BRITISH MEDICAL JOURNAL

LONDON SATURDAY DECEMBER 27 1958

THE THREE PILLARS OF CLINICAL RESEARCH*

BY

H. J. B. ATKINS, D.M., M.Ch., F.R.C.S., Hon.F.A.C.S.

Director of the Department of Surgery, Guy's Hospital; Dean of the Institute of Basic Medical Sciences

Despite its title, this paper is really addressed to surgeons. The message, if there be a message, may be applicable to other groups within the profession, but I am not qualified to expound on the difficulties met with or the foibles displayed by my medical, gynaecological, paediatric, geriatric, or other colleagues in the course of their clinical inquiries; and so it is to the surgical fraternity primarily that I address myself.

As surgeons, we handle a great deal of material which is properly the substance of clinical research, but I am afraid that in many instances we handle it clumsily so that the final product is uncertain and open to criticism. Moreover, we are unduly addicted to armchair theorizing, and are, for instance, in the habit of extolling a method of therapy which is little more than the creature of our contemplation and has never been subjected to the rigours of experimental trial. Why we should behave like this when the man who said, "Why think? Why not try the experiment?" was himself a surgeon, it is difficult to say. Surgeons started off as anatomists, then they became pathologists, so that thirty years ago it was still possible to quote the aphorism: "Know your anatomy, know your pathology (and you'll know your surgery)." To-day, surgeons are physiologists; to-morrow perhaps they will be scientists, and so, more than two hundred years later, arrive breathless at the summit reached with such easy mastery by John Hunter.

Imagine a surgeon dissatisfied with his present methods of treatment and conscientiously and hopefully thinking how he might improve them. He summons his anatomical, pathological, and physiological knowledge to his aid. "Placed there," he will say, "it will have a better blood supply." "Thus and thus the lymphatics will be more completely removed." "Done so, the acid will be less injurious; radiotherapy, if necessary, more efficacious; the anastomosis more secure; the respiratory excursions less impeded; the enzymes more fruitful"; and so a new operation is born. So far, so good, but in ten years there is found to be a snag and students ask plaintively, "What is this operation 'gastrorrhaphy,' mentioned so enthusiastically in my father's textbook?"

The Stochastic System

The pattern for research should be, first, the idea; second, and if suitable, its trial on the experimental animal; thirdly, its trial on human beings; and fourthly, its announcement in the surgical journals. In much surgical research the two middle steps are omitted. One of the reasons for this is that surgeons are unfamiliar

with the idea, which is a commonplace for physicists, of the stochastic system. An example will serve to show what is meant.

If I am completely uncertain which of two courses to adopt, I might toss a coin and act according to the dictates of what befell. I should regard this then as entirely a matter of chance, but if I knew all the facts this would not be the case. If I knew, for instance, the exact weight of the coin, its size, the strength of the impetus imparted to it, and the site of impaction in relation to its centre of gravity, the temperature and humidity of the atmosphere, the distribution of densities in the coin in relation to its centre of gravity, the distance it had to fall, and perhaps other factors such as the stickiness and temperature of my fingers, I should be able to calculate whether the coin would come down heads or tails. If, that is, I knew what are called the "hidden parameters," tossing a coin would no longer be a matter of chance. The spinning coin is not transgressing the principles of causality, and whether it comes down heads or tails is governed by strict law.

We may contrast such a system with that obtaining inside an unstable atom such as an atom of radium. Sooner or later this atom is going to discharge an emanation and become degraded to radon. On analogy with the above example we might argue that if we knew all the processes going on within the nucleus we could predict exactly when the atom of radium would disintegrate and that it is only the complexity and minuteness of these hidden parameters which prevents us from doing so. This, however, is not the case; no amount of knowledge, however intimate, of what is going on inside the nucleus would help us to determine whether the radium atom would disintegrate in the next ten seconds or in the next ten years. Such systems where the reactions are not individually determined according to the principles of causality are called stochastic systems.

Now, the human body has so many hidden parameters that to work out from the armchair how it is going to behave after a certain operation is a most uncertain enterprise. These hidden parameters are so multifarious that we may for practical purposes regard the human body as a stochastic system; indeed, if we admit the principle of free will, and if I understand him aright *pace* Spinoza's "metaphysical determinism," it is a stochastic system.

Trying Out

It is therefore the number of hidden parameters or even the possible stochastic nature of the system with which we are dealing which makes the application of cybernetic principles to the human organism so chancy.

*A lecture to the Medical Faculty delivered at Queen's College, Dundee, on October 30, 1958.

The trouble with advocating a course of action which in practice turns out, perhaps only in the long run, to be without value is that all the work has to be undone and a start made afresh with the consequent wasting of years of effort and the obstruction of real advance in the meanwhile. Let us take it as an axiom then that, however sound the idea might seem to be and however carefully checked in all its apparent ramifications, before it is publicly advocated it must be tried out so as to "see," as they say, "what happens in practice."

