Comments on the Analysis of the Burlington Randomized Trial

J. B. GARNER*
Research Associate, School of Nursing
McGill University

In view of the continuing importance of the large-scale study described as "The Burlington randomized trial of the nurse practitioner" (Sackett 1974, Spitzer 1974), it is worthwhile discussing some of the difficulties in interpreting the published account of the statistical analysis of this study. These comments are offered for the purpose of clarification and in no way are meant to minimize the considerable achievement of this study.

The purpose of a statistical significance test is to assess the numerical evidence for the inclusion of a new parameter. Following the principle of parsimony of Occam's Razor, which may be taken in the form "entities are not multiplied without necessity" (Jeffreys 1961), one commences with the situation in which all the variation is treated as random. Parameters are introduced one by one which appear to explain a sufficient amount of the variation to merit their inclusion and the remaining variation is treated as random. This latter part never completely disappears. At any stage in this process of successively introducing fresh parameters the present status quo may be described as the "null hypothesis". This is the "working rule" presently achieved. The "alternative hypothesis" next considered usually contains one additional, or new, parameter or, one additional degree of complexity. The evidence for the inclusion of this new parameter is sometimes simply assessed by considering the ratio of the estimate of this new parameter to the standard error of the estimate. However the assessment of the evidence is made, the "null hypothesis" always describes the present working rule and the alternative hypothesis describes a rule of greater complexity, the "thesis" or proposition which the experimenter has in mind whilst designing the experiment. Often the main purpose of the experiment is to compare the experimenters' thesis with the present working rule. If the sample size is very small the estimate of the standard error is often too vague to be able to reject the null hypothesis. When the same size is very large the evidence for the introduction of the new parameter may be

*In discussion with Professor Moyra Allen and M.Sc.(A) students Kiyoko Matsuno, Inez Blackwell and Rosaline Wosu in Nursing 615b, Health Care Education.

overwhelming even if the magnitude of the parameter is too small for any practical purpose in everyday use.

One of the problems in applying classical statistical methodology, as described above, to numerical, replicable observations made in practical nursing situations is that often the present working rule stipulates, or presumes, the existence of a parameter, and the experimenter would like to show that the magnitude of this parameter is less than that presently assumed or, alternatively, that for "practical purposes" its presence may be ignored. Consider a very simple hypothetical example. Suppose that for the purpose of dressing a certain type of surgical wound on patients, the rules of procedure require a nurse to scrub her hands for three minutes and that nurses with considerable experience with this procedure are not at all convinced that three minutes of scrubbing is necessary. It would probably be accepted that some scrubbing is necessary, but, given general state of hygiene prevailing nowadays and the fact that the nurse's hands touch neither the dressing nor the wound, it might be worthwhile to set up a carefully controlled experiment in which varying lengths of scrubbing times were related to the rate of infection of the wounds (taking into account the several concomitant variables). The outcome might be a recommendation that a careful ten-second scrubbing was as effective as an unsupervised three-minute scrubbing. In such a situation the working rule, or null hypothesis, is already assuming that scrubbing has a beneficial effect; the purpose of the experiment is to attempt to measure the magnitude of this effect more carefully. There is no question of any significance test — merely one of more precise estimation.

Consider the Burlington randomized trial on Nurse Practitioners. The problem was to assess the comparative effectiveness (on certain measures) of nurse practitioners (NP) and family physicians (FP) in providing primary health service in a certain situation. Suppose that at the time the trial commenced a Martian had arrived on Earth without prior knowledge of the effectiveness of either NP or FP. Then presumably the Martian would have followed the classical statistical procedure outlined earlier. As a working rule, or null hypothesis, the Martian would have assumed no difference between the two populations, NP or FP, in terms of the measures used. The alternative hypothesis would have been that some effect existed, that is, on a given measure the mean of the NP would have been greater (or less) than that of the FP. If the sample size taken by the Martian were very small, it is highly likely that the Martian would have continued to use the null hypothesis of no difference until further

evidence was available. This is satisfactory since the Martian commenced with the null hypothesis of no difference; the Martian should continue with this belief, or state of mind, until sufficient evidence accrues to the contrary. For a person in the Martian's situation, the classical statistical approach is satisfactory.

