OBSERVER VARIATION IN REPORTS ON ELECTROCARDIOGRAMS

BY

L. G. DAVIES*

From the Department of Medicine, Postgraduate Medical School of London

Received August 7, 1957

It is now well known that the results of almost any examination or investigation may be reported differently by different workers, for example, the interpretation of a medical history (Cochrane *et al.*, 1951), the detection of cyanosis (Comroe and Botelhoe, 1947) or clubbing (Pyke, 1954), the clinical diagnosis of emphysema (Fletcher, 1952), the radiological detection of tuberculosis (Birkelow *et al.*, 1947; Garland, 1950), and the assessment of pneumoconiosis (Fletcher and Oldham, 1949).

Disagreement may be due to true errors of observation where certain abnormalities are wrongly identified or even missed altogether. But in the investigation reported here, as in many others there is a second factor, the observation may be correct, but there is a difference of opinion about its interpretation. The term observer variation used in this paper is defined as including disagreements due to both causes, for they occur together and it is sometimes impossible to tell which is responsible. Because of this, observer variation is a better expression of the problem than the term observer error.

The practical significance of observer variation has often been disputed (Lancet, 1954) and only in radiology is its importance being accepted. With the single exception of a report by Sloan *et al.* (1952) on the detection of the third heart sound, the problem of observer variation in cardiology has not been studied. This is surprising, for an electrocardiogram might be expected to be as difficult to interpret as is a chest film. They are of equal importance in that one is as valuable in the diagnosis of heart disease as is the other in the diagnosis of pulmonary disease; and both are single investigations upon which diagnosis and treatment may largely depend. Yet the assumption appears to be that unlike the chest film, electrocardiography is a test free from observer variation. The purpose of this paper has been to test that assumption, and it has proved to be false. Not only has observer variation been found, but its extent was surprising and it has been possible to identify many of its causes. The practical significance of this variation will be discussed in the belief that it may often contribute to diagnostic errors.

MATERIAL AND METHODS

One hundred tracings were selected from the files of the Department of Cardiology at Hammersmith Hospital. They included 50 that had been reported to show infarction, 25 that had been reported as normal, and 25 that had been reported to show some abnormality other than infarction. The last group included right and left ventricular hypertrophy, myxœdema, pericarditis, the effects of electrolyte upset, and so on. In order to simplify reporting, only tracings that had been taken with the same photographic machine were selected. Each group formed a consecutive series except that tracings from children or those containing less than nine leads (standard and unipolar limb leads with V1, V3, and V5) were excluded. Arrhythmias when present were incidental. These tracings were then mixed together and given in turn to each of the following 10 observers:

Dr. L. G. Davies
Dr. B. D. van Leuven } .. .. Medical Registrars, Hammersmith Hospital.
Dr. T. B. Counihan

* Present address—National Heart Hospital, W.1.

Dr. Max Zoob	Senior Medical Registrar, Hammersmith Hospital.
Dr. J. F. Goodwin	Lecturer in Medicine, Hammersmith Hospital.
Dr. Fulvio Camerini	World Health Organization Fellow, Hammersmith Hospital.
Dr. G. R. Venning	Senior Medical Registrar, Manchester Royal Infirmary.
Dr. William Phillips	Physician, Cardiff United Hospitals.
Professor I. G. W. Hill	St. Andrews.

All the readers had special interest and experience in cardiology, three were consultants and the others were then of registrar or senior registrar grade. Their individual identities have been concealed behind alphabetical symbols (A to I—but not of course in the above order). Included for comparison are the results obtained by a tenth reader, J, who is an experienced physician with a wide knowledge of medicine, but has his special interests outside cardiology.

