Melatonin secretion and anorexia nervosa – a serious type II error

SIR: The paper by Bearn *et al* (*Journal*, March 1988, **152**, 372–376) claims an unwarranted certainty about a negative result. This is an error that recurs time and time again in the medical literature (Rothman, 1978; Freiman *et al*, 1978; Reed & Slaichert, 1981) and it cannot be allowed to stand unchallenged.

The authors attempted to "clarify a controversy as to whether melatonin secretion is related to body weight". Their summary says: "There was no change in aMT6s excretion after weight gain"; and the discussion states: "We conclude that weight gain in patients with anorexia nervosa is not associated with significant changes in aMT6s excretion". Neither this conclusion nor the remarkable assertion of "no change" is supported by the data which they show in their Table I. Among the ten patients there was a mean rise in secretion of 4.026 μg/24 h, which is not statistically significant at $P = 0.05$. That is to say that it is not so unlikely ($P < 0.05$) that these observations arose by sampling from a large population within which there is *no* association between weight and a MT6s excretion that we should abandon the possibility of no association. This is the roundabout but very sound logic of statistical hypothesis testing. This logic allows the researcher to specify a probability that he or she will *mistakenly* reject the 'null hypothesis' of no association. That probability is the 0.05 'type I error rate'. That applies when the test is found to reject the null hypothesis. The complementary logic when the test fails to reject the null hypothesis is equally sound and has been misconstrued by Dr Bearn *et al*.

When designing a trial, one should specify the strength of an association that one would consider worth finding and then must also specify the affordable probability of failing to find such an effect if it were actually present. This new probability sets the 'type II error rate". Such specification is only possible given a foreknowledge of the population standard deviation and, in a first report of this type, is an unrealistic counsel of perfection. Because of this difficulty with formal hypothesis testing the *British Medical Journal* now insists on the presentation of confidence intervals for all results (Langman, 1986; Gardner & Altman, 1986). Confidence intervals provide an indication of size of an effect that might have been missed. If they embrace equality between the means of each group then the possibility of no difference has not been excluded – equivalent to a negative result from the *t*-test. The width of the 95% confidence interval for the data of Dr Bearn *et al* is from −2.18 to +10.24 μg/24 h. It can now be seen that a very large positive association could have been missed (the true difference will lie within the 95% confidence interval in 95% of such tests). This results from the size of the sample and the large variance in the results. With small numbers of cases the confidence intervals themselves are imprecise but their import cannot be ignored. Inspection of the table by eye reveals that one patient showed a rise from 3.5 to 30.35. This is an extreme outlier and the range of differences when this patient was excluded was from −1.64 to +6.46. Even excluding this case, the change is not significant ($t(8) = 1.47$; $P = 0.18$) but the confidence interval falls to a range from −0.84 to +3.82.

A valid summary of the results would be that they show one remarkable outlier who increased her aMT6s excretion by over 8 times between the two readings, while the remaining nine patients showed results compatible with a mean change between a 1 μg loss and a 4 μg gain. This is a much less satisfying and elegant conclusion than the one offered by Dr Bearn *et al*, but it clearly summarises the data more accurately. Given the statement in the paper that aMT6s excretion shows considerable variation between individuals but little within individuals, these results suggest that weight may well affect excretion, but that that effect itself may show marked inter-individual variation. This criticism should not be taken as an attack on what was an otherwise very impressive and interesting study, merely on the analysis and interpretation. I hope the authors will conduct a study of a larger number of subjects in which excretion is measured before and after weight gain and again after a similar period of time at the sustained weight.

C. EVANS

St George's Hospital Medical School
Cranmer Terrace
London SW17 0RE

References

FREIMAN, J. A., CHALMERS, T. C., SMITH, H. Jr & KUEBLER, R. R. (1978) The importance of beta, the type II error and sample size in the design and interpretation of the randomized control trial. *New England Journal of Medicine*, **299**, 690–694.

GARDNER, M. J. & ALTMAN, D. G. (1986) Confidence intervals rather than *P* values: estimation rather than hypothesis testing. *British Medical Journal*, **292**, 746–750.

LANGMAN, M. J. S. (1986) Towards estimation and confidence intervals. *British Medical Journal*, **292**, 716.

REED, J. F. & SLAICHERT, W. (1981) Statistical proof in inconclusive 'negative' trials. *Archives of Internal Medicine*, **141**, 1307–1310.

ROTHMAN, K. J. (1978) A show of confidence. *New England Journal of Medicine*, **299**, 1362–1363.

https://doi.org/10.1192/bjp.153.1.121 Published online by Cambridge University Press