A Method for Evaluating Computer Programs for Electrocardiographic Interpretation

I. Application to the Experimental IBM Program of 1971*

By James J. Bailey, M.D., Samuel B. Itscowitz, M.D., John W. Hirshfeld, Jr., M.D., Leonard E. Grauer, M.D., and Martha R. Horton, B.A.

*The programs studied in these three reports were chosen for two reasons: first, they were available to the National Institutes of Health, and second, they could be implemented on computers readily available to the National Institutes of Health. However the methods for evaluation reported here could be used at any other institution and are not dependent upon a specific program or system. The authors have a continuing interest in automated ECG analysis and may select for study additional programs, the treatment of ECG analysis of which differs significantly from the programs reported here. The results of these studies will be submitted for publication in an appropriate journal. Conclusions drawn from these studies including the present three reports represent the considered opinions of the authors and do not represent any official policy of the Department of Health, Education and Welfare.

SUMMARY

A method for evaluating computer programs for electrocardiographic interpretation is described. This method requires a significant proportion and variety of abnormal tracings, the application of specific fixed criteria, and the separation of disagreements between the computer program and the clinician into those resulting from criteria differences and those resulting from programming errors, viz., pattern recognition failures, mismeasurements, and/or deficient program logic. When applied to the experimental IBM program 1971, staff cardiologists found essential agreement with the program's results in 76% of 1150 unselected tracings. Clinically significant disagreements based strictly on the application of different criteria occurred in 20% of the tracings, whereas disagreements based on program errors were found in only 4%. Although this program requires some system of human overview and quality checking, its potential for clinical implementation is worthy of consideration.

Additional Indexing Words:

Computer analysis
Clinical application

IN THE PAST DECADE numerous computer programs for electrocardiographic (ECG) interpretation have been developed and are now available for routine use. However the reliability of these programs has never been adequately evaluated. Disagreements between ECG interpretations by physicians and computer programs have occurred and the frequency of these disagreements reported in published studies has varied widely depending upon the specific computer program analyzed and the methods of analysis utilized by the authors.¹⁻⁻⁷

For example, in some studies the percentage and variety of abnormalities were small. In other studies the physician-reviewers were allowed to examine computer printouts before making a final judgment and thus were not constrained to specific fixed criteria. In still other studies tables show agreement between computer program and physicians on the existence of an abnormality even though there was a severe disagreement concerning the nature of the abnormality (e.g., infarct vs arrhythmia). In most studies there was no separation of disagreements resulting from the application of different criteria from those disagreements which were a product of some fundamental error in computer programming. Unfortunately, the controversy generated by these conflicts has served only to confuse rather than clarify the problems involved.⁸

The purpose of this paper, the first of a series of three, is twofold. The primary one is to introduce a method for evaluating ECG computer programs which avoids the pitfalls of previous studies, analyzes

From the Laboratory of Applied Studies, Division of Computer Research and Technology and Cardiology Branch, National Heart and Lung Institute, National Institutes of Health, Bethesda, Maryland.

Address for reprints: James J. Bailey, M.D., Building 12A, Room 2053, National Institutes of Health, Bethesda, Maryland 20014.

Received February 8, 1974; revision accepted for publication March 26, 1974.
the program's performance accuracy, and enables the clinician to make a judgment as to the usefulness of a given program for his specific setting and needs. The second purpose is to apply this method to a specific program, namely, the experimental IBM program of 1971.

In the second paper of this series we apply the method described in this paper to two additional programs. In these first two papers, programs are examined from a clinical standpoint and the emphasis is upon clinical accuracy. This accuracy is dependent upon the choice of criteria as well as program errors. In the third paper, programs are evaluated for reproducibility of results, independent of clinical accuracy or criteria choices.

Material and Methods

From November 1971 to June 1972, 1150 unslected electrocardiograms were collected on a Marquette cart at the Clinical Center of the National Institutes of Health (NIH). The leads used consisted of the standard twelve plus the XYZ system described by Smith. Definite abnormalities were found in 55% of ECGs. Each ECG was recorded both on strip chart and on high fidelity FM analog tape. The analog data were digitized. Subsequently the digital data were processed by the experimental IBM program of 1971 at IBM facilities in Mohanias, New York, and the computer printout was returned to the NIH. Each ECG was reviewed in detail by a board of four staff cardiologists, none of whom was aware of the computer interpretation at the time of his review. The cardiologists' and the computer program's interpretations were then compared. As the method of comparison and the data analysis is the keystone in any investigation of ECG computer programs, this will now be outlined in detail below.

