INSTITUTE OF PHYSICS PUBLISHING PHYSIOLOGICAL MEASUREMENT

Physiol. Meas. **25** (2004) 629–643 PII: S0967-3334(04)72667-2

Protocol to assess robustness of ST analysers: a case study

Franc Jager[1,2], **George B Moody**[1] **and Roger G Mark**[1]

[1] Harvard-MIT Division of Health Sciences and Technology, 77 Massachusetts Avenue, Cambridge, MA 02139, USA
[2] University of Ljubljana, Faculty of Computer and Information Science, Tržaška 25, 1000 Ljubljana, Slovenia

E-mail: franc.jager@fri.uni-lj.si

Received 27 November 2003, accepted for publication 2 March 2004
Published 5 May 2004
Online at stacks.iop.org/PM/25/629
DOI: 10.1088/0967-3334/25/3/004

Abstract

This paper proposes principles and methods for assessing the robustness of ST segment analysers and algorithms. We describe an evaluation protocol, procedures and performance measures suitable for assessing the robustness. An ST analyser is robust if its performance is not critically dependent on the variation of the noise content of input signals and on the choice of the database used for testing, and if its analysis parameters are not critically tuned to the database used for testing. The protocol to assess the robustness includes: (1) a noise stress test addressing the aspect of variation of input signals; (2) a bootstrap evaluation of algorithm performance addressing the aspect of distribution of input signals and (3) a sensitivity analysis addressing the aspect of variation of analyser's architecture parameters. An ST analyser is considered to be robust if the performance measurements obtained during these procedures remain above the predefined critical performance boundaries. We illustrate the use of the robustness protocol and robustness measures by a case study in which we assessed the robustness of our Karhunen–Loève transform based ischaemic ST episode detection and quantification algorithm using the European Society of Cardiology ST-T database.

Keywords: ST segment analyser, assessing robustness, critical performance boundaries, noise stress test, bootstrap evaluation of performance, sensitivity analysis

1. Introduction

Interest in the detection and quantification of transient ST segment episodes of electrocardiogram (ECG) compatible with ischaemia during coronary or intensive care

monitoring, during ambulatory monitoring and during stress testing, has grown in the last few years. Assessing the properties of ST analysers as well as predicting their behaviour in the real-world clinical environment is a difficult task. Performance assessment using standard inputs (Jager *et al* 1991, ANSI/AAMI 1998a, 1998b, Jager 1998) can provide much useful information about an ST analyser's behaviour and its expected performance in the real world. These performance measurements typically characterize how the standard inputs are analysed in terms of assessing: (1) the accuracy of detecting transient ST segment episodes; (2) the accuracy of detecting total ischaemic time and (3) the accuracy of measuring the ST segment deviations. However, these tests do not include methods for assessing robustness which is another important issue when evaluating a given ST analyser. While performance measurements typically characterize how the standard inputs are analysed, it is important to understand to what extent performance depends critically on the variation and choice of inputs. An analyser whose performance varies little over a range of inputs may be said to be robust with respect to the variation of input. It is often the case that robustness is achieved at the cost of absolute performance. Robust methods are generally preferred, because they are less likely to fail catastrophically than non-robust methods. There is no generally accepted methodology for assessing the robustness of ST analysers and algorithms. Assessing the robustness of ST analysers and algorithms should answer the following questions:

- To what extent performance depends critically on the variation of the noise content of input signals, or, is an ST analyser robust with respect to the variation of input signals?
- To what extent performance depends critically on the choice of the database used for testing, or, is an ST analyser robust with respect to the distribution of input signals?
- To what extent the analysis parameters are critically tuned to the database used for testing, or, is an ST analyser robust with respect to the variation of its architecture parameters?

An analyser whose performance is not critically dependent on the variation of the noise content of input signals is said to be robust with respect to the variation of input signals. An analyser whose performance is not critically dependent on the choice of the database used for testing is said to be robust with respect to the distribution of input signals. Similarly, if the analysis parameters do not critically affect the performance as they are adjusted within some range, an analyser is robust with respect to the variation of its architecture parameters.

This paper describes methods and a protocol for assessing the robustness of ST analysers and algorithms according to these robustness questions. We illustrate the use of the protocol through a case study in which we present a way to assess the robustness of our two-channel Karhunen–Loève transform (KLT) based transient ischaemic ST episode detection and quantification algorithm (Jager *et al* 1998) using the European Society of Cardiology ST-T Database (ESC DB) (Taddei *et al* 1992) as the test database.

2. Assessing robustness

The *protocol to assess the robustness* of ST analysers and algorithms includes the following procedures:

1. *Noise stress tests*. Determine the critical (minimum) signal-to-noise ratio at which the performance remains acceptable;
2. *Bootstrap estimation of performance distributions*. Determine if the performance is critically dependent on the choice of the database used for testing;
3. *Sensitivity analysis*. Determine if the analysis parameters are critically tuned to the test database.

An ST analyser or algorithm is considered to be robust if the performance measurements obtained during these procedures remain above the predefined *critical performance boundaries*. To assess the robustness, we used common *performance measures* (Jager *et al* 1991, 1994, ANSI/AAMI 1998a, 1998b, Jager 1998). Relevant performance measures to assess the ability of ST analyser to detect ST episodes were selected following: gross and average ischaemic ST episode detection sensitivity, *IE Se*, and positive predictivity, *IE + P*, and gross and average ischaemic ST duration sensitivity, *ID Se*, and positive predictivity, *ID + P*. To assess the ability of ST analyser to quantify ST episodes, robust and informative performance measures in the presence of outliers are needed: discrepant ST measurement percentage, $p_{(100\ \mu\mathrm{V})}$, i.e., the percentage of measurements for which the absolute difference between the algorithm and reference ST deviation measurements differ by more than 100 μV, and the value of error which 95% of measurements did not exceed, $e_{(95\%)}$. Since outliers are likely to be rejected by the ST analyser from the input signals, the percentage of rejected noisy heartbeats while still keeping analyser's performance acceptable, p_{n}, is another necessary robust performance measure.