However, the state of mind of the Martian at the beginning of the trial was not the state of mind of the man in the street, the medical profession, the government, or even, perhaps, of many nurses themselves. In support of this statement we need do no more than consider the governmental expenditure on the Burlington trial itself. If a null hypothesis of no difference was taken a priori then this expenditure could not have been justifiable since what is the purpose of running a long and costly experiment in order to come up with evidence supporting a (null) hypothesis which is generally accepted and unchallenged anyway. For whatever reasons, the FP had already won for themselves an acceptance that, in general, their primary health treatment was more effective than the treatment given by any other group of persons. It is clear therefore that the null hypothesis or working rule under which the Burlington trial commenced was that a definite effect exist and that the FP were at least as good as the NP on all of the measures suggested and superior on some.

The statistical formulation followed in the Burlington study may be summarized by the following quotation: "Since it was our thesis that the outcomes of RNP (nurse practioner) care would be equivalent to those resulting from RC (family physician) care, the hypothesis that the RNP care was effective and safe would be supported if no statistically significant differences could be shown between the outcomes of the RNP and RC groups" (Sackett). The thesis, or alternative hypothesis, stated by the experimenters in the first phrase is one of no difference, or, that "no effect" existed. It will be seen at once that this formulation of the problem does not follow the classical procedure, since the null hypothesis, "an effect exists", is being compared with a less complex alternative hypothesis "no effect exists". The first part of the second phrase, "the hypothesis that the RNP care was effective and safe" is not unambiguous but appears to suggest that under the alternative hypothesis there is either no difference or the NP population has higher (more beneficial) outcome measures than the FP population. Let us call this "the extended alternative hypothesis". *The final part of the quotation suggests that the alternative hypothesis of "no difference" was supported if test statistics constructed on the basis of a null hypothesis of "no difference" were sufficiently small.* The importance of clearly stating the

null and alternative hypotheses before using such a phrase as "statistically significant differences" may be exemplified by considering that an observed (small but non-zero) sample difference which may support a hypothesis of no population difference may even more strongly support a hypothesis of a stated population difference (consider the hypothesis which states that the population difference is precisely that observed in the sample). Furthermore, using the criterion of "no statistically significant difference" between the two collections of outcomes could have led to the experimental thesis being confirmed merely because the sample size was too small relative to the magnitude of the real difference. Despite the careful statement of their thesis by the authors, all subsequent language in their paper appears to indicate that they are using "no difference" (or even the extended alternative hypothesis) as their null hypothesis and a difference favouring the FP as their alternative hypothesis. It would appear that, more correctly, the Burlington problem was one of estimating more precisely the magnitude of an effect which was accepted a priori as being present.

If the magnitude of the difference appears to be sufficiently small then it may be justifiable to consider all possible expenses and benefits to both the patients and to the government, converted to some common unit of "utility", and question whether the government should continue to spend the extra amount necessary to train FP rather than NP for the purpose of providing primary health service.

There are other parts of the statistical analysis which appear open for discussion. The analysis presented in Sackett uses the notion of "the 'beta' level of the test of significance". Beta is the probability of not rejecting the null hypothesis when it is false. In the context of the trial, beta may be interpreted as the probability that on a given measure the outcomes of the two groups differ by less than some multiple of the standard error of the difference given that the measures of the two groups truly differ. The experimenters state that in their evaluation of beta they stipulated "a true difference between the groups of 5% or more". It is not apparent in general on which base the percentage has been taken. However, the statement appears unambiguous in the situation of the comparison of two proportions, which is discussed in Appendix 1. From the analysis given there it will be seen that any calculation of beta requires a statement of the a priori probability distribution of the underlying population parameters and of the number of standard deviations the estimated difference is allowed to vary away from zero. However, if the problem is reformulated as suggested at the close of the previous but one

paragraph, then the introduction of the notion of beta would be superfluous.

The measure of social function (Sackett: Fig. 2) appears to have a very curiously-shaped histogram; even though on such a large sample the approximate normality of the arithmetic mean may be justifiable, the use of this mean for the purpose of comparison of the two histograms is questionable. It is very curious that the histograms for both samples show zero readings in the intervals 0-1 — 0-2 and 0-5 — 0-7 and it would have been very instructive to have been able to examine a histogram resulting from a third collection of patients taken from outside the study group since the shape of the histogram may be a function of the instrument. In fact, what is striking about the results presented in Fig. 1, Fig. 2 and Table 5 of Sackett's paper is not the differences but the great similarities between the two collections of outcomes. Perhaps the careful selection of matched groups had led to less variation between the two groups than might be reasonably expected at random.