1) Cardiologists-Computer Program Agreements: When the readers and the computer made identical interpretations, this was catalogued as an agreement and no further analysis was performed.

2) Cardiologist-Computer Program Disagreements: All diagnostic disagreements were separated into one of three possible categories, namely criteria differences, program errors, or reader errors.

a. Criteria differences. For purposes of diagnostic consistency, the readers were constrained by a fixed set of diagnostic criteria adapted from several sources. The complete set of criteria utilized by the experimental IBM program of 1971 was available. Diagnostic disagreements were studied in detail to determine if the difference was due simply to the application of a different set of diagnostic criteria. When such a basis was found for the disagreement, it was catalogued as a criteria difference. Crucial to the assessment of this or any other computer program is the understanding that diagnostic disagreements based solely on the utilization of different criteria do not indicate any deficiency in the computer program but simply reflect the choice of different ECG criteria.

b. Program errors. A program error is defined here as a diagnostic disagreement based upon a deficiency in pattern recognition, measurement methods, and/or program logic. For example, if a q wave of 20 msec width in lead aVF is mis-measured by the program as having a 50 msec width, the program may erroneously make the diagnosis of inferior infarction. This kind of error reflects a fundamental difficulty in the program's operation which may not be easy to correct.

c. Reader errors. Reader errors, of course, did occur. When a reader disagreed with the program or with the previous reader, he was constrained to refer to the fixed criteria already mentioned. Thus, ultimately it was these fixed criteria which formed the basis for judgment, not the momentary whim of the individual reader.

Many ECGs in the NIH patient population have multiple abnormalities. Therefore, in addition to evaluating each statement of program output, an over-all rating based on the sum of the statements was made. If the readers agreed substantially and in essence with the program's collection of statements for each ECG, then over-all agreement was registered. But if there were disagreement on one or more clinically significant statements, then over-all disagreement was registered.

This over-all rating did not involve minor disagreements, such as borderline undercalls or overcalls, small differences in measurement (e.g., borderline first degree atrioventricular block, axis deviation), minor differences in the severity of infarct or hypertrophy statements, or minor rhythm disturbances such as the presence or absence of a single premature atrial systole.

Results

The results are expressed in bar graphs (fig. 1). The first bar in each graph reflects the sensitivity of the program relative to the readers and the last bar is an indicator of the specificity of the program relative to the readers. From these values for sensitivity and specificity two-by-two tables for our results can be constructed. The total number of cases for each table would be 1150 except in the instance of infarction, where the total is 1180 due to the fact that more than one infarct statement can be made on a given tracing.

Left Ventricular Hypertrophy (LVH)

The readers utilized the Romhilt-Estes point score system in diagnosing LVH. The experimental IBM program of 1971 utilizes a system which is similar but not identical in arriving at the diagnosis of LVH. Both systems have a weak and strong category, but for purposes of constructing figure 1 the statements were merged into a single LVH positive category.

There were no disagreements due to program error. However in four cases the statement of LVH was not made because the program found an unacceptable level of noise and rejected the tracings. Sensitivity
EVALUATION OF ECG COMPUTER PROGRAMS

with respect to the readers was 88.8% (142/160) and specificity was 95.3% (943/990) (see first and last bars).

Myocardial Infarction

The criteria used by the readers and by the IBM program are outlined in the Appendix of the second paper of this series. Both systems utilized three degrees of severity in their diagnostic language. In constructing figure 2, these three degrees of severity were merged into a single positive infarct category.

As might be expected in a field as complex as the electrocardiographic diagnosis of myocardial infarction, disagreements based on criteria differences were not uncommon. Program errors, though infrequent, occurred more frequently than was the case with LVH. Of 178 infarct statements made by the readers, reader-computer disagreements based on program errors occur in only two (1.1%). In both tracings the computer failed to recognize a wide but very shallow q wave and therefore did not make an infarct statement. Of 162 infarct statements made by the program, 11 (6.8%) were based upon program errors. In seven of these 11 a small but definite initial r wave was not recognized by the computer and the S wave was misinterpreted as an initial Q wave. In the remaining four ECGs the q width was mismeasured, falsely triggering an infarct statement.

Sensitivity of the program with respect to the readers was 65.2% (116-178) and specificity was 96% (962/1002) (see first and last bars).

