Author’s response to reviews

Title: Deprived Children or Deprived Neighbourhoods? A public health approach to the investigation of links between deprivation and injury risk with specific reference to child road safety.

Authors:

Paul J Hewson (phewson@devon.gov.uk)

Version: 3 Date: 15 Mar 2004

PDF covering letter
Biomed Central Editorial

Dear Editor

**MS: 7286062812515332: Deprived Children or Deprived Neighbourhoods? A public health approach . . .**

Thank you for the opportunity to respond to additional statistical review of the above submitted paper. I am pleased to note that the referee does not raise any points which question the substantive conclusions of this paper. Rather it would appear that the referee is concerned that the paper may represent an over-elaborate approach to what appears to be a simple problem. I will therefore give an account as to why I feel it is necessary to use well established and extensively studied epidemiological approaches to a specific and growing problem in public health.

As I understand it, the referee raised three specific points:

1. That modelling could not have adequately addressed the issue of confounding by environmental factors
2. That a GLM approach (and therefore a GLMM approach) was not needed
3. That the study should be essentially descriptive and Figures 1 to 3 along with correlation statistics and p-values should be used.

As I therefore understand it, having read the paper (and descriptive statistics have been included) the referee felt that a descriptive approach should arrive at a similar conclusion.

**Confounding**

However much a public health professional may wish for better data on road injury, in many countries, road safety policy is determined by analysis of data which is in essence a by-product of other official processes. This paper presents a case study using data referred to as “Stats 19”, collected by the police. These data can reasonably be considered as a by-product of a process which seeks to determine whether any of the participants in the road collision should be prosecuted. Confounding will remain a problem with these data that can presumably only ever be dealt with optimally by case-controlled studies. However, regardless of how many caveats we may have to attach to any analysis of these data, it remains the “official” data which acts to define what is globally a massive public health problem. It therefore seems appropriate to expend effort in extracting as much information as possible from these data partly to attempt to understand the problem, partly to attempt to identify limitations in the data and inform the debate on the balance of cost-effectiveness between continuing an imperfect “census” of road injury and attempting more case-controlled studies to understand specific problems.

In addition to particular limitations with these specific data, there are many recognised problems in the analysis of observational data generally. Equally, there are many recognised attempts to deal with this. This paper presents one illustration as to how a model can use a proxy variable (urbanicity) to account for more general environmental differences between different types of roads and neighbourhoods (and possibly behaviours). Whilst this proxy variable clearly captures a much more complex set of phenomena, it is well recognised that urbanicity affects traffic-related casualty rates and indeed many other phenomena.
No doubt this is the reason that the Office for National Statistics in the UK make this data available so readily and why the Department for Transport in the UK recently commissioned a literature review and study of attitudes amongst rural residents which may alter their risk-taking behaviour.

The importance of allowing for confounding variables is illustrated by including a proxy variable, “urbanicity”. As a result, we have a substantial improvement to model fit which alters our understanding of the relationship between deprivation and casualty rates. This is an important finding. However, with only the proxy variable it is not possible to state with any certainty why the relationship is altered. Hence the paper offers some speculation as to what may underly the relationship between urbanicity and casualty rates, based on a review of the injury prevention literature and a comparison of models fit to the count of casualties according to the ward where they were injured or the ward where they were resident. These speculations are heavily caveatfed.

Finally, at a conceptual level there is an additional way in which confounding is dealt with in the paper. Extension of the GLM to include random effects by using a Generalised Linear Mixed Model (GLMM) includes an allowance for extra sources of variability in the model. It is possible to consider these random effects as latent variables which in this context can account for unmeasured covariates that haven’t been included in the model. There is a vast literature on the use of latent variables much of which is outside the statistical field, reviewed for example in Bollen KA. (2002) “Latent variables in psychology and the social sciences” Annu Rev Psychol 53:605-34.

**Appropriateness of modelling epidemiological data**

There is a substantial public health / epidemiological literature which develops the applicability of a modelling approach to this kind of data (reviewed for example in P. Elliott, J.C. Wakefield, N.G. Best and D.J. Briggs (2000) “Spatial epidemiology: methods and applications”). We are using data that is very similar to the SMR (here the Standardised Morbidity Ratio) which is this context can be considered as \(\frac{Y_i}{e_i}\) where \(Y_i\) is the casualty count and \(e_i\) is simply the “population at risk” As with the SMR, this can equally be presented as a model: \(Y_i \sim \text{Poisson}(e_i \lambda_i)\), by modelling \(\lambda_i\) we have a generalised linear model (GLM).

Essentially, this paper reports an analysis exactly analagous to an investigation of the SMR. There are a huge number of advantages in formulating the general problem as a model, for example in the SMR proper, “population at risk” can be adjusted for specific risk factors, spatially weighted and so on. However, the specific reason for needing such a model in this context is that this paper considers small areas. These small areas are administrative wards where the relevant “population at risk” is in the hundreds. As a result, the variance of the denominator of the SMR has very high sampling variability. No doubt the referee is aware that there are two established ways of dealing with this high sampling variance, one is to use Minimum Mean Squared Error approaches the other is to use a random effects approach. I do not feel that this is the aproppriate journal to conduct a comparison of the two approaches (or even a comparison with Bayesian approaches) and have therefore presented what appears to be the more usual way of dealing with this kind of data in the public health literature.

**Study should be descriptive**

The referee suggests calculating correlation coefficients. As the referee will no doubt be aware, it is possible to express the correlation coefficient as a regression coefficient, for example in the bivariate normal case we have:

\[
\mu_{y|x} = \mu_2 + \rho \frac{\sigma_2}{\sigma_1} (x - \mu_1)
\]

I therefore feel that the referees suggestion to use the correlation coefficient as a measure of association is conceptually identical to using parameter estimates derived from a model as in this paper. However, there are a large number of advantages in using model coefficients as a measure of association. Firstly, we are modelling the injury count (a discrete variable) on a continuous explanatory variable (the deprivation
score). An empirical measure of correlation (presumably the referee is thinking of a non-parametric measure such as the Spearman rank correlation coefficient) could not adequately account for the systematic variability in the data. Not only do we have the “small-area” issue referred to above, but we also have the Poisson assumption to contend with (which implies that the variance of the response variable increases as the size of the response increases). In addition, considerable power would be lost when disaggregating the data to incorporate a simple descriptive analysis of the effect of urbanicity. Using an extremely well studied and characterised model allows us to estimate a measure of association which takes account of the variance structure that is present in the data, and also provides a standard error for this estimate. Therefore, by using a model, not only do we have a measure of association between the deprivation score and the casualty rate (and a corresponding standard error) as suggested by the referee but we have made the most appropriate allowance for variance in calculating this measure of association. In addition, we have an estimate for the strength of association between urbanicity and casualty rate (and a corresponding standard error) which could not have been estimated from a descriptive study.

Changes to the paper

The referee does suggest the use of p-values. However, my understanding in the public health field (for example this is an explicit requirement of the British Medical Journal and other journals in that family) was that these were eschewed in favour of confidence intervals. I have therefore dealt with the omission of a measure of “statistical significance” by including information on relevant confidence intervals in the discursive part of the results section.

Secondly, as the referee was concerned that the approach may have been over-elaborate I have included a brief preamble in the Methods section to explain why it is necessary to use well-established epidemiological methods with a data set relating to a major public health issue. However, I have not included the depth of information contained in this reply as I am confident that the arguments are sufficiently well rehearsed and familiar in the public health field as to make that unnecessary.

I look forward to receiving comments from all referees in due course.

Yours sincerely

Paul Hewson