2.1. Noise stress test

Ability to reject noise and noise tolerance are important aspects of the behaviour of an ST analyser or algorithm. Records of conventional ECG databases, in general, do not contain enough noise necessary to assess the noise detection logic of a given analyser. Assessing the analyser's ability to reject a variety of severe noises more accurately predicts its performance in the real-world clinical environment. During the development and evaluation of an ST analyser, it is important to have a test of noise rejection, or a test assessing the ability to analyse under difficult circumstances, which is quantitative and reproducible. A technique of adding noise to ECG records is quantitative and reproducible and allows us to determine the effects of noise on the analyser's performance. Synthesized noise does not guarantee the same characteristics (e.g., non-stationarity) as are observed in the clinical environment. The noise stress test (Moody *et al* 1984) is a method that consists of adding real noise (electrode motion artefacts, baseline wander and muscle noise) to 'clean' ECG signals. An implementation of the noise stress test to severely stress the abilities of the analyser is important to evaluate how the performance degrades in noisy data. The noise stress test has already been used to assess the performance of arrhythmia detectors (Moody *et al* 1984).

The noise stress test database (NST DB) (Moody and Mark 1990) contains records with real noises including baseline wander, electrode motion artefacts and muscle noise. The NST DB is available at *Physionet*[3] Website (Moody *et al* 2000, Goldberger *et al* 2000). The noises of the NST DB were obtained by the 'electrode method'. A variety of noises was recorded using ECG electrodes placed on the arms and thighs of subjects such that the ECG signal was not recorded. The subjects were engaged in vigorous physical activity and the electrodes were moved. Varied noises were created and recorded. After that noises were sorted into three main categories: electrode motion artefacts, baseline wander and muscle noise. During the noise stress test, the noises (noise signals) are added to clean ECG signals at different signal-to-noise ratios. Signal-to-noise ratio is commonly expressed as

$$\mathrm{SNR} = 10 \log \left(\frac{P_{\mathrm{s}}}{P_{\mathrm{n}}\, c^2} \right), \tag{1}$$

where P_{s} is the power of the clean ECG signal, P_{n} is the power of the noise signal and c is an adjustable multiplicative constant to obtain the desired signal-to-noise ratio, given P_{s}

[3] http://www.physionet.org/physiobank/database/nstdb/.

and P_n, calculated separately for each pair of clean ECG signal and noise signal. The P_s is defined as a function of QRS amplitude and the P_n as the noise power measurement. The P_s is defined as the square of the mean peak-to-peak amplitude divided by 8 of the first 300 normal QRS complexes given the ECG record, while the largest 5% and the smallest 5% of the measurements are discarded. To determine the P_n, the first 300 s of the noise record are divided into 1 s chunks. The mean amplitude and the root mean squared difference for this mean are computed for each 1 s segment, while the largest 5% and the smallest 5% of the measurements are discarded again. The P_n is then defined as a square of the mean of these measurements.

Higher noise levels cause higher number of extracted heartbeats by a noise detection procedure of an ST analyser, if the ST analyser extracts noisy heartbeats. The percentage of excluded heartbeats while still maintaining the analyser's performance acceptable, p_n, is largest for the most robust analyser. An important concept when characterizing the results of a noise stress test is the lowest signal-to-noise ratio at which the analyser can still operate acceptably, i.e., the critical performance threshold (Moody *et al* 1984), characterizing the behaviour of an analyser for which noise has been added to its input at various signal-to-noise ratios. The critical performance threshold is smallest for that analyser which is most robust with respect to noise. The analyser is robust if its performance is still above the critical performance boundaries and is therefore relatively insensitive to the variation of the noise content of input signals.

2.2. Bootstrap estimation of performance distributions

In the domain of evaluation of ST analysers and algorithms, second-order gross and average performance statistics are particularly relevant fo predicting real-world performance. Gross statistics models behaviour of the analyser on a large number of events, while the average statistics models behaviour of the analyser on a randomly chosen record. To best predict the analyser's performance in the real-world clinical environment, it is necessary to evaluate it on the basis of records which were not used for development. Many ECG databases (including the ESC DB) are not divided into development and test sets. They are a good representation of the problem domains they represent. Consider a given database which is the 'best' representation of the problem domain. A random division of such a database into two subsets, and developing an analyser for each subset independently, would result in equal architectures of these two analysers and their equal performance. This fact yields an idea of 'bootstrap'. During bootstraping, we can use, for each randomly selected set from the original set, the same analyser architecture, i.e., that which was derived using the original database. The bootstrap statistical procedure (Efron 1979) does not require any assumptions about the distribution of the data to which it is applied, but does assume that the database used for bootstrapping is a well-chosen representative subset of the population of examples for a given problem domain. A bootstrap procedure has been successfully used for assessing the robustness of performance statistics for arrhythmia detectors (Albrecht *et al* 1988).

The bootstrap method estimates the lowest performance which can be expected from a certain database chosen at random (and with replacement) from the original database. Given a set of observations $\{X_i, i = 1, \ldots, L\}$, the bootstrap allows one to estimate the distribution of any statistic $\Psi(X_1, X_2, \ldots, X_L)$. The bootstrap procedure can be presented in four steps:

1. Choose at random and with replacements L elements from the original observations $\{X_i, i = 1, \ldots, L\}$ to form a hypothetical set of observations $\{X_i^*, i = 1, \ldots, L\}$.
2. Calculate the statistic $\Psi(X_1^*, X_2^*, \ldots, X_L^*)$ using the hypothetical set of observations $\{X_i^*, i = 1, \ldots, L\}$.

Table 1. Performances of transient ST episode detectors developed and tested using the ESC DB. A: Taddei *et al* (1995), B: García *et al* (2000), C: Stadler *et al* (2001), D: Jager *et al* (1998), E: Maglaveras *et al* (1998), F: Silipo and Marchesi (1998), [g]: gross, [a]: average, *IE Se*: ischaemic ST episode sensitivity, *IE + P*: ischaemic ST episode positive predictivity, *ID Se*: ischaemic ST duration sensitivity, *ID + P*: ischaemic ST duration positive predictivity.

