A Method for Evaluating Computer Programs for Electrocardiographic Interpretation

III. Reproducibility Testing and the Sources of Program Errors

By James J. Bailey, M.D., Martha Horton, B.A., and Samuel B. Itscowitz, M.D.

SUMMARY

A simple method for testing reproducibility in ECG computer program performance results from using two digital representations of the same analog ECG tracing. Each digital representation is separated from the other by one millisecond in time. When the digital representations are processed by the Mayo Clinic program (1968), the diagnostic statements are identically reproduced in only 60% of 33 tracings. When the method is applied to version D of the PHS program and to the newly released IBM program of 1973, identical reproducibility is 43.5% and 76.0%, respectively, of 217 tracings. After analog filtering these figures are improved to 49.8% and 97.7%, respectively. These results show that reproducibility is most affected by a program’s algorithms for pattern recognition, measurement, consistency checking, and noise handling. Reproducibility is less affected by attenuation of high frequency noise at the analog level. The relationship of reproducibility to program error rate in previous studies is discussed. Hence poor performance on this test obviates the need for a more time-consuming clinical evaluation. The need for human overview and quality checking is re-emphasized.

Additional Indexing Words:

Computer Analysis

In the first two papers of this series, we made an effort to separate the disagreements between computer programs and readers into criteria differences and program errors. We have emphasized the importance of this point, namely, that if the program error rate is high, then the program has serious deficiencies and is unacceptable regardless of the validity of its criteria. However, the studies described in these two papers are tedious and time-consuming; it is generally not feasible for the average potential user to pursue such a study when he contemplates implementing a specific program. The purpose of this paper is, therefore, to describe a simple method by which a potential user can evaluate a program’s performance, independent of its clinical accuracy or the validity of its criteria.

In order to understand the sources of program errors, it is necessary to outline the major phases of a computer system for ECG analysis as follows:

I. Signal Conditioning
   - Analog Filtering
   - A/D Conversion
   - Digital Filtering

II. Intermediate Processing
   - Pattern Recognition
   - Measurement Algorithms
   - Consistency Checks
   - Noise Handling

III. Diagnosis
   - Logic Structure
   - Criteria Thresholds

The accuracy of an ECG computer program depends upon the performance of all three phases; however, reproducibility depends mainly upon the first two phases. (Phase III is involved only in the rare circumstance where a contour measurement is virtually right at some threshold value, thereby allowing a very small variation in that measurement to alter a contour statement.)

Methods and Materials

Reproducibility was tested with two digital data sets, each extracted from the same analog ECG and each separated from the other by a one millisecond phase lag in sample points by the following process:

First, the analog ECG is digitized at 1000 samples per sec-
The top trace in figure 1 is a representative plot of such data. Then the first data set is made up from all the odd samples, 1, 3, 5, etc., for each lead, and hence represents a sampling rate of 500 per second. The result of this process is plotted in the second trace of figure 1. The second data set is made up from all the even samples, 2, 4, 6, etc., for each lead, and also represents a sampling rate of 500. It is plotted in the third trace of figure 1.

This digital extraction was performed on a series of 217 unselected tracings. Of these tracings, 85 were interpreted as “normal” by one of the authors (JJB). For a given ECG, each of the two digital data sets was separately submitted to an interpretative program and the qualitative statements were compared. This was accomplished for both the PHS program (version D) and the newly released IBM program (1973).*

The effect of filtering was tested in the 217 tracings above as follows: The entire procedure was repeated exactly the same except that the analog signal was passed under a 6 pole, 50 Hz Butterworth filter just before digitizing. The typical result of such filtering is plotted in the fourth trace in figure 1.

The Mayo program (1968) was tested with 33 other unselected, unfiltered tracings in the same way. All three programs were implemented on the IBM 370 system at the Division of Computer Research and Technology, NIH.