The composition of the series was explained and the readers given the choice of reporting each tracing as either normal, abnormal, or showing infarction. In view of the selection of the tracings this was thought to be reasonable and the observer was not asked to locate the infarct or identify the nature of the abnormality. Clinical details were not given as they might have caused bias and the investigation was concerned with the diagnostic value of a single electrocardiogram in its own right. The amount of time spent in reporting was left entirely to the reader and ranged from 45 minutes to 2 hours. After an interval of not less than a fortnight the readers were asked to report again on the same electrocardiograms. It was found unnecessary to alter the order of the tracings at this second reading for although an occasional one was familiar, the previous report could not be recalled.

It is assumed that no less time and care were given to reporting than would have been used in normal practice.

Results

The gross results are illustrated diagrammatically in Fig. 1. The number of tracings reported normal by the experienced observers ranged from 13 to 29, abnormal from 20 to 44, and infarction, from 36 to 63. These discrepancies were considerable and there was equal disagreement over each diagnosis. Observer J found twice as many infarcts the second time as he did the first but the

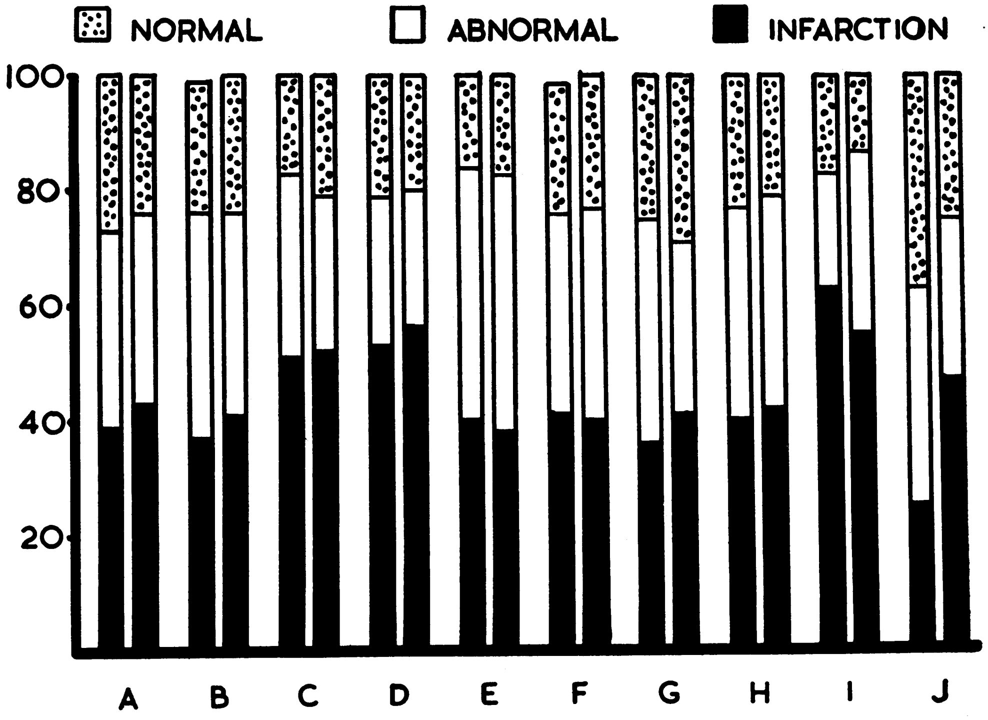


FIG. 1.—Each observer reported twice on the 100 electrocardiograms and this diagram shows the proportion reported each time as normal, abnormal, or showing evidence of infarction.

experienced observers appeared to be fairly consistent. This, however, was not the case, for instance the diagram does not indicate that the 14 tracings reported normal by E on his first reading are not all included in the 15 he reported normal on his second reading.

In fact after both readings by the 9 experienced observers there were only 29 tracings on which a unanimous report had been given. In 49 there was agreement by a majority arbitrarily defined as two-thirds. As each tracing had been reported on twice by the 9 experienced observers, each had collected 18 reports and a majority opinion was therefore 12 or more. It was usually much more, for in this group of 49 tracings there were 14 where there was only 1 disagreeing report, 6 where there were 2, 9 where there were 3, 8 where there were 4, 7 where there were 5, and 5 where there were 6 disagreeing reports.

