lem here is not film quality or size but film reading. When, upon two readings of a series of films, few readers can be consistent with themselves or with others, further investigation of the causes of such variation is imperative. This so-called personal equation which operates as an error-factor in film interpretation demands the most serious consideration. Until the problem is solved, it has been suggested that all survey films be read independently by at least two interpreters. Furthermore, an evaluation of teaching methods employed in departments of roentgenology might suggest profitable changes that would produce improvement in roentgenologic accuracy. Certainly, awareness of error is the threshold to correction.

X-ray examination of the chest, however, compares favorably with any other diagnostic method in the field of medicine, and it must be pointed out with emphasis that the lesions interpreted in these studies are subtle in character and minimal in extent. Almost all gross pathology is visualized on the roentgenogram. Since the great majority of lesions observed in mass X-ray surveys are minimal, these studies of abstruse lesions assume added significance.

The field of roentgenology is fortunate in that the evidence produced by its material and tools is measurable and demonstrable. There are many fields of knowledge in which error is known to exist, but so intangible is the evidence, that the magnitude of that error cannot be demonstrated. It remains for us now to continue our investigations and seek out the approaches to perfection.

FRANCIS J. WEBER,
Medical Director Chief, Tuberculosis Control Division.

STATISTICAL PROBLEMS IN ASSESSING METHODS OF MEDICAL DIAGNOSIS, WITH SPECIAL REFERENCE TO X-RAY TECHNIQUES [1]

BY JACOB YERUSHALMY, Ph. D.

The process of medical diagnosis involves the application to a specific case of the knowledge accumulated from a large number of similar cases. This knowledge may have been derived by systematic observation and detailed analysis of case records, or it may have been obtained intuitively in the course of the physician's experience. In either instance the process is statistical in nature in that it consists of abstracting from a multiplicity of factors and conditions those which are pertinent to specific cases. Moreover, since no two cases are exactly alike, the resulting diagnoses are not absolute but involve some uncertainty and might better be thought of in terms of probabilities. These probabilities may, in some instances, be very high; in others

[1] From Field Studies Section, Tuberculosis Control Division. This paper and the discussion by Professor Neyman, which follows it, were presented before the Institute of Mathematical Statistics, at the twenty-eighth annual meeting of the Pacific Division of the American Association for the Advancement of Science, in San Diego, Calif., June 18, 1947, while the author was visiting professor of biostatistics at the School of Public Health, University of California.

they may have lower values. One of the fundamental objectives of improved medical care is to increase the probabilities of correct diagnoses.

This objective is being accomplished by the increasing employment of various diagnostic aids such as X-ray and laboratory investigations. However, even with the best techniques it is not always possible to raise the level of diagnosis to absolute certainty in all cases of a certain disease. Medical diagnosis of an individual case is often assisted also by the fact that it is not based on a single observation. Repeated studies of the same case allow for integration of a number of related observations and for continuous revision of tentative diagnoses. The chance of error is thus reduced and the probability of correct diagnosis is increased.

There are, however, occasions when the benefit of repeated and continuous case studies is lacking, and at least a preliminary diagnosis must be made on the basis of a *single observation*. This is especially true in the field of public health when examinations for specific purposes are made on relatively large samples of the population. Two main types of public-health activity may be noted where the "single event" form of diagnosis is employed.

In problems of the first type, the information required is a knowledge of the *total number* of positive [2] persons in the group but not the identities of the individual positive cases. In activities of the second type the main purpose is the identification of the persons in the population having a certain characteristic.

Examples of activities of the first type are mass examinations of population groups for the purpose of determining an epidemiological index for some infection or other pathological condition. All that is required in such mass examinations is a knowledge of the population examined and the total number of "positives." Because the individuals need not be identified, the numerical value of the index may in some instances approach more nearly the true value, because the "false negatives" (positive individuals who have been diagnosed "negative") are compensated to a certain extent by "false positives" (negative individuals who have been diagnosed "positive").

Examples of activities of the second type are X-ray examinations or serological tests conducted on large population groups as control measures in tuberculosis or syphilis. In these screening examinations the main objective is not to ascertain the total number of "positives" in the sample but to detect the persons possessing certain characteristics as a first step for further examinations and action on these "positive" cases.

Another type of activity in which the diagnosis is based on a single

[2] The term "positive" refers to individuals who possess the characteristic to be detected by the examination.

observation relates to studies on the efficacy of different diagnostic aids. Also in this instance it is not sufficient to know how many persons in the sample are positive but it is also necessary to identify the persons who possess certain characteristics. For example, if it is desired to compare the performance of different tests for syphilis in detecting infected individuals, it becomes necessary to test the same sample of the population by each of the different methods and to compare the findings for each individual of the sample.

The task of such comparative studies, when approached by the investigator, presents a serious dilemma. If an absolute comparison between the different diagnostic aids is to be made, it becomes necessary to identify the individuals in the sample who are positive. However, this identification can be accomplished only through the use of one or the other of the tests. It is, in a sense, a vicious circle.

