Cluster analysis and disease mapping—why, when, and how? A step by step guide

Sjúrður F Olsen, Marco Martuzzi, Paul Elliott

Growing public awareness of environmental hazards has led to an increased demand for public health authorities to investigate geographical clustering of diseases. Although such cluster analysis is nearly always ineffective in identifying causes of disease, it often has to be used to address public concern about environmental hazards. Interpreting the resulting data is not straightforward, however, and this paper presents a guide for the non-specialist. The pitfalls include the fact that cluster analyses are usually done post hoc, and not as a result of a prior hypothesis. This is particularly true for investigations prompted by reported clusters, which have the inherent danger of overestimating the disease rate through "boundary shrinkage" of the population from which the cases are assumed to have arisen. In disease surveillance the problem of making multiple comparisons can be overcome by testing for clustering and autocorrelation. When rates of disease are illustrated in disease maps undue focus on areas where random fluctuation is greatest can be minimised by smoothing techniques. Despite the fact that cluster analyses rarely prove fruitful in identifying causation, they may—like single case reports—have the potential to generate new knowledge.

Public awareness about potential hazards in our environment is growing. With the advent of powerful computing techniques that can be applied to routinely collected mortality and morbidity data, the demand on public health authorities to undertake investigations into geographical patterns of disease has increased. Nevertheless, several basic epidemiological and statistical issues may present obstacles to the satisfactory handling of such data.¹ Although texts are available that cover recent developments,^{2,3} there is no obvious resource for the generalist reader covering methods for investigating disease clusters and clustering and for interpreting disease maps. This paper is intended to fill this gap by presenting a step by step guide to these problems for the non-specialist. Recent reviews of the statistical methods,^{3,4} as well as a critique of the issues involved in epidemiological applications,⁵ can be found elsewhere.

The statistical detection of spatial patterns of diseases

The types of spatial patterns that we shall consider are disease clusters, clustering, and, briefly, autocorrelation.

Definition of cluster—One common definition, by Knox, is "a geographically bounded group of occurrences of sufficient size and concentration to be unlikely to have occurred by chance."⁶ Although this sounds straightforward, several problems are associated with its realisation in the investigation of clusters.

STANDARDISED MORTALITY (MORBIDITY) RATIO AND STATISTICAL TESTING

To understand the problems, we need to refresh our knowledge of some underlying statistical principles. Let us imagine we are studying whether there is an excess of cancer deaths in a local health authority which is one subarea of a larger region that we wish to use as reference. Assume also that we have population based data on deaths from cancer, as well as information on the populations in which these deaths occurred.

Technically we would calculate the ratio between the observed and expected number of deaths in the subarea. The number of expected deaths in the subarea would be calculated by multiplying the rate by which the cancer deaths occur in the larger region, during a specified period, by the size of the population in the subarea. Since age is a strong determinant of cancer, we would usually take into account any differences in the age structure between the subarea and the region by calculating the expected numbers within age strata separately—for example, by five year age groups—and then adding the numbers up to give an overall number of expected cases. The resulting observed:expected ratio is the (age) standardised mortality ratio (SMR). The standardised mortality ratio is an estimate of the rate ratio—that is, the ratio of the disease rate in the subarea to that in the overall region. Standardised mortality ratios are often presented as percentages (without this being made clear), and this can sometimes create confusion.

If we observe a standardised mortality ratio of, say, 2.0 (or 200, if presented as percentage) in the subarea, how would we decide whether or not this excess has occurred by chance? The underlying idea is that we assume that cases—in the absence of anything unusual in the subarea, or, more precisely, in the absence of a particular disease determinant occurring at a higher level in the subarea than in the surrounding areas—would arise with a natural variability that can be described by a mathematical model, usually a Poisson model. The null hypothesis, that the true underlying rate that has generated the cases observed in the subarea is the same as the underlying rate in the overall region, corresponds to a risk ratio (an observed:expected ratio) of 1.0.