In the remaining 22 tracings the dispute was greater than this and in a few of them opinion was almost evenly divided. In this group there was no satisfactory way of marking the results for there was no certain method of deciding who was right and who was wrong. It was impossible to accept the clinical diagnosis as dictating the correct electrocardiographic report for if the patient had rheumatic heart disease, angina, or hypertension, for example, the tracing could be normal or could be abnormal. But in the group of 49 tracings where there was majority agreement it seemed reasonable to assume that a large majority of 12 or more reports out of 18 was probably right and a small minority of 6 or less reports out of 18 probably wrong. Making this assumption, these tracings and the 29 where there was full agreement have been taken for further analysis.

These 78 tracings are then found to contain 19 that are normal or almost certainly normal, 20 that are " true abnormal," and 39 that represent " true infarction." Now each of the 9 experienced observers had reported twice on the 19 " true normal " tracings, a total of 342 reports: 292 of these were reports of normal, but there were 42 of abnormal and even 8 of infarction. So that even in tracings that are normal or almost certainly normal, there is still a 1 in 7 chance of a wrong report being given. This seems a large risk when the initial selection of the 100 tracings is recalled, especially when the 22 found most controversial have been excluded from this analysis. When the total reports on the true abnormal tracings were studied it was found that there was an almost equal chance of a " wrong " report of infarction being given; while a similar proportion of the true infarction tracings had been reported as merely abnormal. This variation is interesting and surprising though it may be claimed that error here is of less practical importance than the disagreement over the normal tracings.

The 156 reports on these 78 tracings show that this considerable dispute had not arisen as the result of large inaccuracies by one or two observers. In Table I each observer's report has been compared with the majority and presumably correct opinion.

TABLE I

DISAGREEMENT OF INDIVIDUAL OBSERVERS

Observer	A	B	C	D	E	F	G	H	I	J
Disagreements	7	19	25	11	24	12	9	16	25	56

For example, observer A disagreed with the majority in 7 reports out of 156 and was therefore probably wrong in these 7 reports. Although some observers did better than others the range is fairly close and all did much better than the inexperienced observer J who was apparently wrong in one report out of every three. Judged by this rough test, the three consultants as a group were no more accurate than the registrars.

It may be argued that the disagreements reported here are not genuine but are largely due to difference in terminology. For instance, one observer may report a tracing as showing infarction while another would consider it to be ischæmic and report it as abnormal. There is a little truth in this, and as will be seen later the observers did differ in their definition of an infarction pattern. But in considering this objection the most important evidence lies in the personal changes of opinion between the first and second readings. These are shown in Table II.

TABLE II

CHANGES OF OPINION IN INDIVIDUAL OBSERVERS

Observer	A	B	C	D	E	F	G	H	I	J
Opinions changed in	11	16	19	7	16	13	10	10	18	33
Normal in one report, abnormal or infarction in the other	3	7	10	5	7	3	4	4	6	16
Consistency, percentage	89	84	81	93	84	87	90	90	82	67

For example, after observer A had reported a second time on the 100 tracings he was found to have changed his opinion on 11 of them. Indeed the experienced observers on average changed their minds over one tracing in every eight. Yet the range is fairly close and all did much better than J who disagreed with himself over one tracing in three. Again the consultants as a group were no more consistent than the registrars.

While it is possible therefore for differences in definition between one observer and another to cause apparent rather than real disagreement, it is obvious that whatever our definitions are they are certainly variable and inadequate, for they failed to prevent each observer from frequent disagreement with himself.

Are these personal changes of opinion of any importance? In Table II we see that in 3 of the 11 tracings over which observer A changed his mind, the report was altered from normal to some other report. Indeed, of the 120 occasions on which the 9 experienced observers changed their opinion, in 40 of them, that is 1 in every 3, the change was between normal and some other report. These changes must surely be highly important, they throw doubt on the diagnostic value of the electrocardiogram and the practical implications are considerable.