Now, the business of "trying out" is the very "bricks and mortar" of the edifice of clinical research, and while more and more flags are being put out from the upper stories of this building I would wish here to explore a few of the foundations, the very pillars upon which the structure of clinical research reposes. The tour which I would ask you to take with me in a very humble and elementary way is not unlike that taken on a grand scale by Bertrand Russell (1919) in respect of the far more complicated subject of mathematics. It was the self-appointed task of Russell and of his collaborator, Whitehead, to delve down rather than to soar on upwards, and by exploring the foundations of mathematics to lay bare its fundamental structure and so strengthen and accelerate the whole of mathematical thinking. It might be helpful if we tried to perform in miniature the same sort of analysis for clinical research.

It would seem to me that the structure of sound clinical research reposes on three pillars—control, mensuration, and analysis—and it is the purpose of this paper to examine these pillars in detail.

CONTROL

It might be regarded as an impertinence to include a description of this "pillar" in a paper addressed to postgraduates, but sad experience in the past fifty years shows that this is unfortunately not the case. Advances in at least two of the major branches of surgery have been bedevilled by the results of widely advertised research improperly controlled. The swings to and away from some much-canvassed methods of therapy, almost it would seem at the dictates of fashion, are, to say the least of it, inelegant in a so-called learned profession, and are really quite inexcusable departures from the demands of a scientific discipline.

Now, there are controls and controls, and it may not be practicable for the investigator dealing with human material to establish the rigid controls possible in the laboratory. Nevertheless, he should arrange for the most accurate method of control available. One common practice that cannot be too sternly condemned is to arrange and carry out a piece of clinical investigation and then when it is completed to look round in the literature or in the past records of the hospital for something with which to compare it. Such so-called "control" is from the scientific point of view worthless, as the two groups so contrasted are not comparable. For instance, to attempt to compare a group of cases treated for ulcerative colitis in 1957 with a group treated in 1937 takes no account of the innumerable advances, perhaps only in the management of incidentals, that have taken place, and which in sum may make all the difference between success and failure. In such a hypothetical instance one could cite improved antibiotics, better x-ray diagnosis, an appreciation of the Rh factor, a more complete understanding of electrolyte and fluid balance, change in the nutritional and environmental circumstances of the population, and so on almost *ad infinitum*. When stated in this way, it would seem almost ludicrous that anyone should dare in a scientific paper to make such comparisons and call them controls, and yet this is being done continually.

Another objectionable practice is to compare a series of cases treated in one locality with a series treated in another. Even if these are contemporaneous, the circumstances are so different, the skills and disciplines involved in making the diagnosis and in applying the routine methods of therapy are so variable, that, without taking into account the well-known habit of disease to behave differently under different climatic and geographical conditions, it should be abundantly clear that a comparison of two such groups has practically no scientific value.

It might be objected that humane considerations demand that a measure which appears promising should be tried out on any patient without having as a control a group which is denied the possible benefits which such a measure might confer. This is rarely a real difficulty in practice. If the measure is so effective that its prescription is vital, as in the instances of penicillin and insulin, then this effect is at once obvious. The group so treated display such a pronounced advantage over those that have just preceded them in the same institution and over those in other hospitals at the same time where the measure is as yet unavailable or unknown, that the most austere critic would never dream of denying to any of his patients the benefits which have been so clearly established.

Where, however, the choice is between one antibiotic and another, between one insulin preparation and another, or over the use of some drug or new operation which "looks as though it might be an improvement on all that has gone before," then it is far better that these measures should be tested scientifically with adequate control, so that ideas which have to be abandoned a few years later are not foisted on the profession, with all the frustration, suffering, and unnecessary expense which this entails.

In many instances, in determining the value of a highly promising method of treatment in an otherwise desperate or lethal disease the embarrassment of having a group of cases as a control may be avoided by using the patients as their own control. Thus, after a preliminary period of stabilization during which the essential investigations are undertaken, the remedy in question is applied and the pre- and post-treatment courses can properly be compared in a number of individuals (Fig. 1).

FIG. 1.—The patient as his own control.

What in fact is being urged is not the insistence on a rigidity of control only possible in all circumstances in laboratory work, nor that any patient's life should be jeopardized in the interests of science, nor that patients should be incommoded to any greater extent than is inseparable, for instance, from the needs of teaching, nor that any doctor should compromise with his conscience; but simply that he should bear in mind the necessity for controls if any real advance is to be made in medicine and that he should apply these with common sense whenever they are applicable.

