FACULTY OF MEDICINE

UNIVERSITY COLLEGE OF RHODESIA

Gambling Aspects of Medicine

by

WIN CASTLE

Occasional Paper
No. 1. 1969

Faculty of Medicine

Occasional Papers No. 1

GAMBLING ASPECTS OF MEDICINE

by

Dr. W. M. CASTLE

M.B., B.S. (Lond.), M.R.C.S., L.R.C.P., F.S.S., A.I.S.

Lecturer in the Department of Social Medicine

University College of Rhodesia

SALISBURY
UNIVERSITY COLLEGE OF RHODESIA
1969

REPRINTED BY A. W. BARDWELL & CO. (PVT.) LTD.
from the
Central African Journal of Medicine, Vol. 15, 1969.

"In preparation for this coming generation much of great literature will have to be rewritten, else it will be incomprehensible.

Always willing to be in the forefront of any new movement, I have made a start on Julius Caesar:

Computers which have captured from the stars
Their role of prophecy, discern a link
'Twixt body-weight and loyalty, such that p
Is meaningful at less than point-nought-five.
Thus would I rather have about me, men
More to the right of distribution curves
Than yon lean Cassius. He thinks too much.
Statistic'ly such men are dangerous."

—*Lord Platt (with permission).*

Contents

Foreword

The ability to measure is part and parcel of modern science. The art of medicine has perhaps been slow to realise this. but within the last 40 years there has been rapid development in this field. Many medical students and medical practitioners have been puzzled by the "rash of statistics" which have appeared in medical journals. My colleague, Dr. W. M. Castle, in writing these short articles on "The Gambling Aspects of Medicine," has done much to simplify what to many has been a puzzling subject. There must be few who, having read this book, will not be more able to understand and criticise intelligently medical work, their own and that of others.

W. FRASER ROSS.

Salisbury.

GAMBLING ASPECTS OF MEDICINE

I. The Course Steward

A friend of mine with much insight once said that the medical profession ought to produce excellent punters. The qualified doctor, having spent six years learning how to follow up clinical hunches, then spends a lifetime developing the art of gambling without being too open about it.

I am sure Sir Derrick Dunlop would agree that physicians gamble every time they prescribe a drug. They believe it will do more good than harm. Surgeons talk openly about "operation risks," not only with their medical colleagues, but actually with the patients and their relatives. "Intervention" is a gambling term introduced by obstetricians. It means that the odds are stacked against a normal delivery and therefore they must decide to back either their knives, their forks or their vacuum extractors.

Off the wards the radiologists, convinced of their clinical acumen, say, "This X-ray shows neither typically Y nor Z, but . . ." and then go on to a conclusion anyway. Pathologists assume they *know* all the answers (provided the rest of us supply them with adequate specimens and reliable history!). Guesswork is unknown to them, but force them into a diagnostic corner and they must stake their shirts like the rest of us.

Outside the hospital the general practitioners gamble all day. Are these symptoms psychological? Is this girl pregnant? Is this new drug any better than its cheaper counterpart? Can I get away without prescribing an antibiotic, and if not, is a sensitivity plate really essential? Answering these speculative questions all day is liable to make a good general practitioner a very experienced punter. Meanwhile, swivelling in his chair, the medical administrator gambles with priorities. (Fortunately for him, with other people's money!) The *nganga* openly and unashamedly toss their bones. Are we not as a profession merely adding touches of sophistication to the same toss-the-coin principles?

Certainly we do not always gamble in the dark. The academic doctors, with their limited number of beds, have the time to work out some of the odds. Sitting astride their clinical hobby-horses like jockeys, they scurry to push their noses through the winning post (the medical journals) before anybody else in their particular field. Yes, spurred on by the drug companies, they quote the odds to the general medical punters who are nearly always willing to listen and stake their bets.

The medical profession is a noble one. It does not, nor should it, let this clinical race meeting proceed without supervision. There is a course steward encouraging the academics to produce reliable odds on the one hand and protecting the punter from foul play on the other. Although there are two sides to his task, they are closely interrelated. The principles involved in enforcing fair play into clinical trials are the same principles involved in protecting the public from recognising the ploys used by chancers. To know how to barge is the first step towards realising when you are being barged. The gentleman here introduced, the course steward, is without his racing guise, the medical statistician.

As a practising doctor you may be about to decide that any articles about statistics are not for you. If in your particular branch of medicine you never take chances and never read medical journals, then these short chats are indeed not for you. Similarly, if you are a capable academic steering your clinical hunches from a good start over a fair course to an honest result unaided, then sidestep these articles. These articles are written to help the doctor in the street with his medical punting. If you are a doctor who hates figures, who skips the articles in the journals as soon as statistics are involved, especially if you feel you may be missing out, I direct these chats to you. They are intended, like a "with-it" mackintosh, partly as protection, but also I hope with some measure of stimulation.

Numbers are put to good use and bad use in medicine. With previous generations the distinction was usually irrelevant. Discoveries of the uses of insulin, B_{12} and sulphonamides did not need statistics to underline their worth. Professor Barnard and the other jockeys in his particular field do not particularly need a statistician at the moment nor, fortunately, do I require the services of Professor Barnard! General medical research is not leaping forward with such obvious strides as in the field of spare part surgery. Knowledge is usually inched forward by careful comparisons between different drugs or surgical methods. Professor Barnard may have made a great surgical stride, but the best technique and immunosuppressive regime will be discovered only after numerical analysis.

Statistics is, therefore, a necessary discipline in medicine today. Unfortunately as a profession we have a bad name in this field so far as other scientific professions are concerned. In my view much of this criticism is ill-deserved. Agriculturalists and biologists can move cows from pasture to pasture and mate hamsters vir-

tually at will. In contrast, doctors are often lucky to get a hospital bed at all, let alone borrow somebody elses, even though it may satisfy worthwhile research. Medical ethics cannot be made the servant of science, and so many research doctors' hands are tied in a way their critics do not appreciate.

However, there are three ways in which we could improve our professional front so far as statistics are concerned. Firstly, some doctors *refuse* to know anything about the subject. Admittedly many books with their heavy arithmetic and algebra, not to mention the Greek symbols, leave many doctors unresponsive even if they are keen to understand the principles. I will avoid falling into this trap in these articles. The attitude of the few who put up the shutters against statistics without trying to understand is indefensible. It is as unethical to prescribe a drug recommended using faulty evidence as it is to submit patients to clinical trials without their consent.

There is more to be said for those who only put up their statistical shutters than for their counterparts, who then from the security of their closed shops go on to hurl abuse at the subject. Medical statisticians, contrary to popular opinion, do not sit along the road of medical progress poking spokes in every passing wheel. As other sciences have advanced, so have the mathematical aspects of medicine. The doctor who thinks current statistics always demand large scale, expensive and completely unpracticable experiments is out of date.

Doctors can finally improve their statistical image by seeking advice *before* they start a clinical trial if they suspect their ability to see the experiment through alone. Medical information is often too expensive and time-consuming to waste. Although statistics can now go a fair way towards fitting in with the practical aspects of clinical trials, the method of collecting the data can make or break it as far as analysis is concerned. Unfortunately very accurate information is no better than shoddy experimentation if it cannot be used. Every statistician I know would perform willingly the arithmetical contortions for a doctor who had taken care to see that the medical and mathematical aspects had been carefully sutured initially.

In conclusion, it seems that as a profession we do take chances and calculated risks. Because this is so there is a definite place in most aspects of medicine for the science of odds or statistics. Therefore the modern doctor must be ready to accept an understanding of the principles of medical gambling. At best, I hope these eight articles make the uses of numbers in medicine more interesting; at worst, I hope they protect you from their misuse.

