An Open-Source Tool to Evaluate Performance of Transient ST Segment Episode Detection Algorithms

F Jager[1,2], A Smrdel[2], R G Mark[1]

[1]Harvard-MIT Division of Health Sciences and Technology, Cambridge, MA, USA
[2]Faculty of Computer and Information Science, University of Ljubljana, Ljubljana, Slovenia

Abstract

Performance measures and evaluation protocols for evaluating the performance and robustness of transient ST segment episode detection algorithms are specific, complex and not trivial to realize. We developed an open-source tool (EVAL_ST) to evaluate and compare performance and robustness of ST episode detection algorithms. The tool supports all standard and other relevant performance measures, aggregate gross and average statistics, and bootstrap statistical procedure to predict real-world clinical performance. The tool (written in C) is compilable an a wide variety of platforms and contains an additional graphic user interface module (LessTif/Motif environment) for use on the LINUX/UNIX operating systems.

1. Introduction

Assessing the *performance* and *robustness* of *ST segment analysers* and *algorithms* as well as predicting their behavior in the real-world clinical environment is a difficult task. Availability of the the *European Society of Cardiology ST-T Database* (ESC DB) [1] gained development of transient ST segment episode detectors and allowed comparison of their performance. Newly developed *Long-Term ST Database* (LTST DB) [2] provides a wide variety of real-world 24-hour ambulatory records with numerous examples of transient ischemic ST segment episodes and transient non-ischemic heart-rate related ST segment episodes. It gained further development and evaluation of transient ST episode detectors. Due to relative complexity of the performance measures and of the evaluation protocol, we developed an open-source tool EVAL_ST, Version 2.0, to objective evaluate and compare the performance and robustness of transient ST episode detection algorithms. We initially adapted previously developed performance measures [3, 4, 5] by adding performance matrices to differentiate between ischemic and non-ischemic heart-rate related ST segment episodes for use with the LTST DB.

2. Performance measures

Evaluation of an ST segment episode detection algorithm should answer the following questions:
- How well are ST episodes detected?
- How well are ischemic and non-ischemic heart-rate related ST episodes differentiated?
- How reliably are ST episode or ischemic ST episode duration measured?
- How accurately are ST deviations measured?
- How well will the ST algorithm perform in the real world?

Transient ST segment episodes (the events of interest) are characterized by: 1) number, 2) length, and 3) extrema deviation. When evaluating multi-channel ST-algorithm performance, the ST annotation stream for all leads must be combined into one reference stream using a logical OR function. The fact that at any given time there is either an ST episode or an interval with no ST deviation implies the use of two-by-two performance evaluation matrices. We further assume that all ST episodes are equally important. Evaluation of ST episode detection algorithms consists of comparing algorithm-annotated episodes with reference-annotated episodes. Algorithm- and reference-annotated episodes may differ considerably in length, there is not a one-to-one correspondence between the episodes from the two groups, nor non-events can be counted.

Sensitivity matrix (see figure 1, left) summarizes how the reference ischemic ST episodes were labelled by the algorithm, i.e., how many of the reference ST episodes were detected, TP_S, and how many were missed, FN. The positive predictivity matrix (figure 1, right) summarizes how many of the algorithm-annotated ST episodes were actually ST episodes, TP_P, and how many were falsely detected, FP. The performance measures to assess ability to detect ST episodes depend on the concept of *matching* [3]. In measuring sensitivity, we declared that matching of a reference ST episode occurs when the period of overlap includes at least one of the extrema of the reference ST episode, or at least one-half of the length of the reference ST episode. In measuring positive predictivity,

Se matrix		*Algorithm*	
		ST epis	*Not epi*
Refe-rence	*ST epis*	TP_S	*FN*
	Not epi	–	–

+P matrix		*Algorithm*	
		ST epis	*Not epi*
Refe-rence	*ST epis*	TP_P	–
	Not epi	*FP*	–

Figure 1. ST episode sensitivity matrix (left) and ST episode positive predictivity matrix (right).

we declared that matching of an algorithm-annotated ST episodes occurs when the period of overlap includes the extrema of the algorithm-annotated ST episode, or at least one-half of the length of the algorithm-annotated ST episode [3].

ST episode detection sensitivity, $SE\ Se$, an estimate of the likelihood of detecting an ST episode, is defined as:

$$SE\ \ Se = \frac{TP_S}{TP_S + FN} \ . \tag{1}$$

The denominator is the number of reference ST episodes. TP_S is the number of matching episodes, and FN is the number of non-matching episodes.

