used only for beat-by-beat comparisons; they are completely ignored in run-by-run comparisons and do not delineate runs. The following terms and abbreviations are used to denote runs of specific lengths:

— Couplet (C) = a run of two consecutive V or F labels

— Short run (S) = a run of three, four, or five consecutive V or F labels

— Long run (L) = a run of six or more consecutive V or F labels

A segment of ventricular fibrillation or flutter marked by “[” and “]” labels is considered to be equivalent to a VE long run for the purposes of this section; any adjacent V or F labels are considered to be part of the same run. Similarly, a segment of atrial fibrillation or flutter marked by rhythm labels is considered to be equivalent to an SVE long run, and any adjacent S labels are considered to be part of the same run.

4.4.3 Run sensitivity summary matrix

This paragraph describes how to derive the VEB run sensitivity summary matrix.

a) The reference annotation file defines the location of all runs. For each reference run, a match window is defined, beginning 150 ms before the time of first beat label of the reference run and ending 150 ms after the time of the last beat label of the reference run.

b) For each reference run, the reference run length is the number of consecutive V or F reference beat labels within the match window.

c) For each reference run, the test run length is the number of consecutive V or F test beat labels within the match window. If more than one detected run occurs during a single reference run, the test run length is determined by the longest detected run within the match window. If there are no V or F test beat labels during a reference run, the test run length is zero.

d) Each possible combination of reference run length and test run length corresponds to a cell in the run sensitivity summary matrix. For each reference run, the count in the appropriate cell is incremented.

To derive the SVE run sensitivity summary matrix, follow the same procedure, replacing each “V” or “F” with “S” in the description above.

4.4.4 Run positive predictivity summary matrix

This paragraph describes how to derive the VEB run positive predictivity summary matrix.

a) The test annotation file defines the location of all runs. For each test run, a match window is defined, beginning 150 ms before the time of the first beat label of each test run and ending 150 ms after the time of the last beat label of the test run.

b) For each test run, the test run length is the number of consecutive V or F test beat labels within the match window.

c) For each test run, the reference run length is the number of consecutive V or F reference beat labels within the match window. If more than one reference run occurs during a single test run, the reference run length is determined by the longest reference run during the match window. If there are no V or F reference beat labels during a test run, the reference run length is zero.

d) Each possible combination of reference run length and test run length corresponds to a cell in the run positive predictivity summary matrix. For each reference run, the count in the appropriate cell is incremented.

To derive the SVE run positive predictivity summary matrix, follow the same procedure, replacing “V” or “F” with “S” in the description above.

4.5 VF and AF comparisons

For devices that are claimed to detect VF, a VF comparison shall be performed. This test requires the production of an annotation file based on the device’s outputs, containing (at a minimum) the times when the device has determined that episodes of VF have begun and ended. Overlap exists during any interval in which both the reference and algorithm annotations indicate that VF is in progress. Each reference episode for which overlap exists is counted as a true positive for purposes of determining VF episode sensitivity; any other reference episodes are counted as false negatives. Similarly, each algorithm-marked episode for which overlap exists is counted as a true positive for purposes of determining VF episode positive predictivity; any other algorithm-marked episodes are counted as false positives.

Measurement of VF duration sensitivity and positive predictivity requires determination of the total duration of reference and algorithm-marked VF and of the total duration of periods of overlap as defined above.

Additionally, the following information shall be disclosed for each record:

a) the section of record used for testing;

b) whether an alarm was generated for the test record;

c) what the alarm was, if one occurred (e.g., asystole, ventricular tachycardia, or ventricular fibrillation);

d) the gradation of alarms, if applicable;

e) the interval between the onset of the arrhythmia to the time the alarm was activated, if one occurred. (This last requirement only applies to devices that perform real-time monitoring.)

In addition, for algorithms that attempt to detect ventricular fibrillation/flutter, any false positive detections that occur on any record in the database shall be reported.

For devices that are claimed to detect AF, an AF comparison shall be performed. This test is performed in the same manner as the VF comparison with the substitution of “AF” for each occurrence of “VF” in the description above.

4.6 ST comparison

4.6.1 For devices that measure the ST segment amplitude, ST segment slope, or detect ST changes, an ST comparison shall be performed. This test requires the production of reference and test annotations. These annotations may be beat-by-beat or per some fixed or variable time. The test annotations shall be based on the algorithm outputs containing numerical measurements of ST amplitudes and/or slopes. The method of generating the reference ST annotations shall be disclosed including the method of generating the reference ST amplitude and/or slope values, the leads used, any data exclusions, and any data processing or filtering.

ST measurement errors (REFERENCE – algorithm) are measured by comparing each of the algorithm’s measurements to the reference measurements on the same signal and nearest in time to the algorithm measurements. The data used and the method for obtaining the reference ST amplitude values and ST slope values shall be disclosed.

4.6.2 For devices claimed to measure the ST segment amplitude, the following data plots shall be generated for all measurements and for all leads that measure ST amplitude:

a) scatter plot of all algorithm ST amplitude measurements versus reference ST values with the line of identity indicated on the plot (figure 1);

b) scatter plot of algorithm measurement error versus reference ST values, with the mean error and standard deviation indicated for all algorithm ST measurements (figure 2);

c) scatter plot of algorithm ST amplitude measurements versus reference ST values over the reference ST amplitude range from -200 microvolts to +200 microvolts (figure 3).

The graphs shown in figures 1 through 3 are used to illustrate the ST performance with a particular database. If the graphs are used for an individual record, that fact shall be specifically stated in the title.

4.6.3 For devices claimed to measure the ST segment slope, the following data plots shall be generated for all measurements for all leads that measure ST slope values:

a) scatter plot of ST slope measurement error values verses reference ST slope values with the mean error and the standard deviation indicated for algorithm ST slope measurements (figure 4);

b) scatter plot of all algorithm ST slope measurements versus reference ST slope values with the line of identity indicated on the plot (example not shown; similar to figure 5 but over a wider range of values on the x-axis);

c) scatter plot of algorithm ST measurements versus reference ST slope values over the reference ST slope range from - 2.0 millivolt/sec to + 2.0 millivolt/sec (figure 5).

4.6.4 Event-by-event comparisons similar to run-by-run comparisons are needed in order to derive ST episode sensitivity and positive predictivity. Overlap exists during any interval in which both the reference and algorithm annotations indicate that an ST change is in progress. Events match for the purposes of measuring sensitivity when the period of overlap includes either the reference-marked extremum or at least 50 % of the length of the reference-marked event. Events match for purposes of measuring positive predictivity when the period of overlap includes either the algorithm marked extremum or at least 50 % of the length of the algorithm-marked event.

Measurement of ST change duration sensitivity and positive predictivity requires determination of the total duration of reference and algorithm-marked ST events and of the total duration of periods of overlap as defined above.

For devices that detect ST changes based on more than one signal simultaneously, the definition of a reference-marked ST event shall be modified so that such an event is considered to be in progress if any signal has been annotated as having an ST change in progress; in such cases, the events match for the purposes of measuring sensitivity that occurs when the period of overlap includes the reference-marked extremum in signal, or 50 % of the length of the reference marked event.