CAUSE OF DISAGREEMENT

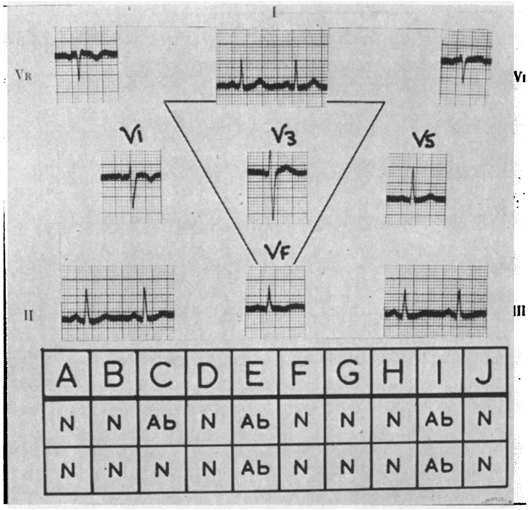
It is convenient to take the 78 tracings already described and consider the minority reports for these are almost certainly wrong. The lead or pattern that had caused disagreement could usually be identified, but this was not always the case and a few opinions were unaccountable. Obviously only a very few tracings can be included as illustrations.

Minor disagreement arose over most of the 19 true normal tracings. In 6 the configuration of aVL was responsible, in 4 the patterns in leads III and aVF (Fig. 2), in 2 the appearance of V3, in 2 the T wave contours, and in the other 2 the cause was not apparent.

In the group of 20 abnormal tracings there were 6 where leads III and aVF gave rise to difficulty and 3 where aVL was responsible. In the group of 39 true infarction tracings, there was disagreement in 9 where the infarct was posterior or postero-lateral and in 5 where it was anterior or antero-lateral. In three tracings there were T wave changes without abnormal Q waves (Fig. 3), and in two there was right bundle-branch block. In a further two tracings, the reasons for disagreement were not apparent.

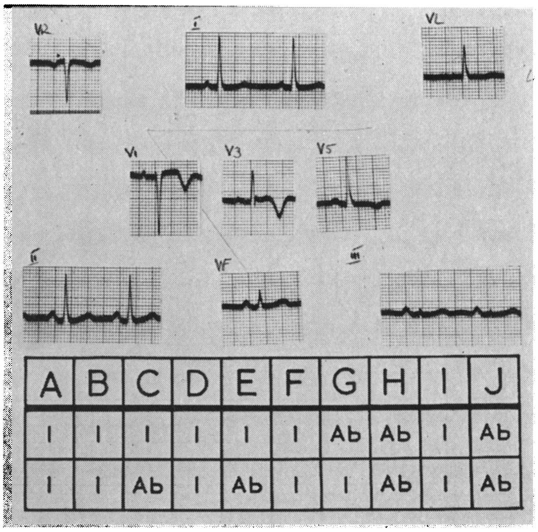
In summary the causes of disagreement are best related to particular leads and the greatest dispute was due to varying interpretation of leads III and aVF; there was also considerable uncertainty about the significance of QR or qR patterns in lead aVL.

It was more difficult to analyse the results in the 22 tracings where disagreement was considerable. In some of these, opinion was fairly evenly divided between normal and abnormal or between



A	B	C	D	E	F	G	H	I	J
N	N	Ab	N	Ab	N	N	N	Ab	N
N	N	N	N	Ab	N	N	N	Ab	N

FIG. 2.—In this and the following figures each observer's reports at the first and second reading are shown. There is some disagreement over this tracing but except for C the reports are consistent. The tracing has been classified as " true normal " on the opinion of a large majority and the contrary reports are almost certainly wrong. The cause of disagreement is the ST–T pattern in leads II, III, and aVF. The patient, a woman of 28 years, was admitted to hospital with no symptoms or signs of organic heart disease. Thyrotoxicosis was suspected but the various tests did not confirm this.