Very few of the thousands of advances that are made each decade in medicine or surgery are of world-shaking import in the way that the discovery of penicillin or insulin might be regarded. Many of them are only modest improvements on what has gone before, and many new ideas which promise so much turn out to be less effective

than established methods. Within this framework, then, which comprises 99% of the field of clinical investigation, rigorous control causes no undue suffering, and is obligatory if we are to know the real value of the measure under discussion. As in navigation, so in clinical surgery, it is highly desirable to know where you started from, if possible, where you are going, and, within certain limits of error, where you are.

Randomization

If two groups are to be formed, one the experimental group and the other the control, then by far the best method of doing this is by randomization. It might be thought that to allot cases to each group alternately might be as good a way as any, but there are certain objections to this method. For instance, the whole series comprising both trial and control groups probably belongs to a certain class of case so that the investigator may have to decide whether any particular case conforms to the class or whether it should be discarded. This in itself may require considerable judgment, as for instance in deciding whether a case of hypertension would be suitable for surgery. If the method of dividing the series into the two groups were by taking alternate cases, then the investigator, if biased, and bias must always be assumed, might be tempted to exclude from the investigation altogether an unfavourable case if he was aware that the next in the series would automatically be placed in the category in favour of which he had a conscious or unconscious bias. Randomization should be undertaken only when the investigator himself, or another person, has decided that the case falls into the class. A card may then be drawn from a box originally containing a specified number of pairs of cards marked "treat" or "control" and the patient is allotted to the "treat" or "control" group, according to the instructions on the card.

Modifications of this method may be devised in order to exclude bias more completely. These methods include the "blind" investigations where the investigator is unaware of which method of treatment he is administering (a device clearly not applicable when comparing the effects of operation with conservative treatment), or he does not know when assessing the patient subsequently from which group he came. These elaborations must, nevertheless, be based on sound methods of randomization in the first instance. With all these precautions, however, loopholes through which bias may intrude still occur, particularly in small series. Thus if a group of six patients were being compared with a control group of six patients, and if, owing to the vagaries of chance, a run of identical cards were drawn early in the investigation, the increased chances of the opposite category being drawn towards the end of the series would probably be known to the investigator. This difficulty can be overcome in theory by keeping the investigator who decides whether the case is suitable in the first instance unaware of the way in which the draw has gone so far; but this is almost impossible in practice, particularly as this person may well be the one best qualified to carry out the subsequent treatment, and, in any event, it is an objection which hardly obtrudes in a big series until the very end, when it is unlikely materially to affect the issue.

A further refinement of this method of randomization may be introduced by arranging that sub-categories are paired together. Thus one box of paired cards may be made to apply only to those under the age of 40; another, to those aged 40 and over, or to any other selection of sub-categories which may lead to a more accurate pairing of like with like.

The Ethical Problem

At this stage we must inquire into the ethics of conducting an investigation involving the randomization of human patients. In the first instance, the rule is extremely simple, and is the fundamental one that if you would allow the choice of treatment to be made on a member of your own family or on yourself by the flip of a coin, then it is

quite ethical to subject a group of patients to a randomized choice in this way. Certain outstanding problems in surgery fit quite comfortably into this category. Thus an investigation into the effect of lumbar sympathectomy on the incidence of gangrene in claudicants, the effect of surgery in a certain class of hypertension, or the relative merits of gastrectomy and gastrojejunostomy in certain types of duodenal ulcer could well be investigated by randomization without disturbance to the conscience.

This matter of conscience, however, can really be overdone. For instance, in trying to organize an investigation into the relative merits of radical mastectomy and the "Edinburgh method" of treatment of early cases of cancer of the breast, it was found that the advocates of radical mastectomy could not be persuaded to subject a control group to the Edinburgh method, because their consciences would not permit them to allow their patients to be subjected to such an inferior measure; the advocates of the Edinburgh school felt the same about radical mastectomy. Nevertheless, it should be possible to find a group of people who are in genuine doubt about this problem and who could therefore properly undertake an investigation of this kind.

The ethical problem becomes more acute as the investigation proceeds and it appears that one method of treatment is showing a clear advantage over the other. The rule here is just as definite, but is not at once so manifest as when the investigation is first promulgated, and the rule is that if it was ethical to start the investigation in the first place it would be unethical to desist until a clear-cut answer has been obtained. This might conflict with considerations depending upon what one would wish for one's own family or oneself, but this should not be so. If a significant result has been obtained, then the investigation is concluded and the difficulty does not arise. If it is simply an impression that one method of treatment is better than the other, then the feeling that you would hope that a member of your own family would, if subjected to the test, draw one card rather than another is little more than a gambler's hunch. To break up the investigation before a statistically valid result has been obtained would be to sacrifice all the value of the investigation hitherto and would be quite irresponsible. If the investigation was a proper one in the first place, it should be continued until at least a 5% level of probability has been achieved, or it has been decided that the two methods of treatment are so nearly comparable that no purpose would be served by prolonging the investigation, in which case the ethical problem does not arise.