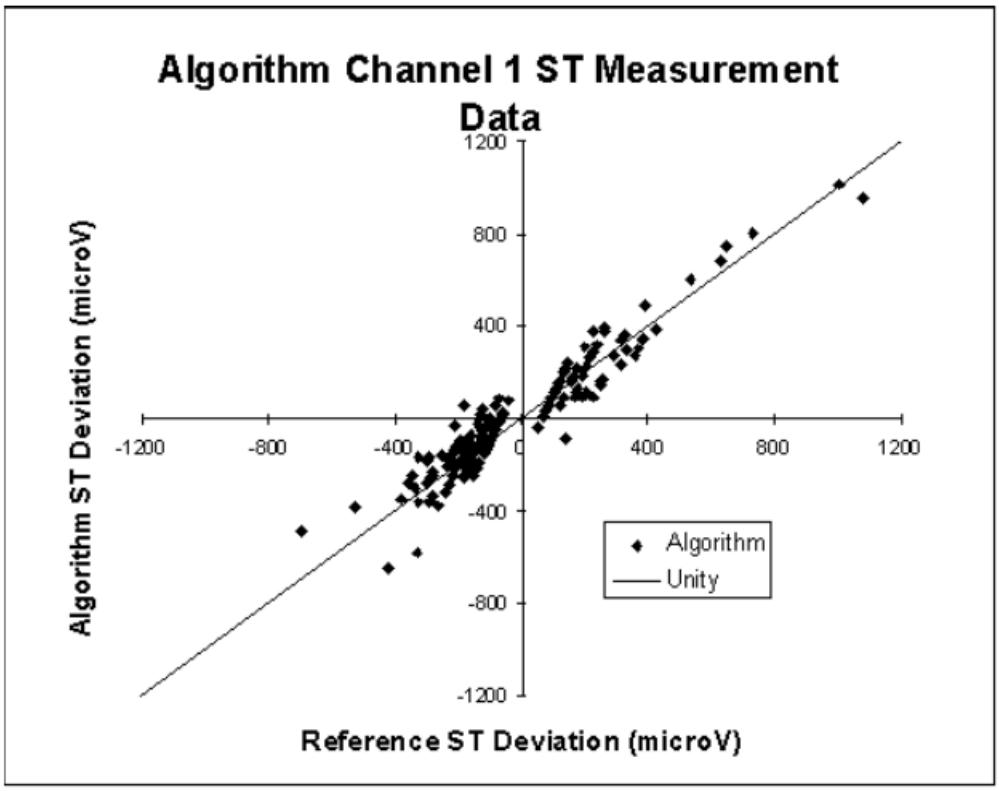

Figure 1—Example of scatter plot of ST amplitude measurement

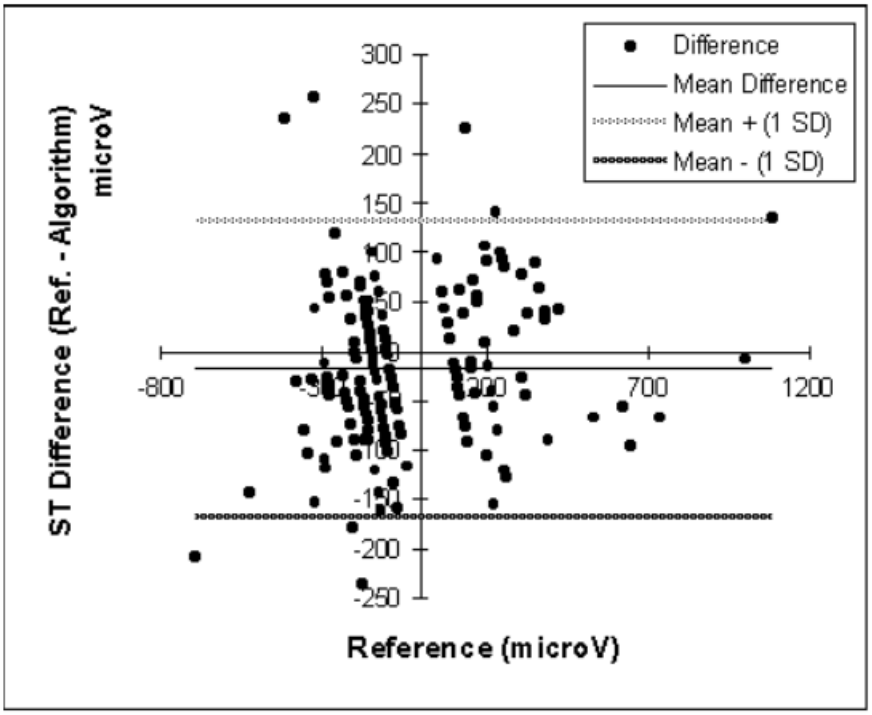

Figure 2—Example of scatter plot of ST amplitude measurement

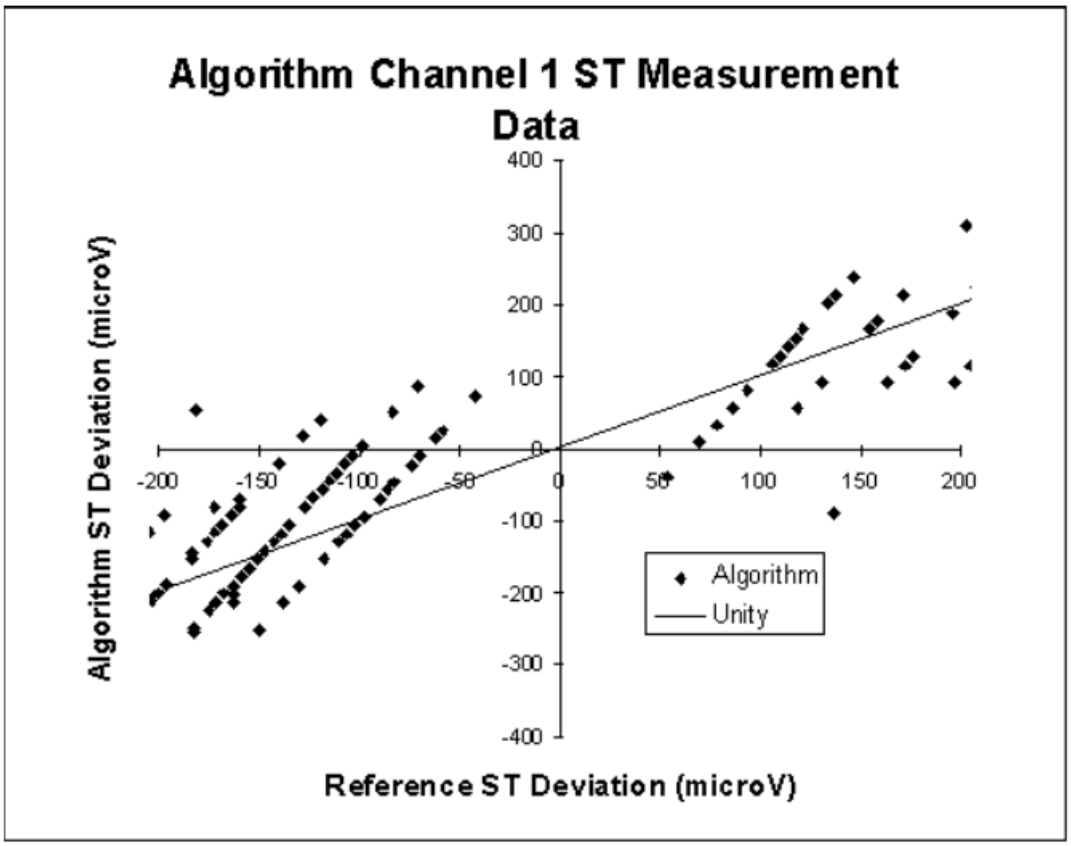

Figure 3—Example of scatter plot of ST amplitude measurement (-200 microvolt to + 200 microvolt reference)

