Proposed New Requirements for Testing and Reporting Performance Results of Arrhythmia Detection Algorithms

John Wang

Philips Healthcare, Andover, MA, USA

Abstract

Several arrhythmia performance measures as specified in the current AAMI recommended practice document do not adequately reflect actual clinical experience in real-time patient monitoring. Additional reporting requirements that are more clinical relevant are thus needed: 1) Due to the large number of QRS complexes that need to be analyzed, even an algorithm with high specificity will generate a large number of false positives, which is not directly reflected by the reported false positive rate. A new recommendation is to report the actual number of false positives. 2) Due to the low PVC prevalence in most monitored patients, the high positive predictive values (PPV) reported using databases with high PVC prevalence are not clinically relevant. A proposed recommendation is to report PPVs at much lower PVC rates. 3) Arrhythmia performance measures as specified in the recommended practice do not include performance bounds. A new proposal is to use the bootstrap method with sample replacement to generate the mean values and 95% confident intervals for all the reported arrhythmia performance measures. Conclusions: Additional performance measures are proposed to further improve the clinical relevance of the reported results. These measures should be considered for inclusion as part of the standard reporting.

1. Introduction

Electrocardiographic monitoring, which allows for continuous non-invasive detection and documentation of cardiac arrhythmia, is one of the most frequently used monitoring procedures for managing in-hospital patients. Current commercial systems are designed to detect most of the ventricular arrhythmias and some of the atrial arrhythmias for patients of all age groups.

In 1987, AAMI published a recommended practice document for testing and reporting ventricular arrhythmia detection algorithm performance [1]. This document was developed to assist in the comparison of arrhythmia algorithm performance using standardized testing procedures and performance reporting. Although other changes, such as testing for some atrial arrhythmias, ST-segment measurement, and heart rate variability, have been incorporated into subsequent revisions, ventricular arrhythmia performance reporting has largely remained unchanged [2]. While the original document seeks to define performance measures that are relevant to the clinical practice, over the years it has become clear that some of the performance measures do not adequately reflect clinical experience in using these algorithms for real-time monitoring application [3,4]. Additional reporting requirements that are more clinical relevant are thus needed.

2. Current performance reporting

2.1. Performance measures

The standard statistical performance measures used in reporting test results and their definitions are summarized in Table 1. As shown in the table, the efficacy of a test is entirely captured by the following four basic measurements: true positive (TP), false negative (FN), false positive (FP), and true negative (TN) (presented in a 2x2 contingency sub-table). From these four basic measurements, all the other statistical measures can then be derived.

Sensitivity (Se) indicates the ability of a test to identify positive cases; a test with high sensitivity has few false negatives. Specificity (Sp) indicates the ability of a test to identify negative cases; a test with high specificity has few false positives. Positive predictive value (PPV) provides the probability of being true positive when the test is positive. Negative predictive value (NPV) provides the probability of being true negative when the test is negative. Likelihood ratio positive (LR+) and likelihood ratio negative (LR–), which combine both the sensitivity and specificity of the test, provide estimates of how much the result of a test will change the odds of being positive and negative, respectively.

Overall accuracy, which combines the true positive and true negative into a single performance measure, is

Table 1. Summary of statistical performance measures and their definitions used in reporting test results.

Test Classification	Reference Classification		Total	Performance Measures
	Positive	Negative		
Positive	True Positive **TP**	False Positive **FP**	All Positive Test Cases **TP + FP**	Positive Predictive Value (PPV) **TP / (TP + FP)**
Negative	False Negative **FN**	True Negative **TN**	All Negative Test Cases **FN + TN**	Negative Predictive Value (NPV) **TN / (FN + TN)**
Total	All Positive Cases **TP + FN**	All Negative Cases **FP + TN**	All Cases **TP+FN+FP+TN**	Overall Accuracy **(TP+TN) / (TP+FN+FP+TN)**
			Prevalence **(TP+FN) / (TP+FN+FP+TN)**	
Performance Measures	Sensitivity (Se) **TP / (TP + FN)**	False Positive Rate = 1 – Specificity **FP / (FP + TN)**	Likelihood Ratio Positive (LR+) **Sensitivity / (1 – Specificity)**	
	False Negative Rate = 1 – Sensitivity **FN / (TP + FN)**	Specificity (Sp) **TN / (FP + TN)**	Likelihood Ratio Negative (LR–) **(1 – Sensitivity) / Specificity**	

not sufficient to describe the performance of a test [5]. For a complete description, both sensitivity and specificity need to be reported. In addition, for most applications, positive predictive value is often included as part of the test performance reporting. Because once a test is positive, one is interested in knowing the predictive value of the test, namely, the likelihood (or probability) that the positive test is indeed a positive case.

2.2. AAMI recommended practice

AAMI recommended practice document [1,2] provides a standardized performance testing and reporting for arrhythmia detection algorithms. The performance measures specified in the recommended practice document for reporting of the PVC detection accuracy are: 1) sensitivity, 2) false positive rate, and 3) positive predictive value. The reason for selecting the false positive rate, instead of specificity, is to make it easier to estimate the actual number of false positives.

