a considerable wedge of bone, the articular surface is altered so that a greater range of dorsiflexion is given without any alteration in the relation between the head of the metatarsal and the base of the phalanx.

The most satisfactory operation for the condition has been that of removal of the base of the phalanx. When this is carried out care should be taken to make certain that at least half of the phalanx is removed at the time of operation; if less than this is taken the usual result is a painful fibrous ankylosis, and the patient may even complain of more pain than previously.

After-treatment is identical with that for hallux valgus; walking is permitted after three weeks, and exercises and massage of the toe lead to an improvement in its range of movement. With a restoration of voluntary control of movement the disability is overcome.

PROBLEMS OF CHANCE IN CLINICAL WORK

BY

DONALD MAINLAND, M.B., CH.B., D.Sc.

PROFESSOR OF ANATOMY, DALHOUSIE UNIVERSITY, HALIFAX,
NOVA SCOTIA, CANADA

A recent article on Bell's palsy in this *Journal*¹ recorded an attempt to find "a criterion whereby we may decide in the early stage [of the palsy] that a given case is unlikely to recover." The reason for the search was that facial nerve decompression should not be undertaken unless the odds against natural recovery were sufficiently high. Even a reader not specially interested in this question will easily recognize that this type of problem is extremely common, and the purpose of the present paper is to use the work on Bell's palsy simply as an example of ideas and methods applicable to many problems. In that work comparison was made between twenty-five cures and eighteen failures (Table I), and the

TABLE I.—Data on Bell's Palsy (Tumarkin¹)

Cases	No Pain, No Loss of Taste	Pain, but No Loss of Taste	No Pain, but Loss of Taste	Pain and Loss of Taste
25 cures ...	18	4	2	1
18 failures	2	6	4	6

conclusion was that "if the patient has both pain and loss of taste the chances are 2 to 1 that he will not recover; but if he has neither of these symptoms the chances are 25 to 1 that he will recover."

Neglect of Chance and of Sampling Errors

This work is a useful illustration of a general statement made two years ago in the editorial section of the *Canadian Medical Association Journal*²—namely, that in much medical literature there is inadequate recognition and treatment of questions of chance, of the allowance to be made for small samples, and of natural variation from sample to sample of the same material, whether that material is composed of blood specimens or cases of paralysis. This statement applies equally to all kinds of unpublished clinical and non-clinical discussions. It is, for example, not uncommon to hear an admission that a certain set of cases is too small to justify conclusions, but it is very uncommon to find any attempt to show what conclusions can be reasonably accepted from the given sample; fairly often the speaker will, in spite of his admission, proceed to generalize and argue from his sample.

To avoid the hasty conclusion that these comments are those of an outsider, a statistical expert unacquainted with clinical problems, it is perhaps desirable to note the stages by which the present writer arrived at this point of view. A medical training in which no attention

was paid to problems of chance and sampling errors was followed by research in embryology and histology in which these problems at once arose. The adequate teaching of anatomy entailed wide and continued acquaintance with clinical matters, and it soon became obvious that the same kind of problem was present everywhere.

Certain other things that became obvious may be summarized thus: (1) the term "statistics" is unfortunate because it suggests either masses of data, sometimes inaccurate, or highly artificial mathematical tests; (2) the numerous books on the subject, including those on medical statistics, give too much and too complicated detail for clinical workers, and commonly fail to show that statistical ideas are of any use to a practitioner; (3) the most important part, even of elaborate statistical technique, is the critical selection of material; (4) for ordinary clinical work it is the ideas underlying the words "chance" and "sampling variation" that are of most importance, while the necessary technical tests are few and simple.

Everyone recognizes that diagnosis and prognosis are based to a very large extent on the weighing of probabilities, even if the physician or surgeon is not thinking in terms of exact numerical odds or chances. For many purposes it is unnecessary to go farther than this, but there are numerous other cases, such as the investigation of Bell's palsy, here discussed, where misleading conclusions are drawn unless the observer and reader of the article understand more clearly how to handle the tool that they are using—namely, the calculation of odds. It is not suggested that the clinical observer should apply elaborate mathematical analysis, but that he should examine more fully the ideas underlying his judgements—ideas that are fundamental in dealing with any living material.