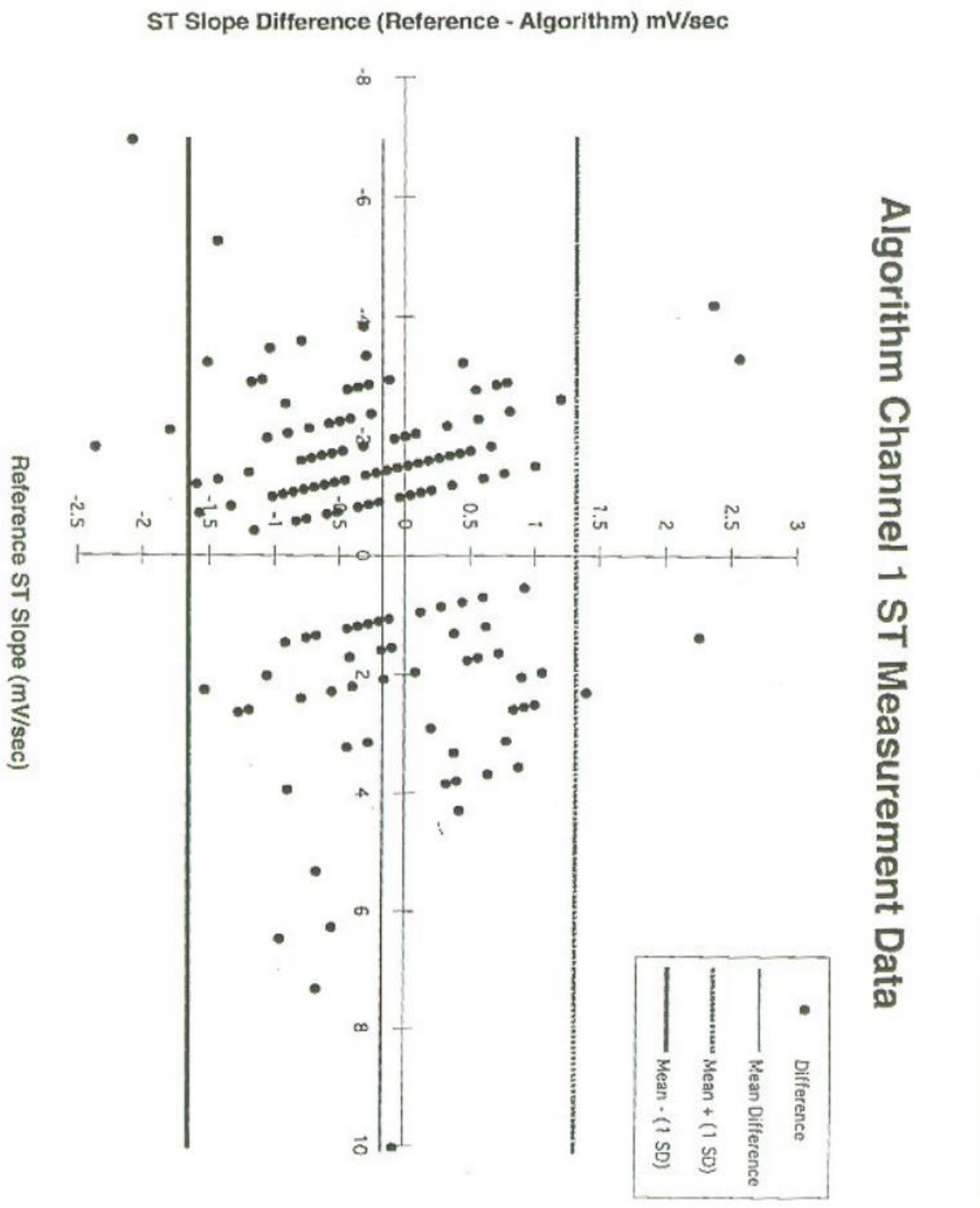

Figure 4—Example of scatter plot of ST slope measurement error

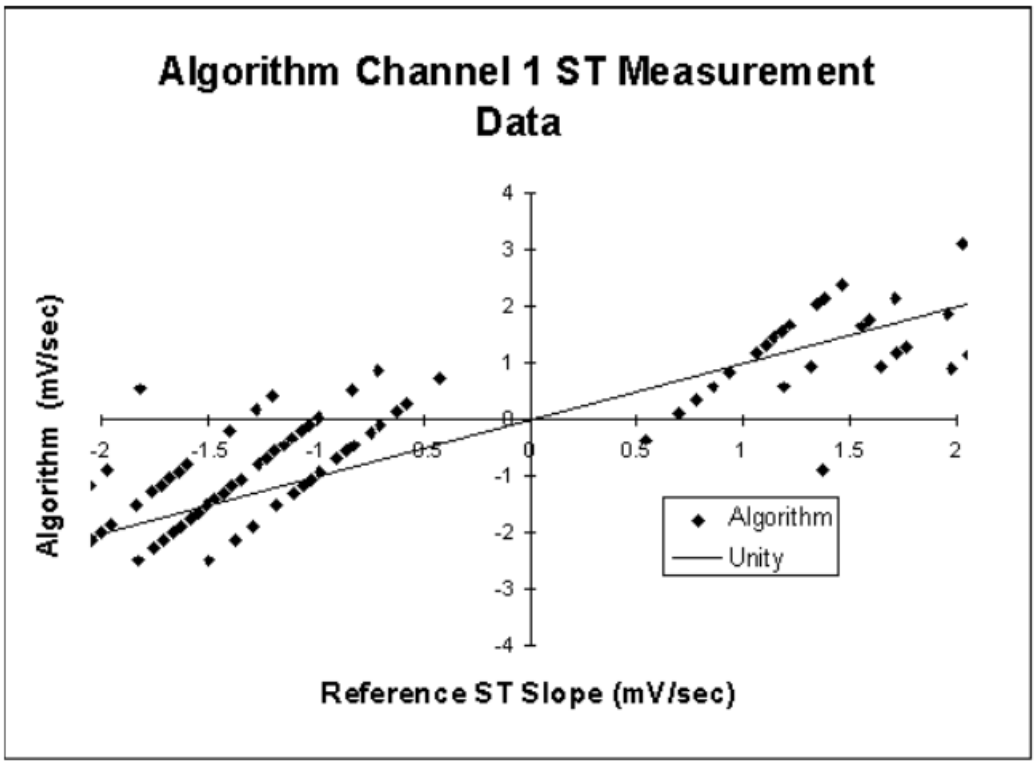

Figure 5—Example of scatter plot of ST slope measurement

Annex A
(Informative)

Rationale and additional guidance

The subclauses in annex A are keyed by paragraph number to the corresponding sections or paragraphs appearing in the normative text of ANSI/AAMI EC57:1998. For example, paragraph A.3.1 contains rationale and additional guidance for section 3.1. Rationale and/or additional guidance is not provided for every section of ANSI/AAMI EC57:1998.

A.1 Scope

No rationale or additional guidance is provided for section 1.

A.2 Definitions of abbreviations

No rationale or additional guidance is provided for section 2

A.3 Algorithm testing

A credible evaluation must be reproducible. For this reason, evaluations of these devices shall be performed without human intervention, i.e., a strictly reproducible “hands-off” evaluation is required. (With human intervention allowed, perfect results are achievable in principle for any device that provides “full-disclosure” output. Thus, evaluations that allow human intervention measure only the persistence and expertise of the operator and are of no value in assessing the performance of the device; for this reason, such evaluations are neither required nor encouraged.)

Full disclosure of the procedure for generating annotation files enables an independent (third-party) evaluator to use the procedure, thereby permitting verification of test results when the same test data are used. It also permits the use of additional test data of the evaluator’s choice as such data become available.

The evaluation methodology of section 4 requires the combination of the device with its interface. In principle, the interface might include significant analytical components when processing the outputs of the device, thereby “improving” its apparent performance. Full disclosure will provide a disincentive for having the interface do anything other than straightforward translations of the device’s normal outputs into standard annotation files.

A.3.1 Databases

As performance is highly dependent on the characteristics of the particular ECGs that are analyzed, evaluations shall be performed using standard recordings so that the results of those evaluations have value for purposes of comparison among devices or against a performance standard.

The exclusion of records with paced beats is permitted only for devices that are not designed to analyze paced analog ECG recordings made without pacer artifact detection or enhancement, because the original analog tapes do not reproduce pacemaker artifacts with fidelity sufficient to permit use of common techniques for recognition of such artifacts in “live” signals.

Most devices need a certain amount of time to learn the underlying rhythm. For this reason, a 5-min learning period is allocated at the beginning of each record and is excluded from calculated performance statistics. If the long version of the AHA DB (containing 2.5 h of unannotated signals per record immediately preceding the 30-min test periods) is used, only the final 35 min of each record (equivalent to the standard version) may be presented to the device under test.

Table A.1—Records to be included in a complete test

Database	Record ID	Description	Number of records
AHA database	1201–1210	No VEBs	10
(included)	2201, 2203–2210	Isolated uniform VEBs	9
	3201–3210	Isolated multiform VEBs	10
	4201–4210	Bigeminy	10
	5201–5210	R-on-T VEBs	10
	6201–6210	Ventricular couplets	10
	7201–7210	Ventricular tachycardia	10
	8201–8204, 8206–8210	Ventricular fibrillation	9
	AHA records in complete test		78
(excluded)	2202, 8205	Paced beats	2
MIT–BIH database (included)	100, 101, 103, 105, 106, 108, 109, 111–119, 121–124	Records selected at random	20
	200–203, 205, 207–210, 212–215, 219–223, 228, 230–234	Records selected to include less common but clinically important arrhythmias	24
	MIT–BIH records in complete test		44
(excluded)	102, 104, 107, 217	Paced beats	4

NOTE—The AHA record ID numbers given refer to the 35-min version of the AHA database. The second digit in the ID numbers is “0” (rather than “2”) for the corresponding 3-hour records. Only the last 35 min of the 3-hour records (equivalent to the 35-min records) may be presented to the algorithm as part of a complete test if the 3-hour records are used.