The final point that should be made is that of the variability in effectiveness and safety in primary health services provided by members of the FP population themselves. It is quite possible that if two FP randomly selected from the population of FP were compared in circumstances similar to that of the Burlington trial, apparent differences would occur. The same statement could be made about random selections taken from the population of NP. Once this point is recognized and accepted it becomes clear that if the results of the Burlington trial are to have applicability or relevence to the comparative effectiveness of the populations of NP and FP then an assessment is required of the variability of both these populations in circumstances similar to that of the study. Some analysis of this point is given in Appendix 2.

ACKNOWLEDGEMENT

The author wishes to acknowledge helpful discussions with Moyra Allen and Kiyoko Matsuno.

REFERENCES:

Sackett, D. L. Spitzer, W. O., Gent, M. and Roberts, R. S. The Burlington randomized trial of the nurse practitioner: Health outcomes of patients. *Annals of Internal Medicine* 80:137-142, 1974.

Spitzer, W. O., Sackett, D. L., Sibley, J. C., Roberts, R. S, Gent, M., Kergin, D. J., Hackett B. C., and Olynich, A. The Burlington randomized trial of the nurse practioner. *New England Journal of Medicine* 290:251-256, 1974.

Jeffreys, H. *Theory of Probability.* Oxford: Oxford University Press, 1961.

Appendix 1
The Evaluation of 'Beta' When Comparing Two Proportions

Suppose the first population has a true proportion of satisfied clients p_1, observed sample size n_1 and an observed number of satisfied clients x_1, with a similar notation for the second population. Assuming binomial sampling the probability that a pair of values (x_1, x_2) will occur knowing n_1, n_2 and the true values p_1, p_2 is

$$f(x_1,x_2|p_1,p_2) = \binom{n_1}{x_1}\binom{n_2}{x_2} p_1^{x_1}(1-p_1)^{n_1-x_1} p_2^{x_2}(1-p_2)^{n_2-x_2} \qquad [A]$$

$$x_1 = 0,1,2,\ldots,n_1 \; ;$$
$$x_2 = 0,1,2,\ldots,n_2 \; .$$

Let N denote $n_1 + n_2$ and m denote $x_1 + x_2$ then for large samples the customary chi-square test suggests that the null hypothesis $p_1 = p_2$ is *not* rejected if

$$N\{x_1(n_2-x_2) - x_2(n_1-x_1)\}^2 < (N-m)m\, n_1 n_2\, K_\alpha \qquad [B]$$

where K_α is a specified constant dependent upon the selected level of significance, alpha. [If alpha = 0.05 then $K_\alpha = 3.84$].

The *usual* definition of beta is the probability of occurrence of a pair of values (x_1,x_2) satisfying inequality B when $p_1 \neq p_2$. If p_1, p_2 are known the probability of any pair (x_1,x_2) occurring is given by expression A. Therefore the value of beta is found by calculating the expression A for each pair (x_1,x_2) satisfying B and then summing the values of these expressions. Thus, symbolically, if p_1 and p_2 are known and unequal

$$\text{beta} = \sum_B f(x_1,x_2|p_1, p_2) \qquad [C]$$

where the summation is over all pairs (x_1 x_2) satisfying B. Beta is the probability of coming to the conclusion that $p_1 = p_2$ is satisfactory when in truth $p_1 \neq p_2$. The closer beta is to zero the less likely we are to make this type of false inference. Unfortunately in general and for the case of the Burlington trial in particular, the true values of p_1 and p_2 are not known and so expression C *cannot be found.* Usually statisticians attempt to make informed guesses about the population of interest based on the information available a priori and in this manner calculate rough values for beta enabling them to suggest useful sample sizes for survey samples and experimental designs. For a precise value of beta we are able to calculate only the (weighted) average value for beta found by averaging the value of beta given by expression C over all the possible values of p_1 and p_2 as follows. Let

$\pi(p_1,p_2)$ denote the a priori probability (density function) of the values (p_1,p_2) when the hypothesis $p_1 \neq p_2$ holds, then

$$\text{beta} = \iint \{\sum_{B} f(x_1,x_2|p_1,p_2)\} \quad \pi(p_1,p_2)dp_1 \quad dp_2 \qquad [D]$$