Results

Tables 1 and 2 show the results of the PHS program operating upon unfiltered and filtered data respectively. These results are tabulated by type of statement; e.g., “myocardial infarct” and “left ventricular hypertrophy” are registered as QRS statements. “Repeated” means the statement appeared in both sets of outputs without alteration. “Altered” means the statements had essentially the same diagnostic meaning but with a different level of severity; e.g., “consider infarct” versus “possible infarct.” “Omitted” means a statement in one set was omitted in the other set.

A number of tracings had identical output except for a single slight alteration of contour statement; e.g., a change of “LV by voltage criteria” to “left ventricular hypertrophy” or of “vertical axis” to “right axis deviation.”

The PHS program places each ECG in one of four categories: normal, atypical, borderline, and abnormal. Hence the number of cases where category assignment was not reproduced is also tabulated. In many of the cases which were not normal, the category assignment remained the same, even though there was a clinically significant change in a contour statement.

*The experimental IBM program of 1971 is essentially the same as the program IBM released in 1973. (The latter is a 370 version of the former, which was written for the 1900). A test of the 1800 version has been reported.* (personal communication from Dr. R. E. Bonner).

![Figure 1](http://circ.ahajournals.org/)

The top trace is a plot of the digital data at a sampling rate of 1000 samples per second. The second trace is a plot of the odd samples, and the third trace, the even samples; both these traces represent digital data at 500 samples per second. The bottom trace is the digital data at 1000 samples per second but filtered under 50 Hz just before digitizing.

Tables 3 and 4 show the results of the IBM program on unfiltered and filtered data respectively. The IBM program makes no category assignment; therefore, that item does not appear. Otherwise, the items are the same as in tables 1 and 2.

Table 5 shows the reproducibility of the Mayo program on 33 unfiltered ECGs. This program gives out three categories: normal, borderline, and abnormal.

Reproducibility of the Mayo Clinic program (1968) was not pursued beyond table 5 because it has been replaced at the Mayo Clinic by a new program.* However, other workers using consecutive analog tracings have shown similar results. Bruce et al.*

### Table 1

<p>| Reproducibility of Results in 217 Unfiltered ECGs When Two Separate Digital Data Sets from the Same Analog Tracing are Processed by the PHS-D Program |
|---------------------------------|---------|---------|--------|----------|</p>
<table>
<thead>
<tr>
<th>Statements</th>
<th>QRS</th>
<th>ST-T</th>
<th>P</th>
<th>Rhythm</th>
</tr>
</thead>
<tbody>
<tr>
<td>Repeated</td>
<td>149</td>
<td>98</td>
<td>38</td>
<td>96</td>
</tr>
<tr>
<td>Altered</td>
<td>18</td>
<td>16</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Omitted</td>
<td>82</td>
<td>51</td>
<td>52</td>
<td>23</td>
</tr>
<tr>
<td>Total</td>
<td>249</td>
<td>165</td>
<td>91</td>
<td>120</td>
</tr>
</tbody>
</table>

### Table 2

<table>
<thead>
<tr>
<th>Category Changes</th>
<th>27</th>
<th>27 Different Leads Rejected</th>
</tr>
</thead>
<tbody>
<tr>
<td>Identical Readings</td>
<td>94/217 (43.3%)</td>
<td></td>
</tr>
<tr>
<td>Normal</td>
<td>75/85 (88.2%)</td>
<td></td>
</tr>
<tr>
<td>Not Normal</td>
<td>19/132 (14.4%)</td>
<td></td>
</tr>
<tr>
<td>Slight contour change only</td>
<td>43</td>
<td></td>
</tr>
<tr>
<td>Reasonably Reproduced</td>
<td>137 (63.1%)</td>
<td></td>
</tr>
</tbody>
</table>
Table 2

Reproducibility of Results in 217 ECGs Filtered Under 50 Hz Just Before Digitizing When Two Separate Digital Data Sets from the Same Analog Tracing are Processed by the PHS-D Program