A.3.2 Testing Requirements

The incidence and variety of arrhythmias and ectopic beats in the 90 records of the ESC DB are insufficient to allow that database to serve as a substitute for the AHA and MIT–BIH databases for the purposes of assessing QRS detection and classification performance. An evaluation using the 90 records of the ESC DB and the same beat-by-beat and run-by-run comparison protocols, however, can supplement the required AHA and MIT–BIH database evaluation. Such a test can be particularly useful for assessing the robustness of QRS detection and classification performance in the presence of ST-segment and T-wave changes.

The AHA, MIT–BIH, NST, CU, and ESC databases are not accompanied by reference heart rate variability (HRV) values. The accuracy of the HRV calculation is best evaluated from controlled inputs for which the exact reference HRV parameters can be predicted. The databases do provide a set of defined QRS times and labels that can be used as common, realistic, easily available, standard input sequences for HRV algorithms. If just the HRV results were available from two different HRV algorithms, comparisons of equivalence could be made. Where discrepancies are observed, the discussion of differences in algorithm implementation or differences in index definitions could begin with a real focus. Over time, a consensus set of correct results for every well-defined index of HRV should evolve.

This recommended practice cannot address all measures of HRV that might be in use at the time of this document’s publication or that could be invented in the future. The test methods and reporting requirements described here, however, are expected to be useful for these other indices as well.

The diagnostic utility of HRV analysis, if any, remains to be determined. The requirements of this recommended practice with respect to HRV analysis are not to be construed as definitions of criteria for diagnostically useful measurements. The sole purpose of these requirements is to establish a standard methodology for assessing the numerical accuracy of specific device outputs and not to impute any diagnostic value to those outputs. Such diagnostic value, if any, can only be determined on the basis of clinical studies that are beyond the scope of this recommended practice.

The incidence and variety of VF in the AHA and MIT–BIH databases are insufficient to allow those databases to serve as substitutes for the CU DB for the purposes of section 4.5. An evaluation of VF detection using the 80 records of the AHA DB and the 48 records of the MIT–BIH DB should supplement the required CU DB evaluation, as the CU DB does not contain a sufficient sample of signals likely to provoke false VF detections.

A.3.3 Test environment

No rationale or additional guidance is provided for section 3.3.

A.3.4 Multiple-lead analysis

No rationale or additional guidance is provided for section 3.4.

A.3.5 Requirements for the evaluation report

There are four possible outcomes of an experiment in which a detector is presented with an input that is either an event or a nonevent. A correctly detected event is called a true positive (TP); an erroneously rejected (missed) event is called a false negative (FN); an erroneously detected nonevent is called a false positive (FP); and a correctly rejected nonevent is called a true negative (TN). In many detection problems, nonevents cannot be counted, so that the number of true negatives is undefined. In such problems, the commonly used detector performance measures are sensitivity (Se, the fraction of events that are detected) and positive predictivity (+P, the fraction of detections that are events):

$$\mathrm{Se} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \qquad +\mathrm{P} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}$$

A.3.5.1 Required statistics

It is useful, particularly when the total number of events is small, to define aggregate statistics that describe the performance of a detector on an entire database as a whole. Two types of aggregate statistics are commonly used: gross statistics, in which each event or detection is given equal weight, and average statistics, in which each record (subject) is given equal weight. If the incidence of events and detections were equal in all subjects, these statistics would be equivalent.

When considering detection statistics for persistent events (such as episodes of fibrillation or ST deviation), it is of interest to know how many episodes are detected as well as the total duration of the detected events. Event statistics give equal weight to each episode, irrespective of length. Duration statistics give weight to each event or detection in proportion to its duration. Thus, event statistics for persistent events are roughly analogous to average statistics for discrete events, and duration statistics are similarly analogous to gross statistics.

Reporting requirements: Although the MIT–BIH DB has been available since 1980, and the AHA DB since 1982, it remains a difficult task to determine minimal acceptable levels of performance for ECG analyzers. Users should understand clearly that diagnostic outputs of these devices cannot be accepted uncritically. Given that review is necessary in any case, what constitutes “acceptable” performance depends to a significant extent on how much effort the user is willing to devote to assessing the accuracy of a device’s outputs. (The effort required of the user will, in turn, depend on the quality of the review and editing facilities provided by the device, if any.)

Performance is often characterized in terms of aggregate statistics, which provide a convenient summarization of device performance on many records. To extrapolate from an aggregate statistic to a prediction of real-world performance is difficult, because the selection criteria used by database developers vary, as do subject populations among clinical practices. It might be expected that average statistics, in which each record is equally weighted, would be better predictors of real-world performance than gross statistics. The record-by-record statistics on which average statistics are based are often unreliable, however, as the number of events in each record may be small. As a result, average statistics can be extraordinarily sensitive to single errors and are usually less robust estimators of performance than are the gross statistics, which are based on larger numbers of events. For this reason, most of the reporting requirements are specified as gross statistics, and reporting requirements for statistics, such as average VEB positive predictivity, have been omitted intentionally.

The distribution of record-by-record statistics is a somewhat better basis for predicting real-world performance to the extent that the records studied are representative of the subject population in clinical practice. Informally, it is clear that performance on a previously untested subject can be predicted with more confidence given a narrow distribution of performance on tested subjects than given a wide distribution. These distributions are rarely normal (Gaussian), however, and classical parametric models (e.g., measures such as sample variance) are inadequate for characterizing or comparing them. Bootstrap estimation is a nonparametric method for determining confidence limits on performance, which has been applied to this problem; it is also useful when comparing the robustness of different statistics.

Other aspects of performance: Several issues cannot be addressed adequately using existing test methodology. Automated P-wave detection, though desirable, is beyond the current state-of-the-art for ECG analyzers that rely on body-surface leads alone. The MIT–BIH DB includes five records with annotated nonconducted P-waves; no other P-wave annotations are present in any of the available databases. Similarly, T-wave annotations are wholly absent, except for annotations that indicate possibly significant changes in T-wave morphology in the ESC DB. Conduction disturbances exist and are annotated in nine records of the MIT–BIH DB and in two records of the European ST-T DB, but it is not clear how accuracy in analysis of conduction disturbances can be confidently measured with a sample of this size. Similar concerns arise with respect to junctional rhythms (annotated in three MIT–BIH DB records) and SVTA (annotated in seven MIT–BIH DB records and three ESC DB records). Major concerns are evaluation of arrhythmia detectors in the context of paced beats and the corollary issue of evaluation of pacer function analysis algorithms and pacer malfunction detectors. A modern database of high-fidelity pacer recordings, including examples of pacer malfunction, is needed in order to address these issues.

A.3.5.2 Requirements for all arrhythmia algorithms

QRS sensitivity and positive predictivity: Using the beat-by-beat comparison matrix definitions from 4.3, QRS sensitivity and positive predictivity are derived as follows:

$$\begin{aligned}QTP = {} & Nn + Ns + Nv + Nf + Nq + \\ & Sn + Ss + Sv + Sf + Sq + \\ & Vn + Vs + Vv + Vf + Vq + \\ & Fn + Fs + Fv + Ff + Fq + \\ & Qn + Qs + Qv + Qf + Qq\end{aligned}$$

$$\begin{aligned}QFN = {} & No + Nx + \\ & So + Sx + \\ & Vo + Vx + \\ & Fo + Fx + \\ & Qo + Qx\end{aligned}$$

$$\begin{aligned}QFP = {} & On + Os + Ov + Of + Oq + \\ & Xn + Xs + Xv + Xf + Xq\end{aligned}$$

$$QRS\,Se = \frac{QTP}{QTP + QFN} \qquad QRS + P = \frac{QTP}{QTP + QFP}$$

VEB and SVEB Sensitivity, Positive Predictivity, and False Positive Rate: Using the beat-by-beat comparison matrix definitions from 4.3, VEB sensitivity and positive predictivity are derived as follows:

$$VTP = Vv$$

$$VFN = Vn + Vs + Vf + Vq + Vo + Vx$$

$$VFP = Nv + Sv + Ov + Xv$$

$$\begin{aligned}VTN = {} & Nn + Nf + Nq + Ns + \\ & Sn + Sf + Sq + Ss + \\ & Fn + Ff + Fq + Fs + \\ & Qn + Qf + Qq + Qs + \\ & On + Of + Oq + Os + \\ & Xn + Xf + Xq + Xs\end{aligned}$$

$$VEB\,Se = \frac{VTP}{VTP + VFN} \qquad VEB + P = \frac{VTP}{VTP + VFP}$$

$$VEB\,FPR = \frac{VFP}{VTN + VFP}$$

Note that VTP and VFP do not include Fv or Qv; thus, a detector is neither penalized nor rewarded for its treatment of ventricular fusion beats and ambiguous beats.