Measure (%)	A	B	C	D	E	F
[g] *IE Se*	81	–	79.2	85.2	85.0	–
[g] *IE + P*	76	–	81.4	86.2	68.7	–
[g] *ID Se*	–	–	–	75.8	73.0	–
[g] *ID + P*	–	–	–	78.0	69.5	–
[a] *IE Se*	84	84.7	81.5	87.1	88.6	77
[a] *IE + P*	81	86.1	82.5	87.7	78.4	86
[a] *ID Se*	–	75.3	–	78.2	72.2	–
[a] *ID + P*	–	68.2	–	74.1	67.5	–

3. Repeat steps 1 and 2 many times.
4. Use the estimates of Ψ from step 2 to form the estimate of the distribution of Ψ.

When the distribution is known, it is possible to estimate the lowest expected performance (the 5% confidence limits) of the performance of the ST analyser, and thus to predict its performance in the real world. The bootstrap cannot make the original database more representative than the population from which it is taken. No estimation technique is capable of removing bias in the original sampling procedure from the set of original observations.

The ST analyser is robust if its lowest expected performance is still above the critical performance boundaries. The bootstrap is also useful for comparing robustness of different performance statistics, as well as for comparing performance and robustness of different analysers. The narrower the distribution for a statistic, the more robust the corresponding performance statistic. Narrower distributions of performances as estimated by the bootstrap (e.g., narrower intervals between the 5% confidence limits and the raw statistics, or, small standard deviations of expected performances), indicate a more robust analyser. Such an analyser yields nearly the same performance given very different circumstances (the distribution of input signals) and is therefore relatively insensitive to the choice of the database used for testing.

2.3. Sensitivity analysis

Sensitivity analysis addresses the question of how performance varies given small changes in analysis parameters. During the development of an ST analyser or algorithm, optimal detection thresholds were typically determined using the analyser performance characteristic curves by summarizing the relationship between sensitivity and positive predictivity of the ST episode and ST duration detecting by varying detection thresholds continuously between their possible largest and smallest values. Variations of these optimal parameters from their optimal values lead to a drop in the performance. The performance of a robust analyser should not deteriorate below acceptable levels if such changes are made. If this does happen, it suggests that the design of the algorithm may be tuned to the database used for testing.

3. Results: case study

Table 1 comparatively summarizes the performances of those transient ST episode detectors which were developed and tested on the ESC DB using its *original reference*

Table 2. Performance of the KLT-based ST episode detection and quantification algorithm obtained during the noise stress test in a variant when noise was added prior to ARISTOTLE's analysis and using the ESC DB. Performances that do not meet the goals (critical performance boundaries) are boxed. Goal: performance goals (critical performance boundaries), SNR: signal-to-noise ratio, Raw: raw statistics, $p_{(100\mu V)}$: discrepant ST measurement percentage, $e_{(95\%)}$: value of error which 95% of measurements did not exceed, p_n: percentage of rejected noisy heartbeats.

		SNR						
Measure	Goal	6 dB	12 dB	18 dB	24 dB	30 dB	36 dB	Raw
[g] *IE Se* (%)	>80	78.0	83.2	85.6	84.0	83.6	83.6	85.2
[g] *IE* + *P* (%)	>80	49.5	59.0	77.0	82.7	84.4	85.7	86.2
[g] *ID Se* (%)	>70	74.0	71.9	75.3	74.9	74.4	74.5	75.8
[g] *ID* + *P* (%)	>70	37.9	55.5	70.3	76.1	77.8	78.4	78.0
[a] *IE Se* (%)	>80	83.4	84.0	86.3	85.5	85.1	85.4	87.1
[a] *IE* + *P* (%)	>80	53.2	67.2	79.1	84.6	86.0	87.3	87.7
[a] *ID Se* (%)	>70	76.2	74.7	77.4	76.5	76.2	76.4	78.2
[a] *ID* + *P* (%)	>70	43.5	56.2	66.8	71.7	73.6	75.0	74.1
[g] $p_{(100\mu V)}$ (%)	<20	48.4	33.2	23.1	16.0	12.5	11.4	9.8
[g] $e_{(95\%)}$ (μV)	<200	795	365	205	140	135	125	115
[g] p_n (%)	<50	61.5	51.9	32.6	21.2	16.3	14.4	13.7

annotations. These systems incorporate time-domain analysis (Taddei *et al* 1995, García *et al* 2000, Stadler *et al* 2001), KLT approach (Jager *et al* 1998), neural network approach (Maglaveras *et al* 1998) and a combination of the KLT and neural network approach (Silipo and Marchesi 1998). The ESC DB contains 368 ischaemic ST episodes as annotated in each single ECG lead, or 250 lead-independent ischaemic ST episodes if the episodes are combined in the sense of logical OR function. The published sensitivities and positive predictivities in detecting transient ischaemic ST episodes (*IE Se* and *IE* + *P*) of these systems are about 80% or 85%.

We demonstrate the use of the robustness protocol by assessing the robustness of our KLT-based two-channel ischaemic ST episode detection and quantification algorithm (Jager *et al* 1998) using selected performance measures and the ESC DB as the test database. Since no performance requirements have been previously published, we set critical performance boundaries (goals) based on performance measurements obtained from a variety of current analysis algorithms and from our KLT-based analysis algorithm (see the second column in table 2). These levels of performance are possible but not trivial to achieve, and in our estimation represent the standard of performance to be expected of clinically useful ST analysis algorithms at present.

3.1. KLT-based ST episode detection and quantification algorithm

Next, we briefly describe the KLT-based ST episode detection and quantification algorithm. The algorithm was devised as a post-processor to the ARISTOTLE arrhythmia detector (Moody and Mark 1982) and operates simultaneously on two ECG leads. The algorithm is composed of preprocessing procedures, noise detection procedure, and feature-vector trajectory recognition procedures to detect transient ST segment episodes.

