Upsala J Med Sci 84: 37–46, 1979

Evaluation of a Computer Programme for Interpretation of 12-lead Electrocardiograms

Johan Landelius and Lars Nordgren

From the Department of Clinical Physiology, University Hospital, Uppsala, Sweden

ABSTRACT

A 12-lead electrocardiogram (ECG) interpretation programme (Cardionics, Brussels) was tested for computer diagnosis of ECG. The computer performance was evaluated on an ECG population from hospital patients by a method giving results which will allow the clinician to judge the usefulness of the computer diagnosis in clinical practice. Diagnoses made by experienced ECG readers were in essential agreement with the computer diagnoses in 83.5 % of 493 ECGs. Clinically significant disagreements due to differences in criteria occurred in 6.0 % of the tracings,whereas such disagreements due to programme errors were found in 10.5 %.

INTRODUCTION

Several computer programmes for interpretation of electrocardiograms (ECG) are available today. Evaluation of such programmes is time-consuming and in some aspects difficult, especially concerning their performance in the clinical situation. Firstly it is essential to define the composition of the tested population of ECGs in terms of the prevalence of certain abnormalities. The ideal would be to have access to an archive of ECGs accepted generally by investigators and including ECGs with abnormalities due to diseases of a known type and at a known stage, as well as ECGs from normal subjects. The performance of a computer programme could then be determined in relation to a particular ECG abnormality, to a certain disease, and to a normal as well as a mixed ECG population.

Secondly, it is important to analyse disagreements between the computer diagnosis and the correct diagnosis, that is the diagnosis made by manual interpretation, and to classify them into those due to differences in criteria and those due to programme errors. The latter will consist either of mismeasurements or of deficient programme logic. As pointed out by Bailey et al. (1), criteria differences do not indicate any technical deficiency in the programme.

Thirdly, it is also necessary to consider the possibility of human reader errors and failures in the recording quality in studies of this kind.

MATERIAL AND METHODS

A material of 510 unselected ECGs was obtained at random from patients sent to our department from different wards and outpatient units of the University Hospital in Uppsala. The tracings were collected from consecutive cases examined in one of four routine ECG stations between June and September, 1974. The tracings represent a hospital population and many of the patients had cardiopulmonary diseases. Each recording consisted of the standard 12-lead system (I, II, III, aVR, aVL, aVF, V_1-V_6) and the orthogonal vectorcardiographic leads X, Y and Z as described by Frank (2).

A Cardionics cart was used for the recordings, which were made both on a strip chart and on high fidelity FM analogue tape. The tape recordings were sent to Cardionics, Brussels, where they were digitized and computerized by the Mount Sinai Hospital programme (1973), whereafter the computer printout of each ECG diagnosis as well as a D-A-converted printout of the same ECG tracing were returned to our department. Six ECGs had to be excluded - two of them were lost for unknown reasons and four were lost in the technical process. Of the 504 ECGs remaining for further analysis, 10 were excluded because of faulty lead connections, as discussed later, and one was discarded because it could not be classified. This left 493 ECGs, obtained from 493 patients (281 male and 212 female). The age range of the men was 19 to 85 years (mean 58; S.D. 14) and of the women 17 to 89 years (mean 54; S.D. 17).

Each ECG was interpreted in detail by the authors, neither of whom was aware of the computer interpretation or of the interpretation of the other reader at the time of the first reading. The two manual interpretations of each ECG were then combined by the two readers to fit a set of ECG criteria common to the department and in agreement with international rules (3). This combined "departmental" interpretation is referred to in the following as the "reader diagnosis". The individual reader error appeared to be small but was not further analysed. The vectorcardiographic leads were not considered by the readers in their manual interpretation but were used by the computer programme, mainly to determine loop areas and vector angles.

The reader diagnosis was then compared with the diagnosis produced by the computer. The comparison was made essentially along the lines proposed by Bailey et al. (1). Their definitions for agreement and disagreement were modified as follows.

1. Reader-computer agreement.

"Agreement" was defined as an identical interpretation by the reader and the computer (group A). No further analysis was performed.

2. Reader-computer disagreement.

"Disagreement" was divided into (a) minor disagreement probably of no clinical importance (group B); (b) major disagreement probably of clinical importance (group C); and (c) major disagreement definitely of clinical importance (group D).

