HOW CAN WE PREDICT REAL-WORLD PERFORMANCE OF AN ARRHYTHMIA DETECTOR?

George B. Moody and Roger G. Mark

Massachusetts Institute of Technology, Cambridge, Mass., and
Beth Israel Hospital, Boston, Mass., USA

Summary

The availability of annotated ECG databases, though a considerable benefit to the process of developing software for arrhythmia monitors, has not solved the problem of determining the clinical acceptability of such devices. It appears that adequate performance on annotated databases is not a sufficient basis for predicting real-world performance. We discuss the uses and limitations of annotated databases for development and evaluation, and describe a new application for them in the assessment of long-term algorithm stability. Finally, we introduce techniques for deriving useful predictors of real-world performance from unannotated tapes.

What is real-world performance?

Ultimately, real-world performance is judged by the clinical user, and is best interpreted as user confidence. The user's evaluation of an arrhythmia detector will be based primarily on data observable by the user, although published performance evaluations may be available. Inevitably, the user's perception of real-world performance will be colored by the nature of the user/machine interaction. Real-world performance is also application dependent; a device which is expected to sound alarms in a CCU will be judged with respect to false positives in a manner utterly different from that appropriate for a high-speed tape scanner.

Factors affecting user confidence

Certainly user confidence in an arrhythmia detector ought to be influenced by sensitivity for clinically significant events. This variable is quite difficult to assess in the clinical environment, and published evaluations rarely present the data necessary for an informed judgement. The emphasis in published evaluations is (and should be) on measures of performance which can be established reliably on the basis of sufficient statistics. Thus the traditional focus has been on measures such as QRS and PVC sensitivity. The clinician's priorities, however, place high value on sensitivity for rare events (e.g., ventricular tachycardia), for which reliable detection statistics are difficult to obtain from annotated database evaluations. Thus the user must establish these performance measures from personal experience. If no formal effort (such as parallel analysis using a technique known to be accurate) is made, user confidence is likely to collapse at the first evidence of missed events.

The factor complementary to sensitivity is false detection rate, and to the user this will usually be the most obvious aspect of deficiency in real-world performance. As noted above, the impact of false detections on user confidence is application-sensitive. In the CCU, false alarms above a certain level are intolerable and likely to result in the disabling of the alarm system, a dangerous situation which does little to promote user confidence. Tape-scanning systems, on the other hand, are controllable by laboratory staff, who act as false detection filters and keep the clinical user confident irrespective of the actual false detection rate.

A third important factor is verifiability of summary data. Again, the naive user may simply take such data at face value until confronted with evidence that errors have been committed. An effort made by the user to establish the accuracy of summary data can promote confidence by allowing the construction of an informal model of the detector's analysis process. We believe that knowing what decisions are made by the detector with respect to individual beats is indispensable to such an effort.

Annotated arrhythmia databases

The availability of the MIT/BIH[1] and AHA[2] arrhythmia databases is of considerable benefit to developers of arrhythmia detectors. The use of annotated databases provides for quantitative, reproducible, beat-by-beat performance measurements.

The databases guide and stimulate algorithm development and improvement. It is possible using an annotated database to perform in a matter of hours an experiment which might have taken several weeks of tedious manual analysis. An iterative process of refinement becomes possible as a result. Improvement in performance can be quite rapid, especially at the early stages of this process. It

0276-6574/83/0000/0071$01.00©1983 IEEE

has been suggested, however, that this successive refinement can have an undesirable outcome as the program becomes "tuned" to the database. Thus a conflict arises over the proper use of a database: should a part of it be used only for evaluation?

Our experience with two independently developed ECG analysis programs and two databases tends to complicate the question further. Using the MIT/BIH database for development, both programs achieved reasonably acceptable performance measures; when we first tested them with the AHA database, we found that performance measured on the AHA database was far better than on the development (MIT/BIH) database. Table 1 summarizes these results. For 17 out of 21 pairs of measurements, including measures designed for robust estimation with respect to a variation in the sample, AHA database results are significantly higher than corresponding MIT/BIH database results. Three exceptions were measures based on small numbers of events, for which expected estimation errors are large; the fourth exception showed equal, near-perfect performance on both databases. Clearly, tuning was not responsible for the differences. The conclusions one can draw from this experience are that databases vary in difficulty, and that one cannot predict performance on database X based on performance measured on database Y, unless one can identify the conversion factors. In particular, performance on database RW, the real world, is not predictable on the basis of these experiments alone.

