EVALUATION OF AUTOMATED ARRHYTHMIA MONITORS USING AN ANNOTATED ECG DATABASE

R.G. Mark and G. B. Moody

Biomedical Engineering Center for Clinical Instrumentation
Harvard-MIT Division of Health Sciences and Technology
Massachusetts Institute of Technology
Cambridge, Massachusetts 02139

ABSTRACT

This paper documents the design and development of the MIT/BIH arrhythmia database. This two-channel, digital database which contains a wide variety of ventricular and supraventricular arrhythmias is suitable for algorithm evaluation and development. Procedures are suggested for its use, and examples of a variety of output formats are presented.

INTRODUCTION

Automated arrhythmia monitors are complex instruments which are called upon to perform very important and sometimes critical roles in patient care. The design and improvement of such systems has been a major focus of engineering effort for at least two decades. Although much creative energy has been expended in algorithm development, the quantitative evaluation of arrhythmia detectors remains a difficult and controversial problem. Reproducible testing of arrhythmia systems, using a universally accepted "yardstick" is a need widely recognized by system designers and prospective users (Ripley et al., 1977; Schluter et al., 1980; Hermes et al., 1980). In response to this need, at least two annotated arrhythmia databases have recently been developed and are now generally available. One, developed under the auspices of the American Heart Association (Ripley et al.,1978), is distributed by the Emergency Care Research Institute in Plymouth Meeting, Pennsylvania. The second was developed by us at Massachusetts Insti-

C. Marcher (ed.), *Ambulatory Monitoring*

tute of Technology and Beth Israel Hospital in Boston, Massachusetts (Mark et al., 1982). These databases are important evaluation tools which permit detailed and quantitative beat-by-beat algorithm testing. In this paper we will review the development of our database, and propose methodology and policies for its use in algorithm evaluation and development. It should be stated at the outset, however, that a beat-by-beat evaluation using annotated databases, while a necessary part of completely assessing an arrhythmia analyzer, is not sufficient. Evaluation of performance must also be done using long-term data (8-10 hours or more) and clinical trials.

DEVELOPMENT AND SPECIFICATION OF THE MIT/BIH DATABASE

The database consists of 48, half-hour, two-channel ECG records containing a wide variety of ventricular and supraventricular arrhythmias, conduction abnormalities, and artifact. Pacemaker rhythms are included as well. The records were selected from a library of over 4,000, 24-hour ambulatory ECG recordings obtained by the Arrhythmia Laboratory at Beth Israel Hospital in Boston, Massachusetts from both in-patients (60%) and out-patients (40%). Original tape-recordings were made on Avionics Model 445 2-channel recorders using modified Lead II and modified Lead V_1 configurations.

Twenty-three of the database records (the "100 series") were selected randomly from the library to provide a typical cross-section of ECG waveforms and artifact. A table of random numbers was used to select particular tapes and half-hour segments. Data was used so long as at least one of the two channels was interpretable.

The remaining twenty-five ECG records (the "200 series") were selectively chosen to represent more rarely occurring but clinically important

arrhythmias. Episodes of ventricular tachycardia, flutter, and other complex rhythms were included in this set. Several records were specifically selected on the basis of their expected difficulty for computer analysis due to background rhythm, QRS morphology variation, or artifact.

The two-channel analog tape epochs were digitized at a sampling rate of 360 Hz with 11-bit resolution over a 5 mV range. The samples were stored as 8-bit signed differences. The overall bandwidth was 0.1 Hz to 100 Hz, which was considered adequate for testing arrhythmia analysis programs.

The digitized data was next annotated by a simple QRS detector program which arbitrarily labeled each beat as "normal". This preliminary label was used only as an aid in the human annotation protocol.

Each ECG record was then edited, beat-by-beat, by two independent experts. They were each given 30-minute, 2-channel strip chart recordings of the data, where each beat recognized by the simple QRS detector had been labeled "N". All beats that were missed by the QRS detector, or falsely detected, or not NORMAL were relabeled appropriately by the human annotators. They also identified the underlying cardiac rhythm and its changes, and episodes of artifact and noise.