The general procedure in dealing with this type of problem is to select one of the techniques as a "standard" and to compare the performance of the other techniques with that of the standard. This solution is not entirely satisfactory in many instances. For one thing such a procedure implies that the diagnosis provided by the standard technique is correct in all instances. In addition, it delays the recognition of the superiority of a technique which may be better than the accepted standard.

It is, therefore, desirable in such problems as were mentioned above, as in others of similar nature, to devise methods of examination and analysis which would increase the probabilities of correct diagnosis in mass examinations, where the subjects appear for examination only once. One obvious method is to substitute multiple observations on the single examination for part of the benefits which could be derived if the subjects were available for study. This may be accomplished by the simultaneous employment of a number of related tests or by increasing the number of independent observers or by a combination of both of these methods.

It is the object of this paper to utilize data from a comparative X-ray study in an attempt to develop a method of defining "positive" cases in the test population which is not dependent on any one diagnostic technique or on the interpretations of any single observer. Instead, the information yielded by all the techniques and all observers is used to increase the probabilities of correct diagnosis and thus furnish a more objective method of comparing the different X-ray techniques.

MATERIAL

The material for this study is provided by an investigation on the relative effectiveness for tuberculosis case finding, of various photofluorographic and roentgenographic methods.

Some 1,200 persons comprising the entire population of a Veterans' Administration institution were X-rayed consecutively on four different X-ray techniques providing for each person a 35-mm. photofluorogram of the chest, a 4″ x 10″ stereophotofluorogram, a roentgenogram on 14″ x 17″ paper negative, and a conventional 14″ x 17″ celluloid film.

These examinations provided, therefore, four sets of chest films of different size on the same group of individuals taken at the same time. These sets of films were interpreted independently by five expert radiologists and chest specialists. A more detailed account of the method of interpretation is given in a previous report.[3] For present purposes it is sufficient to mention that a number of meetings of the readers were held before the films were circularized, a method of interpretation was developed and experimented with, in order that the five readers reach as nearly as possible uniformity in nomenclature.

The films were sent to the readers one set at a time. At the completion of interpretation of all four sets and after a lapse of from 2 to 3 months, the 14″ x 17″ celluloid films were again circularized among the readers for a second independent interpretation of the same films. These activities yielded 25 sets of independent interpretations of the four sets of films, which constitutes the material available for analysis.

ANALYSIS

The object of the analysis is to obtain measures for the relative efficiency of the different X-ray techniques in selecting the individuals in the study group who have X-ray evidence suggestive of tuberculosis. The usual procedure is to determine for each of the four techniques two measures:

1. *A measure of sensitivity* or the probability of correct diagnosis of "positive"[4] cases, and
2. *A measure of specificity* or the probability of correct diagnosis of "negative" cases.

The determination of these measures is complicated by two main difficulties in analysis which are present in many comparative studies of this kind. The first results from the fact that it is not known who in the test population is positive and who is negative, i. e., who should be selected in the screening test for further study because of X-ray evidence suggestive of tuberculosis and who is free from such evidence. The second difficulty is due to subjective errors of interpretation. The latter are of two main types: *interindividuals*, or the inconsistency

[3] Birkelo, C. C., Chamberlain, W. E., Phelps, P. S., Schools, P. E., Zacks, D. and Yerushalmy, J.: Tuberculosis Case Finding. J. Am. Med. Assoc., 133: 359–365 (1947).

[4] By "positive" cases are meant persons who have X-ray evidence suggestive of tuberculosis, and "negative" cases refer to persons who have no such evidence.

of interpretation found among different readers and *intraindividuals*, or the failure of a reader to be consistent with himself in independent interpretations of the same set of films. If these subjective errors were small in magnitude, they would cause little difficulty. In the present study, however, they were relatively large. For example, even with the 14″ x 17″ celluloid technique, which is usually taken as a standard, the number of "positives" selected from among the 1,256 films was as low as 56 for one reader and as high as 100 for another with intermediate numbers for the other three readers. In other words, not only is it necessary to select a "standard technique" but also a "standard reader."

That these difficulties may lead to wrong conclusions may be seen by analyzing the material, as is so often done, without reference to these problems. This was demonstrated in the main report of the study [5] in the following way. The 14″ x 17″ celluloid technique was taken as a "standard" and the interpretations of only one expert were considered. Reader N selected on the standard technique 59 cases as showing X-ray evidence suggestive of tuberculosis. Of these, he failed to diagnose tuberculosis in 27 percent of the corresponding 35-mm. films, in 30 percent of the 4″ x 10″ films and in 24 percent of the 14″ x 17″ paper negatives.

From the performance of reader N, one would be tempted to draw two conclusions: first, that none of the test techniques can be considered efficient (the lowest percentage missed is 24 percent) and, second, that if a choice must be made, the 14″ x 17″ paper negatives have an edge over the two miniature techniques.