<table>
<thead>
<tr>
<th>Statements</th>
<th>QRS</th>
<th>ST-T</th>
<th>P</th>
<th>Rhythm</th>
</tr>
</thead>
<tbody>
<tr>
<td>Repeated</td>
<td>160</td>
<td>111</td>
<td>34</td>
<td>104</td>
</tr>
<tr>
<td>Altered</td>
<td>19</td>
<td>10</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>Omitted</td>
<td>63</td>
<td>51</td>
<td>34</td>
<td>12</td>
</tr>
<tr>
<td>Total</td>
<td>242</td>
<td>172</td>
<td>68</td>
<td>119</td>
</tr>
</tbody>
</table>

Category Changes 20 Different Leads Rejected 16
Identical Readings 108/217 (49.8%)
Normal 81/85 (95.3%)
Not Normal 27/132 (20.5%)
Slight contour change only 53
Reasonably Reproduced 161 (74.2%)

reported category changes in 31.8% (5/16). Carlier et al.4 reported complete reproducibility in less than 46.3% (67/145) and category changes in 20.0% (29/145). Table 5 shows category changes in only 10.0%, complete reproducibility in 60.0%, and reasonable reproducibility in 69.7%.

Discussion

This procedure for testing reproducibility is better than using two consecutive analog tracings. It eliminates the problem of a baseline wander, a variation in noise, or an artifact in one tracing but not the other. In the procedure used here there is only one analog tracing; only the digital representations are different.

Except for digital filtering by the interpretative program itself, phase I input is the same for each program. Hence, the procedure used in this study especially tests phase II, the intermediate processing.

Table 3

Reproducibility of Results in 217 Unfiltered ECGs When Two Separate Digital Data Sets from the Same Analog Tracing are Processed by the IBM Program

<table>
<thead>
<tr>
<th>Statements</th>
<th>QRS</th>
<th>ST-T</th>
<th>P</th>
<th>Rhythm</th>
</tr>
</thead>
<tbody>
<tr>
<td>Repeated</td>
<td>131</td>
<td>66</td>
<td>13</td>
<td>233</td>
</tr>
<tr>
<td>Altered</td>
<td>9</td>
<td>0</td>
<td>0</td>
<td>13</td>
</tr>
<tr>
<td>Omitted</td>
<td>25</td>
<td>34</td>
<td>5</td>
<td>9</td>
</tr>
<tr>
<td>Total</td>
<td>165</td>
<td>100</td>
<td>18</td>
<td>255</td>
</tr>
</tbody>
</table>

Tracings Rejected 7 (both times)
Identical Readings 165/217 (76.0%)
Normal 81/85 (95.3%)
Not Normal 84/132 (63.6%)
Slight contour change only 23
Reasonably Reproduced 188/217 (86.6%)

The effect of altering phase I inputs is tested by heavy analog filtering under 50 Hz. This heavy filtering will blunt the QRS waves as shown in figure 1. Therefore, a marked reduction in noise will be accompanied by some loss in accuracy.

Even with heavy filtering, the reproducibility of the PHS program (table 2) does not achieve the performance of the IBM program operation on unfiltered data (table 3). The reason is the following: the main factor affecting a program’s reproducibility lies in its phase II characteristics, some of which are outlined for the three programs in table 6.

The beginning or end of a QRS may appear flat in one lead but show a visible deflection in another simultaneous lead. Therefore, if a human reader uses two or more simultaneous leads to make these determinations, he can be more precise, even if it is not always worth the extra time. The same is true for computer algorithms for detecting onset and offset of QRS or any other wave, except the extra time is not a significant cost. Mayo and IBM programs should show an advantage over the PHS program in this regard. The PHS program was constructed in 1964 when ECG carts which recorded three leads simultaneously were not widely available. In 1974 this is no longer true.

The algorithm to detect QRS onset was one of the major faults in the Mayo program. This algorithm produced varying QRS durations, and in 33 ECGs (see previous study9), a diagnosis of intermittent aberrant conduction when no aberrancy was, in fact, manifested. A number of false diagnoses of infarct and left axis deviation (later forces) was also attributable to this algorithm. Therefore this algorithm will also adversely affect the reproducibility of those statements which depend upon a precisely determined QRS onset.