In each of groups B, C and D, the disagreements were classified into those due to criteria differences and those due to programme error. The complete set of criteria used by the computer programme was available. Most of the abnormal ECGs in this study contained multiple abnormalities. In groups C and D, only those diagnostic statements which carried clinical significance were noted and submitted to statistical analysis. As mentioned above, 10 tracings had faulty lead connections. This was done intentionally by the clinical technologists in order to test the computer. In seven of these, two precordial leads had been exchanged and as a result the computer gave a false statement of anterior myocardial infarction. In the remaining three, extremity leads had been exchanged. Two of these were correctly identified as "lead wire inversion" but in the third case the computer report was: "unusual electrical axis, compatible with left ventricular hypertrophy". Thus the programme seemed to have difficulty in cases of faulty lead connections. This emphasizes the need for correct recording of the ECG, and also calls for supplementation of the programme logic.

STATISTICS

The following formulas were used, as proposed by Rautaharju et al. (4).

a=true positives, b=false positives, c=false negatives, d=true negatives
Sensitivity (SE) = 100 a/(a+c), specificity (SP) = 100 d/(b+d), accuracy of positive tests (AP) = a/(a+b), accuracy of negative tests (AN) = d/(c+d), error ratio = (b+c)/a,

$$\text{over-all diagnostic accuracy (DA)} = \sum_{I=1}^{N} P(I)\ SE(I)$$

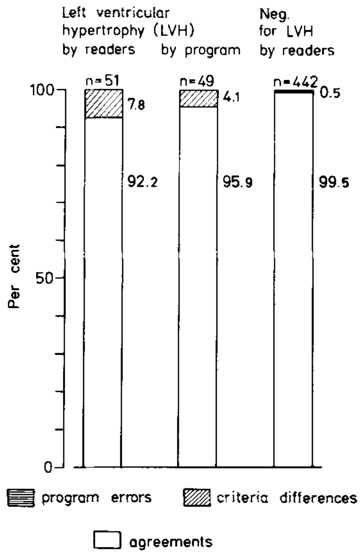
N = number of diagnostic test categories (I)
P(I) = fraction of statements in category (I)
SE(I) = fraction of correctly diagnosed statements in category (I)

None of these indices gives a fully representative and relevant picture of disagreement and failure. The basic limitation is that all events are classified as of equal importance, the more advanced concepts of statistical decision theory being disregarded. However, a more thorough discussion of statistical principles is beyond the aim of this study.

RESULTS

Out of 493 ECG tracings, the reader diagnosis in 234 ECGs comprised one or more clinically significant abnormalities (47.5 % abnormal and 52.5 % normal). The exact number of ECGs in each kind of reader statement is given in the Figures. The results are expressed in bar graphs. The first bar in each figure shows the sensitivity of the computer diagnosis in relation to the reader diagnosis, and the third bar shows the specificity. The total number of diagnoses for each graph was 493, except for the graphs of myocardial infarction and arrhythmia, where the totals were 506 and 529 respectively, due to the occurrence of more than one statement for a given ECG. In the final statistical analysis, groups A and B were both regarded as agreements, and groups C and D as disagreements.

1



2

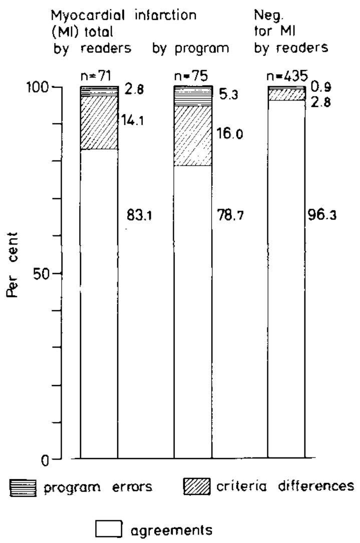


Fig. 1. Left ventricular hypertrophy (LVH). In each bar graph the number (n=) written above the first bar represents the total number of times the readers made the diagnostic statement listed. The number (n=) over the middle bar represents the total number of times the programme made the diagnostic statement in question. The number (n=) over the third bar represents the total number of tracings in which the readers did not make the diagnostic statement under consideration. Each of the three bars may be subdivided into three segments. The bottom segment (no lines) indicates the per cent of cases in which the readers were in agreement with the computer programme. The middle segment (oblique lines) indicates the per cent of cases in which the readers and the programme disagreed due to criteria differences. The top segment (horizontal lines) indicates the percentage of cases in which disagreements between readers and computer programme resulted from programme errors (in this figure, none).