The revised annotations of each reader were separately incorporated into the digitized database with the help of a displaying waveform editor. The database format allowed up to eight independent annotation channels to be preserved.

The annotations of the two independent observers were next automatically compared by a program that identified all disagreements. A

third strip chart recording was made to document them. The disagreements were resolved by a third reader, and final corrections were incorporated into the database to produce a set of "consensus annotations". The fiducial location of each beat annotation was not strictly controlled, except that it was within the QRS complex in all cases, and usually near the maximum peak.

The consensus annotation channel was then checked by an "auditor" program that verified the logical relationship between QRS complex labels and rhythm labels. For example, all complexes within a run of ventricular tachycardia were examined to see if they had a "PVC" label. Omitted beats were detected by listing the longest R-R intervals and their times of occurrence. After the consensus annotation channel was audited and corrected, it became the final "truth" annotation channel.

The completed ECG database was recorded on 9-track magnetic tape in a format compatible with the American Heart Association database for ventricular arrhythmias. The additional annotation codes provided by our database have been added as a superset to specify atrial arrhythmias, conduction defects, changes in predominant rhythm, and episodes of noisy data. A detailed list of annotation codes is given in Table 1 together with AHA equivalents. Table 2 lists the rhythm onset annotations.

TABLE 1: MIT-BIH Database Annotation Codes
(Showing AHA Database Equivalents)

Mnemonic	MITBIH Code	AHA	Description
NORMAL	01	N	normal QRS
LBBB	02	N	left bundle branch block
RBBB	03	N	right bundle branch block
ABERR	04	N	aberrantly conducted beat
PVC	05	V	ventricular Premature Beat
FUSION	06	F	fusion beat ****
NPC	07	N	nodal premature beat
APC	08	N	atrial premature beat
SVPB	09	N	nodal or Atrial premature beat
VESC	10	E	ventricular escape beat
NESC	11	N	nodal escape beat
PACE	12	P	paced beat
UNKNOWN	13	Q	cannot identify, QRS like event.
NOISE	14	O,U	beginning of NOISE *
QUIET	15	O	end of NOISE *
SPIKE	16	O	single QRS-like artifact
P	17	O	P-wave **
Q	18	O	Q-wave **
R	19	O	R-wave **
S	20	O	S-wave **
T	21	O	T-wave **
COMMENT	22	O	comment (text) annotation ***
FIRST	23	O	first annotation (optional) **
LAST	24	O	last annotation (optional) **
BBB	25	N	left or right bundle branch block beat
PACESP	26	O	pacemaker spike without capture
AXIS	27	O	axis shift
ONSET	28	O	rhythm onset (text) annotation ***
OFFSET	29	O	offset comment **
LEARN	30	O	learning **
FLWAV	31	O	ventricular flutter wave
VFON	32	[	onset of ventricular flutter/fibrillation
VFOFF	33	]	end of ventricular flutter/fibrillation
AEB	34	N	atrial ectopic beat
NEB	35	N	nodal ectopic beat
MISSB	36	O	missed beat
BLAPB	37	O	blocked APB
AXLMX	38	O	legal codes are < AXLMX **

* Annotation codes 14 and 15 are used in pairs. If the subtype <1>, no beats are labeled until the next code 15.

** These codes are reserved for future use, and should be ignored.

*** This annotation has a text string, terminated by a null <0>.

**** In the context of paced rhythm (tapes 102, 104, 107, and 217, annotation code 6 is used for pacemaker fusion beats.

TABLE 2: Rhythm Onset Annotations

Rhythm onset annotation (MIT-BIH annotation code 28) include an ASCII string which begins with a "(":

(AB	atrial bigeminy
(AFIB	atrial fibrillation
(AFL	atrial flutter
(AT	atrial tachycardia
(B	ventricular bigeminy
(BI	first degree heart block
(BII	second degree heart block
(BIII	third degree heart block
(IVR	idioventricular rhythm
(N	normal sinus rhythm
(NOD	nodal (A-V junctional) rhythm
(PAT	paroxsysmal atrial tachycardia
(PREX	pre-excitation (WPW)
(SBR	sinus bradycardia
(SVTA	supraventricular tachyarrhythmia
(T	ventricular trigeminy
(VFIB	ventricular fibrillation
(VFL	ventricular flutter
(VT	ventricular tachycardia