The precision of a measurement algorithm which

Table 4

Reproducibility of Results in 217 ECGs Filtered Under 50 Hz Just Before Digitizing When Two Separate Digital Data Sets from the Same Analog Tracing are Processed by the IBM Program

<table>
<thead>
<tr>
<th>Statements</th>
<th>QRS</th>
<th>ST-T</th>
<th>P</th>
<th>Rhythm</th>
</tr>
</thead>
<tbody>
<tr>
<td>Repeated</td>
<td>143</td>
<td>70</td>
<td>13</td>
<td>219</td>
</tr>
<tr>
<td>Altered</td>
<td>8</td>
<td>0</td>
<td>0</td>
<td>19</td>
</tr>
<tr>
<td>Omitted</td>
<td>30</td>
<td>20</td>
<td>2</td>
<td>13</td>
</tr>
<tr>
<td>Total</td>
<td>181</td>
<td>90</td>
<td>15</td>
<td>251</td>
</tr>
</tbody>
</table>

Tracings Rejected 1 (both times)
Identical Readings 173/217 (79.7%)
Normal 83/85 (97.6%)
Not Normal 90/132 (63.0%)
Slight contour change only 23
Reasonably Reproduced 198/217 (90.3%)
uses only one beat depends on how well the program selects a "typical" beat (a task at which the human reader is far superior). A number of program errors in the PHS and Mayo programs were traced back to the fact that the program had selected a nonrepresentative beat, one affected by baseline wander, variation in noise, or artifact.

In greater than 90% of the tracings, the PHS program selects a different beat in one or more leads to do its measurements which further degrades reproducibility of the diagnostic statements.

Averaging measurements over several beats obviates the risk of a poor choice of beat; furthermore, it suppresses the effect of noise and respiratory variation in a way that simple analog filtering does not approach, as is reflected in the results of tables 1-4.

The PHS program has checks for reversed arm leads and misplacement of lead V1. It has practically no other checks for consistency of measurement among leads. For example, it determines the axis from two leads and does not check with the other four. It produces up to 12 values for each of P-R, QRS, and Q-T intervals, one value for each lead measured.

Table 6

A Comparison of Corresponding Characteristics in the Intermediate Processing Phase of Three Programs

<table>
<thead>
<tr>
<th></th>
<th>PHS</th>
<th>Mayo</th>
<th>IBM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pattern Recognition</td>
<td>Separate Leads</td>
<td>Simultaneous Leads</td>
<td>Simultaneous Leads</td>
</tr>
<tr>
<td>Measurement Algorithms</td>
<td>Single Beat</td>
<td>Single Beat</td>
<td>Average of Beats</td>
</tr>
<tr>
<td>Consistency Checks</td>
<td>Weak</td>
<td>Not Required</td>
<td>Extensive</td>
</tr>
<tr>
<td>Noise Handling</td>
<td>Poor</td>
<td>None</td>
<td>Sensitive and Appropriate</td>
</tr>
</tbody>
</table>

These values vary considerably and are not checked against one another. The program does not indicate which of the 12 values is the accurate one.

The IBM program organizes the 12 standard leads into four lead sets each of which contains three simultaneous leads. For each beat, onset and offset of each wave is sought across the three simultaneous leads. The P-R, QRS, and Q-T intervals are averaged for all the beats in a given lead set, then the average of these intervals for all the lead sets is computed. This method has the advantage of suppressing random noise (e.g., patient muscle noise) which improves its accuracy and its reproducibility. Additionally, the

These values vary considerably and are not checked against one another. The program does not indicate which of the 12 values is the accurate one.