II. The Jargon

WHAT THE DOCTOR SAW

Statistics in Medicine are rather like slimming diets—a nuisance at the time but not without their long-term benefits.

To be effective a slimming diet needs reasonableness on the part of the designer and willingness to co-operate on the part of the over-endowed. Similarly, a successful medico-statistical allegiance requires reasonableness on the part of the statistician and willingness on the part of the doctors. I have always found that slimming diets (and I have considered many!) either demand that the victim carry a scale everywhere or limit the victim to "acceptable helpings" of certain specified foodstuffs. The first idea I abandon as completely impracticable, the second approach I find never works. There is a breakdown in communication, as my interpretation of an "acceptable helping" obviously differs from that intended.

Medico-statistical jargon often means different things to doctors and statisticians. It is true that words like "bias" and "significance" convey the same broad ideas to both groups, but the shades of meaning and implications are important and cause confusion. Politically, terms like "apartheid", "communism" and "one-man-one-vote" can be defined by all shades of political opinion. Although the definition of, say, "one-man-one-vote" is obvious, the implications can range from moral perfection to economic chaos. The interpretation of these terms is as vital as are their definitions. Even when conversing in the same language I wonder whether politicians with different viewpoints really understand what the other side is talking about. I think the difficulty is as acute in medical statistics and as great a need exists for a medico-arithmetical settlement.

The implications of words like "significance" have to be conveyed without misunderstanding between research doctor and statistician. From here the proper meaning must be transmitted via the journals to you, the practising doctor. Each link in the chain must preserve the true message. Unfortunately, "send reinforcements we're going to advance" intended at the beginning of the chain is all too often interpreted eventually as "send 3/4d. we're going to the dance."

I shall avoid discussing stodgy definitions and chat about their more interesting implications for after all, articles in the journals are trying to convey a message rather than test your knowledge. The articles generally fall into two broad categories. Firstly, there is the "anecdotal type"—for example, a rare case of chronic Smith-Wilson's syndrome reported in two Africans in Enkeldoorn. This type of article causes little trouble and is generally very interesting. The second involves reports on series of cases and potentially causes confusion originating from three main sources; the actual results, the method of their collection and the conclusions drawn from them. Each of these sources will be highlighted in turn in these articles, starting this month with the actual results.

The *mean* or average and the *standard deviation* play innocent and prominent parts—like virgins in a promiscuous society. They sit rather smugly in print in their various disguises (such as $\bar{x}$ for the mean and s or σ for the standard deviation) knowing that they meet with the full approval of the editors of the journals. They serve different purposes but both are very suitable for use in subsequent statistical tests. Imagine a drunk in casualty walking along a straight line. The line lies at the middle of his staggers and is like the mean in the middle of the results. The average amount the drunk deviates from his line is the standard deviation. They both hold a key to the test—too far from the mean as measured by the standard deviation—and you're in trouble. These characters give an article some tone and apart from their frightening appearance cause little trouble.

Ratios, proportions and *percentages* are more treacherous, although more sinned against than sinning. Occasionally, they are mistaken for each other—for example, one may read that the ratio of A to B is 10 per cent. instead of 1:9. This disputed paternity should never reach the journals. I read the other day that "it was particularly interesting to see that as the proportion of a particular type of egg increased, the proportion of the others decreased." Like salary cheques, if the *proportion* of income tax rises, the *proportion* of earnings for spending *must* drop. The egg finding is inevitable and not " particularly interesting" as quoted.

The next treachery occurs when they are inadvertently used for comparative purposes. Imagine that a gynaecologist notices in his wards that over a period of time the proportion of admitted cases of cancer of the body of the uterus relative to all cases of cancer of the uterus has increased. He wonders why, and decides to investigate his admitted cases of cancer of the body of the uterus in much greater depth. Is this right? Unfortunately, he has forgotten that because he has diagnosed some cases of cancer of the cervix very early by employing Papanicolaou

smears, he has decreased his number of hospital admissions with this diagnosis. Hence, because the proportion of admitted cervical cancers are decreasing the proportion of cancers of the body of the uterus are bound to increase from this cause alone. So he could think of a very likely answer without requiring to study his cancer of the uterus at all. An increase or decrease in any proportion may well be important, but the change in the other proportions is perhaps the more valuable finding.

The more reliable relation of the ratio-proportion-percentage family is the *rate*. If the gynaecologist had noticed that the rate of cancer of the body of the uterus was increasing, he would have been off on a sounder footing. By rate, of course, I mean

$$\frac{\text{The number of cases of cancer of the body of the uterus}}{\text{Relevant number of women at risk.}}$$

Even so, rates are rather like wolves in sheep's clothing. They are not beyond reproach when used for comparisons. For example, I would drop the *Crude Mortality Rate* from my circle of friends. The *Pocket Oxford Dictionary* defines crude as "lacking finish" and "rude", but the word seems to have little "offputting" impact in this context. You know his type:—

$$\text{e.g. } \frac{\text{The number of deaths in Wankie in 1965}}{\text{The population of Wankie in 1965}}$$

You find him acceptable, you say? Why then is the Crude Mortality Rate of Marandellas three times that of Wankie? Is coal dust beneficial to the lungs, after all? Do women drivers predominate in Marandellas? Are Wankie doctors of higher calibre than those at Marandellas? The answer is this. Wankie people live in mine houses during their employment and then move out on retirement to be replaced by youngsters not yet contemplating the great unknown. They move, in fact, to places like Marandellas with its old people's home where they live their lives to the full until called to take their place in their own individual Crude Mortality Rate. We can readily do without Crude Mortality Rates.

Yet we must replace him. Like heroes, the *Age Standardised Mortality Rate* and his surgical cousin the *Age Corrected Survival Rate* fill the breech well. They originate from a statistical gimmick. For example, imagine a fictitious standard population in a town called Chongololosdorp. If we considered the Wankie mortality picture as applying to the people of Chongololosdorp, and then applied the Marandellas findings to the same artificial population, the different results would be more comparable as there would be no age discrepancy. The results would be "age standardised".

Unfortunately, these characters are unpopular in some circles because they replace genuine deaths with artificialities; corpses, with ghosts!! "Stone dead hath no fellow", quote the surgeons. Indeed, to the family losing its father or the surgeon losing his first prostate-transplant patient, the death in itself may be enough. However, without the artificialities of statistical standardisations and corrections it is grossly misleading to compare the Death Rate at Marandellas with Wankie's and one surgeon's 5-year survival rate with anothers.

A brief word about the life and soul of our party, *Life Expectancy*. He doesn't often appear in the journals, but is my favourite. Of course, you may be interested only in your own life expectancy at this moment. The fun starts. Firstly, your life expectancy is about your death and not your life. Secondly, it refers only to those of your age or older who died or are going to die this year and I hope you are not included. In fact, "your life expectancy" this year has everything to do with the death of others and only a passing relevance to your own life. It is a sobering thought that the only time your life expectancy applies to you is the year you die and this, probably, is the year you are least interested. You may not agree with me that life expectancies are nevertheless fun, but you must agree that they are financial wizards. The biggest buildings in most towns bear witness to their success.

The terms discussed here tell you "what the doctor saw" and they are the least innocuous of the three sources of error. Such characters are most important as they serve to summarise long lists of results so that the writer can tell his story. Indeed, they may be rather motley, but they are more wholesome than the group describing "who the doctor saw", which we will discuss next month.

III. The Jargon

WHO THE DOCTOR SAW

Last month we chatted about the terms used to summarise "what the doctor saw" when series of cases are written up in the medical journals. Today, in discussing the people included in the series, or "who the doctor saw," we meet the biggest stumbling block to a medico-statistical settlement. Who the butler saw may be more interesting, but who the doctor saw, his actual sample, in terms of medical research, is of greater importance.