ST episode detection positive predictivity, $SE\ \ +P$, an estimate of the likelihood that a detection is a true ST episode, is defined as:

$$SE\ \ +P = \frac{TP_P}{TP_P + FP} \ . \tag{2}$$

The denominator is the number of ST episodes annotated by the algorithm. TP_P is the number of matching episodes, and FP is the number of non-matching episodes.

In differentiating ischemic and non-ischemic heart-rate related ST episodes, we assumed that at any given time there is only one type of episode: ischemic, non-ischemic heart-rate related, or an interval without significant ST deviation, which implies three-by-three performance evaluation matrices (see figure 2). Each reference- and algorithm-annotated episode is submitted to the *extended matching test*. The test is the same as defined previously for ST episodes, but extended in the sense that matching of an episode (ischemic or heart-rate related) occurs when the episode is *sufficiently* and *uniquely* overlapped by ischemic or by heart-rate related ST episodes. The status of a reference ST episode (ischemic or heart-rate related) after the matching test is "ischemic" if the episode is sufficiently overlapped by ischemic algorithm-annotated ST episodes. Otherwise, if the episode is sufficiently overlapped by heart-rate related algorithm-annotated ST episodes, its status is "heart-rate related". If the episode is not sufficiently overlapped by ischemic or heart-rate related episodes, its status is "missed". The status of an algorithm-annotated ST episode (ischemic or heart-rate related) is "ischemic" if the episode is sufficiently

Se matrix		*Algorithm*		
		Isch	*HR rel*	*Not epi*
Ref	*Isch*	*a*	*b*	*c*
	HR rel	*d*	*e*	*f*
	Not epi	–	–	–

+P matrix		*Algorithm*		
		Isch	*HR rel*	*Not epi*
Ref	*Isch*	*g*	*h*	–
	HR rel	*i*	*j*	–
	Not epi	*k*	*l*	–

Figure 2. Performance matrices assessing the ability of an ST episode detection algorithm to differentiate ischemic (*Isch*) and non-ischemic heart-rate related (*HR rel*) ST episodes.

overlapped by ischemic reference ST episodes. Otherwise, if the episode is sufficiently overlapped by heart-rate related reference ST episodes, its status is "heart-rate related". If the episode is not sufficiently overlapped by ischemic or heart-rate related episodes, its status is "falsely detected". The sensitivity matrix (figure 2, left) describes how many reference ischemic, a, and heart-rate related, e, ST episodes were correctly detected. b is the number of reference ischemic episodes detected as heart-rate related, and d is the number of reference heart-rate related episodes detected as ischemic. c and f are the numbers of missed ischemic and heart-rate related episodes respectively. The positive predictivity matrix (figure 2, right) describes how many of the algorithm's ischemic, g, and heart-rate related, j, ST episode detections were actually ischemic and heart-rate related episodes. h is the number of the algorithm's heart-rate related episode detections which actually are reference ischemic episodes, and i is the number of the algorithm's ischemic episode detections which actually are reference heart-rate related episodes. k and l are the numbers of falsely detected ischemic and heart-rate related episodes respectively.

Furthermore, if we consider both ischemic and heart-rate related changes together as ST-change episodes of unique type, then the performance matrices can easily be reduced to two-by-two, with: $TP_S = a + b + d + e$, $TP_P = g + h + i + j$, $FN = c + f$, and $FP = k + l$, yielding performance matrices in figure 1. Since the events of clinical interest are the ischemic ST episodes, we can further consider all non-ischemic heart-rate related ST episodes as episodes of no deviation. This consideration yields: $TP_S = a$, $TP_P = g$, $FN = b + c$, and $FP = i + k$, and leads to the ischemic ST episode detection sensitivity, $IE\ \ Se$, and ischemic ST episode detection positive predictivity, $IE\ \ +P$, which are defined as for the $SE\ \ Se$ and $SE\ \ +P$ (equations 1 and 2).

ST episode duration detection sensitivity, $SD\ \ Se$, defined as the fraction of true ST episode duration detected:

$$SD\ \ Se = \frac{SD_{R \wedge A}}{SD_R} \ , \tag{3}$$

and ST episode duration detection positive predictivity, $SD\ \ +P$, defined as the fraction of algorithm-annotated ST episode duration which is true ST episode:

$$SD\ \ +P = \frac{SD_{R \wedge A}}{SD_A}\ , \tag{4}$$

are estimates of the accuracy with which an algorithm can measure the duration of ST episodes within the observation period. $SD_{R \wedge A}$ is the total duration of algorithm-annotated ST episodes which overlaps reference ST episodes, and SD_R and SD_A are the total durations of reference- and algorithm-annotated ST episodes respectively. Similarly, ischemia duration detection sensitivity, $ID\ \ Se$, and ischemia duration detection positive predictivity, $ID\ \ +P$, are defined using total duration of algorithm-annotated ischemia, ID_A, which overlaps reference ischemia, ID_R, and their overlap, $ID_{R \wedge A}$.