Furthermore, although our experience suggested that the AHA database is in some sense less difficult for our programs than is the MIT/BIH database, the availability of the AHA database permitted us to improve our programs by exposing them to a wider range of data. On one tape, low amplitude PVCs caused problems for the QRS detector of one of our programs; the experience led us to revise the detector's second-pass strategy. On another tape, the sudden appearance of noise on one channel with a simultaneous loss of signal on the other channel presented a situation which we simply had not considered in our earlier development work; again, we were able to make use of the knowledge we gained, and corrected a subtle bug.

Should it be possible to identify and measure the conversion factors which transform database X performance measures into those for database Y? Intuitively, it would seem that if the difference between databases can be characterized only in terms of difficulty, that this should be possible. In fact, however, the fundamental difference is that they represent different samples of the real world, and that no existing database is claimed to be a representative sample of the universe of ECGs. Thus notions of "difficulty" are inevitably defined in terms of what is difficult for a specific program under test, and universal conversion factors, or difficulty indices, seem unattainable.

In summary, we would conclude that the interest of developing better algorithms is served by exposing them to the widest possible variety of data; that database evaluations are not necessarily predictive of real-world performance; and that even the effect of tuning, if any, on algorithm performance is not necessarily measurable on the basis of database evaluations, and may not be separable from the substantial effects of differences in difficulty between databases.

Performance Measure	Program A		Program B	
	MIT DB	AHA DB	MIT DB	AHA DB
QRS Se	99.8%	99.9%	99.4%	99.6%
QRS +P	99.7%	99.7%	99.7%	99.8%
PVC Se	92.8%	97.4%	84.1%	91.2%
Est. PVC Se *	90.3%	96.1%	69.1%	80.7%
PVC +P	86.1%	93.3%	95.4%	97.2%
V. Couplet Se	86%	95%	77%	73%
V. Couplet +P	70%	91%	91%	82%
VT Se	79%	95%	45%	48%
VT +P	37%	88%	69%	77%
>5 beat VT Se	65%	83%	69%	38%
>5 beat VT +P	61%	57%	92%	100%

Table 1. Annotated database evaluation results for two (independently developed) arrhythmia detection programs. Se = sensitivity, +P = positive predictivity. All figures are gross statistics, except for (*) estimated PVC sensitivity, calculated in accordance with the method of Hermes and Cox[3]. MIT DB results based on all 48 tapes, excluding 50-beat learning periods; periods of ventricular flutter/fibrillation were ignored for beat-by-beat statistics, and fusion PVCs were ignored for PVC detection statistics. AHA DB evaluations based on all 56 available tapes (1001-1010, 2001-2004, 2006-2010, 3001-3010, 4001-4009, 6002-6010, 7001-7009); five-minute unannotated learning periods preceded the evaluated segments, and fusion beats were ignored for PVC detection statistics. None of the pacemaker fusion beats on tape 2002 were called PVCs by program A; one was called a PVC by program B.

Limitations of annotated databases

Real-world performance is difficult to predict on the basis of annotated database evaluations for several reasons. The major limitation of annotated databases is small size. This limitation implies not only that an insufficiently diverse sample of the universe of ECGs is represented in an annotated database, but also that insufficient numbers of rare but clinically significant events (such as runs of ventricular tachycardia) are present to permit statistically sound assessments of sensitivity and positive predictivity for such events.

A second limitation is that conventional evaluations using annotated databases tend to underestimate the impact of noise and artifact on

the quality of detector output. Again, as a result of the small size of databases, the variety of real-world noise in a variety of contexts is poorly represented. Furthermore, the penalty paid for inadequate noise rejection may not be observable in terms of the detection statistics usually reported for database evaluations. Thus such statistics are of little use in assessing this important aspect of real-world performance. It may be possible to assess noise immunity by adding noise to annotated database tapes, however.

Finally, conventional database evaluations do not test long-term algorithm stability, since annotated segments are relatively short compared to the length of a typical monitoring procedure. Sophisticated programs for ECG analysis adapt to their inputs and can be said to "learn" from the signal what is normal for the patient under observation. This behavior is necessary for any but a rudimentary analysis of the ECG, given the wide range of inter-patient variability. High-speed tape analysis systems may give the human operator some or all of the responsibility for adapting the program to the input ECG. In either case, the time-varying properties of the program's behavior, while contributing to the accuracy of the analysis, are a matter of concern in the evaluation of the program. As in any system containing feedback, a program which "learns" may be unstable.