On the basis of ARISTOTLE's fiducial point $FP(j)$, where j is the heartbeat number, the algorithm derives a five-dimensional KLT feature vector for the ST segment, $\mathbf{s}(j)$, and another five-dimensional feature vector for the QRS complex, $\mathbf{q}(j)$, for each single iso-electric corrected heartbeat. The noise detection procedure and the feature-vector trajectory recognition procedures are fully implemented in the KLT feature space and use the Mahalanobis distance measure, d_N, between feature vectors, where $N = 5$ is the dimensionality of feature vectors. Each next heartbeat is considered noisy if the normalized residual error for the ST segment or for the QRS complex exceeds certain percentage, or if the ST segment or QRS complex feature vector differs sufficiently from those of the past few heartbeats. Sequences of remaining feature vectors are resampled and further smoothed to form equidistant time series of ST segment and QRS complex feature vectors, $\mathbf{s}(k)$ and $\mathbf{q}(k)$, where k is the sample number.

The feature-vector trajectory recognition procedures incorporate: (1) correction of the reference ST segment level to account for the slow ST level drift; (2) detection of significant axis shifts due to postural changes and (3) detection of transient ST segment episodes. The correction of the reference ST segment level is performed by updating the mean reference ST segment feature vector of 'normal' ST segments after each new ST segment feature vector. Using the second-order distance function, d_N^2, between ST segment feature vectors, the new mean ST segment feature vector is updated as the exponentially weighted sum if the new ST segment feature vector is 'close' to the mean feature vector. Significant axis shifts are detected by searching the first-order distance function, d_N, of the time series of the ST segment and of QRS complex feature vectors for simultaneous significant 'step' changes using low-pass first-order differentiation which have to be preceded and followed by 'flat' intervals. ST episodes are detected by sequentially classifying samples of the first-order ST segment distance function after a correction of the reference ST segment level, denoted by $C_N(k)$, as normal and deviating ones. Consecutive samples of the $C_N(k)$ which are 'far' from the class of normal ST segments form ST episodes. Actual architecture of the ST episode detection procedure is more complex. Samples of the $C_N(k)$ are actually classified according to two decision thresholds, i.e., lower and upper decision thresholds, $L_N(k)$ and $U_N(k)$. These two thresholds are equivalent to 50 μV and 100 μV decision thresholds used by human expert annotators of the ESC DB to annotate transient ischaemic ST episodes in the time series of time-domain ST segment level deviation measurements. Due to the non-stationary nature of the $C_N(k)$, the $L_N(k)$ and $U_N(k)$ are adaptive. The two decision thresholds are adaptive only within a predefined region, i.e., within a 'guard zone', of which center, Λ_{cN}, is also adaptive, and is defined by its initial centre, Λ_{c_0N}, and its lower bound, $\frac{2}{3}\Lambda_{c_0N}$.

For details of the architecture of the KLT-based ST episode detection and quantification algorithm see Jager (1994), Jager *et al* (1998).

3.2. Is the algorithm robust with respect to the variation of input signals?

For the noise stress test procedure, noise from the NST DB was added to all the records of the ESC DB. The first 5 min of the ECG signals were left 'clean' for each record, to give the algorithm an opportunity to measure the reference baseline ST segment deviation levels accurately and to learn successfully. Following this period, noise was added throughout the entire records. We wanted to simulate real circumstances so all three kinds of noise: baseline wander, electrode motion and muscle noise were represented equally. Records to which a given type of noise was added were chosen at random. Signal-to-noise ratios were chosen from 36 dB down to 6 dB with a step of 6 dB. An example of input ECG signal of a record of the ESC DB to which muscle noise artefacts were added at different signal-to-noise ratios is