Arrhythmias

The cardiac rhythm was readily apparent in all of the tracings studied with the exception of one, which was therefore excluded. Differences between the readers and the IBM program with regard to distinguishing ventricular premature complexes from aberrant supraventricular complexes occurred very rarely and were catalogued as agreements. The IBM program makes the statement "undetermined rhythm" whenever a tracing fails to meet its logic and criteria for a specific rhythm diagnosis. In this study, such a statement was made in 24 cases, 18 of which were found to be atrial fibrillation, while the remaining six were normal sinus rhythm with first degree atrioventricular block. Thus, with the exception of this "undetermined rhythm" subgroup, criteria differences did not occur in this section of the data analysis. Therefore the bar graphs in figures 3-5 do not contain a segment for criteria difference; i.e., only the categories of agreement and program error are necessary. "Undetermined rhythm" statements are illustrated where pertinent.

Figure 3 summarizes the over-all performance of the IBM program in detecting the presence of a rhythm disturbance, though not necessarily iden-
Rhythm Disturbance. Eighteen (7.4%) of the rhythm disturbances and six (0.7%) of those with normal sinus rhythm were labeled "undetermined rhythm."

Ventricular Arrhythmias

The readers identified 72 tracings with ventricular arrhythmias; 65 with premature ventricular contractions (PVCs), five with electronic pacemakers, and two with bigeminy (fig. 4). In seven of these 65 tracings the program failed to recognize PVCs. In three additional tracings with complex arrhythmias, the rhythm disturbance was improperly identified. For example, one tracing showing sinus arrest with ventricular escape beats was labeled atrial fibrillation.

Thus all ten of these disagreements (14%) are a result of program errors, presumably deficient program logic for rhythm diagnosis. It is important to note also that the program has no logic for calling attention to unique events of great potential interest to the clinician such as the R on T phenomenon or two sequential ventricular premature beats.

The program made 70 statements of ventricular arrhythmia. Eight times the program falsely stated that PVCs were present. In three of these tracings there was a complete bundle branch block (either right or left). One tracing showed atrial fibrillation. Thus sensitivity of the program to ventricular arrhythmias was 86% (62/72) and specificity was 99.3% (1070/1078) (see first and last bars of figure 4).

Atrial Fibrillation

Of 159 cases of atrial fibrillation identified by the readers, the IBM program correctly identified 136 (85.6%). The program rejected two of the remaining 23 cases because of noise, while on 18 occasions it made the diagnostic statement "undetermined rhythm." Erroneous diagnoses were made in the remaining three cases, resulting in a program error rate of 1.9%. In one of these three tracings, atrial fibrillation was present with a perfectly regular ventricular rate of 72 beats/min. The IBM program stated, "low heart rate; suspect heart block." The third tracing had atrial fibrillary waves of moderate amplitude and occasional periods of ventricular regularity, in-

![Figure 3](https://example.com/figure3.png)

**Figure 3**

Rhythm Disturbance. Eighteen (7.4%) of the rhythm disturbances and six (0.7%) of those with normal sinus rhythm were labeled "undetermined rhythm."

![Figure 4](https://example.com/figure4.png)

**Figure 4**

Ventricular Arrhythmias. All false negatives (10) and all false positives (8) resulted from program errors.
terspersed with segments that were clearly grossly irregular. The IBM program stated, "junctional tachycardia." Thus although the computer program was in error on each of three tracings, its diagnostic language was not seriously misleading to the clinician.

The program labeled 138 cases as showing atrial fibrillation and the readers agreed with 136 (98.6%). Two erroneous statements of atrial fibrillation were made (1.4%). In one of these tracings there was a normal sinus rhythm but with PVCs and a P wave of very low amplitude. The other tracing also exhibited normal sinus rhythm but this was punctuated by episodes of sinus arrest with ventricular escape beats. The specificity of the program with respect to atrial fibrillation is 99.8% (989/991) (fig. 5, last bar).

First Degree Atrioventricular (A-V) Block

The IBM program and the readers utilized the same criteria in the diagnosis of first degree A-V block and therefore criteria differences did not occur. As shown in figure 6 (first bar) the program failed to identify nine of 61 tracings which had first degree A-V block. In six (9.8%), the statement "undetermined rhythm" was made. The three IBM errors (4.9%) consisted of two statements of normal sinus rhythm without further commentary, and one diagnosis of atrial fibrillation. Diagnostic failures occurred when the P waves were difficult to identify either because they were diminutive or because they were superimposed upon the preceding T wave. In every case in which the program made the diagnosis of first degree A-V block, the readers were in agreement; hence, the program's specificity in this category was 100% (figure 6, third bar).

Intraventricular Conduction Delays. Two (1.7%) false negatives and two (0.2%) false positives were caused by program errors.

Second Degree Atrioventricular Block

There were eight cases of second degree A-V block. The computer program correctly identified only two of these eight reflecting the difficulty it has in recognizing P waves superimposed upon preceding T waves. No false positive statements were made.

