Editorial

Playing with Statistics *or* Lies, Damn Lies and Statistics

Before statistics, physicians depended upon anecdotes. An anecdote is defined as "a secret or hitherto unpublished narrative or details of history," or "the narrative of an interesting or striking incident or event". We did not know that this was an evil, only used by simple-minded doctors too ignorant to know that anecdotes, even if interesting and historically correct, are an abomination, a sin against medical science. Critics of anecdotes apparently believe that anecdotes have contributed nothing to medical science and that all the discoveries which form the basis of modern medicine sprang full blown from the heads of a few who had never sinned by using anecdotes. No one should pay any attention to the works of the great anecdotalists such as Sir William Osler whose popularity and stature depended on anecdotes. He probably thought they were histories.

To replace the traditional careful histories or medical anecdotes that created modern medicine, the new experimentalists want to use only double-blind controlled experiments. A careful examination of even the most avid double-blind enthusiasts show they are not against case histories when they practise medicine or psychiatry but they are violently opposed to anecdotes. They want to extract bits of information using scales rather than histories and probability tests rather than using one's judgment. In some modern clinical accounts of therapeutic trials no patient is evident, only statistics, and one might wonder what species of animal was being tested.

But – and this may surprise the readers of this journal – the double-blind method has never been validated. It is a method universally used which has not even passed the first test of any new method: does it do what it is intended to do? In science, new methods, new diagnostic procedures, must be evaluated very carefully to meet two main criteria. First, they must be reliable. This means that anyone using the same method will come up with the same answer, like a ruler which should yield the same information regardless of who uses it to measure length. According to Sir Lancelot Hogben,[1] a highly respected statistician, it is impossible to use small samples and hope that they will truly represent the population from which they are drawn. Even with large scale studies this is impossible. For example in the large sample Finnish study on beta carotene and lung cancer,[2] presented as a double-blind, the group of heavy smoking heavy drinking males who were given beta carotene had been smoking one year longer than any of the other three comparison groups. The authors ignored this defect in their design, even though physicians know that the longer a person smokes the greater the incidence of lung cancer. By ignoring this they were able to leave the impression, without actually saying so, that beta carotene had increased the incidence of lung cancer in this group. They concluded that the differences were insignificant but made no attempt to release a formal statement to counter the world press, who ignored the study's conclusion and drew their own that beta carotene was toxic and could cause lung cancer. The difference was slight. It accounted for only one new case of lung cancer (out of one thousand). But, ever since, reports dealing with the Finland study conclude that beta carotene increased the risk of getting lung cancer, and it has even been recommended that beta carotene be taken of the market because it is a carcinogen.

The first criterion has not been examined experimentally but we know from empirical studies that different investigators using the same method arrive at different conclusions. It is impossible to ensure that exactly the same type of patients are enrolled in the double-blind therapeutic trial. But the failure to test the second proposition is much more serious. The second criterion is the validity, i.e. does it do

what it is intended to do? In this case does the double-blind study really remove bias from medical practise, and does it therefore really tell us the therapeutic value of any new or even old treatment?

No one as far as I know, and I have challenged many to show me one published paper, has tested the second principle, does the double-blind method remove bias? I have examined this issue.[3,4,5] In fact it does not, and in addition has many inherent defects which make it only one of the methods available if used very carefully and it should never be depended upon as the sole method. It is not ethical since it forces the treating physician to lie to the patients. If he does not know what the patients are getting, if it is truly double blind, how can he discuss what that drug or compound will do for the patients? Some claim that this problem can be avoided by using different doses of the same compound, i.e. using a very small dose which could not possibly be effective against a large dose, but the same criticism applies, for the doctor must fool the patient into believing that a dose is being used which he knows can be of no help, while letting the patient believe it is being used for therapy.

Even if the double blind were ethical it is almost impossible to carry out. Very few double-blind experiments are really double-blind. It is too difficult to carry on since patients and ancillary staff are always keen on breaking the code and in most cases will do so. Many years ago a physician told me about his double-blind experiment giving intravenous valium to alcoholics. He thought it was great and then added that the nurses giving the injection knew immediately what they were getting because the valium in the syringe swirled differently than did the placebo liquid.

The Montreal studies on vitamin B_3 and schizophrenia were heralded as double-blind experiments using niacin and placebo. But it is impossible to double-blind niacin until someone discovers how to prevent the flush of niacin. Yet these studies are claimed to be models of clinical research and to have been an attempt to repeat our earlier experiments. We used niacinamide as a hidden control when comparing this vitamin against placebo. Six of our experiments, the first in modern psychiatry,[6] were truly double-blind. I doubt there have even been any other double blind studies using niacin. Some double-blind method supporters demand that in every double-blind experiment there must be a method for proving that the experiment was really double-blind.