The example below, based on hypothetical data, shows one way of presenting the information required by this section. Details of formatting the evaluation report are left to the discretion of the tester.

SVEB sensitivity and positive predictivity are similarly defined:

$$SVTP = Ss$$

$$SVFN = Sn + Sv + Sf + Sq + So + Sx$$

$$SVFP = Ns + Vs + Fs + Os + Xs$$

$$SVTN = Nn + Nv + Nf + Nq +$$

Vn + Vv + Vf + Vq +

Fn + Fv + Ff + Fq +

Qn + Qv + Qf + Qq +

On + Ov + Of + Oq +

Xn + Xv + Xf + Xq

$$\text{SVEB Se} = \frac{\text{SVTP}}{\text{SVTP+SVFN}} \qquad \text{SVEB+P} = \frac{\text{SVTP}}{\text{SVTP+ SVP}}$$

$$\text{SVEB} = \frac{\text{SVFP}}{\text{SVTN+SVFP}}$$

Note that Qs is excluded from SVTP and SVFP, so that a detector's treatment of ambiguous beats does not influence its measured SVEB detection performance.

Table A.2—Example of a line-format, beat-by-beat performance report

Beat summary statistics for MIT–BIH database

Record	Nn´	Vn´	Fn´	On´	Nv	Vv	Fv´	Ov´	No´	Vo´	Fo´	Q Se	Q +P	V Se	V +P	V FPR
100	1900	0	0	0	1	1	0	0	0	0	0	100.00	100.00	100.00	50.00	0.053
101	1521	0	0	0	0	0	2	0	0	0	0	100.00	100.00	-	-	0.000
103	1723	0	0	0	2	0	0	35	4	0	0	99.77	98.01	-	0.00	2.102
105	2036	2	1	4	78	27	4	39	7	0	0	99.68	98.04	93.10	18.75	5.422
106	1235	2	0	0	0	452	0	5	1	6	0	99.59	99.70	98.26	98.97	-
Sum	73235	250	450	4104	200	5605	37	95	4018	45	136					
Gross												95.00	95.00	95.00	95.00	0.378
Average												95.00	95.00	95.00	95.00	0.500

Total QRS complexes: 83976 Total VEBs: 5900

Summary of results from 44 records

Table A.2.1—Condensed beat-by-beat summary matrix containing 11 elements

		Algorithm		
		n+ f+q	**v**	**o + x**
Reference	**N**	**Nn´**	**Nv**	**No´**
↓	**V**	**Vn´**	**Vv**	**Vo´**
	F + Q	**Fn´**	**Fv´**	**Fo´**
	0 + X	**On´**	**Ov´**	

Note—The linear format performance (Table A.2) is based on a condensed matrix.

Table A.2.2—Summary table (matrix format) of beat-by-beat comparison

		Algorithm					
		n	**v**	**f**	**q**	**o**	**x**
Reference	**N**	**Nn**	**Nv**	**Nf**	**Nq**	**No**	**Nx**
	V	**Vn**	**Vv**	**Vf**	**Vq**	**Vo**	**Vx**
	F	**Fn**	**Fv**	**Ff**	**Fq**	**Fo**	**Fx**
	Q	**Qn**	**Qv**	**Qf**	**Qq**	**Qo**	**Qx**
	S	**Sn**	**Sv**	**Sf**	**Sq**	**So**	**Sx**
	O	**On**	**Ov**	**Of**	**Oq**	**Oo**	**Ox**
	X	**Xn**	**Xv**	**Xf**	**Xq**	**Xo**	**Xx**

Shutdown Statistics: Shutdown is defined as that period of time when the algorithm is not performing its detection/classification function. The following shutdown statistics are derived using the beat-by-beat comparison matrix definitions from 4.3:

$$\%\,\text{beats missed during shutdown} = \frac{Nx + Vx + Fx + Qx + Sx}{QTP + QFN}$$

$$\%\,\text{N and S missed during shutdown} = \frac{Nx + Sx}{Nn + Nv + Nf + Nq + No + Nx + So + Sx + Sn + Sv + Sf + Sq}$$

$$\%\,\text{V missed during shutdown} = \frac{Vx}{Vn + Vv + Vf + Vq + Vo + Vx}$$

$$\%\,\text{F missed during shutdown} = \frac{Fx}{Fn + Fv + Ff + Fq + Fo + Fx}$$

TOTAL SHUTDOWN TIME is defined as the amount of time during the test period for each record that the algorithm is not performing its detection/classification function. For each record, it is expressed in mins and sec in the format MM:SS.

The example below, based on hypothetical data, shows one way of presenting the information required by this section: a line-format shutdown report. The formatting of this report is left to the discretion of the tester.

Table A.3—Example of a line-format shutdown report

Record	Nx + Sx	Vx	Fx	Qx	% beats missed	% N and S missed	% V missed	% F missed	Total Shutdown Time
AH8006	3	0	0	0	0.26	0.32	0.00	-	16 sec
AH8007	0	0	0	0	0.00	0.00	0.00	0.00	6 sec
AH8008	0	0	0	0	0.00	0.00	0.00	-	4 sec
AH8009	0	0	0	0	0.00	0.00	0.00	-	0 sec
AH8010	0	0	0	0	0.00	0.00	-	-	1 sec
Sum	129	5	0	0	-	-	-	-	136 sec
Gross									
Average									

Summary of results from 78 records

A.3.5.3 Requirements for algorithms with optional capabilities

RMS heart rate error: The RMS heart rate error is derived from the results of the methods of 4.3.3.1. Although HR and HRV measurements depend on RR interval measurements, some algorithms for obtaining these measurements are robust with respect to occasional RR interval measurement errors, while others are particularly sensitive to such errors. The purpose of testing HR and HRV measurements based on algorithm-derived RR intervals is to establish if the measurement algorithms are robust, at least with respect to the particular errors committed by the device under test.

The purpose of testing HRV measurements based on reference RR intervals is to permit direct observation of the effects of RR interval measurement errors on HRV measurements (by comparing the results of this test with those of the same test performed using the algorithm-derived RR intervals).

The purpose of testing HRV measurements based on simulated analog ECG data is to establish the noise floor for these measurements, i.e., the sum of the contributions of analog and sampling noise to errors in these measurements. The purpose of testing HRV measurements based on the simulated (digital) RR interval sequences specified in section 4.3.3.3 is to demonstrate the extent to which these measurements agree with predictions based on the stated measurement definitions and on known statistical properties of the simulations; hence, this test indirectly establishes whether the implementation of the measurement algorithms is likely to be correct.

VF and AF detection: From the counts of true positives, false negatives, and false positives derived according to the methods of section 4.5, VF and AF episode sensitivity and positive predictivity are derived in the usual way.

The VF duration sensitivity and positive predictivity are calculated as:

$$\text{VF duration Se} = \frac{\text{duration of overlap}}{\text{duration of reference - annotated VF}}$$

$$\text{VF duration + P} = \frac{\text{duration of overlap}}{\text{duration of algorithm - annotated VF}}$$

The AF duration sensitivity and positive predictivity are calculated in a similar way.

The example below, based on hypothetical data, shows one way of presenting the information required by this section: a line-format report. Details of formatting this report are left to the discretion of the tester.

