SUSPENDED JUDGMENT
n-of-1 Trials

Stephen Senn, PhD

Medical Department, CIBA-GEIGY Ltd., Basel, Switzerland

> *. . . nor has rhubarb always proved a purge, or opium a soporific to everyone who has taken these medicines.*
>
> David Hume, *An Enquiry Concerning Human Understanding*

Scratch a standard statistical procedure and you will find an additivity assumption, whether expressed in terms of means, log-odds, or some other statistical device. For the statistician either no patient benefits or each does to the same extent: even if, for this to be true, the benefit may have to be measured in terms of some complicated function of the probability of being cured. The physician, however, likes nothing better than to divide the world into responders and nonresponders: his opinion as to which drugs are useful in treating patients is "none for all and all for some." Clearly, any reasonable commentary must concede that the truth lies somewhere between these extremes. It is this that makes single subject or *n*-of-1 trials both interesting and potentially misleading. For if it really is the case that averages calculated from groups cannot be applied with confidence to the individual, then for the purpose of prescribing for that individual there will be value in studying that person's response; but by the same token the results of such individual study will be of little value for patients as a whole [1].

Recently, in this journal, it has been suggested that *n*-of-1 trials "have immense potential for use in the early phases of drug development programs" [2 p. 88]. Elsewhere the single-subject trial has been praised because it "may provide new insight into vaguely defined conditions, improve therapeutic decisions, strengthen the doctor–patient relationship and create a more critical attitude toward drug treatment both among patients and doctors" [3 p. 174], These are strong recommendations but they are not being greeted with universal assent [1]. In my opinion, there are good reasons for caution.

First, we may note in passing that *n*-of-1 trials can be linked to many of the most controversial issues in clinical trials. If we are going to carry out series of *n*-of-1 trials, then presumably we need methods for the sequential

Address reprint requests to: Stephen Senn, PhD, Medical Department, CIBA-GEIGY Ltd, CH 4002 Basel, Switzerland.

Received December 23, 1991; revised May 20, 1992

© Elsevier Science Publishing Co., Inc. 1993

0197-2456/93/$6.00
655 Avenue of the Americas, New York, New York 10010

analysis and meta-analysis of such trials. Since by definition they involve repeat administrations in one patient, they may in principle be affected by the carry-over of effect from one replication to the next. Since they presuppose the possibility of patient-by-treatment interactions, their analysis may have to be based on models with random effects. Since they may be used to help decide on the prescription for an individual patient, then even the diehard frequentist might concede that a Bayesian analysis could be considered [4]. These are issues on which even the wise have disagreed.

Such trials may be controversial but they are not new. The first example in the first book written on the statistical approach to designing experiments is an *n*-of-1 trial: Fisher's example of the lady and the cups of tea. "The subject has been told in advance of what the test will consist, namely that she will be asked to taste eight cups, that these shall be four of each kind, and that they shall be presented to her in random order. . . . Her task is to divide the eight cups into two sets of four, agreeing if possible with the treatments received" [5, p. 11]. For a trial where treatments were repeatedly administered to individual patients to estimate the effect of these treatments on patients in general we have an even earlier example: Cushny and Peebles' [6] investigation of the soporific effects of optical isomers, subsequently quoted (incorrectly) by Student [7] in the paper that introduced the *t* distribution. (See Preece [8] for an interesting discussion of these data.)

Modern commentators have not always matched Fisher's perspicacity regarding masking of treatments. A common mistake (not limited to *n*-of-1 trials but particularly serious where these are undertaken) is to presume that the degree of masking can exceed the richness of the randomization (i.e., the number of possible sequences that the randomization may produce). Given knowledge of the randomization procedure adopted, the probability of guessing all treatments correctly cannot be lower than the reciprocal of the number of possible sequences. In the tea tasting, the lady's probability of guessing all cups correctly is determined by the number of ways in which eight objects can be divided into two groups of four; since there are 70 such divisions, her probability of a perfectly successful guess is 1/70. Had the randomization simply consisted in choosing one of the two sequences ABABABAB or BABABABA at the toss of a coin, the probability of guessing successfully would be 1/2. Fisher wisely eschews the possibility of attempting to deceive the lady about the randomization employed, thus avoiding the problem of having to wonder whether he is cleverer than his subject and has outbluffed her; on the contrary, he explains the procedure to her. I note in passing that if the same standards of honesty and acumen were applied generally in clinical trials, we wouldn't have the nonsense of *placebo* run-ins.