It was shown, however, that neither of these conclusions is justified. For, first, the same reader N missed on his first reading of the 14″ x 17″ celluloid films 22 percent of those which he called positive for tuberculosis on the second reading of the same 14″ x 17″ celluloid films. In other words, the relatively high percentages of missing positive cases do not necessarily measure the limitations of the different techniques, but, to a large extent, the subjective errors of interpretation for reader N.

Secondly, it is not possible to conclude that the paper negatives are more reliable than the miniature techniques because the same results are not found for all the readers. For reader M the score on the 4″ x 5″ is better than on the other two. For readers O and Q the best performance was that on the 35 mm., while reader P had the same score on all three techniques.

This example illustrates that no valid conclusions may be drawn from an analysis which fails to take into account the subjective errors

[5] Loc. cit. p. 361.

of interpretation. Even if the 14″ x 17″ celluloid film is taken as a standard, it becomes necessary to resolve the variance found for the different techniques into its components and to minimize the effect of the subjective errors.

One method which accomplishes this end result consists in basing the analysis on "group opinion" derived from the individual independent interpretations. Briefly, the procedure is to consider as "positive" only those cases which were so called on the standard technique by more than one interpreter. Since the interpretations were all independent, the probability that a positive case will be missed by more than one reader is greatly reduced. Similarly, it is unlikely that a 14″ x 17″ film which shows no X-ray evidence suggestive of tuberculosis will be called positive by several readers, each reading independently of the others.

By this method it is possible to line up the three test techniques and to compare their performances to that of the standard 14″ x 17″ celluloid technique. For example, there were 61 cases which were called positive on the 14″ x 17″ celluloid by three or more interpreters. These are considered the only "positive" cases in the population examined, i. e., these 61 are taken to be the only ones who presumably have X-ray evidence of tuberculosis and who should be selected, by a screening technique, for further study. The films for these 61 cases are then reviewed for each of the other techniques, and a *technique* is considered to have missed one of these cases if at least 3 of the 5 interpreters called it negative for tuberculosis. The result of the comparison by this method in the present study was that each of the test techniques missed the same number of films (10 percent).

This method eliminates the major part of the difficulty resulting from subjective errors of interpretation, but retains the limitation that one of the techniques (in this case the 14″ x 17″ celluloid) is taken as a standard. What may be concluded is that the 35 mm., the 4″ x 10″, and the 14″ x 17″ paper techniques are equally efficient in selecting positive cases. However, the impression obtained from the above that the efficiency of these techniques is approximately 90 percent that of the 14″ x 17″ celluloid technique cannot be accepted without further study.

COMPARISON OF ALL FOUR TECHNIQUES WITHOUT REFERENCE TO A STANDARD

It now becomes desirable, if possible, to remove the last restriction, that of selecting the 14″ x 17″ celluloid technique as a standard, and to devise means whereby the four techniques may be compared on an equal footing.

When the problem of comparing the four techniques is approached

760653—47——2

without reference to a predetermined standard, it is useful to distinguish between false information inherent in the technique and false information due to incorrect interpretation by a reader. A shadow, for example, may be very definite on the technique but missed or misinterpreted by a reader. On the other hand, a distinct shadow which suggests disease may be present on a film, but in reality the person has no disease and the shadow on the film is due to an artefact or to some exaggerated markings; again, a shadow of an existing lesion may not appear on the film or appear so indistinctly that it cannot be clearly visualized. It is the existence of those latter types of false information which argues against the selection of any technique as a standard.

In order to visualize more clearly the interplay of the two types of false information mentioned, it is useful to consider how they are affected by the addition of more techniques and more observers. Suppose, first, there were only two techniques and one reader. Each person examined can then be represented by a twofold table in which the first symbol is the interpretation of the reader on one of the techniques and the second represents the interpretation of the same reader on the second technique. Each person will be represented by one of the following four types of tables:

TYPES

Technique	1	2	3	4
S	−	+	−	+
T	−	+	+	−

Persons represented by tables of the first two types are, on the available evidence, clearly defined. Those of type 1 are probably negative and those of type 2 are probably positive. The status of persons represented by tables of type 3 and 4, however, is undetermined. If the two techniques are given equal weight, it is impossible to say whether the persons are positive or negative. In this case it becomes necessary to select one of the techniques (for example that represented by the first symbol) as a standard. Persons represented by tables of type 3 are then considered "negative," and those represented by table 4, "positive." The false information is then attributed entirely to the test technique. The positive reading in a table of type 3 is considered a "false positive," and the negative reading in a table of type 4 is considered a "false negative" and counted against the performance of the test technique.