It will be seen that the calculation of a precise numerical value for beta requires the value K_α and the form of the test, as given by expression B, and either the true values p_1 and p_2 or statement (and justification) of the a priori probability $\pi(p_1,p_2)$. Neither of the latter are given in Sackett (1974) and so a reproduction of the published beta values is not possible. It is further stated in Sackett (1974) that the evaluation of beta stipulated 'a true difference between the groups of 5% or more'. That would imply that the double integral and the a priori probability $\pi(p_1,p_2)$ in expression D were both taken over only the region $p_1 > p_2 + 0.05$ (since the test is one-sided). This would make the calculation of beta even more difficult and it is possible that this quotation is an editorial misprint for a statement to the effect that the level of significance of the test of the hypothesis $p_1 = p_2$ was taken as 5% (so that $K_\alpha = 3.84$).

Appendix 2

Let us consider one particular measure of client satisfaction, and let x_{ij} denote the value of this measure recorded for the jth client of the ith nurse. If we could find the value x_{ij} for all possible clients of the ith nurse and draw up a frequency table of the results then the mean, μ_i, of this frequency table is the true mean of the values of this measure of client satisfaction for the ith nurse and the standard deviation $\acute{o}_i$ of this frequency table would be the standard deviation of the possible values of this measure of client satisfaction for the ith nurse. Now in reality all we have is a sample x_{i_1}, x_{i_2}, . . . , x_{i_n} which may be considered as being drawn from this hypothethical frequency table of all possible values for the ith nurse. The *average* of the sample values, $\bar{x}_i$, is taken as an *estimate* of the *true mean* μ_i and the *standard deviation of the sample values,* s_i, is taken as an *estimate* of the *true standard deviation* $\acute{o}_i$ of all possible values of this measure for the ith nurse.

If we consider another nurse (nurse k) then the (hypothetical) frequency table of the values x_{kj} of our measure of client satisfaction for the population of all possible clients of the kth nurse will have a true mean, labelled μ_k. For each member of the population of nurses

under consideration we have, in our imagination, a value μ (μ_i for the ith nurse, μ_k for the kth nurse and so on) which measures that particular nurse's mean value of our measure of client satisfaction. Now imagine each nurse having her/his particular μ value. This frequency table of the true mean values for each and every nurse has a mean value ξ and a standard deviation τ. The value ξ represents the mean value of all the true mean values μ (one μ from each nurse). Hence if the particular nurse who was measured in the Burlington trials (labelling her as nurse i once more) could be assumed to be typical of all nurses (theoretically, if she was chosen at random from all possible nurses in the population of nurses under consideration) then her true mean μ_i should be representative of the value ξ in the sense that μ_i is an estimate for ξ (albeit possibly a highly variable estimate). Of course the true value μ_i is not available but we have an estimate of μ_i, namely $\bar{x}_i$, so if the ith nurse is typical then her actual average value $\bar{x}_i$ estimates ξ, the true mean of all possible nurses if they were in her position. However, in order to be able to estimate how much this estimate varies around ξ and how much nurses may vary one to another with respect to their μ values we have to be able to estimate τ. But we have only *one* value (or, strictly, only an estimate of one value) taken from this frequency table of all possible μ values, namely μ_i estimated by $\bar{x}_i$. Therefore, there is no way we are able to estimate the *variability* of the μ values from nurse to nurse *based on the sample evidence alone.*

Commentaires sur l'analyse de l'essai aléatoire de Burlington

Bon nombre des traditions et coutumes bien établies de la profession infirmière sont petit à petit soumises à une évaluation quantitative à l'aide des méthodes d'inférence statistique. Il faut souvent dans ces évaluations trouver l'équivalence entre un traitement nouveau (ou une pratique nouvelle) et un traitement courant (ou une pratique courante). Il n'existe pas encore de technique statistique type pour entreprendre une telle évaluation. Cet article mentionne une fois de plus le cadre classique de l'hypothèse nulle par opposition à l'hypothèse de remplacement; on y montre que, dans un effort pour résoudre le problème déductif de vérification de l'équivalence, l'analyse statistique de l'étude de Burlington a pris comme hypothèse nulle les facteurs ce qu'elle voulait prouver comme hypothèse de remplacement. L'article traite également d'autres problèmes inhérents à l'analyse.