Doing the clinical trial double blind does not ensure that the experiment will be done adequately. No experiment will be perfect since new information accrues steadily but double blinds should at least use the information already known in designing their experiment. This is seldom done. The Finland study was so badly done it was severely criticized. J. Challem[7] found many errors in design, in using the right compounds and in controlling for smoking and drinking.

The results of double-blind experiments are not known until the clinical trial is completed and the code is broken, i.e. someone looks up the secret code and then can separate those who were on placebo from those who were on the drug. This is done after the final clinical evaluation of each patient so that the evaluator will not know what they were getting. Occasionally the experiment will be terminated sooner for cause. The first double-blind we did was on a yeast nucleotide preparation for which had been claimed to be therapeutic for schizophrenia. But after we completed the test on a large sample we did not see even one patient improve. Since we would expect about half the patients to be on the active compound and half on placebo we concluded that there was no activity. If one quarter or even one tenth of the group had shown a response we would not have terminated

the trial. We therefore advised our statistical advisor, Dr. Bud Fisher in Ottawa, of this and suggested we terminate. He agreed.

April 7, 1998, *The Globe and Mail* newspaper reported that the tamoxifen preventive trial which has been running for some six years had been terminated early because the were told, "The results of a ground-breaking study on breast cancer prevention were so convincing that researchers halted the trial more than a year ahead of schedule so that more women could benefit." There are only two reasons why a trial should be terminated. The first is that everyone in the trial got better indicating that the treatment was not necessary since placebo was as good. The second is that so few got better that it is clear the drug has no value since it is no better than placebo. Why then did these investigators terminate one year early? How could they possibly know that their treatment group did better until they had decoded and found out who had been on placebo or on drug? I therefore looked again at the summary provided by *The Globe and Mail*. (Figure 1, below)

Out of about 6,500 women on placebo there were 154 women who developed breast cancer. I think the study had been running about six years. This means that the probability of the women they selected for the study was 154 divided by the total number or 2.4%. From the 6,500 on Tamoxifen 85 developed breast cancer. The probability here was 1.3%. Thus a woman taking tamoxifen over the period of this study would decrease her chance of getting breast cancer from 2.4 to 1.3%. But to make what appears to be a very small clinical advantage a major discovery, the investigators divided one probability by another and came up with a much more substantial statistic. If you divide 2.4 by 1.3 the result is 1.84. People reading this statistic could assume that a much large number of women would be benefitted.

Looking at it another way, in 1,000 women from the Tamoxifen group, cancer would develop in two of them each year. From the placebo group the figure is 3.5. If we now add the complications created by Tamoxifen the picture gets even worse. Adding the uterine cancers and blood clots, both equally serious, the difference in the probabilities become narrower. From the Tamoxifen group, two percent developed either breast cancer, uterine cancer or blood clots. From the control group 2.6 percent developed these same conditions.

Using the reasoning of the investigators who reported in *The Globe and Mail*, women on Tamoxifen had 23% greater chance of getting uterine cancer, and 280% greater chance of getting a blood clot with a ten percent chance of dying from it. The investigators were careful not to use the same reasoning with these dangerous complications as they did when they were trying to justify their conclusion of the value of Tamoxifen. A solution is to increase the therapeutic value of the drug

Figure 1. Results of Tamoxifen placebo double-blind six year study.

Treatment	Number in Study	Cancer Breast	Cancer Uterine	Blood Clots	Total
Tamoxifen	6500	85	33	17	135
Placebo	6500	154	14	6	174

and to decrease its toxic side effects. Already there is a race among companies to see who can come out first. Another suggestion thrown out in a recent TV show was to remove the uterus (hysterectomy) from women over 60 since these are at greatest risk of developing uterine cancer. The doctor could point it out to the patient who then could be freed of the danger of uterine cancer. But would she be willing to undergo the risks of surgery and its consequences? This solution would not apply to blood clots since no one will ever suggest that the organ called blood be removed.

