Comparative Evaluation of the IBM (12-lead) and Royal Infirmary (Orthogonal Three-lead) ECG Computer Programs

PETER W. MACFARLANE, PH.D., DONALD I. MELVILLE, M.B., MARTHA R. HORTON, B.A., AND JAMES J. BAILEY, M.D.

SUMMARY A comparison of two computer programs for ECG interpretation was undertaken. Twelve-lead ECGs from 300 patients with various clinical abnormalities were interpreted at the National Institutes of Health using version 1 of the IBM program and corresponding orthogonal three-lead ECGs were analyzed by the Glasgow Royal Infirmary (GRI) program. Interpretations were compared with respect to the clinical documentation, wherever possible, and with each other directly in the case of diagnostic statements for which non-ECG documentation was not available. The two programs had a similar performance in determining abnormalities such as myocardial infarction and ventricular hypertrophy. However, with respect to conduction defects and ST-T-wave statements, certain discrepancies between the two program performances were revealed. There were 222 disagreements between various diagnostic statements. GRI was judged correct in 119 of these disagreements and IBM in 70. In these 189 cases the disagreement could most often be accounted for by different criteria and/or algorithms in the two programs or by the use of different ECG lead sets. The remaining 33 disagreements had to be classified as inconclusive.

THE ADVANTAGES of computer systems for analysis of the routine ECG are supposed to include:
(1) reduction in physician reading time, especially when the ECG load is heavy; (2) reduced variation in ECG interpretation; (3) enhanced teaching of ECG interpretation; (4) reduced technician, clerical, and storage costs; and (5) wider availability of the ECG reading service.

However, a program that makes too many erroneous measurements or statements cannot save a physician much reading time or enhance the teaching of ECG interpretation. Hence, accuracy remains a key issue, even though accuracy alone is not sufficient for these other economic and professional advantages to be realized.

Unfortunately, most evaluations of ECG programs have tended to rely upon unconstrained human judgment as a standard for ECG diagnosis. Intraobserver and interobserver variation have been well documented by Simonson et al. Therefore, it is not surprising that the results of evaluations should vary widely from one study to the next.

In addition to problems of objective standards, a comparative evaluation of two or more ECG programs imposes several other difficulties. Different programs use different terms, e.g., "repolarization abnormality" vs "nonspecific ST-T change," according to the tastes of local cardiologists. Recognizing semantic equivalents is a major task before diagnostic statements can be compared. Different programs use different criteria, reflecting the character of the patient population used for development; it is useful to know when a disagreement between two programs is due to a difference in criteria alone and not to any technical deficiency. Different programs also use different methods for signal processing, pattern recognition, feature extraction, and measurement; it is important to know when a particular algorithm (method) is consistently producing discrepancies. Finally, different programs may, as in this study, use different ECG lead systems. Occasionally a change may be seen in one lead system and not the other; this must be accounted for in the comparison.

In this study the axial lead program developed at Glasgow Royal Infirmary (GRI) was compared with a 12-lead program, version 1 of the IBM program. To date, no study has compared an axial lead program with any of the widely used Frank lead or 12-lead programs, even though the axial lead system is reported to have some advantages over the Frank lead system.

In this paper, we describe a comparative study that attempted to deal objectively with the difficulties listed above as applied to the GRI and IBM programs in a series of ECGs.

Methods and Materials

ECGs were collected on 300 unselected patients at Glasgow Royal Infirmary, although for technical reasons, 16 patients subsequently had to be excluded. The remaining 284 patients were distributed in diagnostic categories (table 1), according to criteria listed in the Appendix. For each patient, simultaneous XYZ leads of the modified axial ECG were recorded for approximately 20 seconds; the GRI program uses 8 seconds for analysis of morphology and rhythm. The ECG amplifiers used had a frequency response of 0.05-200 Hz, but on replay to the
PDP8E computer, the ECG signal was filtered under 125 Hz and converted to 11-bit digital data at 500 samples/sec. The digital data were then analyzed using methods previously described.4, 8

At the same recording session, the ECG leads were changed so that the 12-lead ECG could be obtained on the patient in four groups of three simultaneous leads, I, II, III; aVR, aVL, aVF; V1, V2, V6; and V4, V5, V6. A fifth lead group required by the IBM program was recorded, consisting of quasi-orthogonal leads (either aVF, and V1 or II, aVL, and V6). Each lead group was recorded for at least 10 seconds on a Hewlett Packard 3907B tape recorder using amplifiers with a frequency response of 0.05–1000 Hz. Analog tapes containing the ECG data were then forwarded to the National Institutes of Health (NIH). The NIH system passed the analog signal under 100 Hz and converted it to 12-bit digital data at 250 samples/sec, which is suitable for the IBM program. The digitized data were processed using version 1 of the IBM program with its optional digital notch filter set at 50 Hz.