A	B	C	D	E	F	G	H	I	J
I	I	I	I	I	I	Ab	Ab	I	Ab
I	I	Ab	I	Ab	I	I	Ab	I	Ab

FIG. 3.—This graph is clearly abnormal but not all observers are convinced that it is diagnostic of infarction. Should infarction be reported in the absence of abnormal Q waves? This problem was an important cause of disagreement between observers and also a frequent cause of personal changes of opinion.

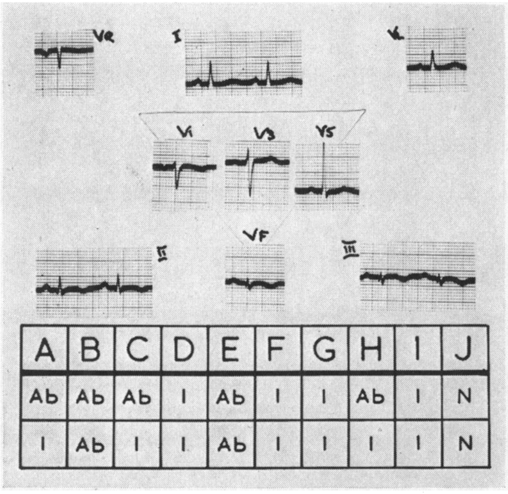
abnormal and infarction, but several tracings received all three alternative reports. Eight tracings came from patients where the clinical diagnosis or necropsy finding was myocardial infarction; in 6 of these 8 the lesion was posterior and difficulty was found in the interpretation of leads III and aVF (e.g. Fig. 4). The clinical diagnosis in this group of patients was supported by only 4 reports in every 7. The same leads caused confusion in three patients with hypertension and a further four with various diagnoses. Once again they stand out as the major cause of uncertainty and disagreement.

The configuration of aVL was the next most important cause, being wholly or partly responsible for dispute in four tracings (Fig. 5). Other causes were QS patterns in V3, the T wave contours in some præcordial leads, the size of the P wave, the P–R interval, and bundle-branch block. Four of the tracings were of rather poor technical quality and this may well have added to the other difficulties.

In fact, this group of 22 tracings was only quantitatively different from the others. The same difficulties were present, but in greater degree and sometimes in combination. This view is confirmed by finding that these 22 controversial tracings caused 52 personal changes of opinion while the remaining 78 caused only 101 changes. As might have been expected, the more likely a tracing is to cause disagreement between different observers, the more likely is it to lead to personal changes of opinion.

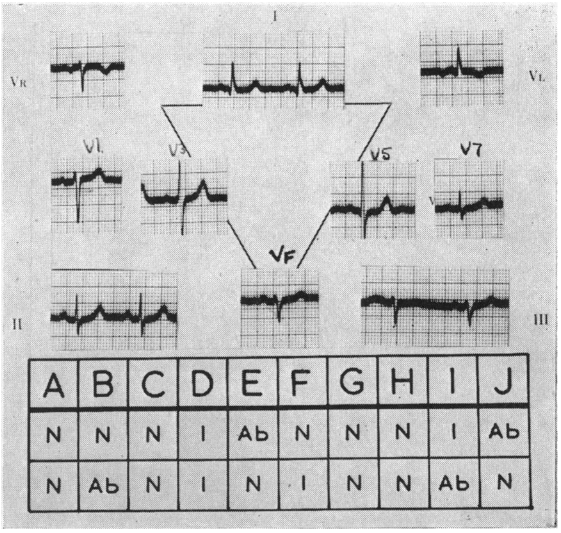
DISCUSSION

As there are now 22 English and American text books on the subject, electrocardiography must be considered an important branch of medicine. Graybiel (1951) writes in one that " a great responsibility is placed on the man who interprets electrocardiograms " and advises that he should



A	B	C	D	E	F	G	H	I	J
Ab	Ab	Ab	I	Ab	I	I	Ab	I	N
I	Ab	I	I	Ab	I	I	I	I	N

FIG. 4.—Leads III and aVF gave rise to disagreement more frequently than did any other leads. There was no general agreement as to what constituted an abnormal Q wave in these leads and criteria based on proportion are difficult to apply when, as in this tracing, the QRS voltage is low. A further difficulty was found to be the interpretation of Q waves in these leads when the T was upright.



