SOME PITFALLS IN GENERAL PRACTICE RESEARCH

E. V. KUENSSBERG, M.B., Ch.B.
Edinburgh

1. *Medical Survey and Clinical Trials*, by L. J. Witts (Oxford University Press, 1959).
2. *How to Lie with Statistics*, by Durrell Huff (Gollancz, 1954).
3. *Design for Decision*, by Irvine D. J. Bross (Macmillan, New York, 1953).
4. *Symposium on Clinical Trials*, by Bradford Hill (C. C. Thomas, Illinois, 1960).

The enthusiast who attempts to collect facts and figures has many traps set for him, traps which beguile him into believing that he has really measured the problem completely, or that he is at the dawn of a new explanation for an old or new phenomenon in medicine.

These pages may throw cold water on your enthusiasm and make you return to your library or university. In general-practitioner research the pitfalls are numerous, because the general practitioner so often tends to work in medical isolation. His contacts with colleagues are short and brief. We in general practice are therefore frequently and disastrously out of touch with the bibliography.

The most common difficulty which the general practitioner has to face in research is the handling of figures in the light of statistical requirements. It is only very recently that a few of the universities have begun teaching their medical students the rudiments of statistical knowledge.

The difficulty about statistics is clearly illustrated by the following example taken from recent medical literature:

> In a maternity hospital during 1948-57 there were 259 labours (0.8 per cent of all deliveries) which lasted more than 48 hours. Among these the uncorrected perinatal mortality rate was 3.4 per cent as compared with the annual overall hospital perinatal rates of 3.8—5.3 per cent during the same years.

If you look at these figures you might be entitled to think that it is safer for the baby to be born after a prolonged labour. This is obviously nonsense. The answer to this figure is of course that

approximately half the perinatal mortality is accounted for by prematurity, whereas hardly any premature babies will be born by prolonged labour. Thus, to get the real answer, the above quoted figures would have to be corrected for prematurity.

The use of percentages is another method of unconscious self-deception, particularly when small numbers are involved. General practitioners have published drug trials where they report an analysis of 32 cases, as follows:

56 per cent completely cured
25 per cent partially cured
19 per cent no effect of the treatment

When one realizes that the 56 per cent only represents 18 patients and 25 per cent—8 patients, and that 6 patients were not improved at all, the emphasis on the 56 per cent is not a very valid deduction. After all, if two additional patients had been recorded as failure, because, for instance, they had indulged in certain self-medication, in addition to what the doctor ordered (but did not want to spoil the doctor's treatment and enthusiasm by telling him about it) the percentages would then be 25 per cent failed—quite different from 19 per cent.

Much could be made of this difficulty which the analysis of small numbers produces. Small numbers can of course be used in a drug trial if they are one hundred per cent (or nearly so) success or failure, though the statisticians are able to produce safe working rules for these problems and should always be consulted for this kind of problem on numbers, *before* the trial is got underway.

One of the pitfalls which few general practitioners seem to be able to avoid is the wrong use of *technical language.*

We all know the patient who talks of his " gastric stomach ". Yet we in turn use such terms as " controls "—" random "—" unselected " and " significant " without realizing the snags. " *Controls* " are a very specialized subject for the research worker, and must only be quoted, selected, or adumbrated after careful consultation with the expert research worker, or statistical adviser. " *Random* " is today a technical phrase which denotes an equal chance of selections from a definite population. Its most expert example is " Ernie " who picks the premium bond winners every month. Don't tamper with " Ernie "—let yourself be advised by the experts when " Random " is truly " Random ". " *Unselected* " in the research jargon just does not exist. If you, as a doctor, choose unselected patients (which of course is a contradiction) you forget

that certainly in general practice in the United Kingdom the patient still chooses his doctor. " *Significant* " is a magic word, which can only be used in research work after a statistically knowledgeable person has performed all kind of rites on a slide rule or electric calculating machine and pronounced p greater or smaller than 1. Keep off this terminology unless you get the " all clear " from those who know what it means to be " significant ".

Good record keeping is one of the most important insurances against pitfalls. How readily reliance on memory can become a pitfall is demonstrated by the following fact. An enquiry into the association of rheumatoid arthritis and cancer, led to the statement that no rheumatoid arthritis patient in our practice ever died of cancer, at least over the past 12 years. However, a prick of conscience instituted a search at the local rheumatic unit hospital records where it was discovered that each year about six rheumatoid arthritis patients died of proved cancer and when the names were scrutinized I recognized three former patients about whom I had completely forgotten. Yet, one felt sure that such an infrequent association as rheumatoid arthritis and cancer would imprint itself on one's memory.

The enthusiasm of the doctor has been mentioned. There is not enough space to discuss fully how enthusiasm can bias results of treatment. The literature on placebo reactors abounds. We must assume that the statement that at least one third of our civilization are placebo reactors is correct. This seems that a third of our patients will say that they are better of a certain condition for a stated time, even if they have received distilled water, or lactose dummy pills, or any other form of inert treatment.

Not only is enthusiasm in treatment and approach to the patient an important factor, but also our enthusiasm in a given cause, our bias towards certain events, or successes, or even types of patients. This is a most complicated subject, but its significance lies in the fact that it is extremely difficult to compare one doctor's work with another, unless it has been done in a laboratory or has sufficient correlating measurements which allow personal bias to be ruled out, or to be measured or assessed.