Errors in Estimation of Chances

With regard to the conclusions drawn from Table I the first criticism concerns the numerical value of the chances. Out of twenty cases with no pain or loss of taste eighteen were cured and two were not cured. Therefore the chances of cure in such cases, so far as they can be estimated from this sample, are 18 out of 20—that is, the odds are 9 to 1 that such a patient will recover. Out of the seven cases with pain and loss of taste one was cured. Therefore the odds, so far as this sample shows, are 6 to 1 against cure.

The odds of 2 to 1, quoted above, seem to have been derived from a consideration of the eighteen failures, six of whom showed pain and loss of taste, and the only conclusion from this fact could be that, among the cases of failure, the odds were 12 to 6—that is, 2 to 1—against a patient who failed to be cured having shown both pain and loss of taste—a deduction that was not required. A similarly erroneous method seems to have been employed in assessing the odds of 25 to 1 in favour of cure when neither pain nor loss of taste is present.

Variation between Samples

It is true that the results of this investigation on facial palsy were put forward tentatively, but before even a tentative argument was based on them a second question should naturally have been asked—namely: how far is one justified in generalizing from these samples—that is, in applying the (corrected) odds, 9 to 1 and 6 to 1, to other patients?—one of the specified objects of the investigation. This question can be expressed otherwise by asking: what evidence is there from the sample of twenty patients (with paralysis but no pain or loss of taste) that, if all the thousands of patients in that class could be examined, the odds in favour of cure would not be 3 to 1 or 2 to 1, instead of 9 to 1 as suggested by the sample? The same kind of question can be asked regarding the sample of seven patients of whom only one was cured. The answer to such questions involves ideas of sampling that seem almost too elementary to mention, but which are often not properly considered before calculation is started.

First, a sample must be representative of its class. Theoretically the only proper way of obtaining such a sample is by random picking from the whole group, so that each individual is given an equal chance of appearing in the sample, but in practice this is seldom possible, and yet precautions have been taken to ensure that the given sample is equivalent to a random sample. The statistical expert may be acquainted with the kinds of errors that can detract from the randomness of the sample, but the clinical observer himself is in a much better position to know these. He must decide whether, in regard to factors that have or may have an influence on recovery from disease—for example, age, sex, severity of attack, and treatment—his sample can be considered as equivalent to a random sample of all the possible patients to whom he wishes to apply the odds estimated from his sample.

This consideration naturally leads one to look into the cases themselves for evidence of factors that may have great weight in the cure or failure to recover. It is obviously useless to find the chances of death from all cases of whooping-cough and apply the information to children under 1 year of age, in whom it is a very dangerous disease. Similarly, among cases of facial paralysis the question must be asked: have age, sex, mode of treatment, and so on, had any appreciable effect in the sample that I am investigating? The tests that may be applied to answer that question are beyond the scope of the present paper, and it will be assumed that, for the samples under consideration, all such detectable influences have been ruled out. Actually, if it were possible to determine the nature and degree of all factors promoting recovery or failure to recover, it would be possible to form a prognosis with the precision of an experiment in physics.

In medicine, prophecy in terms of chances or odds must be used, partly because many of these factors are unknown, and largely because, as in all biological phenomena, results are dependent on a large number of small factors, varying independently of each other from one patient to another. Recovery from disease depends partly on almost innumerable factors, such as the severity of the cause, the detailed anatomical arrangements of the affected parts—for example, vascularity, the vigour of tissue reactions, the history of the patient, constitutional factors, environment, including medical treatment, and many others. If this conception is followed into an apparently theoretical and somewhat mathematical discussion it will reveal the way in which one can answer such practical questions as those given above regarding the reliability of the odds estimated from a single sample.