Suppose now that another interpretation is obtained on the same two sets of films. Each person in the study may now be represented by a fourfold table in which the symbols in the first row represent the two interpretations on one technique, and those in the second

row the interpretations on the second technique. There will now be 16 distinct types of cases. For the present purposes, however, it is sufficient to focus attention only on those cases which previously were represented by tables of type 3 and 4. Among these there will be some in which the second reading on the first technique will confirm the original reading on the second technique rather than that of the first technique. For example, a case which originally was of type 3, i.e., $\left|\begin{matrix}-\\+\end{matrix}\right|$ may take the form of $\left|\begin{matrix}-&+\\+&+\end{matrix}\right|$. In other words, a case which was originally read negative by technique S and positive by technique T, is read as positive on technique S on the second reading. It is obvious that this case has a very strong probability of being positive, and therefore it should have originally been counted as a "false negative" for the standard technique rather than a "false positive" for the test technique. Again a case which originally was of type 4, $\left|\begin{matrix}+\\-\end{matrix}\right|$ may now become $\left|\begin{matrix}+&-\\-&-\end{matrix}\right|$. This case should have originally have been counted a "false positive" on the standard rather than a "false negative" on the test technique.

The addition of a second reading, therefore, undermines our confidence in the simple procedure used when only one reading was available in which all the false information was thought to be derived from the test technique. It is now apparent that a certain part of it must be attributed to the standard. It should be noted than this modification results from the presence of subjective errors attributed to the reader.

Although the addition of the second technique resolves the uncertainty of some of the cases, there still remain undetermined cases of the form $\left|\begin{matrix}+&+\\-&-\end{matrix}\right|$ and $\left|\begin{matrix}-&-\\+&+\end{matrix}\right|$. For these, the evidence provided by the two techniques is conflicting, but the interpretations are consistent for each technique. If no additional evidence is available, the uncertainty can be resolved only by again considering one of the techniques as standard (the technique represented by the two readings in the first row) and the other as a test technique. In other words, persons represented by tables of the first type (those read positive on the standard) are considered to be "positive," while those represented by tables of the second type (read negative on the standard) are considered "negative." The number of cases of the first type will, therefore, furnish a measure of lack of sensitivity of the test technique, while those of the second type will indicate its lack of specificity. It should be noted that since the two interpretations for each technique were in agreement, the discrepancies are more nearly the results of differences in the techniques themselves and not of

subjective or human errors of interpretation. On the available evidence all disagreements are counted against the test technique.

Suppose, however, that a third technique is introduced, i.e., every person has been tested by three different techniques and the films of each technique have been interpreted twice, either by the same or different readers. The persons in the study can now be represented by 64 different types of tables of 3 rows and 2 columns. Again, attention may be focused only on the undetermined cases which were previously represented by tables of the form $\begin{vmatrix} + & + \\ - & - \end{vmatrix}$ and $\begin{vmatrix} - & - \\ + & + \end{vmatrix}$. Some of these may now become $\begin{vmatrix} + & + \\ - & - \\ - & - \end{vmatrix}$ and $\begin{vmatrix} - & - \\ + & + \\ + & + \end{vmatrix}$. In other words, the readings on the third technique confirm not those of the standard (first row), but they agree with the readings of the test technique (second row).

The weight which may be assigned to this additional information as evidence as to whether the cases are likely to be positive or negative may be debatable. To some, the superiority of the standard as compared to any test technique is so completely accepted that it outweighs the combined evidence provided by any number of other techniques. To others, the fact that two different films confirm one another as to the presence or absence of a shadow will carry sufficient weight to cast doubt on the evidence provided by the standard. It is probably not too much to ask of even the confirmed believer in the standard to review again the films for these cases. In actual practice such a review reveals frequently that the combined evidence of the test techniques more nearly represents the probably true facts than does the evidence provided by the standard. This is understandable because the errors responsible for the types of discrepancies represented by such tables as are here discussed are traceable not to interpretation but to such considerations as positioning and physical characteristics of the films. It is, for example, more likely that a shadow will be hidden behind a rib in one film than that a shadow of a non-existent lesion would occur twice on two different films. Indeed, it is the exceptional case where the evidence provided by a single technique comes nearer to revealing the true facts than does that provided by two or more techniques.

From the above it is seen that the addition of a technique or of a reader modifies our confidence in the infallibility of the standard. The modification from the addition of a reader is associated with the presence of subjective errors, while that from the addition of another technique is mainly a result of errors which are inherent in the technique. When both the number of readers and the number of tech-

niques are increased, additional information becomes available which may be used to evaluate for specific cases the probability of their being positive or negative. Although it is not possible to assign *a priori* relative weights to the evidence provided by the different techniques and the different readers, the utilization of all the information provided by them makes possible a more comprehensive analysis of the material, which may lead to a more realistic evaluation of the relative efficiency of the different techniques.