Table A.4—Example of a line-format report

Record	TPs	FN	TPp	FP	ESe	E+P	DSe	D+P	Ref duration	Test duration
231	0	0	0	0	-	-	-	-	0:00.000	0:00.000
232	0	0	0	0	-	-	-	-	0:00.000	0:00.000
233	0	0	0	0	-	-	-	-	0:00.000	0:00.000
234	0	0	0	0	-	-	-	-	0:00.000	0:00.000
Sum	1	0	2	1					1:37.900	1:01.000
Gross					100	67	47	75		
Average					100	50	47	45		

Summary of results from 44 records

VF and AF time to detection and false positive report: The following information shall be disclosed for each record with ventricular fibrillation/flutter waveforms:

— the section of record used for testing;

— whether an alarm was generated for the test record;

— what the alarm was, if one occurred (e.g., asystole, ventricular tachycardia, or ventricular fibrillation);

— the gradation of alarms, if applicable;

— the interval between the onset of the arrhythmia to the time the alarm was activated, if one occurred. (This last requirement only applies to devices that perform real-time monitoring.)

In addition, for algorithms that attempt to detect ventricular fibrillation/flutter, any false positive detections that occur on any record in the database shall be reported.

The examples below, based on hypothetical data, show one way of presenting the information required by this section: a VF detection performance report and a false VF detection report, respectively. Details of formatting these reports are left to the discretion of the tester.

Table A.5—Example of VF performance report

Record	Reference Vfib Segments		Algorithm Labels				Alarm Activity	
ID	Start	Stop	N	V	F	Q	Time	Type
207	00:40.73	00:50.97	1	15	0	0	00:48.39	Run
207	00:54.76	01:00.36	2	16	0	0	00:55.10	VFIB
207	04:02.14	04:06.43	0	0	0	0	04:02.42	Run
207	04:07.89	04:21.45	0	0	0	0	04:12.11	Run
207	04:29.46	04:40.90	0	0	0	0	04:29.82	VFIB
							04:35.87	Run
							04:38.70	Run

Table A.6—Example of false VF performance report

Record	False Vfib Segments		Reference Labels				
ID	Start	Stop	N	V	F	Q	U
8002	32:18.25	32:31.25	0	35	0	0	0
8002	32:36.25	32:40.62	0	13	0	0	0

Couplet and run sensitivity and positive predictivity: The results of run-by-run comparisons (section 4.4) can be used to derive VE couplet and run sensitivity and positive predictivity:

$$CTPs = S22 + S23 + S24 + S25 + S26 \qquad CFN = S20 + S21$$

$$CTPp = P22 + P32 + P42 + P52 + P62 \qquad CFP = P02 + P12$$

$$\text{VE Couplet Se} = \frac{CTPs}{CTPs + CFN} \qquad \text{VE Couplet} + P = \frac{CTPp}{CTPp + CFP}$$

$$STPs = S33 + S34 + S35 + S36 + S43 + S44 + S45 + S46 + S53 + S54 + S55 + S56$$

$$SFN = S30 + S31 + S32 + S40 + S41 + S42 + S50 + S51 + S52$$

$$STP\ p = P33 + P43 + P53 + P63 + P34 + P44 + P54 + P64 + P35 + P45 + P55 + P65$$

$$SFP = P03 + P13 + P23 + P04 + P14 + P24 + P05 + P15 + P25$$

$$\text{VE Short Run Se} = \frac{STPs}{STPs + SFN} \qquad \text{VE Short Run} + P = \frac{STPp}{STPp + SFP}$$

$$LTPs = S66 \quad LFN = S60 + S61 + S62 + S63 + S64 + S65$$

$$LTPp = P66 \quad LFP = P06 + P16 + P26 + P36 + P46 + P56$$

$$\text{VE Long Run Se} = \frac{LTPs}{LTPs + LFN} \qquad \text{VE Long Run + P} = \frac{LTPp}{LTPp + LFP}$$

The example below, based on hypothetical data, shows one way of presenting the information required by this section: a line-format couplet and run performance report. Details of formatting this report are left to the discretion of the tester.

Table A.7—Example of a line-format couplet and run performance report

Record	CTs	CFN	CTp	CFP	STs	SFN	STp	SFP	LTs	LFN	LTp	LFP	CSe	C+P	SSe	S+P	LSe	L+P
AH8004	0	1	1	32	0	4	2	32	0	0	0	21	0	3	0	6	-	0
AH8006	1	1	1	9	2	1	2	6	1	1	2	5	50	10	67	25	50	29
AH8007	41	8	60	2	66	16	91	5	33	17	35	3	84	97	80	95	66	92
AH8008	0	1	1	2	0	0	1	1	0	0	0	4	0	33	-	50	-	0
AH8009	2	2	3	0	2	0	4	0	7	1	4	0	50	100	100	100	88	100
AH8010	0	0	0	0	0	0	0	0	0	0	0	0	-	-	-	-	-	-
Sum	956	54	968	126	400	41	457	101	53	24	61	81						
Gross													96	82	71	22	79	72
Average													75	67	91	53	76	68

Total couplets: 999 Total short runs: 464 Total long runs: 79

Summary of results from 78 records

SVEB couplet and run statistics are similarly defined.

A.3.6 Simulated test patterns

No rationale or additional guidance is provided for section 3.6.

A.4 Automated analysis

A.4.1 Use of standard databases

No rationale or additional guidance is provided for section 4.1.

A.4.2 Use of annotation files

No rationale or additional guidance is provided for section 4.2.

A4.3 Beat-by-beat comparison

A.4.3.1 General description

No rationale or additional guidance is provided for section 4.3.1.

A.4.3.2 Method for beat-by beat comparison

No rationale or additional guidance is provided for section 4.3.2.

A.4.3.3 Heart rate, and heart rate or RR interval variability

The heart rate variability results derive from the testing method of 4.3.3.2. These results shall be reported separately for each HRV (RRV) measurement.

Table A.8—Example of results of HRV program run on MIT–BIH database reference annotations

NOTE: Refer to the definitions from table 5.

Record	Mean	SDNN	SDANN	ASDNN	NN50	pNN50	rMSSD	VLF	LF	HF
100	795	36	15	32	124	5.7	28	191.17	43.04	484.71
101	968	66	42	49	360	19.6	38	691.09	312.46	796.52
102	802	26	0	26	5	5.6	27			
103	866	46	13	42	203	9.8	32	652.49	182.96	598.77
104	787	31	3	32	7	15.2	34	149.81	3.79	1.28
105	701	34	24	24	26	1.2	21	93.42	11.11	360.69
106	954	107	66	92	336	35.3	54	3114.50	482.87	560.89
108	1025	104	51	88	647	40.2	88	1644.35	2147.88	1757.91
109	713	31	11	29	81	3.4	25	73.99	9.31	425.28
Two special cases were noted. Record 102 has only 99 normal beats and 26 minutes of constant pacing, so no FFT spectra were available. Record 107 is entirely paced rhythm, so no HRV analysis could be done.										

Note that the configuration guidelines of 4.3.3.2 are meant to harmonize the calculations performed by algorithms from different developers for the purpose of making the results comparable. This sometimes causes the calculation to produce results that might otherwise be considered clinically suspect. For example, the elimination of an upper limit on intervals means the HRV result would include the effect of a very long interval, such as the 53-sec interval of record AHA8210. However, it would be difficult to establish a universally acceptable upper limit, and it is often the extreme inputs that demonstrate the differences between algorithms most clearly. Another example would be a pattern of trigeminy. Such a pattern of two normal beats and an ectopic beat would produce a set of NN intervals and would thus allow HRV calculations. Clinically, such a result would be highly suspect because of the great amount of ectopy. The NN intervals amount to sampling of the sinus node activity only once every trigeminal cycle, and this might be as slow as once every three seconds. That would make estimates of the respiratory HRV very undersampled. Still, for the purposes of comparison of algorithm function, such patterns are very useful. Therefore, when testing an HRV algorithm, all RR intervals shall be submitted as input to the HRV algorithm. Similarly, when testing an HRV algorithm as part of an ECG analysis algorithm/device, the entire ECG recording shall be submitted as input to the QRS detector. In no case should the evaluator tamper with the input data, but it is entirely appropriate for the algorithm under test to examine its inputs and for it to treat suspect intervals differently than presumably reliable intervals (provided that the user is informed of the exclusion or weighting rule).