Fig. 2. Myocardial infarction. Explanation, see Fig. 1.

Left ventricular hypertrophy (LVH). There were 51 reader diagnoses of LVH. The criteria used by the computer programme were very similar, although not identical, to those of the readers. There were no disagreements due to programme errors. The sensitivity was 92 % (47/51), and the specificity 99.5 % (440/442).

Right ventricular hypertrophy (RVH). A reader diagnosis of RVH was made in only five cases. Only two of these were correctly identified by the computer but the disagreements were due to criteria differences, the programme preferring statements such as "right intraventricular conduction delay". The specificity of the computer diagnoses was 99.5 % (486/488).

Myocardial infarction. The computer programme utilized several degrees of severity in its diagnostic language, based on the amplitude and width of Q waves, the location of Q, combinations of Q, R-wave progression and combinations of Q waves or Q equivalents and ST-T changes in different leads. In the reader diagnosis a two-graded scale of severity was used. This was based on much the same criteria but certain differences did exist. Programme errors, though rather infrequent, occurred more often for myocardial infarction than for hypertrophy. In one case the computer erroneously reported diaphragmatic myocardial infarction, based on a T wave obscured in both shape and amplitude by atrial flutter waves. In another case the computer did not measure S-T segment elevation correctly because S-T displacements were measured by a J point with a fixed time relation to the preceding QRS complex. The sensitivity of the computer diagnosis was 83 % (59/71) and its specificity 96.3 % (419/435). Programme errors were more common than criteria differences regarding diaphragmatic infarctions, while the reverse was true for anterior or antero-lateral infarctions.

Arrhythmia. There were no difficulties in diagnosing arrhythmia in the manual interpretations in any of the ECGs. The computer reported "undetermined rhythm" whenever a tracing failed to satisfy the computer's logic or criteria for a specific rhythm diagnosis. This happened in several cases, mostly cases of atrial fibrillation, but these disagreements were usually classified as group B, as we felt the additional diagnoses in these cases to be clinically more significant and as the computer made correct statements of the latter diagnoses. However, in five cases of a false computer diagnosis of "undetermined rhythm", the report was classified as group C or D on the ground of mismeasurement. To summarize the over-all performance of the computer programme regarding arrhythmia, there were no disagreements due to criteria differences. The sensitivity and specificity of the computer diagnosis, determined by a two-group classification procedure, were 82 % (71/87) and 97.5 % (431/442), respectively. The over-all diagnostic accuracy, determined by a multi-group classification procedure, was 95.1 %.

Atrial fibrillation. Out of 41 tracings of atrial fibrillation, the computer programme correctly identified 31 (76 %). The specificity was 99.8 % (451/ /452). In eight cases the computer stated "sinus rhythm", mostly in combination with "sinus arrhythmia", "sinus arrest" or "supraventricular ectopic

Table 1. Frequency of different rhythm diagnoses made by reader

Sinus rhythm	442
Ectopic atrial rhythm	4
AV nodal rhythm	2
Atrial flutter	4
Atrial fibrillation	41
Supraventricular ectopic beats	12
Ventricular ectopic beats	24
	529

beats" and in two cases it stated "undetermined rhythm". The computer falsely reported atrial fibrillation in one case where the tracing showed a normal sinus rhythm suddenly changing into ectopic atrial rhythm.

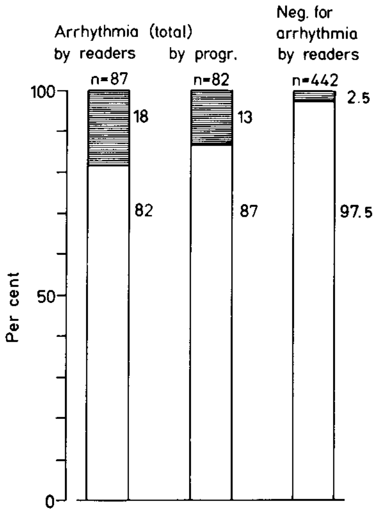
Other supraventricular arrhythmias. The computer programme falsely reported junctional rhythm or junctional tachycardia in eight cases of normal sinus rhythm, where the P waves were not detected. This gave a specificity of 98.4 % (483/491). Two actual cases of AV nodal rhythm were correctly stated as such. Concerning atrial flutter, the computer mismeasured two out of four cases but did not give any false positive answers.