After the ECGs were processed by GRI and IBM programs, copies of the program output were exchanged between GRI and NIH. A comparison of the outputs was independently pursued at each site according to a fixed set of classification and adjudication rules used at both sites. These rules were necessary to decide when the programs were in agreement and whether differences should be classified as significant or minor.

Results

Two hundred eighty-four of 300 ECGs were processable by the IBM program. Two of the 16 rejected were technically satisfactory ECGs with Wolff-Parkinson-White patterns that the IBM program was unable to recognize as coherent ECG. (The GRI program correctly reported the Wolff-Parkinson-White pattern in both of these cases.) The other 14 ECGs were rejected because of noise or other technical problems in the 12-lead ECG. The main source of this noise was a significant 50-Hz power-line interference. In an additional 117 ECGs, the IBM program was able to process the record, though it made notation concerning the presence of noise.

More than half of the processable cases represent coronary artery disease with or without myocardial infarct. Table 2 shows the accuracy of the programs with respect to infarct. The programs show about the same sensitivity, i.e., 81–84% for acute infarct and 76% for old infarct. The IBM program produced more false positives than GRI (12 vs six), and hence, the predictive value of a positive statement was lower for IBM (89% vs 94%) in this series. There were 24 cases of coronary artery disease without documented infarct that could not, therefore, be regarded as definitely positive or negative and were excluded. The IBM program detected 48 of 59 inferior infarcts and GRI detected 52. In four of these cases IBM gave its weakest statement, “consider ... infarct,” whereas GRI reported only ischemic changes. The IBM program detected 56 of 58 anterior infarcts while GRI detected 55. In some cases the location of the infarct was obvious from abnormal wall motion seen in the ventriculogram; in others the location was attributed according to the presence of Q waves with or without ST-T changes in appropriate leads.

There were 79 instances of ST-T-wave changes associated with infarction (four ECGs had more than one change). The GRI program correctly noted these changes in 92% and the IBM program in 63%. In some cases the IBM program failed to incorporate ST-T changes into the infarct diagnosis, even though it had measured them. In other cases the IBM program did not seem to recognize a biphasic T wave or improperly characterized it. In two cases the GRI incorrectly measured the J point and failed to report acute infarct. In two cases of left ventricular hypertrophy, a statement of anterior or anteroseptal infarct was falsely made and the T-wave change accompanying the left ventricular hypertrophy was falsely linked to the infarct statement by GRI.

Seventy patients had clinical evidence of left ventricular hypertrophy, whereas 138 had no such abnormality. The remaining 76 cases were omitted as being

*The predictive value of a positive statement is the number of true positives divided by the sum of the number of true positives plus the number of false positives.
TABLE 3. Accuracy of Hypertrophy Diagnoses by Programs

<table>
<thead>
<tr>
<th></th>
<th>IBM</th>
<th>GRI</th>
<th>Total (n)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Pos (n)</td>
<td>Neg (n)</td>
<td>Pos (n)</td>
</tr>
<tr>
<td>LVH</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>70</td>
<td>48</td>
<td>22</td>
</tr>
<tr>
<td>Negative</td>
<td>138</td>
<td>17</td>
<td>121</td>
</tr>
<tr>
<td>Inconclusive</td>
<td>76</td>
<td></td>
<td></td>
</tr>
<tr>
<td>RVH</td>
<td>24</td>
<td>3</td>
<td>21</td>
</tr>
<tr>
<td>Negative</td>
<td>141</td>
<td>6</td>
<td>135</td>
</tr>
<tr>
<td>Inconclusive</td>
<td>119</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Abbreviations: LVH = left ventricular hypertrophy; RVH = right ventricular hypertrophy; GRI = Glasgow Royal Infirmary; pos = positive; neg = negative.