Assessment of long-term algorithm stability

Ideally, program stability should be assessed by observation of error rates over time as the program "learns". Error rates for a stable analysis should not change significantly during the test, the length of which should be comparable to that of a typical monitoring procedure. Error rates can be determined with confidence only if beat-by-beat annotations are available, though, and the expense of annotating 24-hour recordings makes this approach unfeasible. Using the MIT/BIH or AHA databases permits a close approximation to the ideal procedure, however. The PASTE test (Protocol for Assessment of STability, shown schematically in figure 1) is performed by "sticking together" segments of annotated ECG into 24-hour records and then playing these records into an ECG analysis program. Since all beats in the record are annotated, it is possible to perform an automated beat-by-beat comparison for the entire 24-hour record. An advantage of the PASTE test over the "ideal" stability test is that error rates over the long run should not be influenced by varying levels of difficulty presented by the input data. Thus a stable analysis should produce nearly constant error rates. In our tests, we observed no significant variability in the error rate over time, suggesting that instabilities were probably avoided in the design of the program under test.

The PASTE test may also be quite useful for comparing widely varying techniques for long-term ECG analysis. When beat-by-beat comparison is not feasible, as when evaluating a Holter scanning service, summary data in the appropriate format may be generated from the beat annotations, and compared

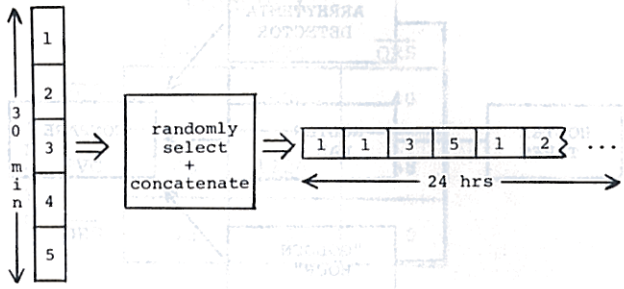

Figure 1. The PASTE test (see text).

with summary data produced by the test. Although it is unlikely that a totally automated analysis would be designed to recognize data fabricated in this way, alert Holter technicians are usually able to do so. For this reason, it seems best to divide an annotated tape into a number of shorter segments, and select them in random order. Segments should be chosen carefully to avoid discontinuities in rate, signal amplitude, or dominant beat morphology. It also seems wisest to avoid using tapes which have a very small number of highly unusual events (e.g., a distinctive combination of ECG and artifact), since the recurrence of such events may make the nature of the data obvious.

The principal limitation of the PASTE test in algorithm evaluation is that of the databases themselves: the data are selected from only a limited sample of the universe of ECGs. Nevertheless, the PASTE test does permit an inexpensive, albeit somewhat artificial, appraisal of long-term algorithm behavior.

Long-term trials

In an effort to resolve some of the questions about real-world performance which cannot be well answered with conventional annotated database evaluations, we have developed techniques for measurement of detector performance on long-term unannotated tapes.

Our experimental procedure is outlined in figure 2. The ECG is recorded on 2-channel Holter tapes for 24 hours as a part of the usual clinical routine. The arrhythmia detector under test analyses the ECG from tape. The outputs of the detector analysis are checked manually. False positives are removed, and a corrected report is prepared based on the findings of the detector. This report contains an hour-by-hour summary of the level of ventricular ectopy coded using a modification of the Lown arrhythmia grading scheme, as shown in figure 3. The grade for each hour is based on the best judgement which can be made on the basis of the detector outputs; specifically, couplets and runs are not counted unless documented in rhythm strips.

Independently, the Holter lab staff prepare a report summarizing, on an hour-by-hour basis, major findings with respect to ventricular and atrial ectopic activity. The report usually includes

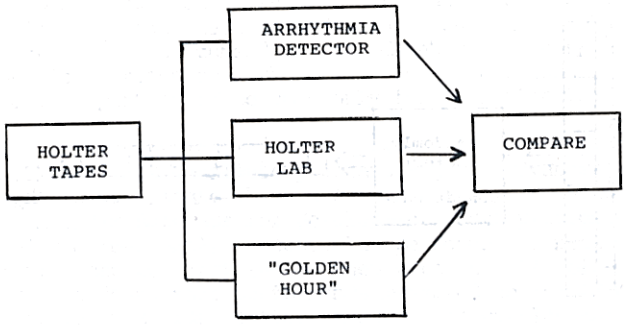

Figure 2. Long-term trials.

documentation of all observed runs of ventricular tachycardia in the form of rhythm strips. (If runs of VT are very frequent, this rule is not followed strictly.) Other arrhythmias are documented as deemed appropriate by lab staff.

