Apparent association between benzene and childhood leukaemia: methodological doubts concerning a report by Knox

J F Bithell, G J Draper

Abstract

A recent study by Knox concludes that cases and "clusters" of two or more cases of childhood leukaemia and non-Hodgkin's lymphoma occur closer to many kinds of industrial installation than to supposedly comparable control locations. It is argued that these findings could be largely or entirely artefactual, the apparent differences arising out of the inappropriateness of the control data. Knox used randomly selected postcode units as controls, a procedure that leads to the comparison of individuals located in areas with typically quite different population densities from those for the cases. The resulting potential for bias is explored and the arguments are exemplified by analysing household data based on postcodes.

(*J Epidemiol Community Health* 1995;**49**:437–438)

The question of whether leukaemias – particularly childhood leukaemias – occur in clusters is a long standing epidemiological issue.[1 2] In general, the accumulated evidence can fairly be described as weak, whether it is addressing a generalised tendency to case aggregation[3] or possible proximity to specific risk sources.[4] A recent paper by Knox, however, purports to show that cases, and "clusters" of two or more cases, occur closer to many kinds of industrial installation than to supposedly comparable control locations.[5] We believe that these findings are probably largely or entirely artefactual: the apparent differences arise out of the inappropriateness of the control data. Knox selected postcode units randomly or quasi-randomly and used their locations as controls for the locations of the cases of leukaemia. This method creates a potential for bias. The resulting issues of sampling theory are of general importance in geographical epidemiology and it therefore seems worthwhile to explore in detail why the method used by Knox is likely to give misleading results.

Our major criticism of Knox's study, which we explain in detail below, is that unless all postcodes contain the same number of children (which they clearly do not), a sample of postcodes is not equivalent to the sample of control children required for the analysis.

Knox's analyses

Knox uses three methods of comparison. In discussing these, we will, for the sake of brevity, refer to tables in the original paper[5] and not repeat the description of the data set used.

CASE-CONTROL COMPARISON OF DISTANCES

In the second, most straightforward, of the analyses the locations of 9406 cases of leukaemia and non-Hodgkin's lymphoma were compared with a set of postcodes randomly selected from the $1\frac{1}{4}$ million residential postcodes in Great Britain. The average distance of a case location from the nearest instance of a particular type of installation, such as a rail yard or power station, was then compared with the average for the postcode locations. For each of the installation types reported in Knox's tables 2 and 3,[5] the mean of the case distances was less than the mean of the control postcode distances – in several instances significantly so.

It is easy, however, to show that, if μ and σ^2 are the mean and variance of the distribution of the population of children in postcodes as a whole, the "index" postcodes containing the cases contain on average $\mu+\sigma^2/\mu$ children (see Appendix). The variance σ^2 will obviously not be zero and the formula therefore demonstrates that the index postcodes have systematically larger populations than postcodes in general. The selection of postcodes through the ascertainment of children uses what is known in some contexts as *weight biased sampling*; this is precisely the same method as that used when ascertaining families from a sample of children, which leads to larger families than from a random sample.

Furthermore, postcodes with larger populations tend to be in areas with a higher population density. Taken together, these relationships imply that the index case locations are on average in places with a higher population density than are the control postcodes and thus will tend to be closer to geographical features located in areas of higher population density; this could therefore lead to precisely the type of finding reported by Knox.

COMPARISON OF "CLUSTER"-CONTROL DISTANCES

The first analysis in Knox's paper (table 1)[5] compounds the above problem by considering "clusters" or groups of two or more cases whose distances apart were less than 0·15 km. Although the algebra is less straightforward in this case, it is clear that requiring a postcode to contain at least two cases favours the se-

Department of Statistics, University of Oxford, 1 South Parks Road, Oxford OX1 3TG
J F Bithell.

Childhood Cancer Research Group, University of Oxford, 57 Woodstock Road, Oxford OX2 6HJ
G J Draper

Correspondence to:
Dr J F Bithell.

Accepted for publication March 1995

lection of even more populous postcodes. Even if the cases in a cluster are in different postcodes, the requirement that they should be geographically close further favours the higher density areas. We might, therefore, expect the difference between index and control distances to be more marked for the cluster-control analysis than for the second, case-control, analysis; comparison of tables 2 and 3[5] bears this out, in that 10 of the 11 features show a greater percentage difference between cluster and control distances than between case and control distances.

CIRCULAR REGIONS AROUND POINT SOURCES

The third analysis (table 4)[5] is of a different character. Here the numbers of cases within particular distances (5, 10, and 15 km) from a putative risk source were compared with the numbers of randomly selected postcodes within the corresponding circles. If all postcodes contained the same population, the expected number of postcodes within a circle would indeed be proportional to the population within the circle, and hence to the expected number of index children under the null hypothesis. However, as argued above and exemplified below, the number of households per postcode is greater in high density regions. It follows that in urban areas the expected numbers of sampled postcodes occurring within a circle will be smaller than the expected number of cases occurring within the circles, even when cases occur at random.

Analysis of Oxfordshire postcodes

To add some numerical detail to the argument, we analysed 1991 census data relating to postcodes in a 1000 km^2 rectangle in Oxfordshire, chosen to ensure that every 1 km square was at least 2 km from the county boundary. It is relatively difficult to determine the number of children living in each postcode unit, so we illustrate the effect described above by counting the numbers of households. The distribution of the number of households per residential postcode had a mean of 10·02 and a variance of 64·72, from which we conclude that a postcode selected by choosing a household at random would on average contain around 65% more households than one selected by simple random sampling. To study the relationship between postcode population size and population density, we counted the number of households in each of the 1000 1 km squares. Excluding squares containing no postcode centres, the correlation between the mean number of households per postcode in a square and the total number of households in that square was 0·486, substantiating the claim made above in relation to children that postcode population size is related to population density. If, as seems likely, Oxfordshire is more homogeneous in its spatial distribution of population than the country as a whole, the effect shown here quite possibly underestimates the bias in Knox's calculations.