This is because the research worker, of necessity, is basing his observations on a few patients and yet everybody is requiring information about all people with disease X. You may conclude that cigarette smoking is beyond reasonable doubt associated with cancer of the lung in your wards, but does this mean all cigarette smokers are courting lung cancer? My advice to the research worker is to stick to talking about the patients in his ward, as there will always be plenty of readers who will think he is discussing everybody. My advice to the reader is to look more carefully at whom the doctor is describing than at the conclusions he draws. If ever you are to be taken for a proverbial ride, omit to read how the people came to be selected for a trial.

Even in the most competent statistical shoes the step from talking about a sample to predicting about the population is as hazardous as is the step from talking about love to actually making a proposal. The sample is the battleground of medicine and statistics. As a medically qualified statistician it makes me schizoid. My statistical left hand covets the biometrician's trump card, "random mating," whereas my medical right hand demands a more conservative if not downright prudish medical compromise. Perhaps the era of the promiscuous society will bring some relief! Even today the situation is not hopeless, a medico-statistical divorce is not inevitable, because in my experience there is always room for some settlement.

Ethics demand that a patient must be no worse off during a clinical trial than he would be in the hands of a capable doctor. Within these boundaries the medical profession morally owes it to society to make their clinical trials worthwhile. Their samples should, like a good photograph, provide a reliable image of the particular population. Just as the surgeon tries to make his biopsy specimen the best practical image of the organ, so the medical statistician tries to make his sample the best practical image of the population. To this end the population must be defined exhaustively and in an almost pedantic manner. If there is uncertainty about whether a particular person is included in the population, then there can be no certainty that conclusions based on a representative sample from that population are referable to such a person.

A good photograph and biopsy specimen must be of sufficient magnitude that conclusions can be drawn from it. One of the commonest requests made of a medical statistician is the number of cases which should be included in a survey. The exact answer to such a query demands knowledge of the variation in the population and the difference which the worker would consider to be of practical significance. The former piece of information is one of the by-products of a pilot or, to coin a term, a mini-survey. The latter piece of information is derived from experience. Without these results a statistician's ideas are no better than a good guess. There is nothing magical about the number 100; in fact, the arithmetic is probably easier with a sample of 101. Moreover, a sample of 100 is only twice as good, statistically, as one of 25, although there is four times the amount of work. Although not trying to underestimate the worth of large samples, the advances in statistics have tended to decrease rather than to increase the numbers required.

Two golfing terms are relevant while talking about sample size. Firstly, the "sudden death" approach is statistically not acceptable. By this I mean that it is wrong to determine sample size (the number of holes played) as the game progresses. This decision, statistically, must be made before the experiment is started. However, there is one exception, namely, the sequential design. Take, for example, a clinical trial of two drugs. The patients are included in pairs, one of each pair receiving one of the two alternative drugs. The patient who shows a better effect wins the hole in terms of the drug used. Initially boundaries are drawn such that the trial ends immediately a boundary is reached, but the particular point on the boundary depends on the information already at hand. This is the match-play type of approach.

One of the problems with medical sample sizes not referable to golf is that of the missing results. Who ever heard of a golfer missing out a hole, particularly the nineteenth? Even though cases are omitted from the results, they should not be forgotten. The reason for their omission, whether it be due to failure to attend the clinic,

refusal to take the treatment or death, must always be stated. You, the reader, can then decide whether their omission is relevant to the topic of the study. Statistically there are means of compensating for their loss, although there is a price to pay in sensitivity.

The missing results are a source of what medical people and statisticians both call bias. They both mean, in terms of sampling, that the sample is off-target so far as the population is concerned. In my view the medical profession usually underestimates the problem while perhaps the statisticians overemphasise it. Either way the statistical price of fair play in sampling is often more expensive than the medical research worker is prepared to pay.

Randomisation is the statistical insurance against bias in sampling. Often the medical man is obliged to abandon the idea of a random sample, particularly in a developing country such as this. As long as he is honest about it, I will generally support him. However, the implications of managing without the ideal are often misunderstood by those reading the journals.

Many think the random sample is the same as the haphazard or "willy-nilly" sample, such as "I walked into the ward and chose the new cases I saw, not knowing their diagnoses" type of sample. I learnt that this was not so in a rather sobering way. While studying statistics in London my class was sent one cool October evening on to the streets in the Tottenham Court Road area to note down individually the last digit of 100 cars that were driven past us. This was to simulate a series of random numbers. A few of us decided to spend the time in a pub unbiasedly imagining 100 numbers or generating them on the darts board. Warm with pride and sustenance, we returned with our cooler counterparts to have our numbers statistically tested for randomness. I do not think the professor thought less of us for our initiative, but he certainly thought very little of our lists of random numbers.

A simple random sample is one into which every member of the population has an equal chance of inclusion. The State Lottery winners constitute a simple random sample. In practice we enumerate each member of the population and use tables of random numbers (similar to those generated by listing the last digits of the car numbers properly), to choose a sample. In fact, randomisation does not prevent bias creeping into a sample, but it enables everybody to measure the chances of this happening. It makes chance work for us.

The random sample is important because most statistical significance tests are based on the laws of chance and it is therefore imperative to allow chance to operate without any restraint. To apply such significance tests without complying with the regulations is rather like treating a disease X albeit similar to a disease Y as though it was disease Y. Sometimes this manoeuvre is performed with surprisingly satisfactory results in both the medical and statistical field, but in both cases it is hazardous.

In conclusion, most of the misuse of numbers is covered in taking the sample. My overall advice to the authors is to be completely honest while trying all practical means to comply with the mathematical models. You, the readers, in assessing the value of the medical literature, are advised to treat omissions of method gloomily. Think more of the doctor who says this is not a random sample than the man who does not comment.

IV. Other Samples

WHO THE DOCTOR COULD HAVE SEEN

Life is not much fun if everything a person wants is either "illegal, immoral or fattening." Medical statistics is not much fun either if everything the article tells you is either incorrect, unethical or biased. So far this may be the impression these articles are creating. I may have persuaded you to view a haphazard medical sample pessimistically and I, myself, have conceded that the possibility of obtaining a reliable random sample, particularly in a developing country, is remote. It now seems necessary to discuss some positive alternatives.

After all, medical statistics is dependent on medical samples for its very existence. Nor is sampling always the poor relation of studying the whole population (in some circumstances such as blood sampling it is the only method!). The early volumes of the *Journal of the Royal Statistical Society* highlighted the hazards faced by medical pioneers when they were intent on evaluating entire populations. Their task was impossibly hard, standards inevitably fell and their results were consequently unreliable. Their problems would have been solved had they limited their evaluation to a representative sample from the population.

Notice that whenever we talk about samples we automatically talk about populations. I previously mentioned that any population must be defined initially in an almost pedantic manner. Suppose our population is "Doctors in Rhodesia." We want, for example, information about fees, or opinions on facets of the curriculum at the medical school. We must define our doctor population and decide whether this population is to include retired doctors on the one hand and those doing preregistration jobs on the other. You may be inclined to exclude medical statisticians! So long as all these decisions are relevant to the particular topic and are taken initially, our population is acceptable. Of course, if we exclude preregistration doctors any conclusions based on our sample only refer to registered doctors.

All samples are a compromise between reliability and convenience. This is particularly true in terms of the size of the sample, a factor which we have discussed previously. It is also the case when we talk in terms of any off-target result due to sampling or bias. Generally, as already stated, the random group of sampling techniques produces a more reliable less biased sample than the haphazard group of samples. Unfortunately, random samples are usually less convenient than others. The chief inconvenience is that up-to-date lists of every member of the population are generally a prerequisite. If these are not available a tremendous amount of time can be spent in producing them. To illustrate these points let us apply three random sampling methods to our doctor population and observe what might happen in practice.