In a third independent analysis, one hour (the "gold hour") of the tape is checked meticulously by visual examination of all QRS complexes. Detailed counts of all ectopic beats and arrhythmias are made. Selection of the gold hour is performed at random on half of the tapes; for the remaining tapes, the hour which is deemed "worst" by subjective appraisal of the Holter lab staff is chosen. This appraisal usually reflects the degree of ectopy, but may also reflect difficulty of analysis as a result of noise, morphology variation, or other factors.

VEA grades:

0 - no PVC's
1 - occasional (< 30/hr) PVC's
2 - frequent (≥ 30/hr) PVC's
3 - couplets
4 - runs

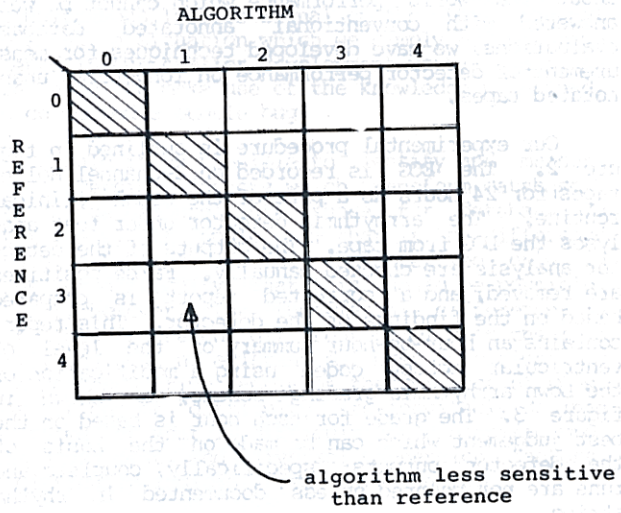

Figure 3. Grading scheme and confusion matrix used for hour-by-hour comparison in long-term trials.

Each technique of analysis may now be compared against the others in several ways. It is quite straightforward to create an hour-by-hour confusion matrix[4,5] for two methods, as shown schematically in figure 3. Elements on the principal diagonal indicate agreement between the techniques; off-diagonal elements show disagreements.

Comparing the Holter lab against 98 gold hours, we found agreement in 97 hours. This suggests that the sensitivity of the Holter lab for VEA is quite high when measured on an hour-by-hour basis, and that the Holter reports can serve as a reliable reference standard. This conclusion has significant implications for the usefulness of long-term trials, since we were able to identify a relatively inexpensive technique of high accuracy which can provide a large amount of data for arrhythmia detector evaluation. Presently there are over 8000 24-hour tapes with Holter reports in our library.

We have compared two arrhythmia analysis programs against Holter reports on an hour-by-hour basis. The results are shown in figures 4 and 5. Several useful measures of performance may be derived from the confusion matrices. The probability of underestimation of VEA for either technique is determined by summing the off-diagonal elements on the appropriate side of the principal diagonal, and dividing by the sum of all elements in the matrix. Although this simple measure weighs all errors equally, it is a useful index of performance, having an intuitive basis which is readily appreciated in clinical terms. From figures 4 and 5, we find that the probability of underestimation of VEA is 11% by program A, 18% by program B, and 3% or 4% for the Holter lab.

PROGRAM A

HOLTER \	0	1	2	3	4
0	40	4	0	0	0
1	5	59	1	0	0
2	0	5	28	4	0
3	0	3	5	34	2
4	0	0	1	5	18

Total hours = 214, 22 patients

-program A underestimates VEA in 24 hrs (11%)
-holter underestimates VEA in 7 hrs (3%)
-hours with ventricular tachycardia
total = 26
detected by program A = 20 (77%)
detected by holter = 24 (92%)

Figure 4. Results of hour-by-hour comparison from long-term trials of program A.

Since all false positives have been removed, we may add all elements to the left of and above a selected element on the principal diagonal to get the total number of hours in which the selected VEA level is known to have been present. This number can be used to calculate sensitivity measures which express the conditional probability that a given level of ectopy, if present in a given hour, will be correctly detected by a given technique. Thus we can estimate that the probability of detection of VT (during an hour in which VT is present) is 77% for program A, 79% for program B, and 96% for the Holter lab (the last figure obtained by adding the figure 4 and figure 5 matrices).

Another source of performance measures from the long-term trials is beat annotations on rhythm strips produced by the detectors. As indicated earlier, our bias is strongly in favor of making such algorithm decisions visible to the user. By manually annotating the beats on each strip and comparing these annotations against those provided by the arrhythmia detector, one may generate a beat-by-beat confusion matrix and calculate QRS and PVC sensitivity and positive predictivity as in an annotated database trial. Clearly, such measures are not equivalent to those obtained in a database trial, since the data selected by the detector for documentation will be richer in both VEA and noise and artifact than the database is. Thus, estimates of positive predictivity will be "worst case" estimates, but they will necessarily reflect the same data on which the clinical user will be judging real-world performance. Figure 6 illustrates the results of this analysis for program A, on the same 214 hours of ECG analyzed on an hour-by-hour basis in figure 4.