At first the discussion can be confined to patients with a certain degree of severity of facial paralysis, denoted by or associated with pain and loss of taste. Let it be remembered that in these patients there are numerous small factors, independent of each other, promoting or retarding recovery. Even among patients who are placed in this class there must be some small differences in the degree of severity or extent of the lesion, perhaps differences that are not detectable by any available method. In some patients the factors for recovery will happen to predominate, in others the factors for failure. Let it be supposed that in a large series, say many millions of people, half recover—that is, the "true odds" * against (or for) recovery are even (1 to 1). If one now takes at random a number of samples with two patients, A and B, in each, then on the average half the A's will be cures and half will be failures. For each class of A, half the B's will be cures and half will be failures. Hence the general result will be:

Both cures: 25 per cent. of samples.

One cure; one failure: 50 per cent. of samples.

Both failures: 25 per cent. of samples.

The same type of reasoning can be applied to samples containing any number of patients, but mathematical

devices are employed to save labour. Thus it can be shown that, with samples of seven patients, if the odds against recovery are 1 to 1, the percentages in Table II should be expected.

TABLE II.—*Expected Samples if Odds For and Against Cure are Even (Seven Patients per Sample)*

Cures	Failures	Percentage Frequency
7	0	0.78
6	1	5.47
5	2	16.41
4	3	27.34
3	4	27.34
2	5	16.41
1	6	5.47
0	7	0.78
Total ...	—	100.

Reliability of Odds Estimated from Samples

With even chances or odds, the commonest samples are naturally those with four cures and three failures, and those with three cures and four failures—that is, they centre round the average value 3.5 cures (3.5 failures), and the original question with regard to the actual sample of seven patients can now be expressed in the form: what right has one to suppose that this sample does not belong to such a series, where the true odds are even (1 to 1)? Table II. shows that samples that differ from the average value by as much as, or more than, the actual sample should be expected in 12.5 per cent. of the total (such samples being bracketed in the table). *Therefore, unless an observer is willing to risk being wrong in over 10 per cent. of his judgements in such cases, he should not claim that a sample of seven patients with six failures (or six cures) indicates odds other than 1 to 1.* Moreover, he should not be much surprised if another sample from the same type of patients showed six cures instead of six failures.

It is, of course, entirely open to any worker to set his standards of accuracy and safety where he chooses, but it is generally felt by workers in the field of chances and probability that an error of not more than 5 per cent. should be allowed in judgements of the type under discussion. On that basis there is no adequate proof that the given sample (six failures and one cure) does not come from a series with "true odds" of 1 to 1. This is expressed by saying that the sample has not shown any "significant difference" from a 1 to 1 ratio—that is, that the apparent difference has no proved meaning or significance beyond what would be expected by pure random sampling—the effects of chance alone. This can be expressed also by saying that there is no adequate proof that the true odds are "unlikely" to be 1 to 1.

Now if there were a sample with seven cures or one with seven failures it would be said to show a significant difference from a 1 to 1 ratio because such samples form only $2 \times 0.78 = 1.56$ per cent. of the whole series in Table II. Then one would be justified in saying that it was unlikely that the true odds were 1 to 1. The danger of accepting the actual sample of seven at its face value will be realized when it is pointed out that similar methods reveal that one is not justified in assuming that the true odds are unlikely to be 63 to 1 against cure, or 9 to 8 in favour of cure.

Attention can now be turned to the group of twenty patients with neither pain nor loss of taste. If there were many millions of such patients, and two-thirds of them were cured, the true odds in favour of cure would be 2 to 1. This could be interpreted as follows. The principal factor in determining cure was the lack of severity or extent of the initial disease or position of the lesion. In addition, however, there were, as always, numerous small independent factors contributing to cure

* For the sake of accuracy it should be pointed out that the term "true odds" is used in this article to signify the odds that would be approached more and more closely by taking larger and larger random samples of the same material.

or failure. In different patients these would be present to different degrees. If numerous samples of twenty patients were taken it can be shown, again with mathematical devices to lighten the labour, that there should be expected the various classes of samples in Table III.