For a discussion of the problem of when to put a stop to an indeterminate investigation the reader is referred to Finney's (1955) book, where this matter is fully analysed.

MENSURATION

Science depends upon measurement, and qualities that can be measured numerically are more readily comparable than those which cannot. If the marking is accurate and you get six marks in your examination and I get five, then you have done better than I. Even if the marking is not quite accurate, the school which averages six marks in the examination has done better than the school which averages five, and the larger the number of candidates the more accurately does this average reflect the difference between the two schools. If, therefore, we are comparing two methods of treatment it is most helpful if the results of the treatments can be compared numerically, by marks if you will. This method of expressing the results has the advantage that it is definite, and, moreover, it generally gives an idea of the degree of difference between two treatments. For instance, if the results of treatment can be expressed in terms of the death rate per thousand or the period of survival, then if, on the average, patients under treatment A survive five months and patients under treatment B survive eight months, and provided sufficient numbers have been taken, it is quite clear that treatment B

is to be preferred to treatment A. Furthermore, the figures give some indication of how much better the one treatment is than the other. There are, however, other equally important advantages of expressing results numerically.

Supposing the relation of treatment to a variable is being considered, then if the results of treatment can be expressed numerically a "scattergram" can be constructed and the relation, if any, established on inspection or by calculation. Fig. 2 shows a hypothetical investigation to determine whether the success of a certain treatment is dependent on the initial weight of the patient. The results of treatment are expressed as time in days before the temperature settles, and the scattergram shows, in a way more clearly than any amount of verbal description would have done, that there is such a relation and that fat people respond better to treatment than thin. For more accurate work an exact coefficient of correlation could have been calculated. Finally, where a numerical assessment of treatment can be

FIG. 2.—Scattergram relating weight of patient to the period of pyrexia.

provided the whole investigation is susceptible to statistical analysis, and extravagant claims which are really quite unwarrantable are not put forward.

When there are all these advantages of precise mensuration it is a pity that surgeons as a whole dislike figures and that this dislike is often so ingrained that they do the most illogical things with them if they have to use them at all, as, for instance, in marking examinations. In nearly all surgical examinations there is a desire to give more weight to one part of the examination than to others. It is usually "the clinical" which is singled out for this special attention. In the Treasurer's Gold Medal in Clinical Surgery at Guy's Hospital, for instance, it has been the practice from time immemorial to mark the papers out of 200 and the clinical out of 300. This of course has no effect at all, because as often as not candidate A, who is slightly better than candidate B in the paper, will get 155 marks to B's 150, and, being slightly worse in the clinical, will get 250 to B's 255, so that their total marks are the same and nothing has been achieved. What is clearly needed in order to add weight to one part of an examination is to increase the range of marking in this part of the examination and never mind about the maximum. The fact that this anomaly is by no means confined to Guy's Hospital, but invades Queen Square, whether it be under the aegis of the Royal Colleges or London University, and is in evidence in the examinations of many provincial universities as well, only serves to show how inexpertly medical men can behave when dealing with figures.

Another type of illogical error is exemplified in *Against the Law* by Peter Wildeblood (1957). There occurs there the following statement, with the so-called evidence which is intended to support it.

"It is, of course, perfectly obvious that only a very small proportion of this 650,000 is noticeably odd. It is equally evident that homosexuality is not, as some propagandists would have us believe, a kind of fashionable vice restricted to decadent intellectuals and degenerate clergymen. This idea is extremely useful to those who wish to influence public opinion against homosexuality, and it succeeds because the British people, on the whole, despise intellectuals and distrust the clergy. But it does not happen to be correct.

In a recent survey of 321 court cases involving homosexuality, it was found that the accused men fell into the categories shown in Table I."

TABLE I.—Categories of Homosexuals (Peter Wildeblood, 1957)

Shop and clerical workers	16%	Schoolmasters	4%
Artisans (factory workers)	15%	Agricultural workers	4%
Transport and Post Office	11%	Clergymen	2%
Unskilled labourers	10%	Mentally deficient	2%
Armed Forces	10%	Independent means	2%
Hotel and domestic servants	7%	Unclassified	11%
Students, trainees, schoolboys	6%		

This table, of course, means nothing unless we are at the same time given the percentages of "shop and clerical workers," "artisans," and so forth in the general population. Indeed, it is not inconsistent with the possibility that all members of the above categories are homosexuals.