Further methodological considerations

A further point that should be made in connection with the first two analyses is that, as has been argued elsewhere,[6] mean distance is a bad measure of the *closeness* of points on a map, since it is heavily influenced by the largest distances, which are also the least important. Although the comparison of index and control distances is not vitiated by using mean distance, this does lead to inefficient analyses and also to doubtful validity of the use of the *t*-test, as in the Knox paper, since the distance distributions will be markedly non-normal. The solution to this problem is simple, namely to use an inverse transformation of the distances or their ranks. Paradoxically, because of the inherent bias described above, the results of using these alternatives might well be to lead to results that are substantially more statistically significant, though they would not thereby become more convincing.

Finally, we draw attention to two other potential sources of error that further weaken the conclusions of the paper. Firstly, the cases were diagnosed between 1966 and 1983 and the addresses for cases relate to this period, whereas the sample of postcodes was presumably chosen from a file relating to addresses current around 1990. Movement of population away from the vicinity of the installations would magnify the bias already identified. Secondly, no information is given about the start up dates of the installations considered and the diagnosis dates of the cases in their vicinity.

Discussion

In his first analysis (though not in the second and third) Knox attempts to control for any peculiarities of his comparison by measuring distances to churches, regarded as indicators of population density. It seems clear, however, that this is not a reliable method of adjusting for the fundamental inappropriateness of the controls. The distribution of population over a geographical area is complex and the way in which the case-control distance comparisons behave is not simply and linearly related to population density.

We conclude that, while the associations reported by Knox could be correct and these associations could be causal, the potential sources of bias and error are such that the findings may be entirely artefactual and that the conclusions of the paper should at least be very much more cautious than those given in the Abstract.

Appendix

We here substantiate the formula given above for the mean number of children in the postcodes containing the index children. Suppose that there are k postcodes, containing $n_1, n_2, \ldots, n_k$ children respectively, with $\Sigma n_j = N$. If a postcode is randomly sampled, then each has a probability of $1/k$ of being selected. Hence the expected number of children in such a postcode is $\mu = \Sigma n_j(\frac{1}{k}) = \frac{N}{k}$. Similarly, the variance is $\sigma^2 = \Sigma n_j^2(\frac{1}{k}) - \mu^2$.

If, however, a child is selected at random, the probability that he or she comes from postcode j is n_j/N. Hence the expected number of children in the postcode containing a randomly selected child is $\Sigma n_j(\frac{n_j}{N}) = \frac{1}{\mu}\Sigma n_j^2/k = \sigma^2/\mu + \mu$, as asserted.

1 Ederer F, Myers MH, Mantel M. Do leukaemia cases come in clusters? *Biometrics* 1964;**20**:626–38.

2 Knox, EG. Epidemiology of childhood leukaemia in Northumberland and Durham. *Br J Prev Soc Med* 1964;**18**: 17–24.

3 Cuzick J, Hills M. Clustering and clusters – Summary. In: Draper GJ, ed. *The geographical epidemiology of childhood leukaemia and non-Hodgkin lymphomas in Great Britain, 1966-83: Studies on medical and population subjects,* **53**. London: HMSO, 1991.

4 Bithell JF, Dutton SJ, Draper GJ, Neary NM. Distribution of childhood leukaemias and non-Hodgkin's lymphomas near nuclear installations in England and Wales. *BMJ* 1994; **309**:501–5.

5 Knox EG. Leukaemia clusters in childhood: geographical analysis in Britain. *J Epidemiol Community Health* 1994;**48**: 369–76.

6 Bithell JF, Stone RA. On statistical methods for analysing the geographical distribution of cancer cases near nuclear installations. *J Epidemiol and Community Health* 1989;**43**: 79–85.

Response by Professor G Knox

The inadequacies of postcode(PC)-based map references were clear from the beginning, a major technical concern both in this and in preceding analyses, and clearly stated as such. Sampling biases arise not only from the varying populations per PC, as Bithell and Draper elaborate, but also from varying numbers of PCs per map reference. If access to the PCs had been permitted, a definitive test of clustering could have compared observed and expected numbers of leukaemia pairs sharing a single PC. The national distribution of delivery points (DP) per PC is known, allowing correction for heterogeneity. However, intractable political/ethical problems blocked this access. The choice then lay between a cessation of all enquiries along these lines, or an exploration which tried to make best use of less than satisfactory data.

The demonstration by Bithell and Draper that areas with high population densities also tend to have larger PCs, confirms what was previously a surmise; but this is not the same thing as a link between PC size and proximities to putative industrial hazards. These last relationships are extremely complex. Many industrial sites are on locally depopulated industrial estates or set in even wider industrial zones, and some distance from the nearest "residential" PCs: while factories located within residential PCs compete with houses for space, creating low rather than high adjacent population densities, and low DP PCs. Linear obstructions to postmen's "beats" cause further distortions.

In the face of these uncertainties a major consideration must be the apparent limitation of the proximity findings to processes whose emissions are causally plausible. It is through such demonstrations, also, that we must see the main prospects for advance. A first necessity, still, is to gain access to data which will settle the question of local geographical clustering once and for all. If it is confirmed then the remaining question is much simplified. It is no long necessary to test the reality of the clusters through seeking proximities to plausible local hazards; hazards there are, and the question now is simply what they might be. For this we now need to examine a greater number and a wider variety of potential sources, preferably with new disease and address data, and using more refined methods of demographic standardisation than are possible using only map references.