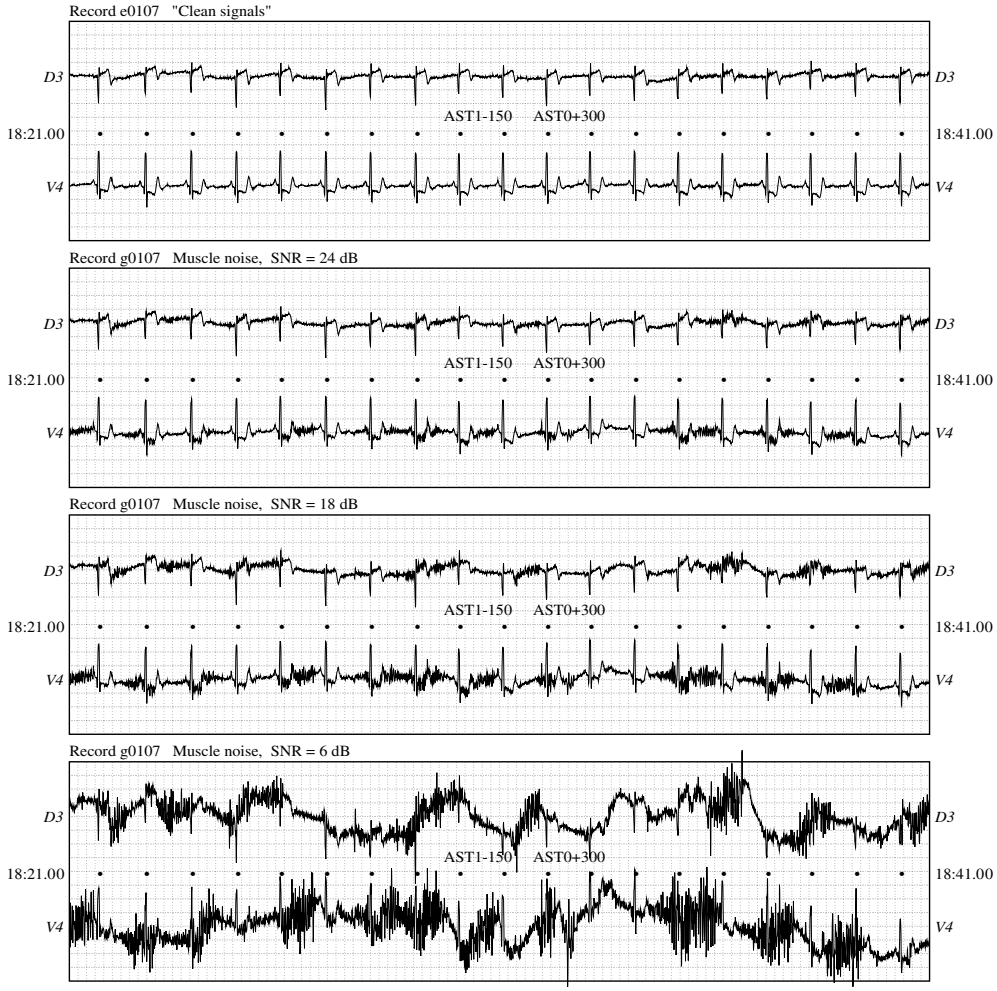

Figure 1. An example of input ECG signal during extrema of ischaemic ST segment episode (record e0107 of the ESC DB) contaminated with muscle noise artefacts at signal-to-noise ratios of 24 dB, 18 dB and 6 dB. Start time: 18:21 mm:ss. End time: 18:41 mm:ss.

shown in figure 1. The noise stress test procedure was performed in two variants. In the first variant, noise was added to signals prior to the ARISTOTLE arrhythmia detector analysis, while in the second variant, noise was added to signals immediately after the ARISTOTLE analysis. Using both variants of test permits us to determine to what extent the performance of ST analysis algorithm may be limited by that of the arrhythmia detector in the presence of noise. The results of the noise stress test for both variants are shown in figure 2. The results suggest that the ARISTOTLE operates accurately and does not influence significantly the performance of the ST segment analysis at all, until SNR $=$ 12 dB and 6 dB. The figure shows stable gross and average *IE Se* and *ID Se*, even for SNR $=$ 6 dB. Gross and average *IE* + *P* and *ID* + *P* stay above the critical performance boundaries until SNR $=$ 24 dB. This signal-to-noise ratio is the critical performance threshold. Performance statistics obtained during the first variant of the noise stress test in comparison with critical performance boundaries are summarized in table 2. Discrepant ST measurement percentage, $p_{(100\mu\mathrm{V})}$, and the value of error which 95%

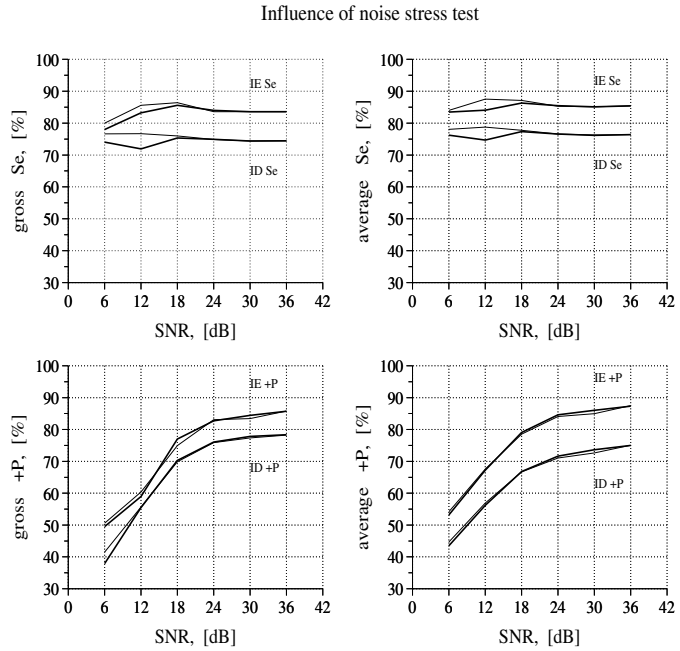

Figure 2. Performance of the KLT-based ST episode detection algorithm obtained during the noise stress test using the ESC DB. Bold lines: a variant when noise was added prior to ARISTOTLE's analysis. Thin lines: a variant when noise was added after ARISTOTLE's analysis.

of measurements did not exceed, $e_{(95\%)}$, both increase as the noise level increases. Reasonably low $p_{(100\mu\mathrm{V})}$ and $e_{(95\%)}$ appear for 36 dB, 30 dB and 24 dB. The percentage of extracted noisy heartbeats, p_{n}, at the critical performance threshold (24 dB) is approximately 21%. These results confirm the performance of the KLT-based algorithm above the critical performance boundaries with respect to the variation of input signals and thus confirms the robustness of the algorithm.

3.3. Is the algorithm robust with respect to the distribution of input signals?

Bootstrap distributions on the basis of 10 000 bootstrap trials of the performance of the KLT-based algorithm using the ESC DB are shown in figure 3. Distributions of gross *ID Se* and $ID + P$ are wider, and thus tend to be less robust performance measures than average *ID Se* and $ID + P$. Significant errors in long ST episodes had a negative influence on gross statistics, but not on the corresponding average statistics. A small number of relatively long but poorly detected ST episodes (poorly overlapped by analyser-annotated ischaemia) resulted in very low gross *ID Se*. Similarly, a small number of very long detections belonging to a relatively short ischaemic ST episodes resulted in very low gross $ID + P$. Table 3 shows the performance statistics obtained when using the bootstrap estimation of performance statistics in comparison with critical performance boundaries. The KLT-based algorithm seems to show good robustness. Its lowest expected performances (the 5% confidence limits) are close to (gross *ID Se* and average $ID + P$) or exceed the critical performance requirements. Differences between the raw statistics and the 5% confidence limits of the performance distributions are from 4.6% to 6.2%, while standard deviations of the performance statistics are from 2.8% to 3.5%. These figures confirm relatively narrow distributions with respect to the distribution of input signals and thus confirms the robustness of the algorithm.