The rhythm annotations make it possible to evaluate algorithm performance in the presence of different underlying rhythms. For example, regions of atrial fibrillation and flutter are indicated on the database, permitting automatic evaluation of atrial fibrillation detectors or evaluation of PVC detectors in the presence of atrial fibrillation.

Table 3 presents a condensed summary of the database showing a rough breakdown of QRS types and rhythm episodes. Because of the complex arrhythmias, changing QRS morphologies, noise and artifact included in the database, we consider it a relatively difficult challenge for automated arrhythmia detectors. A more complete description of the database is available in the tape directory (MIT/BIH, 1980).

TABLE 3 : A Condensed Summary of the Database

a. QRS Types (some sub-types combined)

QRS Type	Number	Percent
Normal	78,591	71.6
Bundle Branch Block	11,804	10.7
SVPB	2,542	2.3
Aberrated SVPB	150	0.14
Other Supraventricular	328	0.30
Paced	7,029	6.4
PVC	7,127	6.5
Fusion/Paced	982	0.89
Fusion/PVC	803	.73
Total	109,790	

b. Rhythm Annotations (partial list)

Rhythm Episode	Records Containing Examples
NSR (incl. Tachy, brady)	45
Paced	4
Bundle Branch Block	8
Atrial Flutter/Fibrillation	7
SVTA	6
Junctional Rhythm	3
2° Heart Block	1
WPW	1
Idioventricular Rhythm	2
VT/V Flutter	13

An analog form of the database has also been prepared, using 4-channel FM tape. The primary reason for this format is economic, since the cost of equipment to play back digital tapes is prohibitive to some users. The analog data is recorded on two of the four FM channels, and the remaining two channels contain annotation data and five second time ticks encoded as EIA RS-232 bit-serial voltage levels compatible with most computers.

The database has been made generally available. To date it has been distributed to approximately 30 industrial and university groups in the United States and abroad.

USE OF THE DATABASE IN ALGORITHM EVALUATION

Creation of Annotation Files

A beat-by-beat evaluation requires that the digital database records be used as input to a version of the arrhythmia algorithm which can produce its own beat-by-beat annotations. To evaluate arrhythmia analysis software running on the same machine that supports the database, it is a fairly simple matter to replace the usual data acquisition routine with a call to a database utility subroutine which reads a sample. Similarly, it is not difficult to arrange for each beat label from the algorithm and its time of occurrence to be recorded by calling a utility routine which writes an annotation file. The result of the process is a file of algorithm annotations in the same format as the reference "truth" annotation file supplied with the database. The two annotation files can then be compared as discussed in the next section.

Frequently the arrhythmia monitor under test is physically separate from the computer managing the database. In this case, the database machine can transmit serial digital data or can be programmed to generate analog ECG output from the digital records. The arrhythmia monitor must be modified to return annotations and their times of occurrence to the database machine which can then produce an annotation file as above. Alternatively, the ECG and fiducial markers (from channel 4) may be obtained from the FM analog tape format of the database.

Comparison of Annotation Files

This stage of the evaluation is of critical importance, since all later results depend upon the proper execution of the comparison. The implementation of an annotation comparator requires resolution of several issues:

1. Annotation mapping: the conversion of one annotation alpha-

bet into another.

2. Beat matching: finding pairs of annotations, one from each annotation file, which are considered to refer to the same QRS complex.
3. Run matching: finding pairs of runs of PVC's, one from each annotation file, which are considered to refer to the same run.
4. Choosing output formats.

The MIT/BIH annotation alphabet allows differentiation of 19 QRS types, and includes a variety of non-QRS labels. The AHA annotation alphabet can easily be mapped into the MIT/BIH code without loss of information, although the reverse is not so, since the AHA code lacks rhythm labels and does not differentiate supraventricular ectopic beats from normals.

