Performance Measures for Algorithms to Detect Transient Ischemic ST Segment Changes

F Jager[2], GB Moody[1], A Taddei[3], RG Mark[1]

[1]Harvard-MIT Division of Health Sciences and Technology, Cambridge, MA
[2]Faculty of Electrical and Computer Engineering, Ljubljana, SLOVENIA
[3]CNR Institute of Clinical Physiology, Pisa, ITALY

Abstract

The availability of the new European ST-T Database makes it possible for the first time to perform quantitative, reproducible performance tests of methods for detecting and measuring transient ischemic ST changes in the ECG. This paper proposes an evaluation protocol and performance measures for use with the database. We describe methods for evaluating the accuracy of ST episode detection, measurement of ischemia duration, and measurements of ST deviation. Bootstrap estimation is used to derive expected lower bounds on performance measures and to assess their utility as predictors of performance. We illustrate these methods by a case study in which we present an evaluation of our two-channel algorithm for automated detection of ischemic ST episodes.

1 Introduction

In the last few years there has been increasing interest in detection and quantification of ischemic ST changes during *ambulatory ECG monitoring* [1, 2] and during *inpatient monitoring* [3]. Evaluation of ST analyzers is a difficult task, lacking a generally accepted methodology. Evaluations have been limited to assessment of an analyzer's ability to detect the presence of an episode [2, 3], or to assessment of ST deviation measurements of dominant normal beats [4]. There are no commonly accepted performance measures for assessing the accuracy of ischemia duration measurements. Until recently, there has been no standard test material to be used as a basis for quantitative, strictly reproducible tests of ST analyzers.

A comprehensive evaluation of an ST analyzer should answer these questions:

1. How well are ST episodes detected?
2. How reliably is ischemia duration measured?
3. How accurately are ST deviations measured?

This paper proposes performance measures and evaluation procedures which can answer these questions.

Detection of *ischemic ST changes* is complicated by the presence of *non-ischemic ST deviations*. These non-ischemic deviations are often caused by position-related changes in the electrical axis of the heart. These axis shifts may cause significant (>100 μV) shifts in the ST level, hence false detections of ischemic ST changes. Although differentiating between ischemic and non-ischemic ST changes is not an easy task for either human experts or automated analyzers, we assert that the clinical importance of making such distinctions is so great that the effort must be made.

2 Test material

The European ST-T Database [5] was created to be a standard set of material for evaluating ischemia detectors. It contains 368 annotated episodes of ischemic ST change (250 episodes of ST depression, and 118 episodes of ST elevation), 11 episodes of ST deviation resulting from axis shift, and 401 episodes of T wave change. Although T wave changes may be independent indicators of myocardial ischemia, we do not address issues related to detection of T wave changes in this paper.

3 Performance measures

There are four possible outcomes of an experiment in which a detector is presented with an input which is either an event or a non-event. A correctly detected event is called a true positive (TP); an erroneously rejected (missed) event is called a false negative (FN); an erroneously detected non-event is called a false positive (FP); and a correctly rejected non-event is called a true negative (TN). In many detection problems,

0-8186-2485-X/92 $3.00 © 1992 IEEE

non-events 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 which are detected), and positive predictivity ($+P$, the fraction of detections which are events).

It is useful, particularly when the total number of events is small, to define aggregate statistics, which describe the performance of a detector on the 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 incidences of events and detections were equal in all subjects, these statistics would be equivalent.

Since the events of interest (ischemic ST episodes) are characterized by number, length, and extreme deviation, performance measures should be designed to assess not only the accuracy of event detection, but also how accurately the duration of ischemia and the extreme deviations are measured. These considerations imply the need for at least three distinct types of performance statistics: sensitivity and positive predictivity for ischemic ST episode detection, sensitivity and positive predictivity for ischemia, and statistics which characterize the amount and distribution of errors in measurement of ST levels.

It may be argued that some ischemic ST episodes (e.g., the lengthiest, or those with the most extreme deviations, or those associated with the occurrence of ectopic beats or rhythms) are more important to detect than others. In principle, one might stratify the episodes by any metric of clinical significance. In practice, however, the total number of episodes in the database is small, and the reliability of stratified performance measures is questionable.

The test signals contain ischemic ST episodes, non-ischemic ST episodes (resulting from axis shifts), and intervals without significant ST deviation. An analyzer must produce a set of annotations which indicate in which of these three categories each segment of the input belongs. Since the events of clinical interest are ischemic ST episodes, we consider non-ischemic ST episodes and intervals without significant ST deviation as non-events. An evaluation of ischemic ST episode detection begins by comparing the analyzer's annotations with the reference annotations, to build a two-by-two matrix summarizing how the reference ischemic ST episodes were labelled by the analyzer. Since there is not necessarily a one-to-one correspondence between reference ischemic ST episodes and detected ischemic ST episodes, we construct a second matrix summarizing how the detected ischemic ST episodes were labelled in the reference annotation files.