On the basis of statistical theory it is now possible to assess the probability by which, under the null hypothesis, we would observe a standardised mortality ratio of 2.0 or one that deviated even more from the null of 1.0; this probability is the P value. We may find that this probability is 0.01, which may be small enough for us to think that it casts serious doubt on the null hypothesis, thereby favouring the alternative hypothesis that the relative risk is truly increased in the subarea—that is, that there is an apparent excess of disease.

Interpretation of such an apparent excess, however, is crucially dependent on whether the investigation was a priori or post hoc (a posteriori). The conventional P value is interpretable only in relation to a priori

Danish Epidemiology Science Centre, Statens Serum Institut, DK-2300 Copenhagen S, Denmark, and Department of Public Health and Policy, London School of Hygiene and Tropical Medicine, London WC1E 7HT
Sjúrður F Olsen, senior research fellow

Department of Epidemiology and Public Health, Imperial College School of Medicine at St Mary's, London W2 1PG
Marco Martuzzi, research fellow
Paul Elliott, professor

Correspondence to: Dr Olsen, Copenhagen.

BMJ 1996;313:863-6

hypotheses—that is, those set up without prior knowledge of the number of deaths from cancer occurring in the subarea of interest. This is discussed further in the next section in relation to the leukaemia cluster near the Sellafield nuclear plant in north west England. Conventionally, a P value smaller than 0.05 is regarded as statistically significant, leading to rejection of the null hypothesis. (An alternative approach to significance testing is to calculate confidence intervals for the estimated observed:expected ratio.⁷)

PROBLEMS IN DEALING WITH REPORTED CLUSTERS

Let us now turn to a different situation, exemplified by a real event. In November 1983 the Yorkshire television programme “Windscale—the nuclear laundry” attracted considerable public attention in the UK as it suggested that there was an excess of childhood leukaemia around the nuclear processing site at Windscale (subsequently renamed Sellafield).⁸ The journalists raised the suspicion that discharges of radioactivity from the plant could have caused the excess. Originally the producer had intended to make a documentary about the health of the workforce at the nuclear site. During his research, however, he heard incidentally from local people of a number of cases of childhood leukaemia in Seascale, a village near the plant, which then became the focus of his programme. The excess rate in the local authority administrative area containing the village turned out to be highly statistically significant compared with the national rate. How should we interpret this finding?

Multiple comparisons—From a statistical point of view, there are at least two problems in such data, both of which relate to post hoc testing of hypotheses. One is that the producer may well have followed up several leads before deciding to pursue the cases of leukaemia. If so, he has de facto undertaken multiple comparisons (even if he probably did not make any significance tests at that stage), complicating the interpretation of the P value, as explained in the next section.

Defining the population and the period: boundary tightening—Another problem that pervades nearly all investigations of reported clusters relates to defining the population from which the cases arose and the time period of the putative cluster. In this example the local administrative area containing the nearby village of Seascale (but not the nuclear plant) was defined as the underlying population, but if the producer had chosen a wider area this would have increased the number of expected cases and thereby reduced the excess. The problem is well illustrated by the example of the Texas sharpshooter who first fires his gun and then draws the target around the bullet hole.⁹ The suspicion of a cluster often begins with the identification of a group of cases:

The identification of a cluster and its source population is all too often like the Texas sharpshooter who first fires his bullet and only then draws the target round it.

only subsequently do we define the underlying population from which the observed (suspected) cluster of cases arose. The more narrowly the underlying population is defined, the less will be the number of expected cases, the greater will be the estimate of the excess rate, and often the more pronounced will be the statistical significance. This process is referred to as boundary tightening, and it represents a major difficulty in cluster investigations: it relates not only to the selection of geographical boundaries and period around suspected clusters but also to the selection of age groups and diagnostic categories.