If masking is considered essential, then it may impose restrictions on the analyses used. Suppose, in an *n*-of-1 trial, that an investigator informs his patient that he will allocate him at random an active treatment, or verum (V) three times and placebo (P) three times. Suppose that the randomization produces the sequence VVPVPP with results 2.94, 2.99, 2.04, 3.11, 2.12, 2.05. A two-sample *t* test will yield the *t* value 16.7, which even with 4 degrees of freedom yields a $p << .0001$. But if the patient, just by chance, produces three high (and similar) responses and three low (and similar responses), reflecting his prejudice about treatment and his hunches about allocation,

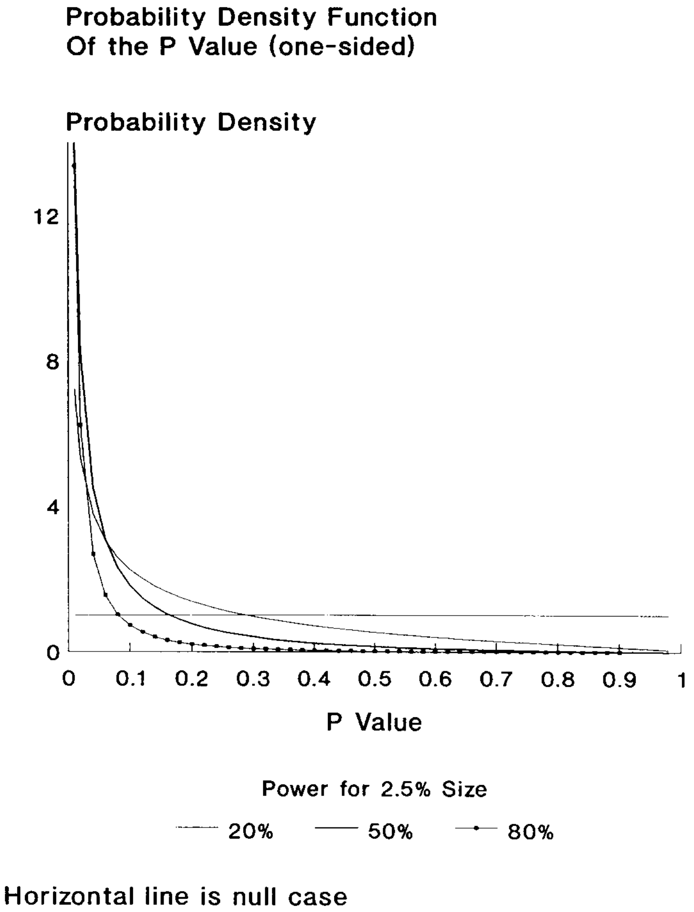

Figure 1 Probability Density Function of the P Value (one-sided)

then the probability that he will produce an extremely impressive *P* value is simply the probability that his hunch is right and this is the probability of dividing six objects correctly into two groups of three, which is 0.05, a rather less impressive result than that given by the *t* test.

There also seems to be a widespread confusion as to the way in which the results from individual *n*-of-1 trials may be interpreted. Peering at *P* values is not particularly edifying but if it is to be done, something about their distribution should be understood. Where treatment effects are additive such distributions are not bell-shaped, as has been claimed in the medical literature [9], but monotonic. In general, where the test statistic has a normal distribution with standard error θ, then a one-sided *P* value, γ, calculated under a null hypothesis, H_0, that the true treatment effect τ is zero (and given the alternative hypothesis that $\tau > 0$) has a probability density function $f(\gamma; \tau, \theta)$, given by

$$f(\gamma; \tau, \theta) = \phi\{\Phi^{-1}(\gamma) + \tau/\theta\}/\phi\{\Phi^{-1}(\gamma)\} \tag{1}$$

where $\phi(z)$ is the probability density function (pdf) of the standard normal distribution and $\Phi^{-1}(w)$ is the inverse cumulative density function, or probit. Clearly, since $f(\gamma; 0, \theta) = 1$, then under H_0 the distribution is uniform. The accompanying figure shows the distribution of γ not only under H_0 but also given values of τ/θ of 1.12, 1.96, and 2.80, which correspond to 20, 50, and 80% power for a one-tailed test with $\alpha = 0.025$. Where the treatment effects are not constant, the pdf of the P value may be obtained by treating τ as random in (1) and applying a mixing distribution.