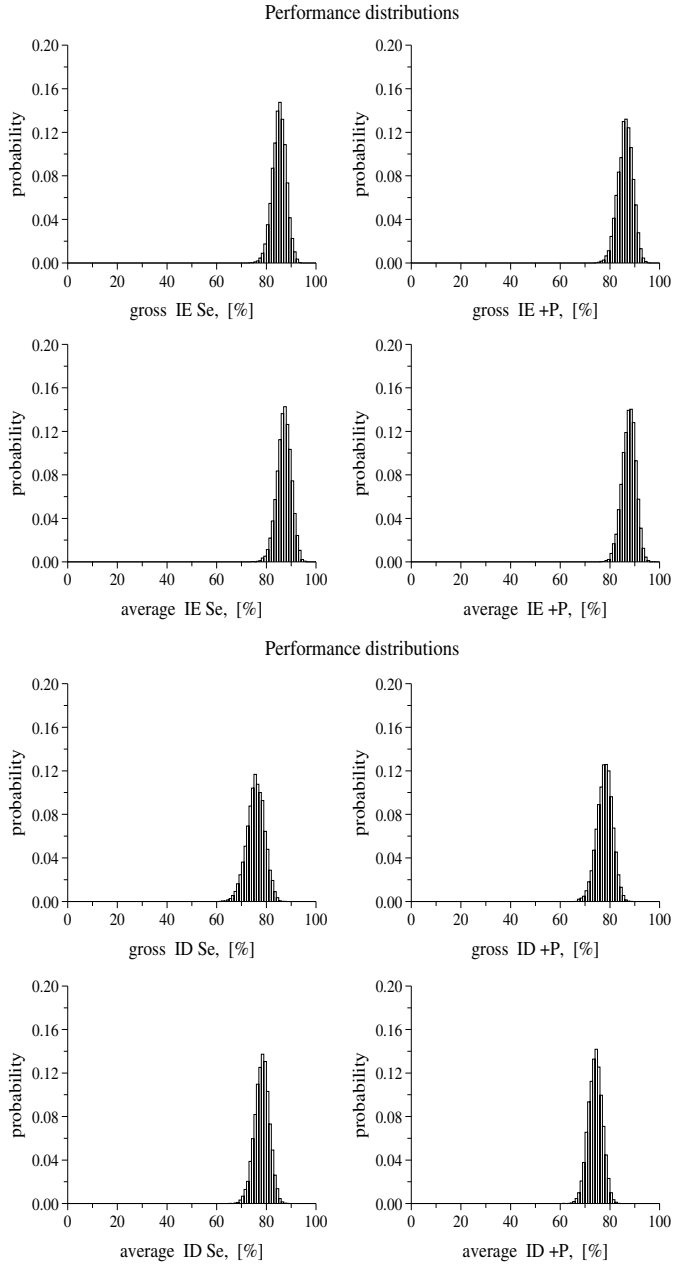

Figure 3. Bootstrap distributions (10 000 trials) of the aggregate performance statistics of the KLT-based ST episode detection algorithm using the ESC DB.

3.4. Is the algorithm robust with respect to the variation of its architecture parameters?

Sensitivity analysis was performed by modifying the most important architecture parameters of the algorithm: (1) the dimensionality, N, of the ST segment, $\mathbf{s}(k)$, and QRS complex, $\mathbf{q}(k)$, KLT feature vectors after the noise detection procedure, where k is the feature-vector sample

Table 3. Performance of the KLT-based ST episode detection algorithm obtained during the bootstrap estimation of performance distribution (10 000 bootstrap trials) and using the ESC DB. The bracketed figures are 5% confidence limits. Performances that do not meet the goals (critical performance boundaries) are boxed. (5%): 5% confidence limits of the performance distribution (lowest expected performance), ΔP: difference between raw performance and 5% confidence limits, Mean: mean of the distribution, SD: standard deviation of the distribution.

Measure (%)	Goal	(5%)	ΔP	Mean	SD	Raw
[g] *IE Se*	>80	(80.6)	4.6	85.2	2.8	85.2
[g] *IE + P*	>80	(81.1)	5.1	86.2	3.0	86.2
[g] *ID Se*	>70	(69.6)	6.2	75.7	3.5	75.8
[g] *ID + P*	>70	(72.5)	5.5	77.9	3.2	78.0
[a] *IE Se*	>80	(82.2)	4.9	87.1	2.9	87.1
[a] *IE + P*	>80	(82.9)	4.8	87.6	2.8	87.7
[a] *ID Se*	>70	(73.2)	5.0	78.2	3.0	78.2
[a] *ID + P*	>70	(69.3)	4.8	74.1	2.8	74.1

number; (2) feature-space boundaries and decision thresholds of noise detection and feature-vector trajectory recognition procedures and (3) by randomly perturbing the exact position of the ARISTOTLE's heartbeat fiducial point, $FP(j)$.

When studying the influence of modifying the dimensionality of feature vectors, we retained the original noise detection procedure incorporating five KLT coefficients. Thus we separated the noise detection and pattern recognition tasks. We wanted to study the influence of feature-vector dimensionality on the quality of feature representation and subsequently on those parts of the algorithm performing the pattern recognition task. We varied the number of KLT coefficients, N, from 2 to 8. The feature-space boundaries and decision thresholds of the trajectory-recognition procedures of the algorithm were in each case recalculated according to the feature space dimensionality ratio. Figure 4 shows the performance obtained when the dimensionality of feature vectors, N, is a parameter. Performance varies significantly, but the variations were smooth. The best performance is obtained when using five KLT coefficients. More significant differences in performance appear when using seven or eight KLT coefficients. Table 4 summarizes the performance statistics obtained during sensitivity analysis. Results are consistent. The performance statistics remained above the critical performance boundaries when the dimensionality of feature vectors, N, varied from 4 to 6.