For many purposes, however, a much smaller annotation set is adequate. In evaluating PVC detection, we have used a set consisting of $\overline{\text{PVC}}$ (sometimes we use the N symbol for this group), PVC, Fusion PVC, and $\overline{\text{QRS}}$. In this mapping, PVC's, R-on-T PVC's, ventricular flutter waves, and ventricular escape beats are considered PVC's. Fusion beats in non-paced records are considered fusion PVC's; and all other beat types are classed as $\overline{\text{PVC}}$'s.

The beat matching algorithm is based on the notion of a "match acceptance window". Since, in general, it is unreasonable to expect precise simultaneity of database annotations and algorithm annotations, the beat matching algorithm selects pairs of annotations, which are separated by a time interval no greater than the match acceptance window. If an annotation has no match in the other file within the match acceptance window, it is paired with a (dummy) "not QRS" annotation. (This would correspond to either a missed beat or a false QRS detec-

tion, depending on which file contained the extra annotation.) If two or more annotations in either file lie within the match acceptance window of an annotation in the other file, the closest match is accepted. (See Figure 1.) If there are no annotations in the other file to correspond with the extras, they are paired with dummy "not QRS" annotations. Thus all beat annotations in both files are paired, and the output of the beat matching algorithm is a stream of paired annotations.

In our experience, a suitable match acceptance window has been 150 milliseconds. This permits a substantial annotation misalignment, as may occur if one annotator labels the peak, and the other the nadir of a PVC, for example. Since the closest matches are always counted, no problems of ambiguity arise.

The clinical importance of ventricular couplets and runs prompted us to develop a run matching algorithm. Throughout the following discussion, isolated PVC's and ventricular couplets are regarded as "runs" with lengths of 1 and 2 beats respectively. The run matching algorithm takes as input

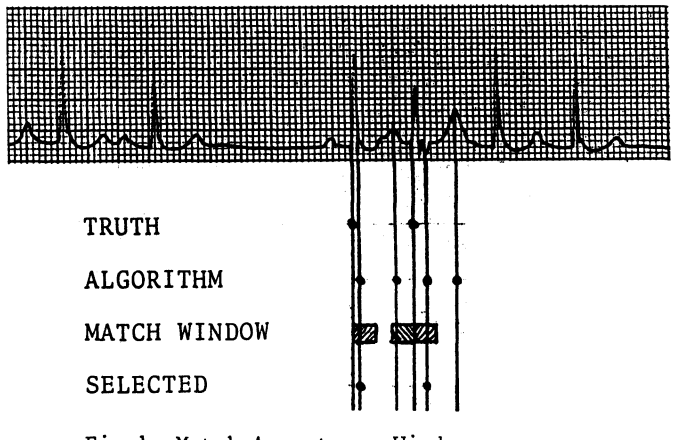

Fig.1: Match Acceptance Window

the stream of paired annotations produced by the beat matching algorithm. The lengths of each run in each file are noted. Whenever a pair of non-PVC's is received from the beat matching routine, any pending runs are compared by length. The result of the comparison is entered into the run-by-run confusion matrix. If two or more runs occur on either channel before a pair of non-PVC's is found, only the longest will be used in the comparison. Thus, the basic criterion for matching runs depends on finding runs which overlap without requiring beat-by-beat agreement for each beat in the run. (See Fig. 2.) We adopted this approach based on the belief that a device capable of recognizing consecutive PVC's, at a time when consecutive PVC's are occurring is performing the most important function of an arrhythmia monitor properly, and the details of beat-by-beat comparisons are less important (and may in any case be determined from other outputs of the comparator). We have arbitrarily declared an annotation pair containing a fusion PVC to be a pair of non-PVC's (i.e. a run terminator). Thus, a monitor is neither penalized nor rewarded for its treatment of fusion PVC's in this context.