The one already mentioned is the *simple random sample*. Each doctor in our defined population would have an equal chance of being included. We would draw the names, using a completely fair method, such as out of the boot of a Rolls Royce or by using a table of random numbers. Surprisingly, the simple random sample is not usually the most reliable method, as the one in a million chance does occur, even if only once in a million times. It is possible though remote that our noble profession's views would be represented entirely by the country's ENT surgeons—or perhaps worse, their anatomical neighbours, the thoracic surgeons! Therefore our simple random sample need not necessarily be unbiased. However, with it we are able to assess the chances of such a biased sample occurring in practice. Moreover, this arithmetic is easy to perform. It is interesting to ponder whether we would allow such a biased sample as the one mentioned or whether we would statistically err by putting all the names back and starting again, pretending such misfortune had not struck. Fortunately I have not yet had such a decision to make.

With foresight the problem could have been avoided by taking instead of a simple random sample a *stratified random sample*. This is more inconvenient in that each doctor would initially have to be stratified according to his or her speciality, e.g.:

Dr. L. S. D.: Psychiatrist;

Dr. B. P.: Physician;

Mr. L. P.: Neurosurgeon;

Prof. P. R.: Academic;

Mr. P. V.: Gynaecologist, etc.

A random sample would then be drawn within each stratum or speciality as the name of the method suggests, so that the different specialities would be represented in the final sample. If we considered the views of the G.P.s to be most relevant to our particular topic, there is nothing to stop us drawing a larger proportion from this stratum, although the subsequent arithmetic is slightly more involved if the proportion varies

between strata. If the factors deciding the strata are relevant to the point under review, this stratified random sample is the most reliable of all. It has the further benefit that as a by-product we could subsequently compare and contrast the views between the different specialities.

If we would need to visit our sampled doctors once we had their names, we would make life easier for ourselves if we had what is called a *cluster sample*. It is less reliable than the stratified sample, but if travelling is a factor then the cluster sample is generally the most reliable per unit cost. Before sampling, we would group the doctors by practices or dorps. Each group or cluster would then be subjected to a random sampling procedure. Once a particular cluster had been selected, every doctor who was a member of the cluster would be included in the sample. This method could be extended, say, to the tribal trust lands, where a map of the area could be gridded and the people in the randomly chosen small areas evaluated.

I suppose it is the ultimate aim of the medical profession to eliminate disease. To measure the size of the disease problem is a first step. It is an ultimate aim of medical statistics to eliminate bias. In the three sampling methods already discussed we have been able to take a first step by measuring the size of the bias problem. In the haphazard methods to be mentioned next, this first step cannot be taken. Yet how convenient our haphazard samples are! What bliss to be ignorant of how biased they may be! Let us take a look at some of these methods used in medicine. As the actress must have said to the bishop, "Let us take a short walk in the dark."

If population lists are available we would be unwise to choose a haphazard sampling method. Having paid the price in terms of inconvenience, we may as well purchase a more reliable method. The most popular haphazard sample is the so-called *systematic*. We could, for example, take as our doctor sample every twentieth name from our up-dated register, yet we may as well go that short next step and use random numbers. The systematic sample is generally unbiased, but we cannot measure its reliability due to the possibility of unsuspected intrusion by its Achilles heel, *periodicity*.

Let us illustrate this problem in terms now of patients in the hospital ward, instead of doctors. We could include the patient in every tenth bed on the particular day chosen for our survey. If the wards tended to be 10-, 20- or 40-bedded, periodicity could well be a problem. Every tenth bed might now tend to include a preponderance of patients situated near to the ward sister's office and may tend to include more than its fair share of seriously ill patients. On the other hand, it may tend to include nobody in this unhappy situation. Periodicity on account of time may be a problem in outpatient clinics or when systematically sampling patients visiting their G.P. The receptionist may squeeze in some ill patients at the beginning of the day. Maybe working men tend to make appointments towards the end of their working day and a higher proportion of these may be evaluated. Overt bias may also occur, as in a drug trial an alert sister may have already made up her mind about which treatment she prefers and may negotiate the candidates for admission accordingly. This problem can be counteracted, of course, not only by allotting the patients using random numbers, but also by running the experiment double blind.

Although everybody knows that *hospital admissions* are not a true reflection of the general population, doctors still base conclusions on them as though they are unbiased. Hospital admissions constitute a very haphazard sample of the population. For a numbers of years some surgeons removed the gall bladder as a contribution to the treatment of diabetes, due to an apparent association *in the hospital* between this disease and cholecystitis. The bias was caused by differing hospital admission rates and was exposed by Dr. Berkson at the Mayo Clinic 23 years ago. Yet some medical research workers still fall into this pit.

Imagine that the hospital admission rates are 90 per cent. for cancer of the lung, 10 per cent. for gonorrhoea and 50 per cent. for correction of squints. Consider that in the general population there are 200 cases of lung cancer, 500 cases of gonorrhoea and 20 per cent. overall prevalence of squint. Moreover, assume that squinting is not in any way associated with either of the other conditions. This means that 100 of the 500 gonorrhoea cases also squint. Of these 100 cases we would admit 55 (10 or 10 per cent. on account of their gonorrhoea and 50 per cent. of the other 90, i.e., 45, due to their squint). We could insert these 55 cases of squinting gonorrhoea into a contingency table below based on the hospital admissions. The table is completed by using the stated facts and applying the same reasoning, i.e.:

	Gonorrhoea	Cancer of the Lung	
Squint	55	38	93
No squint	40	144	184
	95	182	277

Using the hospital admissions, only 20.9 per cent. of cancer of the lung patients squint, whereas as many as 57.9 per cent. of patients with gonorrhoea are unable to look you straight in the face. It is as though the high cancer admission rate pushed squinters into the gonorrhoea beds. One can well imagine a local newspaper reporting: "Hospital records prove that squinting protects you from cancer of the lung but causes gonorrhoea." (!!!)

Matched samples are another popular sampling method in medical research. A patient with a disease is haphazardly matched according to different factors such as sex, age, race and social class. Unless lists are available of all possible mates and they are then chosen at random (I have never heard of this being possible in medical research), matched samples are not randomly chosen and so no measure of reliability is available. Nevertheless matched samples often successfully remove some of the effects of inherent variation between people. Patients are normally matched in pairs. If the matching is unsuccessful in that in fact each member is not more like his partner than those included elsewhere in the trial, this sampling method decreases rather than increases the sensitivity. Moreover, if one of the factors used in matching is directly and unsuspectingly related to the topic under comparison, no association may be detected.

In conclusion, then, there are many different ways of taking a sample which must ideally be from a clearly defined population. All have snags, but some, the random group, enable us to measure the chances of meeting one of the biggest snags, which is bias. Generally, in medicine we take chances in choosing any samples, albeit in good faith—good faith supplemented, I hope, by awareness to the problems.

V. Before the Race

CHECKING THE COURSE

"There was an old woman
Who lived in a shoe;
She had so many children
She didn't know what to do."

"Obviously," adds the precocious child, while toddlers more sympathetic to the old woman's plight admit that she had a real problem. Her situation must be controlled. Lively readers would comment that if she was "that old," then her problem is that of offspring increasing in volume rather than numbers, in which case her lack of space could be controlled best by slackening the shoe lace or finding the other shoe with a view to its occupancy. If she is not "that old," then the control measures come more into our own professional field.