Accuracy of ST-deviation measurement of the extrema of ST episodes is usually summarized by a scatter plot of reference versus test measurements. Other useful summary statistics are: mean error between the algorithm and reference measurements, standard deviation of errors, correlation coefficient, linear regression, the value of error which 95% of the measurements do not exceed, $e_{(95\%)}$, and the percentage of measurements for which the absolute difference between the algorithm and reference measurement is grater than 100μV, $p_{(100\mu V)}$ [3].

Techniques to predict real-world clinical performance are aggregate gross and average performance statistics, and bootstrap estimates of aggregate performance statistics [6].

3. Open-source tool

Advantage of the open-source technology is that a large number of researchers from the world scientific community can read, verify, modify, improve, adapt, and redistribute source code. Characteristics of an open-source software may be summarized following: the software is running on many platforms, it is able to run with other software, and its internal specifications and protocols are public so that others can develop software using the same protocols and environment.

The tool EVAL_ST developed provides first- (record-by-record) and second-order (aggregate gross and average) performance statistics for evaluation and comparison of transient ST episode detection algorithms. The tool allows assessing the accuracy of: 1) detecting transient ST episodes, 2) distinguishing between ischemic and non-ischemic heart-rate related ST episodes, 3) measuring ST episode durations and ischemic ST episode durations, and 4) measuring ST segment deviations. The tool also provides generation of performance distributions using a bootstrap statistical technique [6] for predicting real-world clinical performance and robustness.

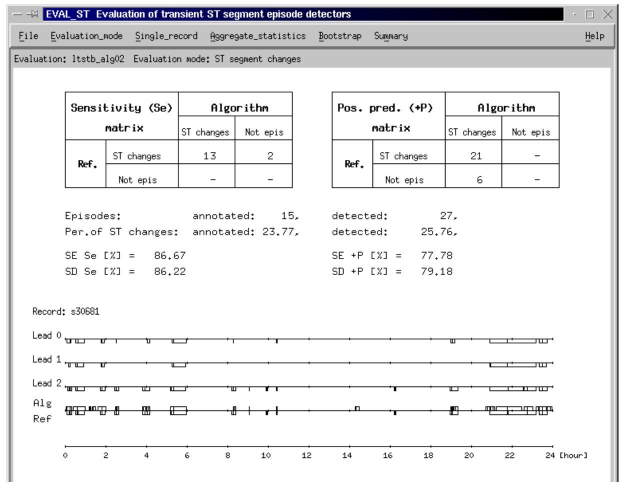

Figure 3. Graphic user interface of the EVAL_ST tool. Evaluation of the ST episode detection algorithm from [8] using the record s30681 of the LTST DB is shown. ST episode detection performance matrices are shown upper, summary statistics are in the middle, while schematically shown ST annotation streams are at the bottom.

The tool is written is C programming language and compilable on any platform running standard C (or C++) compiler. The core module of the tool supports command-line oriented (no graphic display) user interface style thus enabling possible batch processing. Input to the tool are ST segment annotation streams of a reference database (e.g., LTST DB or ESC DB) and ST segment annotation streams of the evaluated algorithms. Evaluation results are stored to output files. An additional graphic user interface module (LessTif/Motif graphic environment) provides graphic display of the evaluation results on LINUX/UNIX operating system. Figure 3 shows the graphic user interface of the tool with an example of evaluation. The main window of the graphic user interface provides display of: overlapping reference and algorithm's ST episode annotation streams of the records, first- and second-order performance statistics, performance matrices regarding ST episode detection and differentiation between ischemic and non-ischemic heart-rate related ST episodes, statistics regarding ST episode and ischemic ST episode duration detection, scatter plot of ST segment deviation measurements, and bootstrap estimates of expected real-world performance with performance distributions. Detailed evaluation results are stored to output files and also graphically displayed in secondary windows of the tool. The tool has been made freely available via the PhysioNet [7] (*http://www.physionet.org*) and via the home page of the Laboratory of Biomedical Computer Systems and Imaging (*http://mimi.fri.uni-lj.si*) of the University of Ljubljana. Besides the source code and the reference annotations of