Fig.2: Run Length Comparison Algorithm

Beat Sequence

Truth: xxVxxVVxxVVVVVVxxVVVVVVVxxxVxxxxxx xx
Algorithm: xxVxxVVxxVVVVxxxxVVVxVVVVxxxxxVVxxVxx

Run Length

Truth	1	2	6	7	1	0	0
Algorithm	1	2	4	4	0	2	1

Output Formats

In presenting evaluation results it is important to indicate exactly what data was excluded from the evaluation if any. We have usually adopted the following policies:

1. Learning beats associated with algorithm start-up are excluded.
2. Regions of ventricular flutter and fibrillation are excluded. All other background rhythms and noisy areas are included.
3. Fusion PVC's are excluded when calculating PVC detector performance.

Lumped results for an entire database may be presented in matrix form. Figure 3 shows evaluation results for a two-channel analysis algorithm (ARISTOTLE) on the MIT/BIH and the AHA databases (Moody et al., 1982). From such matrices it is a simple matter to calculate several commonly used measures of algorithm performance:

		Algorithm		
		$\overline{\text{PVC}}$	PVC	$\overline{\text{QRS}}$
Truth	$\overline{\text{PVC}}$	98,067	935	142
	PVC	400	6,632	79
	FSN	402	387	5
	$\overline{\text{QRS}}$	118	140	

MIT/BIH Database

		Algorithm		
		$\overline{\text{PVC}}$	PVC	$\overline{\text{QRS}}$
Truth	$\overline{\text{PVC}}$	120,372	664	91
	PVC	299	11,963	45
	FSN	614	137	Ø
	$\overline{\text{QRS}}$	33	156	-

AHA Database (55 tapes)

Fig. 3: Evaluation of ARISTOTLE Using two Annotated Databases

$$\text{QRS sensitivity} = \frac{\#\text{ true QRS's detected}}{\#\text{ true QRS's}}$$

$$\text{QRS positive predictive accuracy} = \frac{\#\text{ true QRS's detected}}{\text{total }\#\text{ detected events}}$$

$$\text{PVC sensitivity} = \frac{\#\text{ true PVC's detected}}{\#\text{ true PVC's}}$$

$$\text{PVC positive predictive accuracy} = \frac{\#\text{ true PVC's detected}}{\text{total }\#\text{ events called PVC's}}$$

The calculated values for these parameters for both databases are shown for ARISTOTLE in Table 4. The detector parameter α described by Cox et al.,(1981), is also shown, together with their "average detection probability", $\overline{P}$. Table 5 presents the sensitivity and positive predictive accuracies for couplets, short (3-5 beat) runs of ventricular tachycardia, and long (greater than 5 beat) runs.

TABLE 4: Lumped Results for ARISTOTLE

	MIT/BIH	AHA Database
QRS Sens.	99.79	99.90
QRS Pos. Pred. Acc.	99.76	99.86
PVC Sens.	93.26	97.20
PVC Pos. Pred.	86.05	93.59
Alpha	10.0	25.45
Average detection Prob, $\overline{p}$	90.9%	96.2%
Total QRS's	107,049	134,185
Total PVC's	7,111	12,307

TABLE 5: Couplet and Run Detection

	MIT/BIH	AHA Database
Couplet Sens.	87.26	96.43
Couplet Pos.Pred.Acc.	69.07	91.12
Short Run Sens.	84.85	96.42
Short Run Pos. Pred. Acc.	36.36	87.47
Long Run Sens.	88.24	83.33
Long Run Pos. Pred.Acc.	46.88	46.88

Lumped results do not reflect the substantial variability in algorithm performance on different patient records within the database. Results for individual records may show PVC sensitivities which range from 0% to 100% - an observation not conveyed in the lumped results. Individual tape results may be expressed in tabular form as in Table 6. While this format is quite complete, it is quite awkward. Cumulative distributions of detector sensitivities provide information on individual variability, and have suggested by Cox et al., (1981). An example of such a cumulative distribution is shown in Figure 4, using data from an evaluation of a single patient bedside arrhythmia monitor (Schluter, 1981). In this particular distribution each patient record is weighted by the number of PVC's. The resulting "gross detection ratio" of 84.1% reflects the probability of correctly identifying a PVC from the universe of all beats in the database. Another aggregate sensitivity measure is the "average detection ratio" in which results from each patient record are weighted equally, independent of the number of PVC's involved. The cumulative distribution of sensitivities for the bedside monitor is shown in Figure 5. The average detection ratio of 67.8% is considerably less than the gross sensi-