PROBLEMS WHEN DETECTING CLUSTERS IN DISEASE SURVEILLANCE

Let us now suppose that we are analysing routinely collected data to assess the presence of raised disease rates in a geographical region naturally divided into 20 subareas. As one way of examining the rates, we undertake statistical tests of whether the disease rate in each area differs significantly from the overall rate in the region as a whole, as described earlier. We may conclude that one subarea has a significantly increased disease rate at the 5% significance level. In such circumstances it is less straightforward than in the situation described above to conclude that the data suggest an excess of disease in this subarea, or, in other words, a cluster of cases. The problem is that we have undertaken 20 tests for clusters at the same time, one for each area: having chosen a 5% significance level for our P value, provided that the null hypothesis is true, we would by definition expect 1 in 20 independent tests to be statistically significant at the 5% level. This problem, again one of multiple comparisons, is often encountered in disease surveillance, where researchers are typically undertaking many comparisons without any prespecified hypotheses and where hypotheses may be generated during scrutiny of the data.

CLUSTERING

Rather than focusing on clusters as isolated phenomena, another way of dealing with such data is to test for clustering as a general feature of the disease pattern in the whole region. If we again deal with data categorised into subareas the null hypothesis is still that all subareas have the same rate. This again corresponds to the situation where the cases occur at random across the subareas, with frequencies proportional to the numbers at risk in each subarea and with a variability described by a Poisson distribution.

But we can calculate a summary measure for the total variability of the disease rates across the subareas. For example, we could calculate the χ^2 value, which is the sum of the quantity $(O-E)^2/E$ across all the subareas, where O is the observed number and E the expected number of cases in each subarea. The more the counts of the observed cases differ from their expected values, the larger the χ^2 value; and it is possible to assess the probability (P value) of observing a variability that deviates as much or more from the null situation.

If the P value is less than 5%—that is, it is statistically significant—we may conclude that the observed data are not quite compatible with the null hypothesis; instead data seem to suggest that the variability in the underlying disease rates across the areas is greater than that allowed for by the Poisson model—that is, that there is a tendency for clustering within one or more subareas. Besides being called a test for clustering, the test is also called a test for heterogeneity, χ^2 test, and test for extra-Poisson variation.

AUTOCORRELATION

The above test for heterogeneity tells us nothing about the underlying spatial distribution of the disease, only that some areas have higher rates (while others will

have lower rates). Let us again imagine a map of a region divided into subareas. We can see some variation across the subareas, and there may be a tendency for the rates to decline from one to another part of the map. How should we test for such evidence of spatial organisation? The usual tests for trend can take account of only one dimension; here we have two. Since every subarea has contact with one or more other subareas, one way would be to examine adjacent areas pairwise, and see if their rates tend to be correlated: this would be the case if there were a gradient across the map. In the same way as when we test for correlation between two variables, x and y , we now test for correlation between the pairs of disease rates in adjacent areas: since the variable is being correlated with itself this is called autocorrelation, and the statistic we get is the autocorrelation coefficient.

Although autocorrelation seems to be an intuitive concept, it is often complicated to interpret in practice. One problem is that areas lying close to each other may tend to have similar rate estimates for reasons other than those of interest in the study. Thus, areas with large populations, which are concentrated in and around cities, tend to have more stable rates; spurious autocorrelation may therefore be observed due to uneven population distribution, even under the null hypothesis.

Map of occurrences of adult leukaemia in West Midlands Region, England 1974-86: (a) unsmoothed, (b) smoothed by Bayesian methods.

Furthermore, disease rates based on routine registers may vary according to geography because of differences between areas in recording systems (diagnostic thresholds among general practitioners, coding practices in local hospitals or discharge registers, etc), and this may be less pronounced for areas lying close together. Neighbouring areas also tend to share socioeconomic characteristics with which disease rates are often quite strongly associated. Another complicating factor is that the autocorrelation coefficient can be based on a variety of measures of "closeness," such as distances between population or geographical centroids, rather than merely on whether or not the subareas share a common boundary.