3.1 Ischemic ST episode detection

It is important to be able to estimate the likelihood that an ischemic ST episode was detected, and the likelihood that a given detection was actually an ischemic ST episode. Since the locations of the beginning and end of an ischemic ST episode are difficult or impossible to determine precisely, performance measures for assessing the accuracy of ischemic ST episode detection should be designed to be relatively insensitive to small discrepancies in these locations as given by the reference annotations and by the detector. The performance measures we define in this section and the next depend on the concepts of "matching" and "overlap", respectively. Overlap exists during any interval in which both the reference and detector annotations indicate that an ST episode is in progress. Matching of episodes occurs when the period of overlap includes either the extrema or at least 50% of the length of the episode according to the defining annotations.

The ischemic ST episode sensitivity ($ST\ Se$), an estimate of the likelihood of detecting an ischemic ST episode, may be defined as:

$$ST\ Se = \frac{STP}{STP + FN} .$$

The denominator is the number of reference ischemic ST episodes. STP is the number of matching episodes, and FN is the number of non-matching episodes, where the defining annotations are the reference annotations.

The ischemic ST episode positive predictivity ($ST + P$), an estimate of the likelihood that a detection is a true ischemic ST episode, may be defined as:

$$ST\ +P = \frac{PTP}{PTP + FP} .$$

The denominator is the number of ischemic ST episodes annotated by the detector. PTP is the number of matching episodes, and FP is the number of non-matching episodes, where the defining annotations are the detector annotations. (Note that the definitions of STP and FN are equivalent to those of PTP and FP if reference and detector annotations are swapped.)

3.2 Duration of ischemia

An important clinical index of ischemia is the total duration of all ischemic episodes within the monitoring period, often expressed as the percentage of the time during which ischemia occurred. As estimates of the

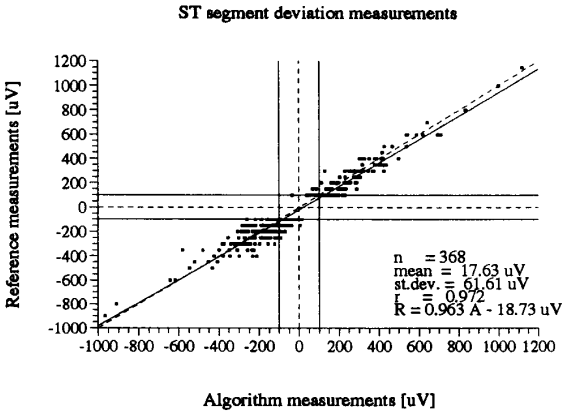

Figure 1: Scatter plot of ST deviation measurements. Line of identity, lines indicating 100μV deviations and results of linear regression analysis are shown.

accuracy with which an ST analyzer can measure this index, we define the ischemia sensitivity ($IS\ Se$, the fraction of true ischemia which is detected), and the ischemia positive predictivity ($IS\ +P$, the fraction of detector-annotated ischemia which is true ischemia):

$$IS\ Se = \frac{IS_{D \wedge R}}{IS_R}\ , \qquad IS\ +P = \frac{IS_{D \wedge R}}{IS_D}\ .$$

IS_R is the total duration of reference-annotated ischemia. IS_D is the total duration of detector-annotated ischemia. $IS_{D \wedge R}$ is the total duration of detector-annotated ischemia which overlaps reference-annotated ischemia.

3.3 ST deviation measurement

The ST deviation is defined for a given waveform as the difference in ST levels between that waveform and the reference waveform for the record. For each of the 368 reference annotations which include an ST deviation measurement (the extremum of a reference ischemic ST episode), the algorithm's measurement should be reported. The purpose of this detailed report is to permit a user to identify the types of waveforms which may cause difficulty for a given ST analyzer. The detailed report should be supplemented by a scatter plot, as in figure 1, in which the ordinates are the reference measurements and the abscissas are the analyzer's measurements. Such a plot allows visual assessment of any systematic measurement bias, nonlinearity, or domain of unreliable performance due to the ST deviation measurement algorithm.

	ST Se	*ST +P*	*IS Se*	*IS +P*
Gross	81.2%	83.9%	70.9%	74.8%
Average	83.8%	87.1%	78.7%	73.4%

Table 1: Aggregate sensitivity and positive predictive accuracy of the ischemic ST change detector evaluated on the ST-T Database.

These results may be summarized by the standard measurements of the mean difference (the bias) between the algorithm and the reference measurements, the standard deviation of the differences, and the correlation coefficient. These statistics, however, are highly sensitive to the presence of outliers. A more robust and informative statistic is the percentage of measurements for which the absolute error is clinically insignificant (below 100 μV); a related statistic is the 98% confidence limit on the absolute error.

4 Case study

To illustrate the use of the evaluation protocol described above, we evaluated our two-lead ST analysis algorithm [6]. An issue which immediately arises in such an evaluation is how to deal with information relating to two signals.