Studying the influence of modifying the most important feature-space boundaries and decision thresholds of the KLT-based algorithm involved architecture parameters of the procedures for noise detection, for correcting the reference ST segment level, for detecting significant axis shifts and for detecting ST episodes. Changing these parameters (while keeping the dimensionality of feature vectors unchanged, $N = 5$) one-by-one up to $\pm 10\%$ did not influence the performance of the algorithm significantly. Results are summarized in table 4. The most sensitive parameter is the initial centre of the guard zone, $\Lambda_{c_0 N}$, i.e., a parameter of the algorithm responsible for classifying ST segment feature vectors as normal and deviating ones. Changing the decision threshold $\Lambda_{c_0 N}$ by -10% resulted, in the worst case, in a change of gross *IE + P* by -6.8% from raw statistics and a change of gross *ID + P* by -6.6%. Changing the decision threshold $\Lambda_{c_0 N}$ for +10% resulted, in the worst case, in a change of gross *IE Se* by -8.0%, and of gross *ID Se* by -8.5%. Otherwise, the performance remained strictly above the critical performance boundaries. Other architecture parameters are less sensitive. Next sensitive parameter is lower decision threshold, $\frac{2}{3}\Lambda_{c_0 N}$. A change of this parameter for +10% leads to a drop in the performance for less than 4% (gross *ID + P*).

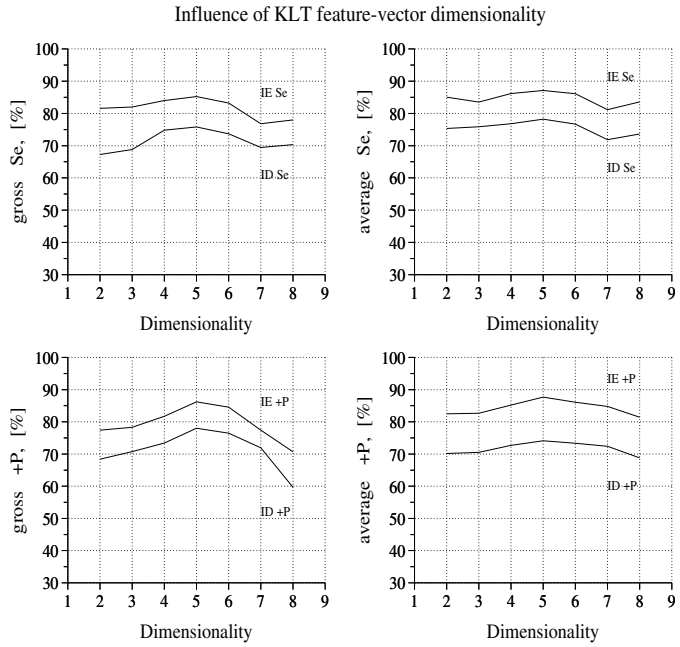

Figure 4. Performance of the KLT-based ST episode detection algorithm using the ESC DB for the modified dimensionality, N, of ST segment and QRS complex feature vectors.

Table 4. Performance of the KLT-based ST episode detection algorithm obtained during the sensitivity analysis and using the ESC DB. Performances that do not meet the goals (critical performance boundaries) are boxed. N: dimensionality of feature vectors, Λ_{c_0N}: initial centre of the 'guard zone', $[\pm n]$: interval to generate uniformly distributed fiducial point jitter.

		N		Λ_{c_0N}		$[\pm n]$	
Measure (%)	Goal	4	6	−10%	+10%	$n = 2$	Raw
[g] *IE Se*	>80	84.0	83.2	86.8	77.2	84.8	85.2
[g] *IE + P*	>80	81.7	84.6	79.4	88.3	78.1	86.2
[g] *ID Se*	>70	74.8	73.7	76.0	67.3	72.2	75.8
[g] *ID + P*	>70	73.4	76.5	71.4	80.6	63.9	78.0
[a] *IE Se*	>80	86.2	86.1	87.7	81.0	87.5	87.1
[a] *IE + P*	>80	85.2	86.1	83.5	89.3	82.7	87.7
[a] *ID Se*	>70	77.8	76.7	77.8	72.1	77.8	78.2
[a] *ID + P*	>70	72.7	73.3	70.7	77.0	70.8	74.1

To study the influence of fiducial point jitter, simulated random, uniformly distributed jitter in the interval $[-n, n], n = 0, \ldots, 8$, original signal samples around the ARISTOTLE's fiducial point, $FP(j)$, was introduced. The jitter was introduced immediately before applying the KLT basis functions to pattern vectors. Such a situation may be expected as a result of inaccurate time alignment of the pattern vector or of a suboptimal procedure for determining the position of the fiducial point. Other sources of jitter may be expected due to measurement noise in the signal or randomly varying waveforms. Fiducial point jitter in the pattern vectors affects the representational power and the classification performance. The study showed that

sensitivity and positive predictivity steadily decrease as the jitter, n, increases. Results of the study are summarized in table 4. The performance of the KLT-based algorithm remained close to (gross $IE + P$ and $ID + P$) and above the critical performance boundaries when introducing a jitter in the window of ± 2 original signal samples (± 8 ms) around the fiducial point. A significant drop in performance occurs at $n = 4$.

These measurements confirm that architecture parameters do not critically affect the performance as they are changed within some range around their optimal values and suggest that the KLT-based algorithm is robust with respect to the variation of its architecture parameters.

4. Discussion and conclusions

Robustness of ST analysers is an important evaluation question and need to be assessed in the light of various procedures. In this paper, we presented principles and methods for assessing the robustness of ST episode detection and quantification analysers and algorithms. We defined an evaluation protocol, performance measures and procedures to assess the robustness. In the case study, we demonstrated how robustness of an ST analyser can be quantitatively evaluated using a standard database. Other systems to detect ischaemic ST episodes (Taddei *et al* 1995, García *et al* 2000, Stadler *et al* 2001, Maglaveras *et al* 1998, Silipo and Marchesi 1998) have been evaluated in terms of performance, but not in terms of robustness, therefore their robustness cannot be compared.

The KLT-based ischaemic ST episode detection and quantification algorithm showed good noise immunity. The algorithm is not critically dependent on the variation of the noise content of input signals. Adding noises to input signals in two variants, prior to and after ARISTOTLE's analysis, tested ARISTOTLE's stability and immunity to noise and exposed its influence to overall robustness. ARISTOTLE has practically no influence on performance. Performance characteristics are almost equal in both variants.