Following the testing methods of 4.3.3.3, results shall be reported separately for each HRV (RRV) measurement.

Table A.9—Example of device measurements of synthetic test patterns

NOTE: Refer to definitions from table 5.

HRV index	noise floor	35 ms	70ms	280 ms	140 ms
SDNN	4.8 ms	25	49	197	99
SDANN	0.5 ms	0	0	2	98
ASDNN	4.1 ms	25	49	197	14
rMSSD	6.1 ms	29	31	123	1
pNN50	0%	0	0	79.9	0
TINN	24 ms	55 ms	89 ms	300 ms	155 ms
VLF	0.04 ms^2	0	0	39106.82	4.64
LF	0.13 ms^2	0	2438.36	7.86	0
HF	1.30 ms^2	579.45	0.17	0.29	0

Table A.10—Example of predicted ideal values for synthetic test patterns

NOTE: Refer to definitions from table 5.

HRV index	noise floor	35 ms	70ms	280 ms	140 ms
SDNN	0 ms	24.75	49.50	197.99	98.99
SDANN	0 ms	0.00	0.00	0.00	97.87
ASDNN	0 ms	24.75	49.50	197.99	14.00
rMSSD	0 ms	29.77	31.25	125.87	0.28
pNN50	0%	0.0	0.0	87.0	0.0
VLF	0 ms^2	0.0	0.0	39200.0	0.0
LF	0 ms^2	0.0	2450.0	0.0	0.0
HF	0 ms^2	612.5	0.0	0.0	0.0

Table A.11—Example of choice of test patterns

Test pattern	1	2	3	4	5
Variation magnitude (ms)	0	35	70	280	140
Period of variation	N/A.	4 sec	10 sec	30 sec	1 hour
Frequency of variation (Hz)	N/A.	0.25	0.1	0.033333	0.000278
Frequency range assignment	N/A.	HF	LF	VLF	VLF
Average interval (sec)	1	0.800	1.000	3.000	1.500
Beats per minute	60	75	60	20	40

The magnitudes of the test patterns were chosen to represent a useful range of real values. The magnitude of 0 ms was chosen to show the noise floor. The magnitude of 70 ms was chosen, because the predicted SDNN value would be 50 ms, a popular choice for a possible clinical cut point. The correct assignment of positive and negative test results depends particularly on the accuracy of near cut points. The magnitude of 35 ms was chosen to be a small value representing the range below 70 ms.

The magnitudes 140 and 280 ms represent large values of HRV. To make a reasonable prediction for the HRV indices, however, it is necessary that the variation in intervals be small compared to the intervals themselves. Large deviations from the average would cause the sampling of each sine wave cycle of variation to be more asymmetric, with many more short intervals during the low half cycle than long intervals in the other half cycle. The average intervals were chosen to be at least ten times longer than the variations. This means for a variation magnitude of 280 ms, the average interval must be almost 3 sec (20 BPM). To avoid more unrealistic low average heart rates, larger magnitudes of variation are not tested. The average intervals were rounded up slightly to produce test patterns that repeat every minute.

The largest magnitude of variation was applied to test pattern 4 instead of test pattern 5, because test pattern 5 might not be applicable for many algorithms due to the long duration of data required to test such a low frequency. It is desirable that all algorithms be evaluated by the maximum test magnitude of pattern 4.

Test pattern 1 is intended to be applied through the complete signal path of the instrument. In other words, test pattern 1 is produced as an analog ECG waveform (see 4.3.3.3, parts a–d), recorded, digitized, and processed by the QRS detector. The noise floor measurement thus reveals the contributions due to sampling effects, phase lock loops, arithmetic precision, and perhaps other effects.

Test patterns 2 through 5 are expected to be applied in the digital domain post QRS detector/classifier (see 4.3.3.3, parts e–j). This is to test the validity of the arithmetic in the absence of effects characterized elsewhere and to avoid the need to build an analog waveform simulator of the required complexity.

HRV frequencies for test patterns 2 and 3 were chosen to match the familiar 4-second and 10-second periods of HRV seen in many people and to exercise the HF and LF bands described on page 1047 of the ESC/NASPE special report.* The frequency of test pattern 4 should exercise the VLF band but still be of short enough duration to be useful to most short-term HRV algorithms. Test pattern 5 was designed to exercise an HRV index that senses variations over time periods much longer than 5 min (e.g., SDANN).

Prediction of some HRV indices for the synthetic test pattern QRS sequences: Throughout the following discussion, "intervals" is assumed to mean only those intervals selected for study. In the case of the synthetic test patterns, all beats have the "Normal" label, and all exclusion rules based on interval relationships are disabled, so all intervals are used by the algorithm. RRDEV refers to the zero-to-peak magnitude of the interval variations and takes on the values 0, 35, 70, 140, and 280 ms in the test patterns.

Some HRV indices have strong relationships to other indices. Two easy approximations are worth noting here. Variance is the square of standard deviation and Parseval's theorem relates power to variance. These two relationships are *approximate* but can serve nicely as reality checks. Users of HRV programs should be aware of them.

Because of the similarity to an analysis of variance (ANOVA), the following is true.

$$SDNN^2 \text{ approximately equals } SDANN^2 + ASDNN^2$$

where:

$SDNN^2$ = variance of all intervals

$SDANN^2$ = between-group variance

$ASDNN^2$ = approximates the within-group variance

The above relationship is only approximate because the definition for ASDNN is the average of *standard deviations,* whereas an ANOVA would compute the average of *variances.*

Because of Parseval's theorem, we can relate power computed in the time domain to power computed in the frequency domain. If power can be computed in the frequency domain over all frequencies down to 0 Hz, then that power can be compared to $SDNN^2$. If power can be computed in the frequency domain for only frequencies above 0.00333 Hz (5-min windows), then $ASDNN^2$ may be compared to the sum of the VLF, LF, and HF powers. ASDNN is computed from only 5-min windows in the time domain.

$$ASDNN^2 \text{ approximately equals } VLF + LF + HF$$

The above relationship is only approximate, because the definition for ASDNN includes no detrending and the definition of HF is limited (< 0.40 Hz) to less than the highest frequencies that might be present.

SDNN: *The standard deviation of all intervals (no subgrouping):* The calculation of standard deviation is the same as root-mean-square (rrms) when the mean value is removed. The rrms value for a sine wave is the zero to peak value of the sine wave divided by the square root of two.

$$SDNN = \frac{RRDEV}{\sqrt{2}}$$

SDANN: *The standard deviation of 5-min mean intervals (variation between 5-min subgroups):* Whenever the test pattern repeats every minute (patterns 1, 2, 3, and 4), the average interval for each 5-min section shall be the same. The standard deviation of a set of constant numbers will be zero. Test pattern 5 is the only pattern that should produce a nonzero SDANN. Test pattern 5 produces a sinusoidal interval variation with a period of 60 min. There will be twelve different 5-min averages. The prediction is similar to the SDNN prediction except the 5-min averages apply a low pass filter with a rectangular impulse response. The amplitude response of such a filter is sin(x)/x. Because the period of variation is twelve times longer than the impulse response, the amplitude response is 0.9886 = $\sin(\pi/12) / (\pi/12)$.

$$SDANN = 0 \text{ for test patterns 1, 2, 3, 4}$$

---

* *Heart Rate Variability, Standards of Measurement, Physiological Interpretation, and Clinical Use*, by the European Society of Cardiology and the North American Society of Pacing and Electrophysiology, *Circulation*, 1996; 93:1043-1065.

$$SDANN = 0.9886\left(\frac{RRDEV}{\sqrt{2}}\right) \text{ for test pattern 5}$$

ASDNN: *The mean of 5-min standard deviations of intervals (variation within 5-min subgroups):* When each 5 minutes is the same as every other 5 mins, this result equals the rrms of the test pattern similar to SDNN. For the case of test pattern 5, ASDNN is not easily predicted. The test pattern repeats every hour, so there are twelve 5-min groups. The standard deviation for each twelfth of a sine wave cycle is not easy to predict. But determined numerically, RRDEV/10.1 is the average of the standard deviations of twelve sections from a sine wave cycle.