Intraventricular Conduction Delays

A total of four program errors was made (first and second bars, fig. 7). The cause of these errors varied and included failure to identify a small initial r wave in V1, mistakenly identifying a coarse fibrillatory wave as an initial r, and falsely measuring the intrinsicoid deflection. Sensitivity with respect to the readers was 87% (103/118) and specificity was 97.8% (1009/1032) (see first and last bars of figure 7).

Axis Deviation. Only one false negative in the assessment of axis deviation was due to program error.
Primary ST and T wave changes. Three (1.3%) false negatives and six (0.7%) false positives were caused by program errors.

Axis Deviation

The program's sensitivity with respect to the readers was 89.5% (198/221) and its specificity was 99.5% (924/929) (fig. 8). Only one error was made (first and second bars), probably from misinterpreting an rSR' as a monophasic R and therefore missing the diagnosis of right axis deviation.

Primary ST and T Wave Changes

The program's sensitivity with respect to the readers was 84.4% (194/230) and its specificity was 97.1% (894/920) (fig. 9). Most disagreements were purely a matter of different criteria; very few resulted from program errors.

Over-all Performance

Figure 10 depicts the over-all performance of the program. In 868 tracings (75.5%), the readers and the program were in agreement, while diagnostic disagreements were registered in the remaining 282. Detailed analysis revealed that 231 of these disagreements were based upon the utilization of different diagnostic criteria (20.1% of the total series). Included in this group are 24 tracings in which the program stated "undetermined rhythm" but no other program deficiency or error was apparent. Also in this group are 22 tracings which the program rejected due to noise.

In the remaining 51 tracings (4.4%), major diagnostic disagreements resulted from program errors, namely, pattern recognition failures, mismeasurements, and/or deficient program logic. The origin and nature of such program errors are to be discussed in the third paper in this series.

Discussion

The method of analysis utilized in this study has distinct clinical advantages. First and perhaps foremost, it avoids the sterile and usually futile debate over ECG diagnostic criteria. Rather, the criteria utilized by the readers become a tool which enables the functioning of the computer program to be analyzed in a way which has subsequent clinical relevance. First, diagnostic disagreements are found. These disagreements are then studied in detail to see if they simply reflect the application of different criteria, or whether they are a consequence of program errors due to pattern recognition failure, mismeasurement, and/or deficient program logic. If the latter errors are common, then the program has serious deficiencies independent of its diagnostic criteria and its use in a clinical setting cannot be recommended.

If, however, the frequency of program errors is low, then the proportion of criteria differences becomes an important variable in deciding whether to implement a given computer program in a given institution. If the proportion of criteria differences is high, then the philosophy of approach to ECG diagnosis by the physicians and the computer program may be sufficiently different to limit the value of implementing such a program.

If both the percentage of criteria differences and the program error rates are low, then implementation of such a program may be warranted. It should be emphasized that with regard to criteria, the results for a given program depend upon the nature of the population that is under study. For example, a computer program the criteria of which are superbly designed
for an outpatient hypertension clinic may be grossly inadequate in a coronary care unit.

This method of evaluation when applied to the experimental IBM program of 1971 reveals a fairly low program error rate (4.4%) in our patient population. This corresponds well to the previously reported error rate in this program. At least one study has cast doubt upon whether human readers can maintain a comparably low error rate.

Diagnostic disagreements based upon criteria differences occurred in 20.1% of our cases (fig. 10). Whether this precludes the use of the IBM program for our patient population remains an open question.

It is clear that the IBM program must be subjected to some system of human review, particularly with respect to complex problems such as second or third degree atrioventricular block, the R on T phenomenon, or two sequential premature ventricular beats (PVCs). In these areas the program’s performance is unacceptable. Furthermore the program does not compare serial tracings recorded from an individual patient. At present this important task must be done by the human reader.

References
A Method for Evaluating Computer Programs for Electrocardiographic Interpretation: I. Application to the Experimental IBM Program of 1971*

JAMES J. BAILEY, SAMUEL B. ITSCOITZ, JOHN W. HIRSHFELD, JR., LEONARD E. GRAUER and MARTHA R. HORTON

Circulation. 1974;50:73-79
doi: 10.1161/01.CIR.50.1.73

Circulation is published by the American Heart Association, 7272 Greenville Avenue, Dallas, TX 75231
Copyright © 1974 American Heart Association, Inc. All rights reserved.
Print ISSN: 0009-7322. Online ISSN: 1524-4539

The online version of this article, along with updated information and services, is located on the World Wide Web at:
http://circ.ahajournals.org/content/50/1/73