Clinical medicine is full of controlling devices. The gynaecologists roll their "pills" ostentatiously into the fore while physicians rather reticent in their wisdom use their drugs, for example, to control hypertension. Surgeons are provided with an arsenal of artery forceps for controlling haemorrhage and radiologists keep radiographers busy by taking extra pictures which they often label "control." This extra expenditure is necessary, as X-ray controls give a yardstick against which pathology can be measured. Similarly, in the gambling aspects of medicine, control groups provide necessary yardsticks against which results of trials can be measured. The control in statistics is, for example, that group on a placebo against which the other group on whom a new drug is being used is measured.

By definition the control group is one identical to the experimental group in all aspects save for the particular factor under review. If a clinician is interested in whether long-standing cysticercosis predisposes to epilepsy in adult male Africans, then the yardstick is adult male Africans with long-standing cysticercosis, but without epilepsy. It is essential in clinical medicine that each research situation offers not only a control group, but the correct yardstick against which to measure the evidence.

A doctor set out to compare two diets for children. He took as one group well-nourished children attending a local school and arranged for them to eat only diet A during a school term. Over the same period he gave diet B to a group of undernourished children in his hospital wards and he gave these children an antibiotic as well in case some concurrent infection predisposed to malnourishment and thereby biased his results. You accept this? His control group should have been an identical group to that on diet A. Both groups should have been well-nourished school children or malnourished paediatric patients, and if he really thought that ethically he could prescribe long-term antibiotics, then both groups should have suffered them. As it was, he found that "diet B was significantly better," but of course we do not know whether hospitalisation, the antibiotic or the leeway the children had to make up was the real source of the difference. It could also have been the diet!

The above is not one of my far-fetched examples like the squinting gonorrhea patients last month. Nor is the use of the wrong control group limited to those busier doctors in developing countries where malnourishment is a problem. In fact, this particular doctor had gone so far as to take his samples using random methods, and his article had been accepted for publication by quite a reputable journal. This article was published some time ago and recently this particular error is not made so obviously in the journals. It is still made, however, and it is unfortunate that the wrong control group is used more surreptitiously now, as it takes doctors slightly longer to appreciate this error.

It is topical these days to have six principles—the other Mrs. Castle has! Care with the control group is my first and there are five others. It has been said that medical statistics is "the application of common sense to medical data." I repeat this because every medical person understands the mistakes once they are mentioned, but he often reads the journals only after a hard day's work and is more likely to overlook the fact that these commonsense principles have been misapplied.

The second principle is to back an *experiment* rather than a *survey* whenever they are on the cards together. A wit described an experiment as "an interference with nature," and to the extent that a survey would then be an evaluation of unadulterated nature, the above definition is partially acceptable. In a survey a patient may have experimental gadgets protruding from every conceivable (and non-conceivable) orifice with a highly academic team analysing every body juice —thus interfering markedly with nature—and yet this would still not be an experiment. The distinction is that in an experiment the investigator either allots to, or deliberately withholds from, the patients the drug, diet or factor under investigation, whereas in a survey he only assesses what is already present. The cysticercosis example is a survey—for it to be an experiment

the doctor would have to allot the disease, preferably randomly, to some hamsters or guinea pigs.

The difference between an experiment and a survey is important, as it affects the conclusions. In a dietary experiment, properly controlled, a significant difference may be due to the different diets or to chance. In a survey a significant difference may be due to the factor under review or chance as above, but it could also be due to some other unrecognised factor. Suppose we were reading about a survey on smoking and lung cancer (most clinical research is survey work). A significant association between smoking and lung cancer could be due to smoking and it could be due to chance as in an experiment, but it could also be due to some other common factor—a particular psychological make-up, for example, may predispose both to smoking and lung cancer. There may be a genetic basis or a familial tendency to both. It is often unlikely that some other factor is to blame, but we can never be sure in a survey. This is especially so if we can, as is the case at the moment, only recognise the tip of a pathologico-psychological "iceberg." The second principle may be described as the "bikini" effect of a survey. What is revealed may well be interesting, but what is concealed is much more vital! By teaching hamsters to smoke and by running a smoking-lung cancer experiment rather than a survey we would eliminate this bikini possibility.

We would also improve on a *retrospective* or backward-looking situation by converting it into a forward-looking study. This is my third principle. An experiment with random allotment of the factor can only be prospective. A survey can be prospective or retrospective and the former is better. Retrospective work is so often the poorer relation that it has been called "backward" in both senses of the word. For example, in a retrospective survey it is not possible to incorporate interesting pointers into a design so that one can observe their interrelationships. Moreover, information gained prospectively is usually more reliable, but the studies are usually much bigger and may be impossibly big. For example, in our cysticercosis study, to observe a group of potentially cystercotic African children prospectively would involve following up a very large number. Similarly, to follow youngsters through to see whether the smokers develop lung cancer more frequently than those who limited themselves to other vices would involve a tremendous amount of administration. Often prospective surveys are non-starters in our clinical races, but when they are they should be fancied by the punters.

Prospective studies, therefore, are more reliable and my next principle also has reliability as its prize. This fourth principle involves the criteria used as measurements, the standards of evaluation. Are these standards as high as they could be? Is the information being used with maximum efficiency? I once delivered a paper concerning bilharziasis and intelligence at a symposium (there was collusion in the writing of this paper; I was the smallest and so was bulldozed into presenting it). We were interested to see whether those children with bilharziasis had a lower I.Q. than those clear of the disease. This is a qualitative criterion so far as the disease is concerned, the children being classed as either having the disease (the sheep) or being free (the goats). The paper was fairly criticised for failing to use the information more efficiently and in a quantitative manner. To do this we could have used the egg counts in those with bilharziasis to indicate the severity of the disease. Similarly with examination results there is much more information in the quantitative score, 5 per cent., than in the qualitative grouping, "Fail." Basically it is not so much qualitative information such as "mini" and "not mini" which is required, but "how mini"! Given a quantitative measurement, the information is used with maximum efficiency.

One of the reasons why, as a student, I personally enjoyed obstetrics and gynaecology was that in this branch of study criteria were usually precisely defined. A woman's journey into toxaemia was clearly signposted, and even as a student one could tell when she had arrived. This is in contrast to hypertension *per se* where one doctor's criteria of where a hypertensive takes over from a person with a blood pressure reading at the upper end of normality is only subjective. Recently a group of about 50 general practitioners at Pretoria handed in their individual definitions of hypertension. No two people were in complete agreement. The fourth principle, then, is to prefer good criteria—criteria which are measured and precise.

Fifthly, it is important to consider the manner in which the information has been collected. Often in journals or reports the potential pitfall this may cause is not recognised. It is beneficial to attempt to visualise how the information was gained. If in an article "hundreds" of a special type of delivery are discussed, it is fairly safe to assume that more than one person has been involved in collecting and collating the results.

Often the most junior staff do most of this

work, but this may be unavoidable and no one expects a top civil servant, for example, to confirm all the notified diseases before including them in an annual report. Nevertheless these records are best kept by the person interested in the work. If others of necessity become involved it is important that their numbers are limited to minimise variations of definitions and opinions. Certainly the doctor requiring the information should show considerable interest in its collection, as this motivates those recording the data to maintain high standards. Keeping the collectors "in the picture" concerning the uses of the information (is it always used?!) and discussing any interim decisions all increase motivation. Motivation is an efficient catalyst for data reliability.

My sixth and final principle is the most important. As a punter you can afford to ignore the statistical arithmetic or recipes. It suffices that one person knows how to "put in the figures, square them, transfer them to the computer at 100° with formula 007 and allows them to stew for a quarter of an hour." However, a reader can check that the results do answer the question posed.

A colleague once asked how to calculate the experimental error for a particular biochemical measurement. He had already collected results on 100 patients. Unfortunately he had only one result per patient and so, whichever way the numbers were cooked, they could only estimate the range in the population and could show nothing about the required experimental errors. This information could only be derived from repeated estimations on a single specimen.