OBJECTIVE COMPARISON OF THE FOUR TECHNIQUES

If, as a first step, it is assumed that each technique and each reader has equal weight as evidence that the case is either positive or negative, the problem of comparing the different techniques becomes relatively simple. The definition of a "positive" individual can then be made to depend on the number of positive readings obtained for him on all the techniques by all the readers. In the present study, for example, every person may be represented by a table of 4 rows (4 techniques) and 5 columns (5 readers), thus providing for each person 20 different readings. The individuals can be classified in 21 groups depending on the number of positive readings, ranging from 0 to 20. The probability that the last group is positive is obviously very great. Similarly, individuals with all 20 readings negative are, on X-ray evidence, negative for tuberculosis. Cases with one or only a very few positive readings are also very likely negative. As the number of positive readings increases, it becomes more difficult to determine where "false positive" stops and "true positive" begins. The decision as to the dividing line between "positive" and "negative" must, of necessity, be arbitrary. Some may want a relatively large number of positive readings before considering an individual positive; others may want to investigate individuals with a relatively small number of positive readings. The important thing, however, is that once the number is decided upon, it becomes a simple matter to compare the ability of the four techniques in selecting these "positive" cases. All that is necessary is to review X-ray interpretations for all individuals who are "positive" by the accepted definition and to count the number of positive readings yielded by each technique. A technique which is more sensitive should yield a larger number of positive readings on these cases than a less sensitive technique.[6]

This procedure has been followed in the present study, and a comparison has been made in different definitions of "positive," all yielding similar results. (See table 1.) For example, on the basis of a majority (11 or more) of the 20 readings, there were in the present

[6] This discussion limited to an investigation of the sensitivity of the different techniques. The problem of specificity will be considered in another publication. It may be seen from table 1 that there are differences in specificity among the four techniques.

study 62 "positives." The maximum number of positive readings that could be obtained on any one technique for these "positive" cases is 62 x 5, or 310. The actual number of positive readings was 259 for each of the three test techniques and 250 for the 14" x 17" celluloid. Similar results are obtained when "positive" is defined by six or more positive readings.

TABLE 1.—*Number of readings called positive by all readers on each technique when the cases are classified according to the total number of positive readings by all readers on all techniques*

Total number of positive readings by all readers on all techniques	Number of cases	Maximum possible number of positive readings on any one technique	Number of positive readings by all readers on—			
			35-mm.	4" x 10" stereo	14" x 17" paper	14" x 17" celluloid
1- 5	171	855	109	90	49	50
6-10	29	145	54	54	58	47
11-13	12	60	36	36	37	37
14-16	20	100	79	76	75	69
17-19	9	45	39	42	42	39
20	20	105	105	105	105	105
11-19	41	205	154	154	154	145
6-19	60	350	199	199	202	192

If the assumption of equal weight for each technique and each reader is valid, it would be possible to conclude from the above that there is little to choose between the four techniques and that they are all equally efficient in finding cases of tuberculosis in mass survey work.

TESTS FOR THE VALIDITY OF THE METHOD

It is now necessary to investigate the possible limitations of this method of analysis. In the main, the testing must determine whether the method is likely to include an appreciable number of negative cases among those that are defined as "positive" or to include a number of positive cases among the "negatives."

As to the latter, it is possible that the "negative" group contains a number of important positive cases which by the accident of definition were not included among the "positives." There may, for example, be a certain type of lesion which one of the techniques, say the 14" x 17" celluloid, detects but which the other techniques fail to identify. Such a case may even have as many as 5 positive readings on the 14" x 17" celluloid film, but will have fewer than 11 total positive readings and will therefore be included among the "negatives." This obviously is an extreme case, but a tendency in this direction, even though less pronounced, may nevertheless lead to the exclusion of a number of significant cases from the group of "positives."

Among the cases which are "positive" by definition, the method may also fail because of a tendency to overread, which may operate on the test techniques. A technique may accumulate as many

positive readings on the "positive" cases as another technique, not because it is equally sensitive but as a result of overreading on some of the films and underreading on others. It will make up by overreading what is missed by underreading and have the total number of positive readings approximately equal to that of the other techniques.

Whether either of these tendencies exists can be tested by comparing the distributions of the films according to the number of positive readings for each of the different techniques. If it is true that the group of "negatives" contains an appreciable number of positive cases which were selected by the 14″ x 17″ celluloid but which failed to be included in the "positive" group because the other techniques failed to select them, there would be among the "negatives" more cases with three or more positive readings for the 14″ x 17″ celluloid than for any of the others. Similarly, if there is a tendency on the part of a test technique to overread some and underread others of the "positive" films, then the distribution for that technique should be in the form of a U-shaped curve, with many cases having zero or one positive reading, and many with four or five positive readings while cases with an intermediate number of readings would be relatively few.

When the distributions for the different techniques are compared separately on the "negative" cases and the "positive" cases, (see tables 2a, 2b and figs. 1 and 2), it is seen that they are similar for the

TABLE 2a.—*Distribution of the 29 cases with 6–10 positive readings on all techniques according to the number of positive readings on each technique*

Technique	Number of positive readings					
	0	1	2	3	4	5
35 mm	1	13	8	3	4	
4″ x 10″ stereo	2	7	13	7		
14″ x 17″ paper	1	9	12	3	4	
14″ x 17″ celluloid	2	12	11	3	1	

TABLE 2b.—*Distribution of the 41 cases with 11–19 positive readings on all techniques according to the number of positive readings on each technique*