2.3. State-of-the-art performance results

As an example of using the performance measures as specified by the AAMI recommended practice, the PVC detection performance results from several versions of a commercial monitoring system [6] are summarized in Table 2. The tests were done using the publically available non-paced AHA and MIT-BIH arrhythmia databases. While these results represent the current state-of-the-art performance for real-time patient monitoring, many false positives will occur due to the large number of QRS complexes to be analyzed. In addition, the PPV will also be lower due to the low PVC prevalence for most of the monitored patients.

Table 2. PVC detection performance reported using the AAMI recommended performance measures.

Database	Sensitivity	Specificity (F+ Rate)	PPV
AHA (78 records)	95.82-97.24	99.83-99.85 (0.15-0.17)	98.26-98.53
MIT-BIH (44 records)	94.11-94.39	99.73-99.80 (0.21-0.27)	96.19-97.17
Combined	94.11-97.24	99.73-99.85 (0.15-0.27)	96.19-98.53

3. Proposed new reporting requirements

3.1. False positive rate

Due to the large number of QRS complexes that need to be analyzed in continuous real-time monitoring (~100,000 complexes/patient/day at an average heart rate of 70), even an algorithm with very high specificity (or very low false positive rate) will generate a large number of false positives. Currently used performance measure, false positive rate, only indirectly reflects this fact. Table 3 shows the actual numbers of false positive PVCs detected at average heart rates of 70 and 140 for several specificity values and false positive rates. As shown in the table, for an algorithm with specificity of 99.5% (or false positive rate of 0.5%) there will be 500 and 1,000 false positive PVCs detected per patient per day at average heart rates of 70 and 140, respectively.

Table 3. Numbers of false positive PVC detected for several specificity values and heart rates.

Specificity (F+ Rate)	F+ PVCs/Patient/Day	
	HR = 70	HR = 140
95.0% (5.0%)	5,000	10,000
99.0% (1.0%)	1,000	2,000
99.5% (0.5%)	500	1,000
99.8% (0.2%)	200	400

A new proposed recommendation is to report the actual number of false-positives/patient/day at heart rates of 70 and 140. With these reported numbers, it is easy to further estimate the total number of potential false positives a real-time monitoring system will generate by multiplying the reported false positive numbers by the total number of monitored beds in the care unit.

3.2. Positive predictive value

Unlike sensitivity and specificity, which are independent of the prevalence of the condition being tested, PPV depends on the prevalence. Lower prevalence results in lower PPV and thus a less reliable positive test.

Given the performance of a test as specified by the sensitivity and specificity, the PPV as a function of the prevalence can be calculated using the following equation:

$$PPV = \frac{[Se/(1-Sp)]\times[prevalence/(1-prevalence)]}{1+[Se/(1-Sp)]\times[prevalence/(1-prevalence)]}$$

As an example to show the relationship of prevalence and PPV, three separate tests, each with 10,000 cases, are conducted. The TP, FN, FP and TN numbers for all three tests are shown in Table 4. From these numbers, the sensitivity, specificity, PPV, and prevalence are calculated and summarized in Table 5. The results show that while all tests have the same sensitivity (95%) and specificity (95%), the PPVs are very different. The PPVs are 95%, 68%, and 16% for prevalence levels of 50%, 10%, and 1%, respectively. The PPV depends on the pretest prevalence. A low prevalence yields a low PPV.

While for many other applications, one can potentially achieve a higher PPV by selecting only cases to be tested with high pretest likelihood (prevalence) [7]. However, for real-time patient monitoring this may not be an option, since very often all patients are monitored regardless whether they are likely to have arrhythmia or not [8-10].

In Table 2, the reported PPVs are very high when tested using the AHA and MIT-BIH databases. These high values are due to the high PVC prevalence values of the two testing databases. As shown in Table 6, the PVC prevalence values are 9.2% and 7.0% for the AHA and MIT-BIH databases, respectively. In actual clinical application, the PVC rate will be much lower and hence a much lower PPV.

In general, in order to accurately measure the test performance in terms of sensitivity and specificity, the database used in the test must contain sufficient number of positive cases. This high prevalence rate may not always reflect the prevalence of the actual targeted clinical application. To obtain the PPV for the targeted application, the PPV must be estimated using the actual prevalence of the targeted environment.

Therefore, to provide a more accurate PPV for a PVC detection algorithm for real-time patient monitoring application, a proposed recommendation is to report positive predictive values for PVC prevalence at levels of 50, 10, and 1 per hour. Table 7 shows the PPVs for sensitivity of 95% and several values of specificity for the recommended values of PVC prevalence.

Table 5. Summary of test performance calculated from results provided in Table 4.

Performance Measure	Test #1	Test #2	Test #3
Sensitivity	95%	95%	95%
Specificity	95%	95%	95%
Positive Predictive Value	95%	68%	16%
Prevalence	50%	10%	1%

Table 4. Summary of test results for an example of understanding the relationship between positive predictive value and prevalence.