$$ASDNN = \frac{RRDEV}{\sqrt{2}} \text{ for test patterns 1, 2, 3, 4}$$

$$ASDNN = \frac{RRDEV}{10.1} \text{ for test patterns 5}$$

rMSSD: *The root mean square of successive differences of intervals:* The greatest rate of change for a sine wave is crossing through the baseline or average value. Because of the definition of the test patterns, the greatest change will be on the downward stroke of the sine wave, $\sin(\pi)$. We want to find the RR interval value just before and just after the variation function passes through the average interval value. Consider test pattern 4. If the average interval value is 3000 ms, then a first approximation is that there is an RR interval to be computed 1500 ms before and 1500 ms after the $\sin(\pi)$. We learn the RR computed from the variation function at 1500 ms before is 3086.525 ms, which is actually a little longer than our first estimate. After four iterations, the estimates are very similar.

$$rr1 = 3000 + 280\left(\sin\left(1500 * 2 * \frac{n}{30000}\right)\right) = 3000 + 86.525 = \frac{3086.525}{2} = 1543.262$$

$$rr1 = 3000 + 280\left(\sin\left(1543.262 * 2 * \frac{n}{30000}\right)\right) = 3000 + 88.934$$

$$rr1 = 3000 + 280\left(\sin\left(1544.467 * 2 * \frac{n}{30000}\right)\right) = 3000 + 89.001$$

$$rr1 = 3000 + 280\left(\sin\left(1544.501 * 2 * \frac{n}{30000}\right)\right) = 3000 + 89.003$$

$$rr2 = 3000 - 89.003$$

$$\text{maximum\_ successive\_difference is } 2 * 89.003 = 178.0 \text{ ms}$$

The derivative of a sine wave is also sinusoidal. The sequence of successive differences is like a derivative and will be approximately sinusoidal if there are enough intervals per period of the variation function. This assumption is weakest for test pattern 2, which has on average only 5 heart beats per variation period. If we accept the sinusoidal nature of the successive differences and we know the maximum successive difference, then we can estimate the root mean square of all successive differences. It will be the maximum divided by the square root of 2.

$$rMSSD = \frac{\max\_scsv\_diff}{\sqrt{2}}$$

Table A.12—Example of RMS interval differences

test pattern	1	2	3	4	5
magnitude variation (ms)	0	35	70	280	140
max successive difference (ms)	0	42.1	44.2	178.0	0.4
rMSSD (ms)	0.00	29.77	31.25	125.87	0.28

pNN50: *The percentage of successive difference by more than 50 ms (increase and decrease combined):* This is easy to predict for all the test patterns except pattern 4. When the maximum successive difference is less than 50 ms, the pNN50 must be zero. When the sequence of successive differences has a maximum of 178 ms, we need to know what part of the time is the sequence above 50 ms. Consider a quarter cycle of a sine wave going from zero to 178. When does it cross 50?

$$\text{Arcsin}\left(\frac{50}{178}\right) = 0.2847 \text{radians}$$

There are $\pi/2$ radians in a quarter cycle. During each quarter cycle, the sequence spends $0.2847/(\pi/2)$ part of the time below 50 ms and 81.87 percent of the time above 50 ms. All quarter cycles are symmetric, so:

pNN50 = 0.0 for test patterns 1, 2, 3, 5

pNN50 = 81.87 for test pattern 4

VLF: The summed power of frequency components between 0.003 Hz and 0.04 Hz

LF: The summed power of frequency components between 0.04 Hz and 0.15 Hz

HF: The summed power of frequency components between 0.15 and 0.40 Hz

The expected power is very easy to compute for all of the test patterns because of Parseval's theorem, which tells us that the total power under the power spectral density curve is equal to the variance of the time domain signal. The only complication to this is when the spectral estimation technique usually cannot observe enough of the signal to see several cycles of the variation. This can easily be the case for some algorithms with test pattern 5, which requires 1 hour to complete one cycle of heart-rate variation. Algorithms that estimate power from segments of data shorter than 1 hour are likely to respond to test pattern 5 with various results, depending on what detrending strategy is used. Indeed, low responses to test pattern 5 might be considered evidence of good detrending strategies.

$$\text{VLF, LF, HFpower} = \frac{\text{RRDEV}^2}{2}$$

Table A.13—Example of summary of frequency components

	Pattern 1	Pattern 2	Pattern 3	Pattern 4	Pattern 5
	0 ms	35 ms	70 ms	280 ms	140 ms
HRV index	**0 Hz**	**0.25 Hz**	**0.10 Hz**	**0.033333 Hz**	**0.000278 Hz**
VLF power in ms^2	0	0.0	0.0	39200.0	0.0
LF power in ms^2	0	0.0	2450.0	0.0	0.0
HF power in ms^2	0	612.5	0.0	0.0	0.0

A.4.4 Run-by-run comparisons

No rationale or additional guidance is given for section 4.4.

A.4.5 VF and AF comparisons

No rationale or additional guidance is given for section 4.5.

A.4.6 ST comparison

Because it is recognized that data with beat-by-beat reference ST measurements are not available at this time, it has been left to the tester to determine how to best generate appropriate reference annotations for testing purposes and then to clearly disclose the chosen method. Algorithm measurements might not necessarily be reported on a beat-by-beat basis. To facilitate comparison, the generation of annotations for the reference and the test data at least should be approximately contemporaneous.

Summary statistics, such as the correlation coefficient or RMS error, can be ill-suited to the task of describing the accuracy of ST deviation measurements. They are highly sensitive to outliers, and do not distinguish between

systematic errors (resulting from bias or nonlinearity) and nonsystematic errors (resulting from poor noise tolerance or unreliable measurement techniques). A better statistic, because of its robustness in the presence of outliers, is a confidence limit estimate over a focused range and over the entire signal range. Because the confidence limits are based on the standard deviation, the tester shall provide the standard deviation in both the line format and on the scatter plot. Many other statistical methods such as Bland–Altman can then be generated from data provided.

The percentage of discrepant ST measurements does not directly quantify accuracy of ST measurements. Algorithms may have a similar percentage of discrepant measurements but have very different levels of accuracy. Furthermore, any specific definition of discrepancy has different levels of significance in the clinical environment, depending on the amplitude of the reference ST deviation. For example, a 100-microvolt discrepancy at an ST level of - 150 microvolts (1.5 mm of ST depression at standard scale) is much more significant than a 100-microvolt discrepancy at an ST level of - 500 microvolts. A better technique, because it directly measures accuracy, is to measure the mean ST measurement error over both a focused range and over the entire signal range.

The purpose of measuring the mean error and standard deviation over a focused range of reference ST amplitudes and slopes (as well as over the entire signal range applied to the algorithm) is to determine the accuracy of the algorithm in the critical region of ST deviations and slopes where most clinical decisions are made, as well as to determine the overall accuracy of the algorithm.

The purpose of generating the scatter plots of ST measurements and ST errors is to summarize results of all individual measurements in a manner that allows rapid visual assessment of any systematic measurement bias, nonlinearity, or region of unreliable performance that could be exhibited by an ST deviation measurement algorithm. In addition, for any arbitrary definition of discrepancy, a rapid visual estimation of percentage discrepancy may be performed.

ST episode and duration detection: From the counts of true positives, false negatives, and false positives derived according to the methods of section 4.5, ST episode sensitivity and positive predictivity are derived in the usual way.

The ST episode duration sensitivity and positive predictivity are calculated as:

$$\text{ST episode duration SE} = \frac{\text{duration of overlap}}{\text{duration of reference - annotated ST episode}}$$

$$\text{ST episode duration + P} = \frac{\text{duration of overlap}}{\text{duration of algorithm - annotated ST episode}}$$

The example below, based on hypothetical data, shows one way of presenting the information required by this section: a line-format report. Details of formatting this report are left to the discretion of the tester.

Table A.14—Example of a line-format report

Record	TPs	FN	TPp	FP	ESe	E+P	DSe	D+P	Ref. duration	Test duration
E0406	0	0	0	0	-	-	-	-	0:00.000	0:00.000
E0408	0	0	0	0	-	-	-	-	0:00.000	0:00.000
E0509	0	0	0	0	-	-	-	-	0:00.000	0:00.000
E0515	0	0	0	0	-	-	-	-	0:00.000	0:00.000
Sum	1	0	2	1					1:37.900	1:01.000
Gross					100	67	47	75		
Average					100	50	47	45		

Summary of results from 90 records