TABLE 6

Summary of evaluation results for aristotle [truth/aristotle]

Tape	N/N	V/N	F/N	-/N	N/V	V/V	F/V	-/V	N/-	V/-	F/-	QRS Se	QRS +P	PVC Se	PVC +P
100	2221	0	0	0	0	1	0	0	0	0	0	100.00	100.00	100.00	100.00
101	1804	0	0	1	9	0	0	3	1	0	0	99.94	99.78	-	0.00
102	2125	1	0	0	7	3	0	0	0	0	0	100.00	100.00	75.00	30.00
103	2002	0	0	0	30	0	0	0	1	0	0	99.95	100.00	-	0.00
104	2175	1	0	12	0	0	0	6	2	1	0	99.86	99.18	0.00	0.00
105	2397	2	0	23	83	36	0	44	3	0	0	99.88	97.41	94.74	22.09
106	1432	4	0	0	26	514	0	0	0	1	0	99.95	100.00	99.04	95.19
107	2027	2	0	0	0	51	0	0	0	6	0	99.71	100.00	86.44	100.00
108	1651	4	1	36	35	10	1	9	13	0	0	99.24	97.42	71.43	18.52
109	2435	2	1	0	7	36	0	0	0	0	0	100.00	100.00	94.74	83.72
111	2069	0	0	0	2	1	0	0	1	0	0	99.95	100.00	100.00	33.33
112	2488	0	0	0	0	0	0	0	0	0	0	100.00	100.00	-	-
113	1733	0	0	0	12	0	0	0	0	0	0	100.00	100.00	-	0.00
114	1774	0	0	0	7	43	4	0	1	0	0	99.95	100.00	100.00	86.00
115	1899	0	0	0	3	0	0	0	0	0	0	100.00	100.00	-	0.00
116	2211	0	0	2	19	108	0	1	23	1	0	98.98	99.87	99.08	84.38
117	1474	0	0	0	10	0	0	0	0	0	0	100.00	100.00	-	0.00
118	2205	0	0	0	7	15	0	0	0	0	0	100.00	100.00	100.00	68.18
119	1503	0	0	1	2	431	0	0	0	0	0	100.00	99.95	100.00	99.54
121	1808	0	0	0	2	1	0	0	1	0	0	99.94	100.00	100.00	33.33
122	2425	0	0	0	0	0	0	1	0	0	0	100.00	99.96	-	0.00
123	1446	0	0	0	18	0	0	0	0	3	0	99.80	100.00	0.00	0.00
124	1507	2	4	0	9	44	1	0	0	1	0	99.94	100.00	93.62	83.02
200	1720	31	1	0	18	773	0	1	0	6	1	99.73	99.96	95.43	97.60
201	1607	1	2	0	64	197	0	0	41	0	0	97.86	100.00	99.49	75.48
202	2042	0	1	0	20	18	0	0	4	0	0	99.81	100.00	100.00	47.37
203	2332	42	0	28	147	381	1	51	11	17	0	99.04	97.35	86.59	65.80
205	2521	1	9	0	1	69	2	0	1	1	0	99.92	100.00	97.18	98.57
207	1536	72	0	0	85	119	0	5	3	5	0	99.56	99.72	60.71	56.94
208	1500	78	117	7	44	886	247	2	9	6	2	99.41	99.69	91.34	95.06
209	2938	0	0	1	14	1	0	1	0	0	0	100.00	99.93	100.00	6.25
210	2359	3	2	1	32	169	5	3	4	19	2	99.04	99.84	88.48	82.84
212	2677	0	0	1	8	0	0	1	12	0	0	99.56	99.93	-	0.00
213	2593	69	259	0	25	150	103	0	0	1	0	99.97	100.00	68.18	85.71
214	1940	0	1	1	14	251	0	1	2	2	0	99.82	99.91	99.21	94.36
215	3137	1	0	0	13	159	1	0	0	1	0	99.97	100.00	98.76	92.44
217	1927	9	0	1	68	149	0	3	1	3	0	99.81	99.81	92.55	67.73
219	2028	0	0	1	10	61	1	1	1	2	0	99.86	99.90	96.83	84.72
220	1984	0	0	0	13	0	0	0	0	0	0	100.00	100.00	-	0.00
221	1983	0	0	0	6	386	0	0	0	1	0	99.96	100.00	99.74	98.47
222	2405	0	0	0	23	0	0	1	5	0	0	99.79	99.96	-	0.00
223	2062	63	3	0	6	408	11	0	0	1	0	99.96	100.00	86.44	98.55
228	1642	3	0	0	9	347	0	3	2	0	0	99.90	99.85	99.14	96.66
230	2194	0	0	0	10	1	0	0	0	0	0	100.00	100.00	100.00	9.09
231	1516	0	0	0	2	2	0	0	0	0	0	100.00	100.00	100.00	50.00
232	1726	0	0	2	4	0	0	3	0	0	0	100.00	99.71	-	0.00
233	2188	9	1	0	11	808	10	0	0	1	0	99.97	100.00	98.78	98.66
234	2699	0	0	0	0	3	0	0	0	0	0	100.00	100.00	100.00	100.00
Tot	98067	400	402	118	935	6632	387	140	142	79	5				