Incommensurable Qualities

This inability to think clearly in figures is often rationalized by claiming that so many of the results of clinical experiment cannot be satisfactorily expressed numerically. Some qualities of human behaviour are, of course, quite impossible to measure—such qualities, for instance, as bravery or fastidiousness; but this does not mean that inquiry related to these qualities *cannot* be given a numerical expression. For instance, if the 1st Battalion of the East Loamshires won 3 Victoria Crosses and 2 Military Crosses in the last war, whereas the 2nd Battalion won no medals and had a sergeant and four other ranks shot for cowardice in face of the enemy, there is good reason on purely numerical grounds for presuming that the 1st Battalion had a better morale than the 2nd Battalion, although morale is essentially a quality which cannot be measured.

Clearly a passion for mensuration can be overdone, as the following charming quotation from R. E. Dickinson (1958) attests: "Lord Kelvin said that we can know nothing of a matter unless we can measure it. Psychologists braced with their success in giving numerical values to I.Q. are now hoping to measure feminine beauty in precise units. The unit proposed is the milli-helen, the quantity of beauty required to launch exactly one ship."

Such endeavours to express essentially incommensurable qualities by numbers must of course be undertaken with great care. Often, however, if it is reasonable to adopt it, this method shows up in a vivid manner differences and relations which might otherwise be obscured in a mass of verbiage. A particularly valuable method of mensuration devised originally by Walpole and Paterson (1949) is the mean clinical value. This method of giving a numerical value to the effects of treatment is applicable especially to those conditions, such as advanced cancer of the breast, where there are apt to be many manifestations and where these manifestations may respond in a different manner to treatment or respond at a different time.

The working of the system is very simple, and depends upon deciding whether, as a result of treatment, a lesion is better than it was before treatment began or worse than it was before treatment began, or whether the situation is doubtful.

Mean Clinical Value

Before treatment all lesions are accurately assessed—by measurement if they are palpable or visible, and by studying their dimensions or densities if they are radiological. Subjective symptoms are better ignored, but the "presence" or "absence" of pain may well be a valuable index of the effect of treatment. In this way a baseline is made and the treatment is started. The patient is seen at four-weekly intervals thereafter and each lesion is again assessed. If the lesion has improved it is given 2 marks, if worse 0 marks, and if stationary or doubtful, 1 mark. These marks are then added together and divided by the number of lesions to obtain an average. The average is multiplied by the factor 6 to bring the figure to a manageable number, which is the mean clinical value (M.C.V.). It will be seen that the maximum M.C.V. is 12, which implies that all the

lesions are improving. An M.C.V. of 0 means that all the lesions are worsening, and an M.C.V. of 6 that the condition is stationary. This last mark may be due to the fact that all the lesions are stationary, or that half are improving while the other half are regressing, and it indicates, therefore, the response of the patient as a whole. The M.C.V. may have any figure between 0 and 12, and a graph can be drawn of the progress of a patient under treatment. The state of the lesions is compared with their state before treatment began and not with their state at the previous visit, because it is more satisfactory to have one standard than to have an ever-changing standard which is apt to lead to swings on the graph that have no real meaning.

The effect of this, however, is that, when a patient who has for some time responded to treatment begins to deteriorate, this deterioration is not at once reflected in the M.C.V. because the lesions may still be better than they were before treatment began and often there is a somewhat precipitate drop in the graph before death. This is, in practice, only a very slight disadvantage, because, certainly in the case of cancer of the breast, when a patient begins to deteriorate she does so very quickly, and the delay in the drop in the graph from this cause is generally only a matter of one period or at the most two periods. As has been indicated, if there is any uncertainty over whether a lesion is changing, then it is given a mark of 1. By the next visit this uncertainty has usually been resolved, and a lesion which looked as though it might have improved but about which there was some doubt has now unequivocally improved and can confidently be given a mark of 2. All that has occurred, therefore, as a result of the possibly mistaken assessment of the lesion at the visit when it was given the "uncertain" mark of 1 is that the shift in the graph has been delayed by four weeks. Conversely, the same applies to a lesion which seems to be deteriorating but about which there is some doubt, and the whole graph is therefore given a sure and steady delineation which properly represents the course of the patient under treatment, albeit with a slight latent period or what the cyberneticist might call "a threshold-controlled stability."

Radiological factors are assessed at every third visit in order not to multiply the number of x-ray films taken, and the previous mark allotted to this factor is retained in the interval to compute the M.C.V. for those two periods. This again leads to some lag, as the x-ray picture may have changed in the meantime and this is not reflected in the graph until later.

With all these slight and largely theoretical defects, the M.C.V. has proved in practice to be a tool of the greatest value in research. Thus, in comparing two methods of treatment, as for instance adrenalectomy and hypophysectomy, for carcinoma of the breast, the individual M.C.V. charts of all patients treated by one of these methods can be congealed into one average graph and compared with a similar average graph for the other method of treatment (Fig. 3).

FIG. 3.—Response to adrenalectomy and hypophysectomy, taking the average M.C.V.s of all patients at four-weekly periods after treatment.