Involved-looking analyses often reflect a less rational approach and a less reliable result than tests involving simple arithmetic. Each step in an involved analysis involves mathematical assumptions which are only approximately true. The story is told about the two medical academic jockeys running in the same field of research. They were unable to discuss their progress unless their statisticians were there!

In reverse order my six principles relevant to good clinical racing are these:

Answer the question, and simply.
Attempt to visualise how the information was gained.
Prefer criteria which are measured and precise.
Prefer prospective studies.
Prefer experiments to surveys.
Consider the control group.

All of them have a part to play in ensuring that clinical hunches are assessed fairly. Now let us start the racing.

VI. The Race Itself

THE IDEAS BEHIND A MEDICAL SIGNIFICANCE TEST

Firstly, a short story, which is partly true.

A physician and a surgeon made arrangements to play golf together every Thursday afternoon, their work commitments and the weather permitting. They each decided to toss a coin at the "nineteenth" hole. Should both toss heads or both toss tails, they would re-toss until the one doctor threw a "head" and bought the drinks and the other threw the more profitable "tail."

Good fortune smiled on the physician, as on the first three Thursdays he tossed "tails" and the surgeon tossed "heads." The surgeon was Scottish, but nevertheless he bought his colleague drinks very willingly. Even on the fourth occasion the surgeon smiled as he emptied his pockets. On tossing heads for the fifth week the surgeon muttered a non-hippocratic oath to himself.

Week six saw the surgeon toss the unlucky "heads" again. Returning home, he muttered to his wife, *"I'm sure the physician is above board,* but I do think there is something uncanny in his tail-tossing ability." The surgeon's wife reassured her husband: *"It is obviously just bad luck on your part,"* she said; *"ignore it."*

After the seventh game the surgeon begrudgingly transplanted his financial resources to the bar counter. The physician's therapeutic smile did not cheer him and the surgeon arrived home in a very belligerent mood. His wife agreed that *even though there was no proof* the physician was tricking her husband, *for the results could be due to chance,* it was most suspicious. *"The line must be drawn somewhere,"* she said; "if it happens again you should play golf with somebody else."

It did happen again, and on the ninth Thursday the surgeon had a "surgical emergency" and the story and the golfing partnership ends there.

I do not recount this story idly, but because some of the ideas brought out are pertinent to the analysis of results from medical research and to the drawing of conclusions from them. The circumstances are different, but each of the three sections emphasised by the surgeon's wife in the story represent important medico-statistical concepts.

Firstly, the surgeon's wife's initial reaction was "It is obviously just bad luck on your part; ignore it." This same attitude was held in the beginning by the surgeon when he assumed the physician to be "above board." This is akin to the legal approach where a man is innocent until shown to be guilty. Statisticians also assume initially that any discrepancy is due to chance. If a new analgesic is being tried we firstly assume that it is no better than its older competitors. It is up to the results to make us change our minds, and unless the evidence is sufficient we rather retain the view that any difference is due to chance alone. Statistically, this original viewpoint is called the "Null Hypothesis," but this is a sophisticated way of saying to the medical research worker, "I don't believe you until you convince me."

The surgeon's wife hit the next statistical nail on the head with her viewpoint that there could be "no proof" . . . "for the results could always be due to chance." I am often told that statistics never prove anything; that people vary, and any differences between their reactions could be due to their inherent variability. If people were identical in their behaviour we could treat the outcome of trials simply. It is the fact that people do vary which reinforces the need for statistics, the subject developed especially to take into account this variability. Progress would be nonexistent if our initial view is that we need convincing and our second view is that we cannot have proof and hence cannot be convinced. Statistical analysis produces a reasoned measure of doubt concerning our faith in the Null Hypothesis. Very rarely are juries given real proof of a man's guilt, but it is their duty to sift through the evidence with an open mind to determine a man's guilt beyond reasonable doubt. It is true that we cannot prove a new analgesic to be better, but with an open mind we can determine them to be different beyond reasonable doubt. Be cynical if you wish, but people do vary, and to produce a reliable set of odds with an open mind seems to me to be no mean achievement.

The final point is that "the line must be drawn somewhere." A measure of doubt, no matter how reliable, is a dead end if no conclusions can be drawn and no action taken. It is true that chance could always be the cause, but the time must come when the evidence is such that it is more realistic to acknowledge some other factor. Juries reach conclusions of "guilty" and judges act accordingly. Statisticians prefer the innocence of the Null Hypothesis, but given sufficient evidence, they will renounce it and accept that which the evidence presents. The line is drawn and it becomes possible to conclude that the new analgesic is significantly better.

It is essential that the line must be drawn somewhere and that it is drawn with an open mind. In statistics it is called the "significance level," and it becomes the means by which conclusions are made. It is drawn before the results are analysed so that the conclusions are made as fairly as possible. Of course, you could look at the results and decide upon your verdict blindly. As there are significance tests designed to guide you, it seems wiser to me to use them. It is difficult to be certain that your decision is not being influenced by some preconceived notion. For example, some people hold views about the wastage among female doctors and it is of some interest that 85 of Rhodesia's 504 practising doctors are female. Antagonism to female doctors is indicated by the statement that "less than 17 per cent. of all practising doctors in Rhodesia are women." This is true, but it is equally true that "there is more than one female doctor in full-time practice to every five males." This now infers that women's contributions to medicine are fairly substantial. The wording reflects the speaker's personal opinion and is likely to influence your own thoughts on the subject. This is equally true with the results from medical research and it is necessary to measure the significance of the results.

Originally I likened the statistician's work to that of a course steward. Supervising races and reaching reliable decisions by significance testing is only one of the aspects of a medical statistician's work. Many punters at a race meeting have no conception of the preparation beforehand; they only see the actual race. Just as there is more than a lot of nags to a race, there is more than a lot of conclusions to a trial. A significance test is only as worthwhile as the preparation which has gone into collecting the data.

Andrew Lang once accused somebody of "using statistics like a drunken man uses a lamp post—for support rather than for illumination." Significance tests may support a theory, but it is the design of a clinical trial which provides the illumination.

VII. The Racing Results

WHAT THE ODDS MEAN

A surgeon once said in Salisbury:

"Statistics mean nothing to me.
A difference real
Is that what you feel.
To count it is just heresy!"

—Anon.

The days of "impressions" in medical research should have gone. Today we have the wherewithal and an obligation to assign a meaningful figure to the results before drawing conclusions.

In the last chat we mentioned the important ideas in performing a significance test. Initially, we assume that any difference is due to chance or, put in another way, we accept the null hypothesis. Next we assess the evidence, p, in support of this null hypothesis. The letter p actually represents the probability that the null hypothesis is the real explanation, the odds or a likelihood that any difference is due to chance. The bigger the value p, the greater the likelihood that any difference is due to chance, and the smaller p becomes, the greater is the temptation to claim that any difference is real. It is at this point that our line, the significance level—the yardstick against which p is measured—is used. This yardstick enables us to make decisions based on these conclusions. When p is bigger than the significance level, the evidence is only sufficient to support the null hypothesis and insufficient to support a real difference. When p becomes less than the significance level, then "vive la difference!" (Fig. 1).

One of the disguises assumed by the significance level in the medical journals is the probability .05. This means that the odds are 20 to 1 against this difference being due to chance. The sign $>$ means "bigger than," so the lady's attire in Fig. 1 could be described as ".05$>$p," although this is probably not the response she hoped for! Whenever an article concludes that the significance level is greater than p (the probability of the results arising by chance), a statistically significant difference can be claimed.