TABLE III.—*Expected Samples if Odds in Favour of Cure are 2 to 1 (Twenty Patients per Sample)*
(Percentages below 0.01 are omitted.)

Cures	Failures	Percentage Frequency
20	0	0.03
19	1	0.30
18	2	1.43
17	3	4.29
16	4	9.11
15	5	14.57
14	6	18.21
13	7	18.21
12	8	14.83
11	9	9.87
10	10	5.43
9	11	2.47
8	12	0.92
7	13	0.28
6	14	0.07
5	15	0.01
4 to 0	16 to 20	0.00
Total ...	—	100.00

With the criterion of significance used above, the actual sample (eighteen cures and two failures) shows a significant difference from the true odds 2 to 1, because Table III reveals that samples as rare as the actual sample, or more rare, form only $(1.76 + 1.28 =) 3.04$ per cent. of the total. This can be expressed by saying that it is unlikely that the given sample came from a series in which the true odds were 2 to 1. As a further illustration let it be supposed that the sample had contained nineteen cures. Samples as rare as this, or more so, form only $(0.33 + 0.36 =) 0.69$ per cent. of the total expected sets of samples in Table III. This is less than 1 per cent., and it can be stated that such a sample shows a "highly significant" difference from the odds 2 to 1—that is, that it is "highly unlikely" that such a sample came from a series with odds of 2 to 1.

A Simple Method of Determining the Likelihood of Odds

If such elaborate methods were necessary before deciding how much reliance to place on a result, it would not be reasonable to suggest their use in clinical work, but the preceding ideas must be grasped to elucidate a very simple test that is for many purposes sufficiently accurate.

The accompanying graph shows the data from Table II—a symmetrical figure, rising to a peak which corresponds to an average of 3.5 failures to recover per sample of seven patients. A quantity called the standard deviation is calculated, and it can be proved mathematically that, when the numbers in the samples are high, twice the standard deviation, measured from the centre or average, will exclude, at the extreme left and right ends of the series, a total of roughly 5 per cent. of all samples, so that, on the criterion stated above, any sample beyond that distance can be considered as significantly different from the average. The validity of applying this to small samples will be discussed later, but in the meantime the method of using the standard deviation can be shown as follows.

In the given sample of seven patients six failed to recover. The question is: does this show that the odds against recovery are unlikely to be even (1 to 1)—that is,

is this sample significantly different from an average of 3.5 failures per sample? If the odds are even—that is, the chances of failure are 1 out of 2, then the probability of failure, p , is expressed as $1/2$. The probability of cure, q , is likewise $1/2$. The number in the sample, n , is 7. The standard deviation is $\sqrt{npq}$ —that is, $\sqrt{7 \times 1/2 \times 1/2} = 1/2\sqrt{7} = 1.32$. Twice this is 2.64. This quantity, added to the average (3.5) gives 6.14, and, subtracted from the average, leaves 0.86. Therefore, unless a given sample contains more than 6.14 failures or less than 0.86 failures, there is no adequate proof that the true odds are not 1 to 1, and the actual sample contained only six failures.

As a second example, the twenty patients without pain or loss of taste can be used. Eighteen recovered. Does

this show that the odds in favour of recovery are unlikely to be 2 to 1? If the odds were 2 to 1, p , the probability of cure, would be $2/3$, and q , the probability of failure, would be $1/3$. On the average, twenty patients would show two-thirds cured—that is, $2/3$ of $20 = 13.33$. Standard deviation = $\sqrt{npq} = \sqrt{20 \times 2/3 \times 1/3} = 2.11$. Twice the standard deviation = 4.22. The average plus twice the standard deviation = $13.33 + 4.22 = 17.55$. The sample contained eighteen cures, and was further away from the average than this; therefore it is unlikely that the odds were as low as 2 to 1.