PROGRAM B

HOLTER \	0	1	2	3	4
0	279	8	0	0	0
1	95	183	12	13	0
2	0	10	102	5	1
3	2	18	48	211	1
4	2	1	9	2	50

Total hours = 1052, 50 patients

-program B underestimates VEA in 187 hrs (18%)
-holter underestimates VEA in 40 hrs (4%)
-hours with ventricular tachycardia
  total = 66
  detected by program B = 52 (79%)
  detected by holter = 64 (97%)

Figure 5. Results of hour-by-hour comparison from long-term trials of program B.

PROGRAM A

TRUTH \	$\overline{PVC}$	PVC	$\overline{QRS}$
$\overline{PVC}$	6531	500	40
PVC	86	1226	58
$\overline{QRS}$	73	83	0

QRS sens. = 98.8% PVC sens. = 89.5%
QRS +P = 98.2% PVC +P =67.8%

patients = 22 # QRS complexes = 8441
hours = 217 % PVC's = 14.5%
strips = 969
strips per hr = 4.5

Figure 6. Results of beat-by-beat comparison from rhythm strips generated by program A during long-term trials.

Strip production rate, or alarm rate, is also a parameter of clinical interest. Based on long-term trials, we measured strip production rates of 4.5/hour for program A (of which 1.9/hour were false positives) and 7.0/hour for program B (of which 3.5/hour were false positives). These measurements are easy to obtain in the real-world environment as well. It is interesting to note that clinical trials of program B, in which the ECG was analyzed directly without the intermediary tape recording process, showed a significant decrease in false positive strips, to 1.7/hour (while total strip production rate decreased slightly to 6.3/hour). These clinical trials produced higher positive predictivity than was observed in the long-term tape trials for ventricular couplets (82% vs. 67%) and for VT (68% vs. 52%). This experience suggests that the long-term tape trial environment may be more difficult than the real-world environment, at least for real-time analysis. Possible reasons for this may be loss of fidelity and introduction of noise in the recording and playback processes, or better quality control in electrode application (since the performance of the detector provides immediate feedback to aid in electrode placement).

In summary, long term trials as we have described them are a useful supplement to annotated database evaluations. They permit hour-by-hour comparisons to a standard, and derivation of clinically useful indices of performance from these comparisons. If annotated strips are produced by the detector, worst-case estimates of QRS and PVC sensitivity and positive predictivity may be made. False alarm rates can be easily determined. Long-term program stability can be verified. Finally, sensitivity and positive predicitivity for rare, but clinically significant, events such as ventricular tachycardia can be estimated with higher accuracy than is possible using annotated databases, because of the larger number of rare events which

can be observed in the much larger amounts of ECG analyzed in the course of long-term trials.

Conclusions

Quantitative annotated database trials are still the best predictors of real-world performance, despite the limitations of annotated databases which we have discussed. Long-term trials are needed to verify program stability and to assess acceptability of outputs. Clinical trials verify long-term tape trial results and are the only way to assess the quality of user-machine interaction.

Acknowledgements

The authors wish to thank W. Jarisch, who shared his thoughts with us on many of the issues discussed here; and J. Mietus, who prepared the figures.

References

1. Mark, R., Schluter, P., Moody, G., Devlin, P., and Chernoff, D. An annotated database for evaluating arrhythmia detectors. In Frontiers of Engineering in Health Care, pp. 205-210. Proc. 4th Annual Conf. IEEE Engineering in Medicine and Biology Society. 1982.

2. Hermes, R. and Oliver, G. Use of the American Heart Association database. In Ambulatory Electrocardiographic Recording, pp. 165-181. Eds. Wenger, N., Mock, M., and Ringqvist, R. Yearbook Med. Pub. 1980.

3. Hermes, R., and Cox, J., Jr. A methodology for performance evaluation of ventricular arrhythmia detectors. Computers in Cardiology 7:3-8. Long Beach, California: IEEE Computer Society. 1980.

4. Stein, I., Plunkett, J., and Troy, M. Comparison of techniques for examining long-term ECG recordings. Med. Instrum. 14:69-72. 1980.

5. Schluter, P., Mark, R., Moody, G., Olson, W., Peterson, S. Performance measures for arrhythmia detectors. Computers in Cardiology 7:267-270. Long Beach, California: IEEE Computer Society. 1980.