Ventricular arrhythmia. There were 24 reader diagnoses of ventricular arrhythmia, i.e. premature ventricular contractions (PVC). The sensitivity of the computer was 88 % (21/24), the disagreements all being due to programme errors. The specificity was 99.8 % (468/469). In one case the computer measured an artifact and stated PVC. There were no ECGs with series of PVCs or more complex ventricular arrhythmia.

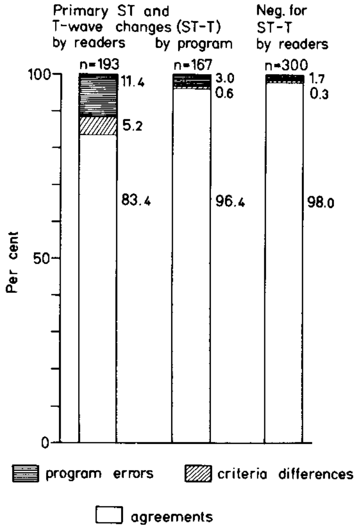
First degree AV block. The reader and the computer used the same criteria for the diagnosis of first degree AV block regarding the time interval. Nevertheless, three computer diagnoses were judged as false positives on the basis of criteria differences, because the computer stated "first degree AV block" in combination with "undetermined rhythm". The logical procedure would have been to supress all AV block statements in the presence of the statement "undetermined rhythm". In one case the computer did not detect normal P waves. The sensitivity of the computer diagnosis was 97 % (37/38), and the specificity 98.9 % (450/455).

Second degree AV block. The computer did not recognize the two cases of atrial flutter. There were no other tracings containing second or third degree AV block.

Left bundle branch block (LBBB). There were 11 reader diagnoses of LBBB. The computer made one false positive and three false negative statements in this respect, giving a sensitivity of 73 % (8/11) and a specificity of 99.8 % (481/482). Two false negative reports were due to programme errors.



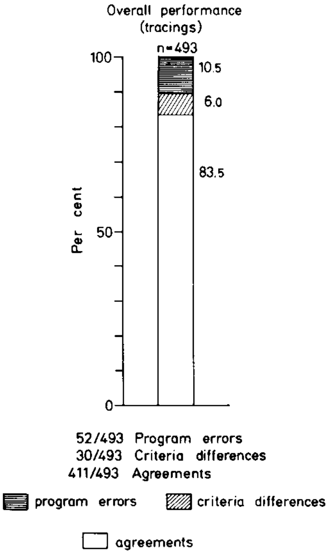
3



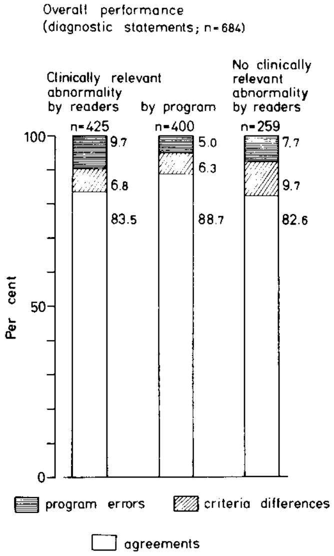
4

Fig. 3. Arrhythmia. Explanation, see Fig. 1.

Fig. 4. Primary S-T segment and T wave changes. Explanation, see Fig. 1.



5



6

Fig. 5. Over-all performance (tracings). Explanation, see Fig. 1.

Fig. 6. Over-all performance (diagnostic statements).
Explanation, see Fig. 1.

Right bundle branch block (RBBB). Out of ten cases of RBBB the computer misinterpreted one as RVH, due to criteria differences, giving a sensitivity of 90 % (9/10). No false positive reports were made.

Intraventricular conduction delay. Due to mismeasurement, the computer reported preexcitation in one tracing where the correct diagnosis should have been intraventricular conduction delay. In 11 tracings out of 12 the reader and computer diagnoses agreed, the sensitivity of the computer diagnosis being 92 % and its specificity 99.6 % (479/481).

Primary S-T segment and T wave changes. The sensitivity of the computer diagnosis was 83.4 % (161/193) and its specificity 98.0 % (294/300). Most disagreements were due to programme errors, where the computer seemed to be inaccurate in detecting the shape of the S-T segment and also abnormal S-T depressions or elevations. The reason for this might have been that judgements concerning the S-T segment are based on deviations from a J point fixed in temporal relation to the preceding QRS complex. The computer also had difficulties regarding the shape of the T wave, especially concerning coupling between this shape and slight changes in amplitude.