It may now be asked: is it "highly unlikely" that the odds were 2 to 1? For purposes such as the present this can be answered by taking three times the standard deviation—that is, $3 \times 2.11 = 6.33$. If this is added to the average, 13.33, the result, 19.66 cures, is farther away from the average than the number, 18, in the actual sample, and therefore there is no adequate proof that odds of 2 to 1 are highly unlikely. (To obtain a test for highly unlikely odds equivalent to the exact one given above, where 1 per cent. of samples was excluded, 2.576 times the standard deviation should be taken, but, besides its simplicity, multiplication by three gives a safer criterion.)

As stated at the outset, the standard deviation method is very accurate where the odds are even, giving a symmetrical series, as in the accompanying graph, and provided that the numbers of items—for example, patients—in a sample are large (one hundred or more). However, by use of exact probability tables,³ from which the numbers in Tables II and III were taken, it has been possible to show that the standard deviation method is sufficiently accurate for the purposes now under discussion, with samples containing as few as three individuals, and with odds of 1 to 1; 2 to 1; 3 to 1; 13 to 3 (roughly 4 to 1); 7 to 1; 15 to 1; 63 to 1; and 255 to 1.

The examples given here have been limited to one disease, but it will be realized how the same methods can be applied to any comparable data. With any given sample it will be obvious how one can by trial find out odds that are unlikely because they are too high, and

odds that are unlikely because they are too low. (A more direct algebraic method is available, but is rather complicated.) The reader should beware of a method sometimes recommended—namely, to calculate a standard deviation from the sample itself and then set off twice this standard deviation on each side of the odds given by the sample, to mark the likely upper and lower limits of the true odds. This method is fairly safe when the odds are nearly even, but is not safe otherwise.

The Number of Cases

It is obvious that, in general, the greater the number of cases in the sample the more accurate will be the odds estimated from the sample, but the bane of statistics in the hands of those unacquainted with ideas of sampling is the trust placed in large numbers, although it does not require profound knowledge to realize that if, for example, one wishes to estimate the incidence of an infectious disease in a whole population, a perfectly random sample of a dozen people or less from the whole population is safer than a sample of several hundreds taken from the slums only.

Various specific questions may be raised regarding the size of samples. A worker who has seven patients of whom six failed to recover from paralysis may ask: how many cases should I observe to prove that the odds against recovery are 2 to 1, or are unlikely to be 2 to 1? The answer is that it can never be proved, without examining all possible patients, what are the exact odds. Even if all existing and past patients were included the result would only be an estimate of the odds when the result had to be applied to new patients. However, by increasing the size of the sample it is quite possible to narrow the margin of error to any desired extent.

Therefore a more specific answer can be given to such a problem as the following.

An observer has a small sample, and suspects that the true odds are 2 to 1. He knows that he cannot prove this, but, assuming that it is so, he wishes to know how large a sample he would require so as to be reasonably sure that the odds are not 1 to 1 or 3 to 1. The answer, obtained by calculation that is rather too elaborate for this paper, is that the observer could expect a sample of 130 to be large enough to show that the odds are unlikely to be 1 to 1, and that he could expect a sample of 450 to be large enough to show that the odds are unlikely to be 3 to 1.

With regard to small numbers the probability tables referred to previously show that, in terms of cure and failure, a sample of less than six patients cannot show that the odds for (or against) cure are unlikely to be 1 to 1, even if all the patients in the sample are cured or all fail to recover. A sample of eight is necessary before it is possible to show that odds of 1 to 1 are highly unlikely, and even then all the patients must be cures or failures before such a judgement is safe.

A Caution Regarding Methods

Although only one type of problem has been discussed in this paper it should be repeated that the idea of variation between samples is essential in all clinical work that involves counting or measuring, and a general caution regarding methods may not be out of place—a caution against the older statistical methods, still often used and often safe enough with large samples, but by no means safe where application to small samples is necessary, as is very common in clinical work. It is to the methods of Fisher,⁴ his students and followers,⁶ that one must turn for safety in such cases.