Other significance levels which are used are .01 and .001, with odds now of 100 to 1 and 1,000 to 1 against the null hypothesis. As p becomes smaller and creeps beyond these small significance levels, so the difference becomes more and more apparent or statistically significant! (Fig. 2). If you see .05$>$p$>$.01, the inference is that the skirt is between the ankle and the knees; there is sufficient evidence to claim a real difference at the .05 level, but not at the .01. If .001$>$p, the answer is very significant!

Fig. 1—"When p becomes less than the significant level, then "vive la difference."

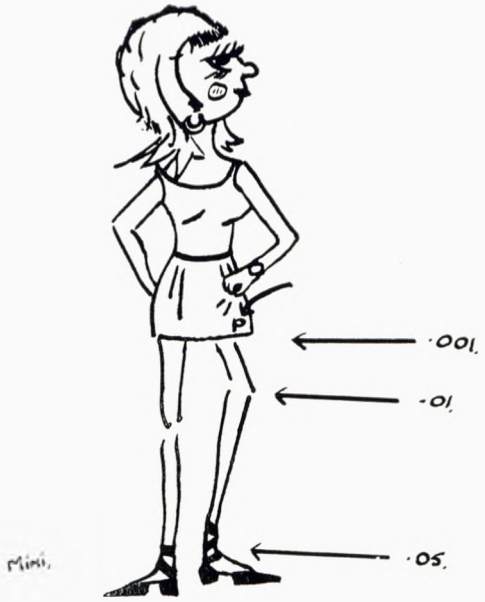


Fig. 2—"As p becomes smaller and creeps beyond these small significance levels, so the difference becomes more and more apparent or statistically significant."

Significance tests are designed so that p, the probability that these particular results could have arisen by chance, can be estimated. Let us see how this works out in practice by referring back to the surgeon's problem. Remember, the surgeon bought the physician drinks each time the physician tossed "tails" and the surgeon tossed "heads" until, after the eighth time, the surgeon quit. The null hypothesis is that the physician's good fortune was due to chance, and when we have chosen a significance level and calculated p, we can give a statistical answer.

Assume we choose .05 as the significance level so that when p is less than .05 we conclude that the odds are longer than 20 to 1 against chance and so advise the surgeon to stop. Ignoring the times when the pair of them tossed the same, the probability, according to the laws of chance, of the surgeon tossing "heads" and the physician "tails" equals the likelihood of the physician tossing "tails" and the surgeon "heads." The chances are 50:50. As probability is measured between 0 (impossible) and 1 (inevitable), the probability of the surgeon tossing "heads" and the physician "tails" the first time is $\frac{1}{2}$, so after one occasion $p=\frac{1}{2}$ or .5. The probability of this happening twice is $\frac{1}{2}\times\frac{1}{2}=\frac{1}{4}$ or .25, and of it happening three times is $\frac{1}{8}$ or .125, etc. These results can be written in a table as follows:

Number of Occasions	p
1	.5
2	.25
3	.125
4	.0625
5	.03125

Notice that even after the fourth throw $p>.05$, the skirt is below the ankle, and so we still accept that the physician's good fortune is due to chance and that he is above-board. However, after five throws $.05>p$, the skirt is above the ankle, and on the basis of our chosen significance level we reject the concept of good fortune and advise the surgeon to stop. We drew the line and the evidence is now sufficient to say the results are statistically significant at the .05 level.

There are some points worth mentioning. Firstly, if we continue the table we would see that after seven throws the results are significant at the .01 level and after ten throws at the .001. This is demonstrating in practice what you would expect in theory. If real differences exist, the more results there are in a trial, the greater the degree of statistical significance. Eventually the difference is so obvious that the need for statistics disappears!

Previously we decided that statistics could *prove* nothing, and now we can go further and admit that occasionally the conclusions are incorrect. Even after ten throws the physician need not have been cheating, although statistically we have concluded that he was. He could have been remarkably lucky. However, statistics measures the chances of this conclusion of guilt being incorrect. A significant difference at the .05 level will be incorrect one time out of twenty. This is because when p equals this value it means that the odds are one in twenty that the results could be due to chance. When p equals .001 and we claim a significant difference, the odds are one in a thousand that chance could be the real factor. There is an error, but we can measure it. Statistics is only a yardstick, and without the significance level, even with this measured possibility of a false conclusion, we could make no decision at all. After all, statistics only claims to provide a "measure of doubt."

There is the other side of the picture. Besides the error in drawing a wrong conclusion of "statistical significance" which can be measured as shown above, there is another type of error which is not so easily measured. This is the error involved when a real difference exists from the start, but the evidence is insufficient to accept it statistically. If the physician was cheating all the time it was a mistake to allow him to continue. It is in order to minimise this type of error that clinical trials and surveys are designed on as large a scale as is reasonably possible. To support the null hypothesis is to say the difference is either unreal or that, if it is real, there is as yet insufficient evidence to be convincing.

The last point is one for which statistics and statisticians cannot be blamed. It seems to me that there is a tendency among editors of medical journals to confuse "statistical significance" with "practical importance." At the end of an analysis one is often tempted to say "So what?" If information is not important, no matter how significant the difference, I would join in the chorus with other practising doctors by saying "So what?"

This is the age of technology and numbers, so the least we can do in medical research is to assign a meaningful figure to the results—to the meaningful results of practical importance, I hope.

VIII. So, How Good a Punter Are You?

My chief aim in this series has been to protect you, the practising doctor, from statistical abuse in the medical journals. I hope you now think statistically, as this is more important than the arithmetical aspects of statistics. Some articles which you will read in the medical journals are excellent, while others are rubbish. In which category would you place the following contribution and why? How good a punter are you?

The references relate to phrases or words in the order in which they arise which are highlighted later as being worthy of comment.

An association between carbohydrate intake and the incidence of ulcerative colitis.

Castle, W. M. & Rittey, D. A. W. University College of Rhodesia.

Introduction

A recent survey from America (Snoopy, Module *et al.*)[1] suggests that patients with ulcerative colitis are thinner than others.[2] In an attempt to follow up this finding,[3] the authors carried out an experiment locally to confirm[4] that the disease was caused[5] by a low carbohydrate intake.

Method

A large team[6] of highly trained dieticians closely questioned[7] 500[8] hospital admissions[9] at each of the local hospitals, i.e., that for the Africans and that for the Europeans.[10] This was to determine the average difference between the carbohydrate intake[11] of the two racial groups.[10] The questioners were particularly interested in the finding that, with the Africans, as the percentage of carbohydrate intake increased, so the percentage of protein intake decreased[12] (see Fig. 1).[13] They reported that this may indicate a need for health education.[14]

Results

Table I[15] shows the 995[16] patients grouped by race and whether they consumed more or less than 50 per cent. carbohydrate in their diet.[17] These results proved without doubt[18] that Europeans eat less carbohydrate[19] than the Africans ($p.>.001$).[20] Meanwhile, Table II[16] shows the

Table I

THE CARBOHYDRATE INTAKE IN THE TWO GROUPS

	More than 50%	Less than 50%	Totals
Africans	409	91	498
Europeans	263	234	497
	670	325	995

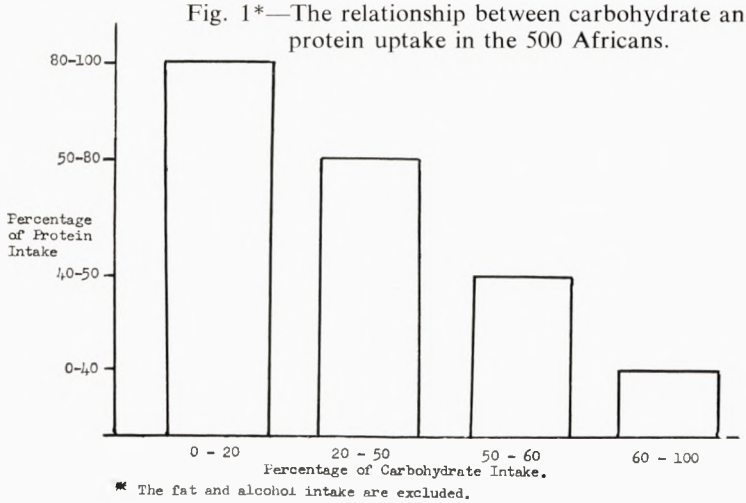
One hundred and eighty per cent. more Europeans than Africans can appreciate the value of a low carbohydrate diet!