Table 3 shows that the programs have about the same sensitivity (66-69%) in determining left ventricular hypertrophy and also about the same specificity (88-91%). The small difference in false positives (17 vs 14) did not significantly affect the predictive value of a positive statement (74% vs 78%). With respect to right ventricular hypertrophy (table 3), the sensitivity of both programs was low (14-17%), and the IBM program produced more false positives (six vs one); however, with such low numbers it is probably not meaningful to compute and compare the predictive values of the positive statement. One hundred nineteen cases were omitted as being neither positive nor negative with respect to right ventricular hypertrophy.

Table 4 shows the ability of the programs to discriminate normal from abnormal cases. There were 45 cases without cardiovascular disease. One of these cases had large QRS voltage and was labeled left ventricular hypertrophy by both programs; another case showed early repolarization in which both programs found “ST elevation.” In three cases IBM commented upon possible intraventricular conduction defect, whereas GRI made statements about ST-T changes. The remaining 40 cases were called normal by at least one of the programs.

The percentages of “true” normals and “false” normals were based on 45 patients without cardiovascular disease and 239 patients with disease, not on “normal” ECGs and “abnormal” ECGs (table 4). The predictive value of a “normal” population in this population is computed simply as the ratio of “true” normals to the “true” normals plus the “false” normals, and was almost identical for both programs.

All statements made by the programs were classified according to the rules mentioned above as follows: agreements, 384 (49.3%); disagreements, 222 (28.3%); and minor differences, 175 (22.4%). The disagreements occurred in the following categories: infarct, 52; hypertrophy, 43; and infarct-associated ST-T change, 32. Other frequently occurring disagreements included: “ischemia,” 29; conduction defects (including bundle branch block), 21; atrial changes (enlargement, overload), 20; and others, 25 (table 5).

All disagreements were examined to determine, wherever possible, which program was correct. In many cases this could be done on the basis of non-ECG clinical data as reflected by the results in tables 2-4. In some cases, atrial changes, for example, the diagnosis was difficult to document from the available clinical data. In seven of 20 statements of atrial hypertrophy, the disagreement had to be classified as inconclusive (table 6).

In cases of rhythm abnormalities or conduction defects, the ECG is the main evidence of the diagnosis. In these cases direct inspection of the tracing may reveal which program is correct. In 15 cases of disagreement, IBM made a statement of conduction defect (or bundle branch block) based on a measurement of a QRS interval that was at least 20 msec more than could be determined manually. These 15 statements were judged incorrect (table 6). In three cases the manual measurement differed from the IBM measurement of the QRS interval by less than 20 msec and these cases were classified as inconclusive.

Table 7 shows program performance with respect to

<table>
<thead>
<tr>
<th>TABLE 4. Accuracy of Normal Diagnoses by Programs</th>
</tr>
</thead>
<tbody>
<tr>
<td>&quot;Normal&quot; cases</td>
</tr>
<tr>
<td>IBM (n)</td>
</tr>
<tr>
<td>---------</td>
</tr>
<tr>
<td>Normals</td>
</tr>
<tr>
<td>Cardiovascular disease</td>
</tr>
<tr>
<td>Predictive value</td>
</tr>
</tbody>
</table>

Abbreviation: GRI = Glasgow Royal Infirmary.

<table>
<thead>
<tr>
<th>TABLE 5. Classification of Other Frequently Occurring Diagnoses</th>
</tr>
</thead>
<tbody>
<tr>
<td>Agree (n)</td>
</tr>
<tr>
<td>-----------</td>
</tr>
<tr>
<td>&quot;Ischemia&quot;</td>
</tr>
<tr>
<td>Conduction defects (incl. bundle branch block)</td>
</tr>
<tr>
<td>Atrial changes (enlargement, hypertrophy)</td>
</tr>
<tr>
<td>Other ST-T changes</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>TABLE 6. Adjudication of Other Frequent Disagreements</th>
</tr>
</thead>
<tbody>
<tr>
<td>Correct program</td>
</tr>
<tr>
<td>GRI (n)</td>
</tr>
<tr>
<td>---------</td>
</tr>
<tr>
<td>Atrial overload</td>
</tr>
<tr>
<td>Conduction defects</td>
</tr>
<tr>
<td>&quot;Ischemia&quot;</td>
</tr>
</tbody>
</table>

Abbreviations: INC = inconclusive; GRI = Glasgow Royal Infirmary.
atrial fibrillation, which was the major rhythm disturbance in this population. The performance of the programs was rather good, with the IBM program yielding three more true positives and one more false positive than the GRI program.