The London School of Hygiene has, for instance, demonstrated that blood pressure readings of patients suspected of raised blood pressure are higher when taken by the doctors who are conducting this investigation than when taken by technicians who do not know the reason why the blood pressure is taken. Furthermore, if the

doctor uses a blind method of reading the blood pressure (a special sphygmanometer where the mercury scale is obscured, and the significant changes in auscultation are denoted when pressing a button), the doctor's bias in expecting a higher reading is undoubtedly revealed.

The knowledge of *background information* is tremendously important—background information about one's own practice. The lack of this constitutes a common pitfall. For instance, a high incidence of whooping cough in a given practice may have several implications. It may be a very young practice population with many young children at risk, and relatively few adults. Post-war re-housing areas for instance, tend to consist of populations where about half are under 30 years of age. Thus you would need to have your age/sex register as an essential background information. However, the high incidence may in fact be due to the absence of any prophylactic immunizations, or you have not had an epidemic of pertussis for some years in that area, and thus there is a relatively larger number of susceptible people exposed to infection.

Similarly, when you report a large incidence of Sonne dysentry, background information as to overcrowding, young families, type of sanitation, and type of housing, and so on, are necessary to allow proper interpretation of your reported observation.

Measurement of severity is an incredibly difficult subject. To demonstrate this there is hardly any need to mention the most difficult of all, measurement of pain. It is wise to seek and accept the advice from those experts who have experience in this. As a working rule it is necessary to establish the beginning and the end of the scale of severity or measurement, and place at least three other grades on this scale in between. It will allow for more fair statistical evaluation, e.g.:

1. Cured
2. Improved
3. Partially improved
4. Hardly improved
5. Failed to cure

When different observers evaluate the results they will be in little doubt about the first two groups, and similarly the last two will make analysis possible. If, however, there has only been one other grade between *cured* and *failed to cure*, such as *partially improved*, different observers would have found it difficult to agree on what is *partially* and what is *failed to cure*, or even *completely cured.*

The *between-observer error* is another common pitfall. Medical

measurements are not sufficiently exact for two or three observers to produce identical results. The results may be sufficiently close to allow them to be used in one series, yet particularly in general practitioner surveys, one finds the " between-observer error " a most stultifying pitfall. For instance, let us consider a trial of pressure bandaging against antibiotic cream in the treatment of varicose ulcers. How firmly is Dr A. going to put on the bandage, and how firmly is Dr B. going to do it, particularly if Dr B. is in a great hurry, and Dr A. has a special theory about how firmly bandaging ought to be done? Dr C. again, is a lady doctor who has a different approach to the patient with her varicose ulcer, who in this case may be a frail old lady over 70. Not that the size of the ulcer could not be measured with a ruler, but then you need to agree what constitutes the healing edge. Is it completely epithialized skin, or is it the dry area?

Enough has been said to emphasize the " between-observer error ", which can only be avoided by accurate definition, frequent personal contact amongst the observers, and what is important, in any lengthy investigation, frequent recalling of these definitions.

This brings me to *consistency of diagnosis and accuracy of diagnosis.* It is surprising how we can gradually shift our premises in an enquiry. It is simple enough to record the onset of measles once you have agreed what constitutes the onset, e.g., the appearance of the rash. If, on the other hand you start recording it from the prodromal symptoms, it is quite easy if there has been a four month gap from your last recorded cases of measles to record it either from the onset of a catarrh or upper respiratory symptoms, or from the injection of the eyes. These events may well coincide, but may in fact, when a final analysis is made, alter the onset of the prodromal stage, by as much as 2 or 3 days, unless the events are stated and measured specifically. Nothing is easier than when there has been a dearth of clinical material for an investigation to forget the exact definitions agreed to, perhaps 9 to 12 months earlier.

Retrospective interpretation of records (and memory) is a chapter of its own. It suffices to say here, that unless it is clearly stated that any investigation is based on retrospective material, the practitioner is really laying himself open to severe and damaging criticism, as all retrospective investigations are subject to the grossest variations owing to defects of recording or memory, unless the recording was done according to a pre-arranged plan with agreed definitions.

Many a pitfall can be avoided by instituting into any enquiry

a system of checks, and frequent study of results as they come in, preferably by a person other than the doctors taking part. Inconsistencies in the results or replies can then be looked at, discussed, and evaluated, rather than leaving it all until the end. This of course really means that it is important to institute what is known as " A Pilot " by which such early results are collected. If alterations to the questionnaire or definitions are necessary, it will be possible to do so without invalidating the whole enquiry, and frequently the pilot can be added to the full enquiry.

In this connection it is important to stress the trap which is laid for the unwary if the reason why certain cases or patients are withdrawn or excluded from a trial is not recorded. It is quite simple, particularly in a drug trial, to withdraw all the most severe cases from a given series of treatment, without realizing this unless the withdrawn or excluded cases are recorded and looked at when evaluating the success or failure, or extent of the treatment.

The College practice of having a recorder for group investigations is, in itself, a possible pitfall if the " bias " and interest of this co-ordinating doctor is so predominating as subconsciously to influence his participation. It should be seriously considered whether in certain circumstances the recorder should not abstain from taking part in an investigation, but act solely as the co-ordinator and finally see the results analysed and evaluated.

The biggest pitfall of course is prepared by nature herself, who likes to play jokes on the medical profession with the greatest regularity. We tend to overlook the natural history of development or degeneration.

Learned observations on greying hair or baldness because we are no longer wearing top hats, without due consideration of increasing age, would be an example.

The research committee of our College, with its advisory function, constitutes a collecting point of the knowledge and awareness of these pitfalls, quite unrivalled elsewhere. Thus it will help you to gain from what has defeated others, by utilizing the research committee's accumulated and ever-growing experience.