Again, by taking an average of all the M.C.V. records of a particular patient, a single figure can be made to represent the overall response of that patient to the treatment, and a table (Table II) can be made out which gives a good idea of the effect of this measure as a whole. By the use of

TABLE II.—Effect of Oestrogen on 94 Cases of Advanced Cancer of Breast, as Expressed by Average M.C.V. Per Case—i.e., 17 Cases Had an Average M.C.V. Between 8 and 12, 39 Cases Between 7 and 4, etc.

Average M.C.V. ...	12-8	7-4	3-0	Unable to take drug
No. of cases ..	17	39	31	7

the "average M.C.V." derived in this way it is also possible to construct scattergrams (Fig. 4) and so determine a relation or lack of relation between the success of treatment and such factors as age, mitotic-cell count, or what you will.

There is one flaw in the M.C.V. system so far practised that is a serious theoretical disadvantage, but which has been so unobtrusive in practice that its full implications have only recently been appreciated.

If we consider a case, for instance, of cancer of the breast with multiple skeletal deposits and multiple skin lesions and in which there are so many lesions that it would be impossible to exploit them all, then it has been the practice to decide arbitrarily how many skeletal lesions and how many skin lesions

FIG. 4.—Scattergram relating M.C.V. to resting-cell count.

should be taken as representatives to act as factors in assessing the M.C.V. Usually the obvious ones or the readily measurable ones have been taken and the others have been ignored. If all the lesions vary in the same way with treatment, then it does not matter very much how many are taken as indices. If, on the other hand, one system improves while the other deteriorates, then it is clearly important how many factors have been taken in each system. For instance, to take an extreme and very unlikely case, if two factors have been taken in the skeletal system which is deteriorating and ten factors have been taken in the skin which is improving, the M.C.V. is 10; whereas if ten factors were taken in the skeletal system and two in the skin, the M.C.V. would be 2, which makes nonsense of the whole method.

In order to repair the flaw in such unusual cases J. Hayward has suggested the following modification of the system which is now being put into operation. The lesions in each system are grouped and an average is taken of these; the mean of the grouped figures is then taken to compute the M.C.V. Thus, in the example cited above, the M.C.V. would be calculated from the fact that the skeletal system scored 0 and the cutaneous system 2, so that the M.C.V. would be $\frac{0+2}{2} \times 6 = 6$, however many or however

few lesions were taken from either system. There is no method of computation which takes care of the case—one perhaps out of several hundred—where there are so many lesions in one system that a selection has to be made of representative factors, but where the lesions common to one system are behaving in a different way under treatment.

Stress has been laid on the defects of a system which in practice runs very smoothly. If a visitor to the breast clinic at Guy's Hospital, where this system has been used now for many years, is asked to give a mark—somewhere between

0 and 12—to the effects of treatment, and basing his opinion on a general clinical assessment, without recourse to exact measurement, it is surprising how closely his guess coincides with the computed value. To state that a patient has an M.C.V. of 9 is a quicker, more useful, and more vivid way of expressing the conclusion that this patient is responding in some degree to treatment and that it is worth while persisting with this treatment, but that not all the lesions are reacting favourably. Furthermore, the alternative is to fall back on the dangerous use of "clinical impressions" which has been productive of such fallacious conclusions in the past.

Recording Measured Data

If we now turn to another aspect of mensuration—namely, that of recording measured data—some consideration must be given to the punch-card system. Smyth and Bainbridge (1957a) have introduced an extremely useful system for tabulating punch-card records which enables the interrelation of a number of factors to be studied readily. The article in which this method was described concerned the interplay of a number of obstetric factors such as the relation between stillbirths and post-maturity, forceps delivery and the duration of labour, and so forth. It included a specimen table consisting in fact of 32 variables, and was so alarming at first sight that it probably dissuaded many readers, who would otherwise have found the system extremely valuable, from pursuing the matter further.

Table III is extracted from their more elaborate table, and for simplicity shows the interrelation of only six of their factors. To quote, with certain modifications, from the paper by Smyth and Bainbridge, this table is constructed as follows:

"The principal cells on the *main diagonal* show the numerical incidence (upper figures) and percentages (lower figures) of 6 factors. The cells in the *bottom left-hand half* show the numerical incidence of patients having the two factors named in the intersecting row and column. (Where these factors are mutually exclusive, as in 'Breech' and 'Occipito-posterior,' the cell is left blank.) The cells in the *top right-hand half* contain two percentages—that is, the number of patients having both of the appropriate row and column factors expressed as a percentage of the total incidence of (1) the *column* factor (upper figures) and (2) the *row* factor (lower figures)."