Because “ischemia” as a pathophysiologic state is difficult to document, it was treated in this study as a descriptive term that was applied to an ST-T change whenever either program used it. In most cases when one program used the term, the other did also or made a similar ST-T comment (table 5). An attempt was made to resolve disagreements in this area by direct inspection of the ECG tracing, which was recorded on an Elema Mingograf 34, which has a flat frequency response to almost 700 Hz. If ST-T changes were clearly discernible in the corresponding leads of the two systems, e.g., in the right precordial leads of the 12-lead system and the anteroseptal lead of the axial system, then the program that failed to comment upon these changes was judged incorrect. The IBM program was judged incorrect in many cases because it seemed not to recognize biphasic T-wave patterns, or measured T-wave inversion but made no diagnostic comment. GRI reported “ischemia” in five cases in which appropriate T-wave changes were not found. This statement resulted from the criterion based on ST-T area and could be triggered when a minor QRS notch or early repolarization caused a premature marking of the ST onset (table 6).

In all, GRI was judged correct in 119 disagreements, IBM correct in 70, and 33 disagreements were left as inconclusive (r = 0.001*). Of this difference, 47 of 49 disagreements were caused by three apparently defective IBM algorithms, namely, the determination of QRS width, the treatment of infarct-related ST-T changes, and the treatment of biphasic ST waves. Excluding these disagreements, there was no significant difference in the performance of the programs in the remaining statements as a whole or in any particular category.

### Discussion

It has long been argued that the proper way to evaluate ECG computer programs is to use cases in which the diagnosis has been established by ECG-independent means. However, this approach, though important, has obvious limitations. In this study, we followed the recommendations of the Tenth Bethesda Conference of the American College of Cardiology and divided ECG diagnostic statements into three distinct types.

Type A statements refer to anatomic lesions and pathophysiologic states (e.g., infarct, hypertrophy) that can be documented by non-ECG tools such as enzyme studies, cardiac catheterization studies, echocardiography and autopsy. Recently, radionuclide studies of the heart have been shown to be sensitive and locally accurate in the diagnosis of myocardial damage and ischemia, and some metabolic and drug-related changes can be documented by serum chemistries.

Type B statements refer to electrophysiologic states or events (i.e., rhythm and conduction disturbances). In contrast to the tools for documentation listed above, the ECG is usually the main evidence for these diagnoses. More recently some of these diagnoses have been “documented” through the use of sophisticated electrophysiologic tools such as surface potential maps or multiple, invasive probes.

Type C statements are ECG descriptors (e.g., “abnormal repolarization”) that may not have diagnostic significance when taken by themselves but may become important when considered in the light of other clinical data.

This study shows about the same level of performance in both programs for the type A diagnoses of infarct and hypertrophy. There was more discrepancy in performance in dealing with infarct-associated ST-T changes that were also treated as type A statements. The IBM program failed to incorporate ST-T changes in some cases even though it measured them (e.g., inverted T wave). This seems to be a matter of the design of criteria. In other cases, the IBM program seemed not to characterize a biphasic T wave adequately. This seems to be a more fundamental problem in pattern recognition. A few GRI errors in this area resulted from incorrect determination of ST onset. Another error resulted because a T-wave change was linked to a false-positive statement of infarct; this is a complex problem possibly involving criteria design and measurement algorithms.

The performance of the programs was very similar in ventricular hypertrophy for the unselected hospital population under study. The specificity for left ventricular hypertrophy was not as high as that reported by Romhilt et al. for the Romhilt-Estes criteria, but the latter was based on a “training,” rather than a test, group of hearts examined postmortem. However, the specificity of the programs is in the same range as the Sokolow-Lyon criteria (87-90%). The sensitivity of the programs (66-69%) is significantly higher than that reported for both Sokolow-Lyon and Romhilt-Estes criteria (54-56%). However, left ventricular hypertrophy is notoriously variable, and these results need to be confirmed in a larger population before a conclusive comparison can be made with data from the literature.

The performance of the programs on type B statements is reflected in the results with respect to
atrial fibrillation and conduction defects. In contrast to the results with atrial fibrillation, there appeared to be a marked difference between the programs with respect to intraventricular conduction defects. The wide QRS intervals measured by the IBM program have also been reported by others. These accounted for the majority of disagreements in this area. The GRI program described axis deviation and/or counterclockwise rotation in most cases of left anterior hemiblock but failed to state left anterior hemiblock because of a minor criterion problem.