The message would not need stating were it not widely ignored: even if there is a treatment effect that is constant for all patients, some n-of-1 trials will show "significance" and some will not. A mixture of significant and nonsignificant results is no proof in itself of the heterogeneity of treatment effects. The graphs, of course, show the expected distributions of P values. In practice we have random variation to deal with as well and this should impose a further degree of caution before we rush to conclude that because we see a lumpy distribution of P values we have a mixture of responders and nonresponders—a mistake that is commonly made.

Furthermore, where our objective is to screen drugs for effectiveness and we fear that patients will respond differently, the solution does not lie in the naive use of series of n-of-1 trials. We run a serious risk of being deceived. A better approach is to undertake a conventional trial but use a method of analysis that makes specific allowance for lack of additivity of treatment effects [10]. If, on the other hand, we wish to use an individual's results to help determine our prescription policy for him, what we need is not only to measure his response but to have designed and undertaken the sort of studies that can effectively resolve the variation observed into its various components. A multiperiod, multipatient crossover is a suitable candidate for a design that allows efficient estimation of treatment by patient interaction. We may perform our analyses on the original or a suitably transformed scale of measurements but the P value transformation is not likely to prove an attractive option. Then, using an appropriate random effects model, we can combine the information a given individual provides with that from other patients to make effective recommendations. How much we lean on individual patient results and how much we discount them using the results of others should depend on the relative importance of treatment by patient interaction and the precision with which it has been possible to measure the various elements of the model [4]. Given an infinity of observations on a given patient, the experience of others is irrelevant: in practice such information will be useful.

The above argument does not mean that individual n-of-1 trials should never be carried out, but simply that best use can be made of their results if they are interpreted against a background of adequate research. Indeed, even single measurements on patients can be used with success given that conventional trials have been performed. For example, Racine and Dubois [11] in a Bayesian analysis showed how with the help of a random effects model appropriate use could be made of group results from a clinical trial and individual measurements to adjust the dose of carbamazepine given to an individual patient. Similar uses could be made of n-of-1 trials, but as Lewis [1] noted, this sort of application is most appropriate to the later stage of drug development. This, *pace* Guyatt et al [2], is my view also. *N*-of-1 trials should be used in the maturity of a drug development and not in its childhood or adolescence: to these ages of growth belong more conventional clinical trials.

REFERENCES

1. Lewis J: Controlled trials in single subjects. 2. Limitations of use. Br Med J 303:175–176, 1991
2. Guyatt GH, Heyting A, Jaeschke R, et al: *n* of 1 randomized trials for investigating new drugs. Controlled Clin Trials 11:88–100, 1990
3. Johannessen T: Controlled trials in single subjects. 1. Value in clinical medicine. Brit Med J, 303:173–174, 1991
4. Senn SJ: Controlled trials in single subjects. Br Med J 303:716–717, 1991
5. Fisher RA: The design of experiments. In Statistical Methods, Experimental Design and Scientific Inference. Bennett, JH, Ed. Oxford: Oxford University Press, 1990
6. Cushny AR, Peebles AR: The action of optical isomers. II. Hyoscines, J Physiol 32:501–510, 1905
7. Student: The probable error of a mean. Biometrika 6:1–25, 1908
8. Preece DA: t is for trouble (and textbooks): A critique of some examples of the paired-samples t-test. The Statistician 31:169–195, 1982
9. Johannessen T, Fosstevedt D, Petersen H: Combined single subject trials. Scand J Prim Health Care 9:23–27, 1991
10. Connover WJ, Salsburg DS: Locally most powerful tests for detecting treatment effects when only a subset of patients can be expected to "respond" to treatment. Biometrics 44:188–196, 1988
11. Racine A, Dubois J-P: Predicting the range of carbamazapine concentrations in patients with epilepsy. Stat Med 8:1327–1338, 1989