Miscellaneous. There were 11 cases with left anterior hemiblock, for all of which complete agreement was found, and there were no false positive reports. This was also the result for 24 cases of marked left axis deviation, two cases of right axis deviation, two cases of counter-clockwise rotation and three cases of clockwise rotation. There were five pacemaker ECGs. In two of these there was considered to be disagreement, on criterional grounds, as the computer did not report the presence of spontaneous beats, which in one case revealed extreme T-wave inversion in precordial leads.

Over-all performance. In 411 cases (83.5 %) the readers and the computer were in agreement, while disagreements were noted in the remaining 82 cases. Detailed analysis revealed that in 6.0 % (30/493) these disagreements were based upon the use of different diagnostic criteria. In the remaining 52 cases (10.5 %) they resulted from programme errors such as mismeasurement, pattern recognition failures or deficient programme logic. Regarding the over-all efficiency of the computer diagnosis, as expressed by a two-group classification procedure, the sensitivity was 83.5 % (355/425) and the specificity 82.6% (214/259). The error ratio was 0.3, the accuracy of positive tests (AP) 0.9 and the accuracy of negative tests (AN) 0.8. When expressed by a multi-group classification procedure, the over-all diagnostic accuracy (DA) was 83.5 %.

DISCUSSION

There are limitations to the use of the present ECG population for evaluating a computer programme. Some diagnoses were very uncommon and some were not represented at all. However, one aim was to test the computer against a representative sample of our own day-by-day routine clinical population of ECGs. The method of statistical analysis of the result also has its limitations, mainly in that no differential weight is given to different ECG diagnoses. Some differentiation in this respect was made in our investigation, however, since the ECGs were initially classified according to their clinical significance in a broad sense. With the present method of analysis, unnecessary discussion about diagnostic criteria is avoided. Discrepancies regarding criteria can of course influence the decision whether to use the services of a computer programme or not. On the other hand, such difficulties can be overcome in cooperation with the computer programme constructor. More serious are disagreements due to programme errors, where the user has to find out firstly, whether the programme is adequate for the kind of subject population in question and secondly, whether the rate of programme errors can be accepted.

Some authors have not only used sensitivity and specificity, as defined above, as measures of diagnostic efficiency, but have also used what is often called mean performance (MP) and association index (AI), where MP = $\frac{1}{2}$(SE+SP) and AI = SE+SP-100. Rautaharju et al. (4) have claimed that all of these concepts are often misunderstood in the sense that they have not been validated against the composition of the test ECG population used. We feel that MP and AI are not of much value in determining the efficiency of the present computer programme as used on the present population.

The sensitivity of the computer diagnosis was the same whether two-group or multi-group classification procedures were used, and the values were fairly satisfactory. Likewise, the specificity was acceptable and the values were within the range commonly reported in other similar studies of computer programme performance. The error ratio (0.3) was not satisfactory, however, as it implied that one-third of the diagnostic statements made by the computer were false. The accuracy of positive tests was acceptable but the accuracy of negative tests was disappointing and reflected an inability to detect arrhythmias and ST-T aberrations, in particular. We consider that the results of this study adequately describe the performance of the tested computer programme as applied to the type of population of ECGs used here.

ACKNOWLEDGEMENT

The authors are indebted to Mr. Wauquez, Cardionics, Brussels, for his kind help in providing us with the equipment and computer services necessary for this study.

REFERENCES

1. Bailey, J.J., Itscoitz, S.B., Hirschfield Jr, J.W., Grauer, L.E. & Horton, M.R.: A method for evaluating computer programs for electrocardiographic interpretation. I. Application to the experimental IBM program of 1971. Circulation 50:73-79, 1974.

2. Frank, E.: An accurate, clinically practical system for spatial vectorcardiography. Circulation 13: 737-749, 1956.

3. Goldman, M.J.: Principles of electrocardiography, Lange Medical Publications, Los Altos, California, 1973.

4. Rautaharju, P.M., Blackburn, H.W. and Warren, J.W.: The concepts of sensitivity, specificity and accuracy in evaluation of electrocardiographic, vectorcardiographic and polarocardiographic criteria. J Electrocardiology 9:275-281, 1976.

Received in final form November 23, 1978

Address for reprints:

Dr Johan Landelius
Dept of Clinical Physiology
University Hospital
S-750 14 Uppsala
Sweden