A	B	C	D	E	F	G	H	I	J
N	N	N	I	Ab	N	N	N	I	Ab
N	Ab	N	I	N	I	N	N	Ab	N

FIG. 5.—Lead aVL was another frequent cause of dispute, as in this tracing. Each of the three alternatives were reported and there were frequent changes of opinion. This is one of the 22 tracings where the results cannot be marked for there is no strong majority opinion to determine the "correct" report. The patient, a woman of 59 with labile hypertension, complained of chest pain that was difficult to evaluate but finally was not accepted as cardiac.

"properly appraise his skill and avoid serious error." No suggestions are made as to how this should be done, nor indeed is any information given about the nature and incidence of errors. The results of this study confirm the validity of this warning yet the other text books contain no suggestions that there is ever any real difficulty in recognizing the different patterns from the descriptions given, and it is also implied that there is unanimous agreement between various authorities about their significance.

Disagreement not only exists, but in this trial was of surprising magnitude. In only a third of this group of tracings was opinion unanimous. In a half there was agreement by a majority of the readers, but this still left a substantial number—one tracing in five—over which there was considerable dispute. Moreover on second reading, the observers disagreed with one in every eight of their original reports. These findings are all the more remarkable when the initial selection of the tracings is recalled and our belief in the diagnostic value of a single electrocardiogram appears to be ill-founded. Other readers would be unlikely to improve on these results which must be taken as illustrating the difficulties of electrocardiographic diagnosis. And although these difficulties appear considerable, the experienced readers all did much better than the single inexperienced reader J.

It is very difficult to arrive at any accurate estimate of the practical importance of this observer variation. The prime use of the electrocardiogram is as an aid in the diagnosis of coronary disease and this is now so common that errors in diagnosis in even a small proportion of cases will affect large numbers of patients. The tracings used in this study were selected with this in mind. Yet in the group of 39 true infarction tracings there was unanimous agreement on less than half of them. Moreover, there were 7 tracings in the controversial group which received many reports of infarction. In none of the seven patients concerned was this thought at all likely, though in only two of them could this diagnosis be rejected with certainty.

It may be claimed that the electrocardiogram, although failing to give the correct answer in every patient with suspected infarction, is nevertheless always accurate enough when taken in conjunction with other evidence. Since diagnosis is made on the history and physical findings, supported by the results of special investigations, errors that might arise in each of these will be cancelled out when the results are considered together at a final diagnostic synthesis. This, at any rate, is the opinion of those who believe that observer variation is of no clinical importance (Lancet, 1954, and Pierce, 1954). This view I believe to be incorrect, for unless we study our diagnostic measures we shall fail to appreciate the extent and nature of the variation that may occur. Also the errors may not cancel out, they may summate, as in the following examples, where misinterpretation of the electrocardiogram contributed to major diagnostic errors.

For instance, in a report by Brumfitt and Rankin (1954) of a patient with chest pain, the electrocardiogram was thought to confirm the diagnosis of myocardial infarction and heparin was given. Two hours later the man died and necropsy showed a large dissecting aneurysm.

In another patient, also with pain believed to be due to myocardial infarction, electrocardiograms showed developing ST–T changes were interpreted as supporting this diagnosis (McCord and Taguchi, 1951). After several days treatment with anticoagulants the patient died and necropsy showed acute primary pericarditis with death due to hæmopericardium and tamponade.

A third example (personal observation) is that of a patient where myocardial infarction was diagnosed on the clinical picture and electrocardiographic changes. After further attacks of pain the correct diagnosis of repeated pulmonary embolism was made; later electrocardiograms—and in retrospect the earlier one—showed changes typical of right heart strain.