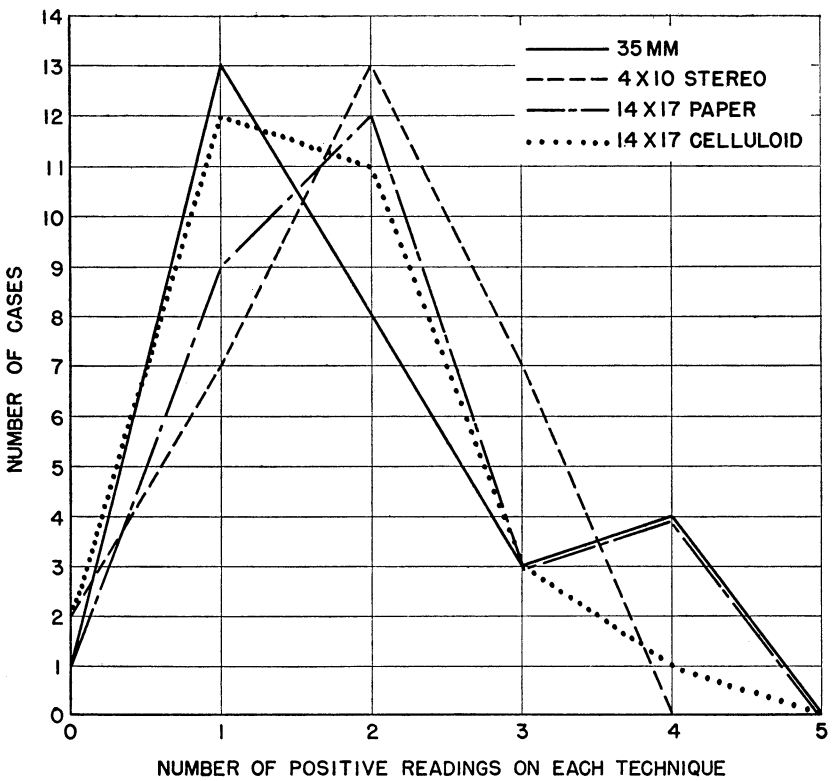
Technique	Number of positive readings					
	0	1	2	3	4	5
35 mm			4	12	15	10
4″ x 10″ stereo			3	13	16	9
14″ x 17″ paper			3	14	14	10
14″ x 17″ celluloid		1	4	14	16	6

four techniques and that such variations as exist from technique to technique are not of the type which would indicate that the limitations mentioned above exist.

It should be pointed out that there undoubtedly are a number of persons who probably have X-ray evidence of tuberculosis but who

are excluded from the group of "positive" cases (11 or more positive readings) because of the failure of one or more of the techniques to identify the lesion on the corresponding films. These persons may be found mainly in the group with less than 11 and more than 5 positive readings. The important consideration, however, is whether this happens more often with one technique than with another. A second test that needs to be made, therefore, consists in the following: If the 4 techniques contribute equal numbers of positive readings

FIGURE 1. DISTRIBUTION OF THE 29 CASES WITH 6 - 10 POSITIVE READINGS ON ALL TECHNIQUES ACCORDING TO THE NUMBER OF POSITIVE READINGS ON EACH TECHNIQUE.

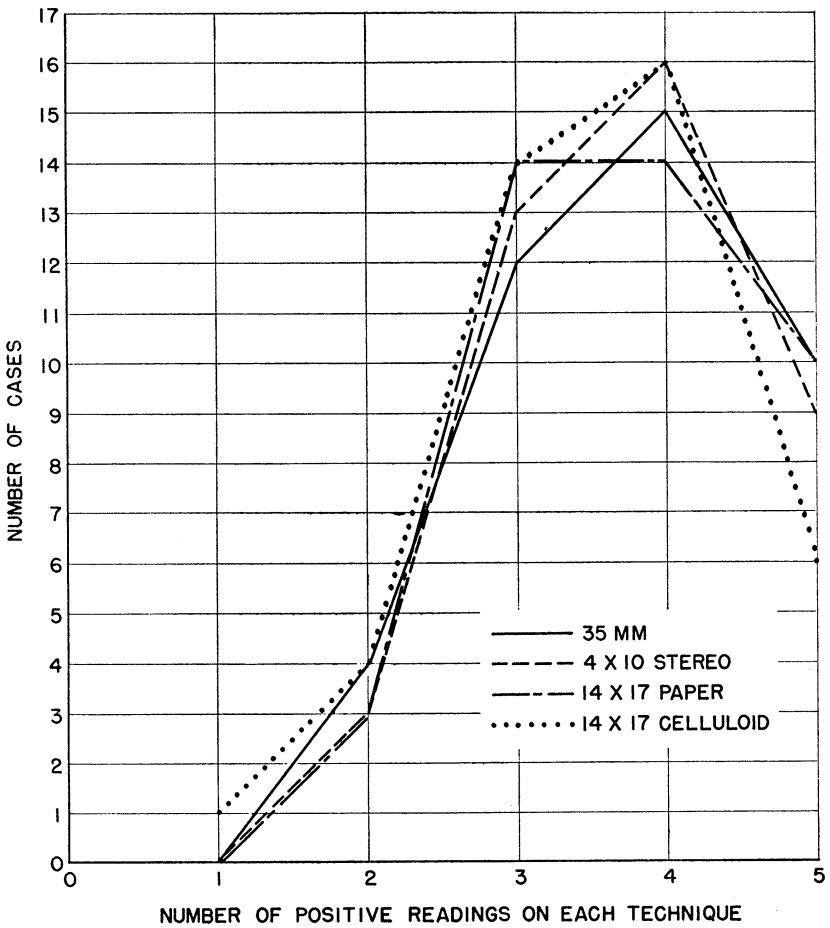


when "positive" cases are defined in terms of a majority of positive readings (11 or more), the same should hold true if the definition of positive cases is broadened to include borderline cases with perhaps 6 or more positive readings. In the present study this was the case. (See table 1.)