Total QRS: 107049
Total PVC: 7111
alpha = 9.9995

Gross detection ratios: 99.79 99.76 93.26 86.05
Average detection ratios: 99.79 99.77 88.46 54.03
Estimated detection ratio: 90.91

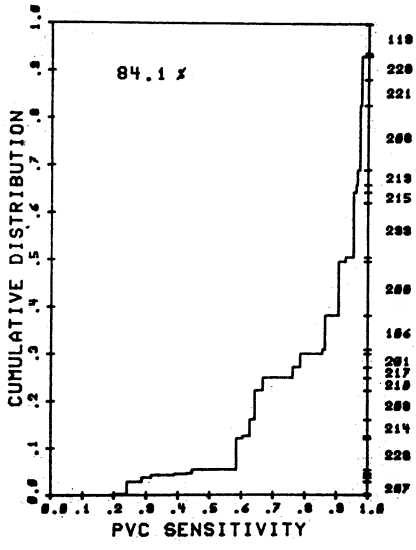

Fig.4: Records Weighted By PVC Count

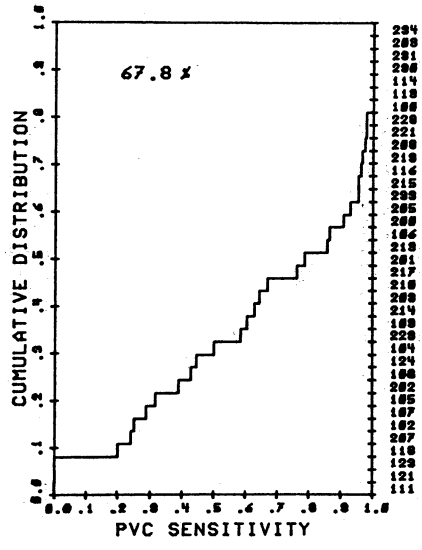

Fig.5: Records Weighted Equally

tivity and is the expected value of PVC sensitivity for a single patient selected at random from a population comparable to the database. Cox et al., (1981) have proposed a stochastic model for the PVC detection process and have developed an alternative estimate of detector performance by fitting the cumulative distribution with a degenerate beta distribution of the form p^{α}. The "average detection probability" then is given by $\bar{p} = \frac{\alpha}{1 + \alpha}$. While it appears that these measures are not database independent, they may be useful is expressing algorithm performance in a more robust manner. Table 4 includes calculations of the alpha parameter and the average detection probabilities for ARISTOTLE.