These examples illustrate the importance of observer variation in the interpretation of electrocardiograms in single clinical problems. But observer variation may be important in another way and in a wider field. In a discussion at the Royal Society of Medicine (1954) it was pointed out that the mortality in 11 separate series of patients with myocardial infarction varied between 13 and 45 per cent. The debate was over the value of anticoagulant treatment and it is not surprising that completely opposite opinions were expressed. It seems highly probable that differences in the criteria for diagnosing infarction could account for some of this wide scatter. The point is brought out in this study for when the extent of the disagreement emerged, each observer was asked to give his working definition for the recognition of infarction. These were surprisingly varied. One observer said he had reported infarction on S–T segment and T wave changes but most observers required the presence of abnormal Q waves. Some thought that large Q waves alone were certain evidence of infarction while others demanded T wave inversion as well. And then there was disagreement over the vital question of what constituted an abnormal Q wave, especially in leads III, aVF, and aVL. This dispute may seem surprising, but it seems clear that the distinction between infarction tracings and those showing lesser abnormalities is purely arbitrary; this point is not widely recognized and should be emphasized.

It is a matter of some interest that the observers had equal difficulty at the other end of the scale, in distinguishing between normal tracings and those showing minor abnormalities. The clinical importance of observer variation here is that misinterpretation of the electrocardiogram may lead to a diagnosis of heart disease in a person whose heart is in fact normal. Rosenbaum (1951) reported two instances where this had happened and his opinion was that one patient " may never recover," while Prinzmetal (1955) has referred to the existence of a syndrome that he calls " heart disease of electrocardiographic origin." The practical problem is the assessment of the chest pain so frequently present in Da Costa's syndrome, for slight changes in the S–T segment or T waves could, if wrongly interpreted, be taken to support an incorrect suspicion of heart disease. There are no figures that tell us how frequently this mistake is made, perhaps not as often as one wrong diagnosis in every seven subjects with healthy hearts; but this was the frequency of incorrect opinions on the true normal tracings. The objection that is always raised is that clinical error will not arise if the history, physical findings, and the tracing are considered together. But Evans (1952) has found the patient's description of his pain less reliable than the electrocardiogram, some infarcts are painless (Lancet, 1954), and most patients with angina show no abnormal physical signs. Again electrocardiograms are now being demanded by some insurance companies as part of the medical examination. If the examiner believes he has found an abnormal tracing he is unlikely to accept the applicant's denial of symptoms. The advocates of the effort test (Master *et al.*, 1942, and Wood *et al.*, 1950) recognize that this diagnosis is often difficult and that a single electrocardiogram may be indecisive.

It follows that we should not expect the electrocardiogram to provide a clear division between normal and abnormal patterns; the range of both overlap and this to a greater extent than is commonly recognized. Distinction is least clear in leads III and aVF but there is also considerable difficulty with lead aVL. Tracings from the intermediate zones, where normal and abnormal or abnormal and infarction overlap are particularly liable to more than one interpretation and are most likely to give rise to changes of opinion. In clinical practice the interpretation would be very susceptible of bias due to the assessment of the history and other findings, and the dangers inherent in this mechanism have been illustrated. We should not expect too much of the electrocardiogram and it is time that we recognized that some tracings are of little diagnostic value.

Summary

The purpose of this paper was to show whether reports on electrocardiograms were subject to observer variation. A test series of 100 tracings was selected: half had been reported routinely to show infarction, a quarter to be normal, and a quarter to show various abnormalities other than infarction.

Nine experienced readers reported their opinions of these electrocardiograms on two separate occasions. They were allowed the choice of one of three reports—normal, abnormal, or infarction.

Complete agreement was reached in only one-third of the 100 tracings, majority agreement in half, but there was considerable dispute about one tracing in five. After the second reading, it was found that on average, the readers disagreed with one in eight of their original reports.