The IBM program organizes the 12 standard leads into four lead sets each of which contains three simultaneous leads. For each beat, onset and offset of each wave is sought across the three simultaneous leads. The P-R, QRS, and Q-T intervals are averaged for all the beats in a given lead set, then the average of these intervals for all the lead sets is computed. This method has the advantage of suppressing random noise (e.g., patient muscle noise) which improves its accuracy and its reproducibility. Additionally, the
IBM program checks axis determinations from each of three pairs of limb leads, one against the other. If they are not consistent, the program reads out "undetermined axis." If the program finds severe inconsistencies, it may reject the tracing altogether.

The Mayo program treats a single lead set of simultaneous XYZ leads; hence it does not need interlead consistency checks of the type just described.

Furthermore, the Mayo program was designed with the assumption that the quality of the tracing will be checked by a human before processing by computer. Therefore, it has few or no algorithms to handle noise. It will process a tracing of random noise as though it were an intelligible ECG.

In contrast, PHS has various algorithms for rejecting leads because of noise or artifact. A common mistake in this program is attributable to the triggering of a rejection algorithm by the F waves of atrial fibrillation. Several QRS diagnoses were missed this way because the critical lead was rejected. On the other hand, occasionally a lead containing noise is accepted. The program does not modify the conclusions from this lead. Note in table 1 that this rejection algorithm is not very reproducible; in 27 cases a lead rejection was not duplicated. Heavy filtering under 50 Hz improves this performance somewhat.

The IBM program is sensitive to noise as shown by the rejection of seven tracings in the unfiltered data, and only one of the filtered tracings. In this series of 217 tracings, the rejection algorithm was reproducible in both sets of digital data. In the previous study there were several tracings where the IBM program issued a statement of uncertainty concerning the presence or absence of an infarct because of F waves or noise in one lead (typically II, III, or aVF).

Figure 2 shows the digital data from the analog record of a battery powered square wave generator, a device which is regularly used to check the quality and frequency transmission of the data collection system, from cart to computer. The data was fed to the program which produced the output shown in figures 3 and 4. This rather bizarre example serves to illustrate many of the program characteristics which have just been discussed.

In the previous studies roughly half of the program errors were in contour statements and were principally attributable to problems in phase II processing.

The other half of the program errors were in rhythm diagnoses where the situation is more complicated. Tables 1-4 show that about 10% of the time the rhythm statement is not reproduced; this is mainly a

---

**Figure 4**

*Analysis of the square wave data by the Mayo Clinic and IBM programs.*

*Circulation, Volume 50, July 1974*
problem in phase II processing. However, some inaccuracies in rhythm analysis are a result of inadequate logic in phase III. For example, the PHS program does not distinguish between atrial and ventricular premature systoles. It does not seek ventricular beats when atrial fibrillation has been diagnosed. These are clearly programming deficiencies of a logical nature.

When the rhythm is atrial fibrillation, the Mayo program occasionally reads out atrioventricular dissociation. It is not clear whether this kind of program error is purely a phase II or phase III error or a combination of the two. The IBM program has the singular advantage that when a tracing fails to meet the logic and criteria for a specific rhythm diagnosis, then the program reads out undetermined rhythm.

Therefore, if a program employs advanced techniques — namely, treatment of simultaneous leads, averaging of measurements over several beats, checks for consistency and quality, and appropriate handling of noise — then phase III processing as well as reproducibility will be improved. The percentage of program errors will be reduced as shown in the previous studies\(^1^2\) of 1150 tracings. In the opinion of Bruce et al.\(^4\) no program should be allowed to operate without routine overview and quality checking by a human reader. The results of this reproducibility study and program error rate in 1150 tracings previously studied\(^1^2\) confirm the need for human overview and quality checking.

In summary we have shown a simple method for testing a program’s performance independent of its clinical accuracy or the validity of its criteria. A poor performance on this test obviates the need for more tedious and time-consuming clinical evaluations.

**References**


A Method for Evaluating Computer Programs for Electrocardiographic Interpretation: III. Reproducibility Testing and the Sources of Program Errors

JAMES J. BAILEY, MARTHA HORTON and SAMUEL B. ITSCOITZ

Circulation. 1974;50:88-93
doi: 10.1161/01.CIR.50.1.88

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/50/1/88