"To See if these Percentages Differ Significantly From the Means.—Taking the lower percentage only, subtract it from the percentage in the principal cell on the same *row* (or vice versa); compare the answer with the figures in columns B to E on the same row; observe how many patients are needed for this difference to be significant; lastly, examine the principal cell in the same *column* to see if that number of patients occurs.

"Column A shows the minimum number of patients required in the *column* principal cell for the test to be applicable to that row."

Thus all the statistical analysis is worked out beforehand and all the clinical investigator has to do to see whether his results are significant is to use a table no more complicated,

as the authors point out, than a mileage chart in the *A.A. Handbook*. For instance, by following these rules it will be found that manual removal will be associated with breech delivery in a highly significant fashion, whereas relative disproportion and toxæmia are not so associated.

The construction of columns A, B, C, D, and E is a simple statistical exercise which a friendly colleague with an elementary knowledge of these matters may undertake in a very short time by reference to the instructions contained in the original article and subsequent correction (Smyth and Bainbridge, 1957b). The figures given in columns B to E are simply twice the standard deviation.

To help in using this method, Garside (1958) has published a table for ascertaining whether differences between percentages are statistically significant, and the result can be picked out by little more than simple inspection.

Another technique for presenting results, which may be very valuable where three variables are involved, and which is much neglected by workers in clinical fields, is the three-dimensional graph. A good example of this taken from a paper by Skoog (1957) is shown in Fig. 5, where the effect

FIG. 5.—A specimen three-dimensional graph relating growth of cells in tissue culture to the concentration of kinetin in the medium. Reproduced by permission of the Editor of the *Journal of the National Cancer Institute*.

of kinetin in different concentrations on the growth rate of cells in tissue culture is beautifully shown in a manner which graphically reveals the relevant points. Having inscribed these enrichments on the capital of the second pillar (I trust they will not be regarded as excrescences), we must now proceed to the third.

ANALYSIS

By analysis we mean, of course, statistical analysis. There are those good practical men with no nonsense about them, for whom the word "statistics" is anathema, who quote

TABLE III.—Method of Recording of Dependent and Independent Variables from Smyth and Bainbridge (1957a)

	Forceps	Manual Removal	Toxaemia	Relative Disproportion	Breech	Occipito-posterior	A No. of Patients required for Valid Result	Significant Difference at 5% Level Patients			
								1,000	500	100	10
Forceps	213 9.16	6.6 19	7.5 6.8	15.9 53	13.6 63	11.7 42	54	1.38	2.29	5.64	18.2
Manual removal ..	14	75 3.23	12 3.8	1 2	8 13	3 3	158	0.84	1.40	3.46	11.15
Toxaemia	16	9	234 10.06	4.3 16	1.3 6.5	2.6 10	50	1.44	2.38	5.89	18.98
Relative disproportion ..	34	1	10	64 2.75	5 7	17 18	180	0.78	1.30	3.20	10.32
Breech	29	6	3	3	46 1.98		250	0.66	1.10	2.73	8.79
Occipito-posterior ..	25	2	6	11		60 2.58	194	0.76	1.26	3.10	10.00

the tag that "statistics can prove anything." Unfortunately, statistics have got to be used in nearly all clinical research, and it is better that they should be good statistics than bad. There is no need for the research worker in clinical medicine to have a knowledge of statistics, but it is an enormous advantage if he has, even though his knowledge of this subject is not very profound. To be able to do these things oneself, however, adds immensely to the interest and value of clinical research. Nevertheless, for the amateur, it would be wise to have the results checked by an expert, because it is fatally easy to make a gross error in a field in which one has not been bred.

Finney (1956) has said: "That a high proportion of those whom I have broadly classed as experimenters should have a knowledge and appreciation of statistical principles is of great value to the progress of their science, but perhaps not so much because they are thereby enabled to do their own statistical work as because they understand better what needs to be done and what conclusions can reasonably be drawn."

Statistics have two roles in clinical research: they act either as a brake or as a headlamp. In the first role they prevent the experimenter from claiming as a consequence a result which may have emerged entirely due to chance—they prevent him, that is, from making a fool of himself. The most mistrustful cannot but applaud such an effect of a study which on the whole they view with suspicion.

Sometimes, however, statistics act like a headlamp and reveal a truth which could not have been apparent without such an analysis. When a result arrived at by a statistical analysis conflicts with common sense, then the statistics are probably wrong; when it conflicts with emotion, as in the relation between cigarette smoking and lung cancer, then this is not necessarily the case.

It must be admitted, however, that many research workers in the clinical field whose contributions are of great value have such an antipathy to manipulating the relatively simple formulae which the practice of statistical analysis demands that they must rely entirely on a professional statistician to do this for them. To get the best out of their results, however, they would be wise to plan their investigation in the first place in consultation with a statistician, one of whose qualifications is that he has mastered the science of the design of experiments. In this way material will not be wasted, and expense and labour will be saved.