Bootstrap analysis showed that the performance of the KLT-based algorithm is not critically dependent on the choice of database used for testing. Relatively narrow distributions of the statistics suggest that the algorithm is marginally robust. Apart from comparing the robustness of different algorithms, the bootstrap is also useful for comparing the robustness of different performance statistics. Distributions of average performance statistics are narrower than those of gross statistics. Thus, the average statistics appear to be more robust estimates of the performance of the algorithm than the corresponding gross statistics, particularly for ischaemic ST duration statistics. Gross ischaemic ST duration statistics are extraordinarily sensitive to single errors and thus less robust estimators of performance than average statistics. Significant errors in a small number of long episodes have a disproportionately negative influence on gross statistics, but not on the corresponding average statistics.

Sensitivity analysis proved that the algorithm is not critically tuned to the database used for testing (ESC DB). The optimal choice for the dimensionality of feature vectors to distinguish between noisy and non-noisy events in the domain of ST episode detection during ambulatory ECG monitoring has been estimated previously (Jager 1994, Jager *et al* 1998). The five KLT coefficients for the ST segment and the five for QRS complex feature vectors were estimated as the optimal (sufficient and necessary) choice for this task. When changing the dimensionality of feature vectors in this study, we retained the dimensionality of feature vectors (five) for the noise detection procedure. Besides assessing the robustness, the sensitivity analysis allowed us to assess the influence of the modified feature-vector dimensionality on the quality of ST segment and QRS complex morphology-feature representation, and subsequently on those parts of the algorithm performing pattern recognition. The present study suggests that

a dimensionality of five is also the optimal choice for feature representation and pattern recognition part of the KLT-based algorithm. Furthermore, sensitivity analysis showed that changing feature-space boundaries and decision thresholds of the algorithm does not influence the performance of the algorithm significantly. The influence of fiducial point jitter was not critical.

Primary motivation for including sensitivity analysis into the robustness protocol was to make an objective comparison of ST analysers possible. Knowing which architecture parameters of a given ST analyser are relevant to be modified in order to 'force' the analyser's sensitivity to the same value as of the other ST analysers, and then comparing their positive predictivities, allows direct comparison of the performance of the analysers.

The evaluation protocol, procedures and performance measures to assess the robustness of ST analysers and algorithms appear to be practical, useful and informative. We conclude that authors would need to publish data that reveal how fragile their ST analyser might be. Adopting the robustness protocol and measures together with a standard database across groups of investigators makes comparison of robustness of ST analysers and algorithms possible. We finally conclude that comparable robustness estimates should be made for many other computer applications in medicine.

References

Albrecht P, Moody G B and Mark R G 1988 Use of the 'bootstrap' to assess the robustness of the performance statistics of an arrhythmia detector *J. Ambulatory Monit.* **1** 171–6

Association of the Advancement of Medical Instrumentation/American National Standard Institute 1998a Ambulatory electrocardiographs ANSI/AAMI EC38 Arlington, VA, USA

Association of the Advancement of Medical Instrumentation/American National Standard Institute 1998b Testing and reporting performance results of cardiac rhythm and ST segment measurement algorithms ANSI/AAMI EC57 Arlington, VA, USA

Efron B 1979 Bootstrap methods: another look at the jackknife *Ann. Stat.* **7** 1–26

García J, Sörnmo L, Olmos S and Laguna P 2000 Automatic detection of ST-T complex changes on the ECG using filtered RMS difference series: application to ambulatory ischemia monitoring *IEEE Trans. Biomed. Eng.* **47** 1195–201

Goldberger A L, Amaral L A N, Glass L, Hausdorff J M, Ivanov P C, Mark R G, Mietus J E, Moody G B, Peng C K and Stanley H E 2000 PhysioBank, PhysioToolkit, and PhysioNet components of a new research resource for complex physiologic signals *Circulation* **101** e215–20

Jager F 1994 Automated detection of transient ST-segment changes during ambulatory ECG-monitoring, PhD Thesis, University of Ljubljana, Faculty of Electrical and Computer Engineering, Ljubljana, Slovenia

Jager F 1998 Guidelines for assessing performance of ST analysers *J. Med. Eng. Techn.* **22** 25–30

Jager F, Moody G B, Divjak S and Mark R G 1994 Assessing the robustness of algorithms for detecting transient ischemic ST segment changes *Comput. Cardiol.* 229–32

Jager F, Moody G B and Mark R G 1998 Detection of transient ST segment episodes during ambulatory ECG monitoring *Comput. Biomed. Res.* **31** 305–22

Jager F, Moody G B, Taddei A and Mark R G 1991 Performance measures for algorithms to detect transient ischemic ST segment changes *Comput. Cardiol.* 369–72

Maglaveras N, Stamkopoulos T, Pappas C and Strintzis M G 1998 An adaptive backpropagation neural network for real-time ischemia episodes detection: development and performance analysis using the European ST-T database *IEEE Trans. Biomed. Eng.* **45** 805–13

Moody G B and Mark R G 1982 Development and evaluation of a 2-lead ECG analysis program *Comput. Cardiol.* 39–44

Moody G B and Mark R G 1990 The MIT-BIH arrhythmia database on CD-ROM and software for use with it *Comput. Cardiol.* 185–8

Moody G B, Mark R G and Goldberger A L 2000 A research resource for studies of complex physiologic and biomedical signals *Comput. Cardiol.* 179–84

Moody G B, Muldrow W K and Mark R G 1984 A noise stress test for arrhythmia detectors *Comput. Cardiol.* 381–4

Silipo R and Marchesi C 1998 Artificial neural networks for automatic ECG analysis *IEEE Trans. Signal Proc.* **46** 1417–25

Stadler R W, Lu S N, Nelson S D and Stylos L 2001 A real-time ST segment monitoring algorithm for implantable devices *J. Electrocardiol.* **34** 119–26

Taddei A, Costantino G, Silipo R, Emdin M and Marchesi C 1995 A system for the detection of ischemic episodes in ambulatory ECG *Comput. Cardiol.* 705–8

Taddei A, Distance G, Emdin M, Pisani P, Moody G B, Zeelenberg C and Marchesi C 1992 The european ST-T database: standard for evaluating systems for the analysis of ST-T changes in ambulatory electrocardiology *Eur. Heart J.* **13** 1164–72