This considerable observer variation affected the normal, abnormal, and infarction tracings equally; it was much larger than had been expected and must represent the unrecognized difficulties of electrocardiographic diagnosis. Nevertheless the results obtained by these readers were all much better than those obtained for comparison by a single inexperienced observer.

The reasons for this large disagreement have been examined and the most important single cause was difficulty with the QRS–T pattern in leads III and aVF. There was also much uncertainty about the significance of QR patterns in aVL, and many minor causes.

From the standpoint of electrocardiographic diagnosis it is an illusion to believe there can be any arbitrary line between normal and abnormal tracings or between abnormal and infarction tracings. The ranges of each overlap and do so more widely than is generally realized; distinction is least clear in leads III and aVF. It is apparent that tracings from the intermediate zones are of little or no diagnostic value, but are very likely to be interpreted according to the clinical bias. In this way observer variation may add to diagnostic error. The clinical importance of this variation is debatable, but it is so large that in the absence of reliable information to the contrary, its importance cannot be denied.

I am greatly indebted to the readers for very willing co-operation and to Dr. C. M. Fletcher for much advice and encouragement.

References

Birkelow, C. C., Chamberlain, W. E., Phelps, P. S., Schools, P. E., Zacks, D., and Yerushalmy, J. (1947). *J. Amer. med. Ass.*, **133**, 359.
Brumfitt, W., and Rankin, N. E. (1954). *Lancet*, **2**, 792.
Cochrane, A. L., Chapman, P. J., and Oldham, P. D. (1951). *Lancet*, **1**, 1007.
Comroe, J. H., and Botelho, S. (1947). *Amer. J. med. Sci.*, **214**, 1.
Evans, W., and McRae, C. (1952). *Brit. Heart J.*, **14**, 492.

Fletcher, C. M. (1952). *Proc. Roy. Soc. Med.*, **45**, 577.
—— and Oldham, P. D. (1949). *Brit. J. Indust. Med.*, **6**, 168.
Garland, L. H. (1950). *Amer. J. Roentgenol.*, **64**, 32.
Graybiel, A., (1951), see *Clinical Heart Disease*, by S. A. Levine, London.
Lancet Annotation (1954). **1**, 87, and **2**, 799.
McCord, M. C., and Taguchi, J. T. (1951). *Arch. intern. Med.*, **87**, 727.
Master, A. M., Friedman, R., and Dalk, S. (1942). *Amer. Heart J.*, **24**, 777.
Pierce, J. W. (1954). *Lancet*, **1**, 737.
Prinzmetal, M. (1955). Cardiac Society Meeting, San Francisco.
Proc. Roy. Soc. Med. (1954), **47** 317.
Pyke, D. A. (1954). *Lancet*, **2**, 352.
Rosenbaum, F. F. (1951). *Ann. intern. Med.*, **35**, 542.
Sloan, A. W., Campbell, F. W., and Steward-Henderson, A. (1952). *Brit. med. J.*, **2**, 853.
Wood, P., McGregor, M., Magidson, O., and Whitaker, W. (1950). *Brit. Heart J.*, **12**, 363.

Editorial Comment

If anyone thinks that a report on a single electrocardiogram can always decide whether it is normal or abnormal or whether there is or is not evidence of cardiac infarction, this valuable paper will show that he is wrong. No one, on reflection, does think this, but many write a routine report as if they did.

Those asked to report were given only the three choices and not allowed to give more borderline reports, and this could be criticized. It may be argued that some doubt about the validity of the report is tacitly assumed by everyone, but this is not so. Everyone should be more aware that although some electrocardiograms are obviously normal and others prove cardiac infarction, there are many that could be normal but should raise a suspicion of coronary disease, many that prove some myocardial disease and raise a suspicion of cardiac infarction, and many with something unusual that might be due to heart disease or biochemical changes. It seems clear from this paper that more such qualifications should be added to reports on electrocardiograms whenever these are needed.