Although ST episodes are indicated in the reference annotation files for each lead separately, they usually occur in both leads simultaneously. We define the onset of each reference ST episode as the time it was first annotated in either lead, and the end as the time it was last annotated in either lead. This process of combining reference ST episodes from the two leads yields a total of 250 ischemic ST episodes and 9 non-ischemic ST episodes (resulting from axis shifts). Although this process is required in order to evaluate a two-lead ST analyzer such as ours, it is inappropriate for evaluation of a single-lead analyzer.

For the entire database, our algorithm obtained STP = 203, PTP = 214, FN = 47, FP = 41. The gross statistics for ischemic ST episode detection in table 1 may be derived from these figures. These figures were also tallied for each record separately, to obtain record-by-record measurements of sensitivity and positive predictivity. The means of these measurements are the average statistics presented in table 1. Bootstrap estimates [7] of the distributions of these statistics, based on 10000 trials, were used to obtain the 5% confidence limits given in table 2.

	ST Se	*ST +P*	*IS Se*	*IS +P*
Gross	75.2%	77.8%	61.7%	69.2%
Average	78.6%	82.4%	73.6%	68.8%

Table 2: The 5% confidence limits for the minimum expected performance statistics of the ischemic ST change detector evaluated on the ST-T Database.

ST deviation measurements were performed for each lead separately. The results are summarized in figure 1. For 89.1% (328/368) of the measurements, the absolute error was less than 100μV, and for 97.8% (360/368) of measurements, less than 150μV.

5 Discussion and conclusions

Instead of considering non-ischemic ST episodes as non-events, as we did, it is possible to count them separately. Taking the reference annotation files as the "gold standard", we found 3 FN and 12 FP non-ischemic ST episodes. No ischemic ST episodes were misclassified as non-ischemic ST episodes by the algorithm. The algorithm correctly labelled six non-ischemic ST episodes noted as such in the reference annotation files, misclassified another one as an ischemic ST episode, missed two others, and detected twelve additional non-ischemic ST episodes. Of this last group, none occured during or adjacent to reference-annotated ischemic ST episodes; nine appeared to be non-ischemic ST episodes which were not noted in the reference annotation files, and the remainder resulted from algorithm errors. Although we consider it highly important to distinguish between ischemic and non-ischemic ST episodes, one may group them together. In that case, we would have $STP = 210$, $PTP = 221$, $FN = 49$, and $FP = 52$.

To predict performance in clinical practice, it is important to model how well a detector behaves on a randomly chosen recording. For this reason, one might expect that average statistics, in which each recording is equally weighted, would be better predictors of "real world performance" than gross statistics. In particular, gross statistics for ischemia duration may be unreliable since significant errors in a small number of cases can have a disproportionate negative influence on these statistics. With respect to episode detection, however, the first-order (record-by-record) statistics on which average statistics are based are themselves unreliable since the number of events in each record is typically less than 5; hence the second-order (average) statistics, though appealing for the reasons outlined above, are likely to be no better as predictors than the first-order gross statistics, which are based on much larger numbers of events.

Although space limitations prevent us from presenting our record-by-record statistics or the list of 368 ST deviation measurements here, we recommend that regulatory agencies require disclosure of these results in formal evaluations. This level of detail is needed by the conscientious user in order to identify the types of signals which might be expected to pose difficulties for a given analyzer.

Acknowledgement

This research was supported in part by the Research Society of Slovenia, Ljubljana, Slovenia.

References

[1] Akselrod, S., Norymberg, M., Peled, I., Karabelnik, E., and Green, M.S. Computerised analysis of ST segment changes in ambulatory electrocardiograms. *Med. & Biol. Eng. & Comput.* **25**:513–519 (1987).

[2] Shook, T. L., Valvo, V., Hubelbank, M., Feldman, C. L., and Stone, P.H. Validation of a new algorithm for detection and quantification of ischemic ST segment changes during ambulatory electrocardiography. *Computers in Cardiology* **15**:57–62 (1988).

[3] Oates, J., Cellar, B., Bernstein, L., Bailey, B. P., and Freedman, S. B. Real-time detection of ischemic ECG changes using quasi-orthogonal leads and artificial intelligence. *Computers in Cardiology* **16**:89–92 (1989).

[4] Burns, M. P., and Downs, W. G. Clinical evaluation of a bedside ST-segment monitor. *Computers in Cardiology* **16**:97–100 (1989).

[5] Taddei, A., et al. The European ST-T database: development, distribution and use. *Computers in Cardiology* **17**:177–180 (1990).

[6] Jager, F., Mark, R. G., and Moody, G. B. Analysis of transient ST segment changes during ambulatory monitoring. *Computers in Cardiology* **18** (1991).

[7] Efron, B. Bootstrap methods: another look at the jackknife. *Annals of Statistics* **7**:1–26 (1979).