In addition, it is necessary to review carefully each of the cases which may have been excluded from the "positive" group (11 or more positive readings) by virtue of the failure of one or more of the

techniques. The number of such cases and their distribution for the four techniques is shown in table 3. In this table are shown all the cases for which there were at least 3 positive readings recorded on 1 of the 4 techniques but for which the total number of positive readings was less than 11.

FIGURE 2. DISTRIBUTION OF THE 41 CASES WITH 11 - 19 POSITIVE READINGS ON ALL TECHNIQUES ACCORDING TO THE NUMBER OF POSITIVE READINGS ON EACH TECHNIQUE.



It is important to note, first, that each of the 4 techniques can claim a number of possibly positive cases which were excluded from the group of 11 or more positive readings by virtue of the failure of the other techniques. Actually there were fewer such cases which the 14″ x 17″ celluloid technique detected than there were for each of the other techniques. In view of the possibility, however, that "positives" detected by the 14″ x 17″ celluloid technique may be

760653—47——3

more likely to be positives than those detected by the other techniques, it is necessary to investigate in more detail the 4 cases in the "negative" group (less than 11 positive readings) which have 3 or more positive readings on the 14″ x 17″ celluloid films. Of these, 1 had only 3 additional positive readings on the other techniques (out of a total of 15), 2 had 5 additional positive readings each, and 1 had 7. It is likely that some, if not all, of these cases may actually represent positive individuals. It is, however, remarkable that even with the small amount of material available in this study, it was possible to find four other cases which were almost identical to these four, except that the 14″ x 17″ celluloid readings are interchanged with one of the test techniques.

Table 3.—*Number of additional positive readings on the other techniques for cases which were called positive by 3 or more readers on a specified technique for all cases with less than 11 positive readings on all techniques*

Called positive by three or more readers on—	Number of Cases	Number of additional positive readings on other techniques							
		0	1	2	3	4	5	6	7
35 mm	9	1	------	1	------	2	1	4	------
4″ x 10″ stereo	9	------	1	1	2	1	1	2	1
14″ x 17″ paper	8	------	------	1	2	1	1	1	2
14″ x 17″ celluloid	4	------	------	------	1	------	2	------	1

The actual readings on all the techniques for these two sets of four cases are presented in table 4. A review of these tables indicates similar situations for the 14″ x 17″ celluloid and for the other techniques. For example, case number 542 had four positive readings on the 14″ x 17″ celluloid, three positive readings on the 35 mm., two on the 4″ x 10″ stereo, and no positive readings on the 14″ x 17″ paper. It is almost certain that the individual represented by this table is positive and that this case is not included in the "positive" group (11 or more positive readings) because of the failure of the 14″ x 17″ paper technique to select it. However, case number 939 shows almost an identical picture except that the paper recorded four positive readings while the celluloid failed to identify it. When the two sets of four films for these two cases were reviewed by the radiologists, it was agreed that both represent positive cases, and the failure of the specific technique to select them was assigned to the physical characteristics of the film or to the positioning of the patient. The important consideration is whether these technical or chance misses happen more frequently with the test techniques than with the standard. In the present study, as was shown above, this was not the case.

It may also be of interest to supplement the above test by means of a hypothetical binomial distribution. The observed distribution according to the number of positive readings may be compared with a

TABLE 4.—*The actual readings on the 4 cases in the "negative" group (having less than 11 positive readings on all techniques) with 3 or more positive readings on the 14" x 17" celluloid, and for 4 similar cases selected by 3 or more readers on other techniques*

Techniques	Readers																	
	M	N	O	P	Q	M	N	O	P	Q	M	N	O	P	Q	M	N	O	P	Q
	Case No. 519					Case No. 641					Case No. 542					Case No. 569				
14" x 17" celluloid	−	+	+	−	+	−	+	−	+	+	+	+	+	−	+	+	+	−	−	+
35 mm.	−	+	−	−	+	−	−	−	−	+	+	+	−	−	+	+	+	+	−	+
4" x 10" stereo	−	−	−	−	+	+	+	−	−	+	+	−	−	−	−	+	+	−	−	+
14" x 17" paper	−	−	−	−	+	−	+	−	−	+	−	−	−	−	−	+	+	−	−	+
	Case No. 530					Case No. 848					Case No. 939					Case No. 964				
14" x 17" celluloid	−	−	−	−	−	−	−	−	−	+	−	−	−	−	−	+	+	−	−	+
35 mm.	−	+	+	−	−	+	+	+	−	+	−	−	+	−	−	−	+	−	−	+
4" x 10" stereo	+	−	+	−	+	−	−	−	−	+	−	−	+	−	+	+	+	+	−	+
14" x 17" paper	−	−	−	−	−	−	+	−	−	+	+	+	−	−	+	−	+	+	−	+