The calculation necessary in applying the standard deviation test is very small. If many extractions of square roots are to be done Barlow's tables⁷ are to be recommended, but various smaller tables are often sufficient, and the direct arithmetical method is not very difficult. Finally, if a worker still hesitates to apply any tests to his results, he will, it is hoped, be prompted to observe the following rules:

1. To be sure that one's sample is the equivalent of a random sample of the material to be studied.

2. Never to express one's results as percentages without giving at least the totals from which *each* percentage was calculated.

3. Not to argue from samples unless they have been adequately tested by oneself or someone else. (An adequate test may, of course, entail only competent inspection of the data and results.)

Summary

In medical work there is serious neglect of highly important problems of chance and of variation between samples, exemplified in a recent estimate of chances of cure in facial paralysis where odds are wrongly estimated, and too great reliance is placed on samples.

What is required in clinical work is not elaborate mathematical tests, but an understanding of the meaning of chance, and adequate precautions that the samples, however small, are unbiased.

Recovery from disease, like most phenomena in life, commonly depends not only on a few important factors, but on numerous small independent factors, varying from patient to patient, and therefore, whatever may be the true odds for or against recovery, samples of patients must vary in the proportion of recoveries, and often give misleading ideas of the true odds. The simple standard deviation method, properly applied, is commonly adequate, even with a sample of three patients, to show what are likely to be the limits between which the true odds must lie.

The influence of sample-size is illustrated, and a caution is given against the use of old statistical methods that are specially unreliable with small samples.

I desire to acknowledge my indebtedness to Dr. R. A. Fisher, F.R.S., for inspecting this paper before it was offered for publication, and particularly for providing an alternative method of obtaining the numbers in the third paragraph of the section on the number of cases.

REFERENCES

- ¹ Tumarkin, A.: *British Medical Journal*, 1936, i, 580.
- ² Mainland, D.: *Canadian Med. Assoc. Journ.*, 1934, xxxi, 656.
- ³ Warwick, B. L.: *Probability Tables for Mendelian Ratios with Small Numbers*, Texas Agriculture Experiment Station, 1932.
- ⁴ Fisher, R. A.: *Statistical Methods for Research Workers*, fifth edition, Edinburgh and London, 1934.
- ⁵ Idem: *The Design of Experiments*, Edinburgh and London, 1935.
- ⁶ Snedecor, G. W.: *Calculation and Interpretation of Analysis of Variance and Covariance*, Ames, Iowa, 1934.
- ⁷ Barlow's *Tables of Squares, Cubes, Square Roots, Cube Roots, and Reciprocals*, third edition, London, 1930.

THE VISSCHER-BOWMAN TEST FOR PREGNANCY

BY

GLADYS H. DODDS, M.D., M.C.O.G.

ASSISTANT IN THE OBSTETRIC UNIT, UNIVERSITY COLLEGE HOSPITAL;
PHYSICIAN, CLAPHAM MATERNITY HOSPITAL

(From the Obstetric Unit, University College Hospital)

The introduction of a new method in the diagnosis of pregnancy, particularly of simple chemical tests which can be performed by any practitioner and will give an immediate result, is always of great interest. The disadvantages of the Aschheim-Zondek test and its modifications are that it requires a laboratory with a stock of animals, and the result is not available for at least forty-eight hours. In 1929 Voge¹ described a chemical test, the bromine test, which was simple, cheap, and rapid, but had 25 per cent. error among pregnant patients and 13 per cent. among non-pregnant patients.²

More recently Visscher and Bowman³ have described another chemical test, for which they claim great accuracy. The test is performed as follows. To 1 c.cm. of urine are added one drop of 1 per cent. hydrogen peroxide, five drops of a 1 per cent. phenylhydrazine hydrochloride solution in water, five drops of a 5 per cent. methyl