The criteria for type C statements are not subject to tests for sensitivity and specificity as are type A statements. Perhaps that is why the criteria, and even the terms themselves, vary so much from one medical center to another. Review of disagreements over “ischemia” was useful in revealing problems with pattern recognition and measurement; however, because this review depended upon human judgment as to whether an ST-T change was definite, equivocal or absent, it revealed nothing about the criteria or the utility of the term itself. Because reviewers at each site used a different ECG lead system, it was not feasible to fix criteria for “ischemia” as Bailey et al. recommended for evaluations of this type.

In summary, this method of evaluation when applied to two or more programs can reveal their relative strengths and weaknesses. The reason for a program failure can often be pinpointed to a specific defect in pattern recognition, measurement algorithm or design of criteria. Occasionally, the ECG itself fails to reflect disease and in such cases the programs cannot be faulted.

The results of this study could not be used to establish superiority of the 12-lead ECG over the axial lead ECG, or vice versa. This tends to support the premise of some authors that the XYZ (Frank or axial) lead system has as much clinically significant information as the 12-lead ECG. This appears to be true if one simply looks at sensitivity and specificity, for example, of infarcts (table 2). But the true positives for each program did not completely overlap, i.e., GRI detected some infarcts which IBM didn’t and vice versa. The same was true for hypertrophy (table 3).

If, as these results might suggest, different information is available in different lead systems, it would make sense to devise a new system that would incorporate all this information. This idea has led to the development of a hybrid 15-lead system (XYZ + 12 leads) and a system consisting of nine “optimally” placed chest leads. Such multilead systems are now more readily attainable because recent advancements in electronics and microprocessors allow the acquisition and rapid processing of data from multiple leads.

The results of this comparison study support the suggestion that the next major development should be computer programs that make use of advanced mathematical tools, such as multivariate analyses and orthogonal transformations, to extract more completely the information available in such multilead systems.

Appendix

Criteria for Non-ECG Documentation

Myocardial Infarction

Typical clinical-history plus typical myocardial enzyme changes (CPK, SGOT) or abnormal wall motion by ventriculography.

Coronary Artery Disease

At least 50% narrowing of one or more major arteries by angiography.

Left Ventricular Hypertrophy

Overload of the left ventricle (e.g., hypertension, aortic stenosis, aortic regurgitation, etc.) plus one other feature (e.g., cardiomegaly, congestive failure, palpable thrust, etc.).

Right Ventricular Hypertrophy

Overload of the right ventricle (e.g., ≥ 2:1 left-to-right shunt, mitral stenosis, pulmonic or tricuspid disease, pulmonary disease, etc.) plus one other feature (e.g., palpable right ventricular heave, cardiomegaly, signs of right ventricular failure, etc.).

Valvular and Pulmonary Disease

All valvular disease was determined by complete cardiac investigation including cardiac catheterization studies. Pulmonary disease was confirmed by x-ray and pulmonary function studies.

References


27. Pipberger HV, Bialek SM, Perllof JK, Schnaper HW: Correlations of clinical information in the standard 12-lead ECG and corrected orthogonal 3-lead ECG. Am Heart J 61: 34, 1961


32. Kornreich F: The missing waveform information in the orthogonal electrocardiogram (Frank leads). I. Where and how can this missing waveform information be retrieved? Circulation 48: 984, 1973

Comparative evaluation of the IBM (12-lead) and Royal Infirmary (orthogonal three-lead) ECG computer programs.
P W Macfarlane, D I Melville, M R Horton and J J Bailey

doi: 10.1161/01.CIR.63.2.354

Circulation is published by the American Heart Association, 7272 Greenville Avenue, Dallas, TX 75231
Copyright © 1981 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/63/2/354

Permissions: Requests for permissions to reproduce figures, tables, or portions of articles originally published in Circulation can be obtained via RightsLink, a service of the Copyright Clearance Center, not the Editorial Office. Once the online version of the published article for which permission is being requested is located, click Request Permissions in the middle column of the Web page under Services. Further information about this process is available in the Permissions and Rights Question and Answer document.

Reprints: Information about reprints can be found online at:
http://www.lww.com/reprints

Subscriptions: Information about subscribing to Circulation is online at:
http://circ.ahajournals.org//subscriptions/