British Journal of Psychiatry (1992), **161**, 712–724

Correspondence

Editor: Ian Pullen

Contents: Schizophrenia following prenatal exposure to influenza epidemics between 1939 and 1960/Ventricle–brain ratio in schizophrenia/Cost–benefit analysis of the Daily Living Programme/Insight and illusion/The meaning of insight/Benzodiazepines with ECT/ECT for young people/Community care and suicide/$5HT_2$ receptors, hallucinations, and dementia/Clinical use of moclobemide in Kleine-Levin syndrome/Lay theories of schizophrenia/Probing dexfenfluramine/Carbamezepine and episodic dyscontrol/Psychophysiological investigations of patients with unilateral symptoms in the hyperventilation syndrome/Longevity and Down's syndrome/Neuroleptic malignant syndrome/Self videotaping during psychosis.

Schizophrenia following prenatal exposure to influenza epidemics between 1939 and 1960

SIR: I have a couple of questions about the paper by Sham *et al* (*Journal*, April 1992, **160**, 461–466) which I would be grateful if the authors could respond to. The first would be trivial were it not for the magnitude of the effect that depends on the answer; were the logarithms used for the model natural logarithms (i.e. to base e) or to the base 10? In one paragraph the authors state that the estimate for the effect of influenza would increase the number of schizophrenic births by $e^{0.000014}$ per death from influenza, but in the next they say this corresponds to a 1.4% increase in such births per thousand deaths. It seems that either the first figure should read $10^{0.000014}$ or else the second figure should be reduced by a factor of ln (10) to a 0.61% increase per thousand influenza deaths. The subsequent estimates of the actual magnitude of this effect would then also need to be reduced to less than a half, so that the contribution of influenza to the spring excess of births would be around one-sixth. I can find no clue to the answer in the text, although it is clear from the description of the scaled deviance that the authors at least sometimes refer to natural logarithms simply as logarithms.

My second question is: what do the authors regard as being the true significance value for the influenza effect? They begin by testing for an effect individually in ten different analyses with the number of influenza deaths being counted from zero to ten months previously, and in two of these analyses the influenza effect reduces the scaled deviance by 4.98 and 5.33. They take these values as χ^2 statistics with one degree of freedom and say each has a significance of 0.02. However, surely we should either consider that there are two degrees of freedom (the magnitude of the effect and the time interval before it becomes manifest) or we should regard this as an example of multiple testing and apply a relevant correction? They then go on to add the deaths for these two months (the second and third) together and find that this reduces the scaled deviance by 7.46 and claim that with one degree of freedom this has a significance of 0.007. However, if we consider that there are nine ways to choose two consecutive months out of ten, then we can apply a Bonferroni type of correction to this value and obtain a significance value of 0.055. Since the monthly figures are not independent, one could argue that this was rather conservative, but one could also argue that the authors were not constrained to use a two-month window for the effect, but could have used any number of months from one to nine to produce the most impressive result, in which case the correction might be seen as excessively liberal.

It seems to me that given the sophisticated model-fitting carried out it should be possible to devise some method of analysis which could give a more reliable assessment of the evidence that an association between influenza deaths and schizophrenic births exists at all. One analysis that I would consider more robust would be to look for an association between births and the total number of influenza deaths occurring within the previous nine months. From the graph it is clear that there are very marked year by year fluctuations in the number of influenza deaths, and this variability seems so large that it is difficult to imagine that an association of any interest would not be detected by this approach. Does such an association exist, or are we left with a seasonal variation (which is convincingly

demonstrated by this paper) for which we have very little in the way of convincing explanations?

DAVID CURTIS

St Mary's Hospital Medical School
Academic Department of Psychiatry
St Mary's Hospital
Praed Street
London W2 1NY

AUTHORS' REPLY: The first question raised by Dr Curtis is concerned with the interpretation of the parameter estimates of the log-linear model. We did not make explicit in our paper that natural logarithm (base e), rather than common logarithm (base 10), was used. On the other hand, the use of natural logarithm in log-linear models is almost universal, and this would be consistent with our statement that "The estimate for influenza corresponds to an increase by a factor of $e^{0.000014}$ in the number of schizophrenic births for every death attributed to influenza two to three months previously". What this means is that, suppose a certain month would have 'experienced' 100 schizophrenic births without the influence of influenza, then a single influenza death 2–3 months earlier would have increased the number of schizophrenic births, on average, to $100 \times e^{0.000014}$. The value of $e^{0.000014}$ is approximately 1.000014, so that the number of schizophrenic births would be increased, on average, to 100.0014. For 1000 influenza deaths, this becomes approximately 101.4, which corresponds to a 1.4% increase.

The second question on multiple testing touches on some fundamental and controversial issues in statistics. The usual advice in such situations is to make an adjustment, such as the Bonferroni correction, to each separate comparison, or to perform an 'omnibus' test which covers all the comparisons, and consider individual comparisons only if the overall test is significant. However, others believe that no adjustment for multiple testing is necessary. For example, Rothman (1986) states "Since no problem calling for any adjustment seems to exist unless the positive results from a large number of comparisons are reported without any information about the total number of comparisons, and since even then it appears that adjustments in the results only make them more difficult to interpret, the best course of action for the epidemiologist to take when making multiple comparisons is to ignore advice to make such adjustments in reported results."

The rationale for the Bonferroni correction is that, in making m independent tests, the probability of at least one significant result (at the conventional level of $P<0.05$) when in fact all the null hypotheses are true is not 0.05, but $1-(0.95)^m$, or approximately $0.05\,m$. Thus, to limit the probability of erroneously rejecting *any* null hypothesis to less than 0.05, the P-values of the individual tests are multiplied by m. However, by making the test more stringent, this procedure also increases the probability of not rejecting a false null hypothesis.

In a complex data set, should the evidence for any particular hypothesis depend on how many other hypotheses are also examined? If the answer is yes, then one can destroy the evidence for almost any hypothesis by increasingly thorough analysis of the data. Conversely, the investigator who analyses and publishes his hypotheses one-by-one will run into no such problem. Clearly this is absurd. We believe that it is not the number of tests which is crucial, but whether there are compelling reasons, based on existing knowledge, for making the predictions. A hypothesis supported by a priori knowledge should be considered in its own right, regardless of how many other tests are performed. However, when there is no previous support for a hypothesis, any positive evidence should be considered tentative, even if only one test is performed. There is a subjective element in deciding whether there is an a priori basis for hypothesis, but this can be incorporated into the modern Bayesian framework of scientific inference (Howson & Urbach, 1988).

Having said this, it is often true, when an investigator makes a large number of tests, that there is no a priori basis for any of the hypotheses. Indeed, under some circumstances, there is an inverse relationship between the number of tests and the a priori probability of each hypothesis. For example, consider two random loci which are being tested for allelic association with a disease. There should be no difference in the prior probabilities that each locus is in linkage equilibrium with a disease locus. However, a chance association may be more likely with the more polymorphic locus, which contains more alleles for comparison. An adjustment for multiple tests will then reduce the risk of becoming over-confident about any 'significant' results.

The apparent need for an adjustment for multiple tests is a reflection of the current emphasis placed on the P-value, of a single study, in determining the truth of a hypothesis. Unless the effect or sample size is enormous, this is rarely possible when there is substantial random variation. The truth of a hypothesis must then be assessed by repeated, independent studies; an assessment that should consider all the studies, positive or negative, but that need not be influenced by whether other hypotheses are also addressed in these studies. Chance findings