Table II

CASES OF ULCERATIVE COLITIS IN THE TWO GROUPS

	Ulcerative Colitis	Other Diagnosis	Totals
Africans	1	469	470
Europeans	27	455	482
	28	924	952

Fig. 1*—The relationship between carbohydrate and protein uptake in the 500 Africans.



* The fat and alcohol intake are excluded.

incidence[21] of ulcerative colitis.[22] The difference in incidence[21] between the hypocarbohydrating[23] Europeans and the hypercarbohydrating Africans is very nearly significant .057 (c.f. .05).[24] When one considers that the single[25] African case on further close questioning[26] was found to be a school teacher eating a mainly[27] European diet, it becomes obvious[28] that the low carbohydrate diet is undoubtedly[18] causing[5] the higher disease incidence[21] (Fig. 2).[29]

Discussion

This strictly controlled[10] experiment[30] confirms[4] that the incidence[21] difference can be firmly associated with the lower carbohydrate diet of the Europeans.

It is true that for the purposes of weight accretion[23] there might be variations in daily carbohydration[23] by the Europeans, but these fall well within recognised parameters.[23] It is very significant[28] that the only African case was a school teacher on a hyperproteinised diet.[31] This confirms the view of Soap, J.[1] (1968), who in a controlled study of ulcerative colitis in urbanised Eskimos and rural aborigines[10] concluded that a hypercarbohydration programme is valueless without a concomitant hypoproteinisation diet with penicillinisation[23] of the intestinal flora affecting endoabsorption.[23]

The authors report that they themselves intend to carry out further experiments[30] to determine whether a high carbohydrate diet protects the African or the low carbohydrate diet in the European causes the disease.[32]

Conclusion

In the authors' view this is a lot of old rhubarb and is the worst article that they have ever seen published in any medical journal.

Summary

Good medical punters would agree with the majority of these references.

References

1. An international reputation does not guarantee the worthiness of the comment.
2. "Others" ought to be defined. They could be hypertensives or arthritics who one would expect to be fatter.
3. Unless the "finding" of Snoopy *et al.* is more clearly defined, it is impossible to "follow up."
4. "Confirm" suggests that the local authors have already decided what the answer is. Suspect them of bias.
5. The authors are probably mistakenly thinking that association and causation are synonymous; statistics help us to test associations, but not causations. Remember, lung cancer is *associated* with smoking; it does not *cause* people to smoke!
6. The larger the team the greater the variation in interpretation of answers. One motivated worker would be better.
7. The answers to questions are subjective and very liable to bias.
8. A large sample cannot correct badly designed research.
9. "Hospital admission" should be more closely defined. Are maternity cases and paediatric cases included, for example? Moreover, hospital admissions are notoriously unreliable and should certainly never be used for comparisons between hospitals.
10. Wrong control group—the correct control would be a group as similar as possible to one of the racial

Fig. 2—The relationship between diet and ulcerative colitis in the African patient.

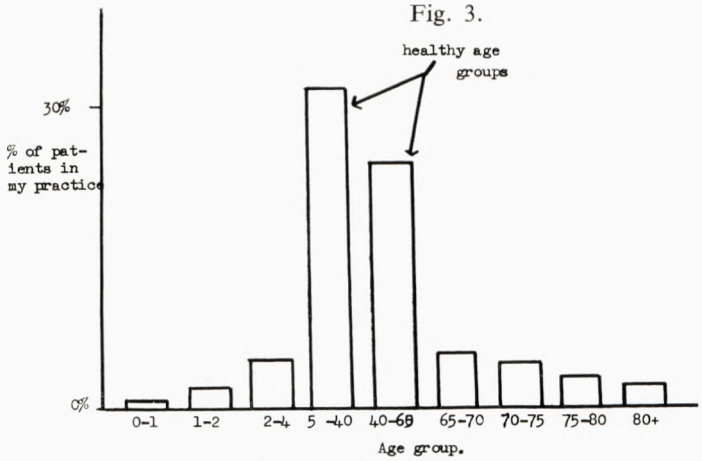
groups but without ulcerative colitis, preferably not in a hospital. Using two racial groups only confuses the issue.

11. This measurement would be very acceptable if contrasting those with the disease and those without.
12. This finding is true with the Europeans also. It is inevitable and not particularly interesting, as is demonstrated in——
13. This misleading figure highlights two further points. (a) Deliberately excluding the effects of fat and alcohol intake seems to be a major point worthy of discussion in the text. (b) Varying ranges along the horizontal axis can mislead (see Fig. 3).
14. !! See 12.
15. Besides the arithmetical errors in this table, the conclusion is stated in an unnecessary "loaded" fashion. Also
16. There are "missing results" not accounted for. Are they perhaps ulcerative colitis patients on a high carbohydrate diet? We need to know whether their absence is related to the topic under discussion.
17. See 11. The original measurement was more valuable than this dichotomy.
18. Statistics never "prove without doubt," but measure doubt.
19. We have no information about the *actual amount* of carbohydrate.
20. Presumably $.001 > p$ is the conclusion. As it stands, p could be any value bigger than the minute .001, e.g., .99.
21. Incidence is the term answering the question "how often?" The term prevalence is required here, as this answers the question "how common?"
22. The disease was not defined fully. The criteria of diagnosis must be comprehensively stated.
23. There is no need for "hyper-intellectual" words. Most of them are meaningless.
24. The purpose of the significance level .05 is to enable us to decide whether we conclude that results are either significant or not significant. There is no such conclusion as "nearly significant," particularly when it is "very nearly significant."
25. "Single" is an ambiguous word. Is he unmarried or is he "the one"?
26. See 7. It is inconceivable that "further close questioning" is needed to reveal the type of diet when the patient has been "closely questioned" already. This suggests that the original replies are unreliable.
27. Words like "mainly" are very subjective. You should be told exactly what this patient eats, as his presence is so important (?) to the authors' case.
28. He has already been included in the analysis and to mention him again does not cause the results to "become obvious." Remember, we are not yet aware whether he eats more or less than 50 per cent. carbohydrate; perhaps his "mainly European diet" is sandwiches for lunch and cakes for tea!
29. A graph from one person! Notice he is a hypercarbohydrater and that the authors were rather sneaky not to show results below 50 per cent. on the vertical axis.
30. This "experiment" is "a survey." This means that a significant result could be due to chance, the factor under review or some associated factor—such as race!
31. This is not true. See Fig. 2.
32. Ah, well! In the introduction to *this* survey the authors said they were "to confirm that the incidence of the disease was caused by a low carbohydrate intake." Better luck next time!

Acknowledgments

You may wish to blame Professor Michael Gelfand, Editor of the *Central African Journal of Medicine,* for suggesting that I write these eight articles about "Gambling Aspects in Medicine." You are certainly indebted to my colleagues, Dr. D. A. W. Rittey and Mr. D. Thompson, who read the drafts and many redrafts until they understood the manuscript in this form—a statistically significant achievement! I am grateful that this text must now be foolproof!

Fig. 3.



Locum wanted: Minimal paediatrics and geriatrics, so very few night calls!