Would any person, seeing this data and finding that their risk of getting the disease is minimally reduced while at the same time the risk of side effects is increased, take this drug? I suppose it depends on which condition you would prefer. And one would have to consider other factors such as premature menopause, osteoporosis, and the additional costs of the medication and the necessary checks with their doctor every six months. The investigators claim that it has no effect on osteoporosis. Perhaps in a short study this is true, but suppose they took the drug for ten years. Even now women are advised not to take the drug more than five years. Two of my breast cancer cases developed very severe osteoporosis after they have been on tamoxifen for some time. The only reasons why more cases will not develop is that on standard treatment alone too few women survive more than five years.

Nearly everyone misinterprets the meaning of statistical significance and confuses it with clinical significance. When the sample size is large enough even a very minor clinical difference will become significant statistically. That is because the only test the statistician has to face is the five percent probability rule. Could these results have occurred by chance? Out of 20 trials a probability of 0.05 means that only in one trial out of twenty would these results be obtained by chance alone. The larger the sample the easier it is to obtain these significant statistical figures. But is it clinically relevant when the differences are so slight that any reasonable person looking at the data would remark "so what"? Unfortunately in cancer treatment results are so dismal that only statistics can be depended upon, and if the effects were clinically important we would not need statistical analysis to decide for us what is relevant or not.

The use of biostatistical analysis was questioned by Hogben who pointed out that sampling techniques would not provide samples that truly represent the population. The double blind clinical trial has never been validated. It does not eliminate bias but it does grossly distort the doctor-patient relationship. It has never been needed to produce new classes of therapeutic chemicals such as antibiotics, hormones and vitamins, and very often shows drugs to be of no value when further tests find they have great value. An example is l-dopa for treating Parkinson's disease. The statistic used can be manipulated to provide meaning to results when there is no meaning to them. In my opinion their greatest value is that it allows investigators to provide enormous amounts of data to the authorities so that they can decide on the basis of the 0.05 percent point when they should be released.

I think there is a better method, the method mankind has used for centuries. It is the trial and error method. Our ancestors thousands of years ago must have suffered enormously by sampling various plants and animal species as food. We no longer need to do so since we profit from their excursions into pharmacology. The best way is: (1) To establish that any new compound is safe. This is where governments can come in, to ensure safety and purity and quality. (2) Once the compounds has been shown to be safe it can be distributed to health care providers, MDs, DOs, NDs and chiropractors. They

will be given all the information available.

The final test would be the response of patients of thousands of doctors. Toxicity would be monitored by compulsory reporting of all adverse reactions. For example we are treating disease A. If one thousand doctors started using it and were still using it one year later we can be reasonably sure that it has some therapeutic value. The competition of patients who want to get well, who will shop around and find a doctor who will do the best job for them will eventually determine what is the most effective available compound. If the drug has no value few doctors will still be using it at the end of the year. If it has great value almost all will be using it. The proportion of doctors using it would be a measure of its therapeutic value. The importance of the placebo has been greatly exaggerated. It works best in the short term but I doubt that many people will take placebos month after month. As a rule the placebo effect is most powerful shortly after the therapeutic factor is initiated. I can not recall more than one case when it was still operative one month later.

A. Hoffer M.D., Ph.D., FRCP(C)

References

1. Hogben L. *Statistical Theory: The Relationship of Probability, Credibility and Error.* Allen & Unwin Ltd. London, 1967.
2. The alpha tocopherol, beta carotene cancer prevention study group: The effects of vitamin E and beta carotene on the incidence of lung cancer and other cancers in male smokers. *N Engl J Med.* 1994; 330: 1029-1035.
3. Hoffer A & Osmond H: Double blind clinical trials. Neuropath 2:221-227, 1961.
4. Hoffer A: Double blind studies. *Can Med Assoc* J 111:752 only, 1974.
5. Hoffer A, Osmond H: Some problems of stochastic psychiatry. *J Neuropsychiatry,* 5:97-11, 1963.
6. Shapiro AK, Shapiro E: *The Powerful Placebo.* The Johns Hopkins University Press, Baltimore and London, 1997.
   These authors grudgingly admit that we were first but suggest that the definitive paper published in the Menninger Bulletin with the first author, J. Clancy, was under my direction as Director of Psychiatric Research. I was the second coauthor of this paper. (Clancy J, Hoffer A, Lucy J, Osmond H, Smythies J & Stefaniak B: Design and planning in psychiatric research as illustrated by the Weyburn Chronic Nucleotide Project. Bull Men Clinic, 18:147-153, 1954). –AH
7. Challem J: Alcohol, Synthetic Beta-Carotene, and Hasty Conclusions may have made for Study Fiasco. *The Nutrition Reporter.* Vol 8, Jan 1997.