"Analysis of Contaminants"

Granted all this, there is still one aspect of analysis which must be done by the investigator himself, and it is this operation which I have termed the "analysis of contaminants."

It must be presumed that in any investigation great care has been taken to avoid sources of error, but, however careful one may be, these are likely to creep in, and, having analysed the results, it is now necessary to go back over the sources of error and see how these may have influenced the conclusions. In describing a system for an analysis of contaminants or sources of error, I have discussed elsewhere (Atkins, 1957) the reasons for doing this at this stage, but it will be necessary to recapitulate some of these reasons here, because this part of the work can be done only by the investigator himself and cannot be left to the statistician.

If, for instance, we are investigating the effect of pregnancy on breast cancer in some primitive community, we would compare a series of women who had been pregnant with a series who had never been pregnant, in relation to the incidence of cancer. We will suppose that it is found that the incidence of breast cancer is higher in the "non-pregnant" series. In such an investigation three contaminants might be considered. The first arises because unmarried women might fail to reveal that they had been pregnant. This means that the allegedly "non-pregnant" series may have been salted, if that is the right word to use, with women who had in fact been pregnant. The effect of this contaminant would be to make the two

series more alike than they should be; it would tend to favour a negative result to the investigation, and on that account it is referred to as a *negative* contaminant.

Furthermore, we might suppose, in a primitive community for instance, and indeed in this country up to a few years ago, that the possession of children ensured that the mother would be supported in her old age and would be more likely to live long enough to develop cancer of the breast. This contaminant, tending to lessen the positive difference between the "non-pregnant" and the "pregnant" series in relation to cancer, would also be a *negative* contaminant.

On the other hand, the fact that pregnant women are more likely to have suckled, and assuming, for the sake of argument, that suckling protected a woman from cancer of the breast, this contaminant would increase the difference between the two series and would on that account exaggerate the positive difference, and is therefore known as a *positive* contaminant.

Thus, a contaminant which tends to increase the greater or to decrease the less is a *positive* contaminant, and one which tends to do the reverse is a *negative* contaminant.

In the investigation which we have been considering we have assumed that analysis has revealed a *positive* result—namely, a significant difference between the two groups. It will be observed that the *negative* contaminants have tended to mask this significance, whereas the *positive* contaminant has tended to exaggerate it. If, on the other hand, an analysis had failed to reveal any difference between the two groups and that a *negative* result to the investigation was implied, then the *negative* contaminants would have strengthened this view and the *positive* contaminant would have weakened it. From this we may derive the general rule that contaminants bearing the opposite sign to the conclusion tend to conceal the conclusion and are, on that account, known as *inhibitory* contaminants; those bearing the same sign as the conclusion tend to exaggerate its significance and are on that account known as *illusive* contaminants. For a full discussion of this matter the reader is referred to the article quoted above, but it will be seen at once that the *illusive* contaminants are the treacherous ones, as they tend to make the investigator think that his results are more significant than they really are.

A paper giving an account of a clinical investigation, from which sources of error of this sort are inseparable, should contain—following the analysis—a formal statement of the *inhibitory* and *illusive* contaminants together with those which it is assumed do not affect the issue one way or the other and are known as *neutral* contaminants. In this manner the reader is able to assess the degree of confidence which can be placed in the investigation as a whole.

We have now completed our tour of the foundations of this edifice—clinical research. To establish these three pillars firmly on a bed of rock and to build on them requires a little extra thought and a great deal of additional careful work, but I believe in the end it will save future generations much labour and will prove far more enduring than the erection of a papier mâché façade to house some colourful exhibition on a foundation of sand.

I am grateful to Sir Charles Symonds for his helpful criticism.

REFERENCES

- Atkins, H. J. B. (1957). *Guy's Hosp. Rep.*, 106, 251.
- Dickinson, R. E. (1958). A letter in *The Observer*, February 23, 1958.
- Finney, D. J. (1955). *Experimental Design and its Statistical Basis*, p. 117. Cambridge Univ. Press, London.
- (1956). *J. roy. statist. Soc.*, Ser. A, 119, 1.
- Garside, R. F. (1958). *Brit. med. J.*, 1, 1459.
- Russell, Bertrand (1919). *Introduction to Mathematical Philosophy*. Allen and Unwin, London.
- Skoog, F. (1957). *J. nat. Cancer Inst.*, 19, 578.
- Smyth, C. N., and Bainbridge, M. N. (1957a). *Brit. med. J.*, 2, 619.
- (1957b). *Ibid.*, 2, 1437.
- Walpole, A. L., and Paterson, E. (1949). *Lancet*, 2, 783.
- Wildeblood, Peter (1957). *Against the Law*, p. 29. Penguin Books, Harmondsworth.