theoretical one which is obtained from the assumption that all the techniques are equally sensitive and that the probability of missing a positive film is constant for the five readers. While this test can be made on the entire distribution, including also the probability of missing a negative film, the methodology becomes too complicated. For the present purposes it was thought sufficient to test only the sensitivity of the different techniques, since in practical work sensitivity is more important than specificity. The test can, therefore, be made only on that part of the distribution embracing 11 or more positive readings. Furthermore, there was a relatively large number of cases (21) in which all 20 readings were positive. It was obvious from a knowledge of the approximate probability of missing a positive case (higher than 0.20) that no binomial distribution which would approximate the rest of the distribution would yield as many as 21 cases with 20 positive readings. It was therefore necessary to modify the hypothesis and to assume that there are a certain number of cases with very obvious lesions (approximately 21) whose probability of detection is 1, i. e., that no technique and no reader will miss them. It is then possible to compare the theoretical distribution obtained from the binomial with that part of the observed distribution ranging from 11 positive readings to 19. The average probability of missing a lesion on these cases was found to be .276. A distribution given by $N(.276+.724)^{20}$ will, therefore, test the hypothesis that with the exception of some 21 cases whose lesions are so obvious that no reader and no technique will miss them, there is an equal probability for each of the techniques and each of the readers to miss the other "positive" cases. The value of N which yielded the smallest value of χ^2 was 41.

TABLE 5.—*Chi Square test for goodness of fit using the theoretical values derived from the expansion of the point binominal 41 (0.276+0.724)*^{20}

Number of positive readings on all techniques	Observed frequency (number of cases) O	Theoretical frequency T	$\frac{(O-T)^2}{T}$
19	3	.45	7.243
18	3	1.65 (2.10)	
17	3	3.85	.188
16	5	6.36	.291
15	9	7.91	.150
14	6	7.71	.379
13	5	5.99	.164
12	4	3.80	.011
11	3	1.96	.551
		$\chi^2 = 8.977$	
		$p = .18$	

The expected frequencies according to this binomial distribution and the observed frequencies are shown in table 5. It may be seen that the fit is not too bad, the chi-square test gives for P the value 0.18. In other words, the test provides no reason for rejecting the

hypothesis. Incidentally the value of 41 obtained for N together with the 21 cases with 20 positive readings yielded a total of 62 positive cases, or the same number obtained by using 11 or more positive readings according to the arbitrary definition adopted.

SUMMARY AND CONCLUSIONS

In this paper an attempt is made to compare the effectiveness, for tuberculosis case finding, of various photofluorographic and roentgenographic methods. Some 1,200 individuals were examined consecutively on four different machines yielding for each a 35-mm. photofluorogram, a 4″ x 10″ stereophotofluorogram, a roentgenogram on 14″ x 17″ paper negatives and a 14″ x 17″ celluloid film. These were interpreted independently by five expert radiologists and chest specialists. A second independent interpretation was obtained from each reader on the 14″ x 17″ celluloid films.

A method was devised by which the comparison was made objectively without a predetermination of any of the techniques as a standard. Instead, the evidence yielded by all five readers on all four techniques was utilized in defining "positive" cases, and a comparison of the four techniques was based on their ability to detect these "positive" cases. Several tests of the validity of this method are presented.

The results of this analysis justify the conclusion that, strictly from the point of view of their ability to find cases of tuberculosis in mass survey work, none of the techniques, not even the 14″ x 17″ celluloid, is superior to any of the others.

OUTLINE OF STATISTICAL TREATMENT OF THE PROBLEM OF DIAGNOSIS [1]

By J. Neyman, *Professor of Mathematics and Director of the Statistical Laboratory University of California, Berkeley, California*

(1)

Every attempt at a mathematical treatment of phenomena must begin by building a simplified mathematical model of the phenomena. In studying the problem of diagnosis as presented by Dr. Yerushalmy,[2] we have to consider a population (or a universe) U of individuals which we shall imagine divided into three exclusive categories.

(i) individuals *entirely free* from the given disease, whose proportion is α.

(ii) Individuals *moderately affected* by the disease, whose proportion is β.

[1] This paper was presented before the Institute of Mathematical Statistics, at the twenty-eighth annual meeting of the Pacific Division of the American Association for the Advancement of Science, in San Diego, California, June 18, 1947 as a discussion of the foregoing paper by Jacob Yerushalmy.

[2] See "Statistical Problems in Assessing Methods of Medical Diagnosis, with Special Reference to X-ray Techniques" by Jacob Yerushalmy, this issue.