DISCUSSION

Automated cardiac arrhythmia detectors are very complex systems which are representative of a general trend toward stand-alone "intelligent" clinical instruments. Large amounts of physiologic data are processed to provide the physician with highly "digested" results. The task of evaluating such devices is enormous, very complex, and may be even more difficult and expensive than the initial development of the instrument.

Annotated ECG databases are a critical ingredient of the evaluation process, a fact recognized by many. The availability of universally accepted databases will be of great benefit to both system developers and potential users. The ability to make meaningful system comparisons should enhance inter-institutional communication and should facilitate progress in algorithm research.

The major limitation of annotated databases is their relatively small size. It is not yet clear how many half hour ECG records are required to adequately represent the "universe". We have found that 48 records is a large enough set to make it difficult to "tune" an algorithm successfully to the entire database, but it is a small enough set that developers may be unreasonably swayed in their judgement by a single record. The availability of the AHA database is a major addition to our resources. Based on our experience, it would appear that the MIT/BIH database is a bit more difficult than the AHA database for automated arrhythmia analyzers. The combination of the two databases provides developers and evaluators with a rather extensive and quite flexible resource. It is likely, however, that still more database development will be required to more adequately represent supraventricular arrhythmias and pacemaker rhythms.

Ideally, the evaluation of an algorithm should be done using a database different from that used during development. If the database were large enough, and a good enough sample of the "real world", this objection would be of less significance. The definition of "large enough" and "good enough" are at present elusive, however.

The correlation between laboratory evaluation using an annotated database and performance in the actual clinical setting remains to be established. Although database trials will be an important part of system evaluation, they are not in themselves a totally adequate measure of system performance. Evaluations using long term ECG data are of major importance. It appears, for example, that database records do not sufficiently represent the impact of noise and artifact on total system performance. While extensive laboratory testing of algorithms is a must, the success of a clinical instrument may depend equally upon such features as effective man-machine interfaces, and hardware and software robustness. Hence, careful multifacted clinical evaluations must supplement database trials.

REFERENCES

Cox, J.R., Hermes, R.E. and Ripley, K.L. 1981. Evaluation of performance. In "Ambulatory Electrocardiographic Recording". (Eds. N.K. Wenger, M.B. Mock and I. Ringqvist). (Yearbook Medical Publishers, Chicago). pp. 183-198.

Hermes, R. and Oliver, G. 1980. Use of the American Heart Association Database. In "Ambulatory Electrocardiographic Recording". (Eds. N.K.Wenger, M.B. Mock, and I. Ringqvist). (Yearbook Medical Publishers, Chicago). pp. 165-181.

Mark, R.G., Schluter, P.S., Moody, G.B. et al. 1982. An annotated ECG database for evaluating arrhythmia detectors. In "Frontiers of Engineering in Health Care - 1982". Proceedings of 4th Annual IEEE EMBS. pp. 205-210.

"MIT-BIH Arrhythmia Database: Tape Directory and Format Specifications". 1980. Biomedical Engineering Center for Clinical Instrumentation. Tech. Report No. 010. (MIT, Cambridge, Mass. 02139).

Moody, G.B. and Mark, R.G. 1982. Development and evaluation of a 2-lead ECG analysis program. Computers in Cardiology (in press).

Ripley, K. and Oliver, G.C. 1977. Development of an ECG database for arrhythmia detector evaluation. Computers in Cardiology. pp. 203-209.

Ripley, K.L., Geselowitz, D.B. and Oliver, G.C. 1978. The American Heart Association database: A progress report. Computers in Cardiology. pp. 47-54.

Schluter, P.S., Mark, R.G. and Moody, G.B. et al. 1980. Performance measures for arrhythmia detectors. Computers in Cardiology. pp. 267-270.

Schluter, P.S. 1981. "The Design and Evaluation of a Bedside Cardiac Arrhythmia Monitor". Ph.D. Thesis. Dept. of Electrical Engineering and Computer Science, (MIT, Cambridge, Mass. 02139).