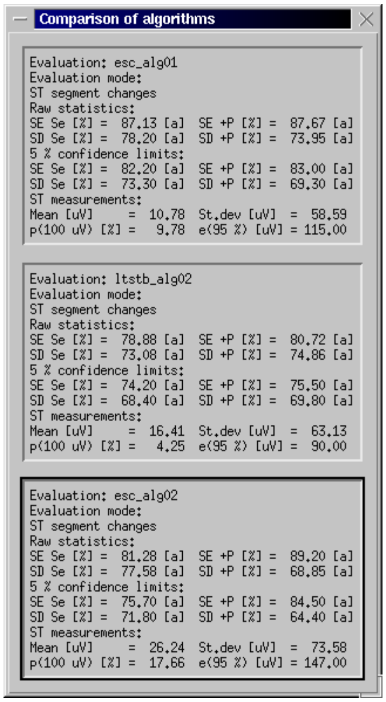

Figure 4. The “Comparison” window of the EVAL_ST tool. Evaluation results shown are of the ST episode detection algorithm from [9] using the ESC DB as development set (top), of the ST episode detection algorithm from [8] using the LTST DB (annotation protocol B) as development set (middle), and of the ST algorithm from [8] using the ESC DB as test set (bottom).

the LTST DB and ESC DB, the annotations of our two ST episode detection algorithms [8, 9], developed and tested using the LTST DB and ESC DB, are also provided.

4. Case study

Figure 4 shows “Comparison” window of the tool summarizing and comparing evaluations performed. Performances shown are of our two ST episode detection algorithms [8, 9] using the LTST DB and ESC DB.

5. Discussion and conclusions

We adapted previously developed performance measures to evaluate transient ST episode detection algorithms to the new ST episode annotation protocol of the LTST DB by adding performance matrices to assess ability of ST episode detection algorithms to differentiate between ischemic and non-ischemic heart-rate related ST segment episodes; and developed graphically supported evaluation tool. The tool is efficient, usable and allows objective evaluation. Command-line oriented user interface style provides maximum flexibility for investigators and developers regarding compiling and use in batch processing, while graphic user interface of the tool provides easy and user-friendly use.

We successfully used the tool for evaluating our two ST episode detection algorithms [8, 9] using the LTST DB and ESC DB. Availability of the tool to the world community simplifies as well as promote easy and unique use of specific and complex performance measures and evaluation protocol, helps to easily compare different algorithms, encourages the use of unique performance measures and evaluation protocols in the field, and gain the consistent use of standard performance measures.

References

[1] Taddei A, Distante G, Emdin M, Pisani P, Moody GB, Zeelenberg C, Marchesi C. The european st-t database: standard for evaluating systems for the analysis of st-t changes in ambulatory electrocardiography. Eur Heart J 1992;13:1164–1172.

[2] Jager F, Taddei A, Moody GB, Emdin M, Antolič G, Dorn R, Smrdel A, Marchesi C, Mark RG. Long-term st database: a reference for the development and evaluation of automated ischaemia detectors and for the study of the dynamics of myocardial ischaemia. Med Biol Eng Comput 2003;41:172–182.

[3] Jager F, Moody GB, Taddei A, Mark RG. Performance measures for algorithms to detect transient ischemic st segment changes. In Computers in Cardiology 1991. Los Alamitos: IEEE Computer Society Press, 1992; 369–372.

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

[5] Jager F. Guidelines for assessing performance of st analysers. J Med Eng Techn 1998;22:25–30.

[6] Albrecht P, Moody GB, Mark RG. Use of the “bootstrap” to assess the robustness of the performance statistics of the arrhythmia detector. J Amb Monit 1988;1:171–176.

[7] Goldberger AL, Amaral LAN, Glass L, Hausdorff JM, Ivanov PC, Mark RG, Mietus JE, Moody GB, Peng CK, Stanley HE. Physiobank, physiotoolkit, and physionet components of a new research resource for complex physiologic signals. Circulation 2000;101:e215–e220.

[8] Smrdel A, Jager F. Automated detection of transient st-segment episodes in 24h electrocardiograms. Med Biol Eng Comput 2004;42:303–311.

[9] Jager F, Moody GB, Mark RG. Detection of transient st segment episodes during ambulatory ecg monitoring. Comput Biomed Res 1998;31:305–322.

Address for correspondence:

Franc Jager
University of Ljubljana / Faculty of Comp. and Inf. Science
Tržaška 25 / 1000 Ljubljana / Slovenia
tel./fax: +386-1-4768-780/4264-647
franc.jager@fri.uni-lj.si