Test Result	Test #1 (10,000 Cases)		Test #2 (10,000 Cases)		Test #3 (10,000 Cases)	
	Disease Positive (5,000 Cases)	Disease negative (5,000 Cases)	Disease Positive (1,000 Cases)	Disease negative (9,000 Cases)	Disease Positive (100 Cases)	Disease negative (9,900 Cases)
Positive	4,750	250	950	450	95	495
Negative	250	4,750	50	8,550	5	9,405

Table 6. Characteristics of the non-paced AHA and MIT-BIH arrhythmia databases.

Database Characteristics	Database		
	AHA	MIT-BIH	Combined
No. of Records	78	44	132
Total Duration (Hrs)	39	22	61
No. of QRS	176,722	100,731	277,433
No. of PVC	16,261	7,008	23,269
PVC Prevalence	9.2%	7.0%	8.4%
PVCs / Hr	417	319	382

Table 7. PVC detection performance for several levels of algorithm performance and PVC prevalence

PVC Prevalence	Algorithm Performance		PPV
	Sensitivity	Specificity	
1.20% (50/hr, or 1200/day)	95.0%	99.0%	53.6%
	95.0%	99.5%	70.3%
	95.0%	99.8%	85.2%
0.24% (10/hr, or 240/day)	95.0%	99.0%	18.0%
	95.0%	99.5%	31.0%
	95.0%	99.8%	53.0%
0.024% (1/hr, or 24/day)	95.0%	99.0%	2.2%
	95.0%	99.5%	4.4%
	95.0%	99.8%	10.2%

3.3. Performance bounds

Contrary to performance reporting in many other applications, the arrhythmia performance measures as specified in the AAMI recommended practice do not include performance bounds [11]. While the required patient-by-patient reporting provides some measure of inter-patient performance variations, it does not allow for easy comparison of different algorithms. Although the mean performance measures can be used for direct comparison, due to the lack of performance bounds they cannot be used to determine whether the differences in performance are statistically significant or not.

One straightforward method commonly used for generating performance bounds is the bootstrap method [12]. Unlike other classical methods, which require assumption of the sample probability distribution to estimate the sample variations, bootstrap method is an empirical method using only the sample at hand to estimate the desired statistics.

A new proposed recommendation is to use the bootstrap method with sample replacement to generate the mean values and 95% confident intervals for all the reported performance measures as specified in the AAMI recommended practice document.

4. Conclusion

Several statistical measures used for reporting arrhythmia detection algorithm performance as described in the AAMI recommended practice do not adequately reflect the actual clinical experience in real-time patient monitoring. Additional performance measures are thus proposed to further improve the clinical relevance of the reported performance results. These new measurements should be considered for inclusion as part of the standard performance reporting of arrhythmia algorithms.

References

[1] AAMI ECAR:1987. Recommended practice for testing and reporting performance results of ventricular arrhythmia detection algorithms.

[2] ANSI/AAMI EC57:2012. Testing and reporting performance results of cardiac rhythm and ST segment measurement algorithms.

[3] Moody GB, Mark RG. How can we predict real-world performance of an arrhythmia detector? Computers in Cardiology 1983;10:71-6.

[4] Wang JY, Helfenbein ED. Long-term performance evaluation of arrhythmia algorithms using unannotated databases. Computers in Cardiology 1988;15:573-6.

[5] Alberg AJ, Park JW, Hager, BW, et al. The use of "overall accuracy" to evaluate the validity of screening or diagnostic tests. J Gen Intern Med 2004;19:460-5.

[6] Wang J, Yeo CL, Aguirre A. The design and evaluation of a new multi-lead arrhtyhmia monitoring algorithm. Computers in Cardiology 1999;26:675-8.

[7] Scherokman B. Selecting and interpreting diagnostic tests. The Permanente Journal 1997;1 No.2: 4-7.

[8] Drew BJ, Califf RM, Funk M, et al. Practice standards for electrocardiographic monitoring in hospital settings. Circulation 2004;110:2721-46.

[9] Funk M, Winkler CG, May JL, et al. Unnecessary arrhythmia monitoring and underutilization of ischemia and QT interval monitoring in current clinical practice. J Electrocardiol 2010;43(6):542-7.

[10] Henriques-Forsythe MN, Ivonye CC, Jamched U, et al. Is telemetry overused? Is it as helpful as thought? Cleve Clin J Med 2009;76:368-72.

[11] Greenwald SD, Albrecht P, Moody GB, Mark RG. Estimating confidence limits for arrhythmia detector performance. Computers in Cardiology 1985;12:383-6.

[12] Efron B. Bootstrap methods: another look at jackknife. Ann Stat 1979;7:1–26.

Address for correspondence:

John Wang
Philips Healthcare, MS-0455
3000 Minuteman Road
Andover, MA 01810-1099, USA
john.j.wang@philips.com