Mapping of diseases

Mapping of disease is an activity closely related to disease surveillance and cluster detection. It is widely used for descriptive purposes to identify patterns of geographical variation in diseases and to develop new ideas about the causation of disease. Interpreting disease maps is, however, often far from straightforward.

Problems in interpreting rates—We can express the disease occurrences in the different areas as standardised rates. The drawback of such a map is that the smaller the underlying population the more the rate estimates are influenced by random (Poisson) variability. Small populations will therefore tend to give rise to the most extreme rates, even if the true disease rates are similar across the areas, focusing viewers' attention on these areas when they scrutinise the map. As an alternative, maps have been published based on the P values from tests of whether the rate in each area differs significantly from, for example, the overall rate. This is of little help, however, since now the areas with the largest populations will tend to dominate the map.

Smoothing—One way of getting around this problem is to use "smoothed" estimates of disease rates.¹⁰ With this technique the rates are adjusted by combining knowledge about the rate in each area with the knowledge about the rates in surrounding areas. When the underlying population of an area is large, and consequently the statistical error of the rate estimate is small, high credibility is given to the estimate, and the adjusted rate will be close to the observed rate. When the underlying population is small, and the statistical error correspondingly large, little credibility is given to the observed rate and it is "shrunk" towards the mean of the surrounding areas. If there is evidence for gradients in the map (for instance, as indicated by a positive test for autocorrelation) the rates can be adjusted towards averages of neighbouring rates rather than a value representing the overall mean of the map. The techniques mentioned here are based on quite complex calculations. The "smoothed" risk estimates are often referred to as Bayesian or empirical Bayesian estimates.

These ideas are illustrated in figure 1. It shows the incidence of adult leukaemia across electoral wards in the West Midlands region in England in 1974-86. The map on the left shows the unsmoothed standardised incidence ratios of each of the 832 wards. The map appears to be dominated by high standardised ratios in the large rural wards at the periphery, where there are areas of apparent clustering. As these wards tend to have the smallest populations, an excess risk may be generated by just one or two cases. With this large component of random variability removed by smoothing (resulting in the map on the right), the map becomes much "flatter" and all the extreme values disappear. While there is evidence of variability in incidence rates ($P < 0.01$), there is no evidence of spatial clustering.

Technical issues—Disease mapping involves a number of choices.¹¹ It is crucial how the variables are

Key messages

- Cluster analyses of geographical data are often difficult to interpret because of their post hoc or data dependent nature
- Investigations instigated by reported disease clusters are typically characterised by a vague definition of the population from which the cases arose, with the inherent danger of overestimating the disease rate through "boundary shrinkage"
- The problem of "multiple comparisons" in disease surveillance may be avoided by testing for "clustering" and spatial autocorrelation as general features of the data
- Maps of disease rates tend to focus attention on the less densely populated areas, where random fluctuation is greatest, a phenomenon which can be avoided by "smoothing"
- Cluster analysis has proved inefficient in identifying disease causes but remains important in addressing public concern with the possible adverse effects of environmental pollutants on human health

categorised. It is recommended to use no fewer than four and no more than eight categories since the eye cannot readily assimilate more than about eight shades of colour or greytone. One has also to decide on criteria for categorisation, and whether one should apply a transformation of the disease measure. Choice of techniques to illustrate variation is also crucial: one can use scales of colours or of greytone, but a particular problem may arise when colour maps are photocopied. The maps can be rendered meaningless if dark colours are used for both the highest and lowest rates since they may become indistinguishable when copied, a point exemplified in figure 1. Continuous scales of colours or greytone are not recommended because the eye will too easily focus on even small variations. Nor should different patterns of hatching or stripes be used. Irrespective of which display techniques are being used, the map will never become better than the underlying data: the validity of the data will always be the limiting factor.

Scientific and public health importance of cluster investigations

Investigations of suspected or alleged disease clusters are often instigated in response to public concern and may help to allay the high levels of anxiety that often accompany such allegations. On the other hand, only a small fraction of such investigations are likely to lead to the identification of preventable causes of human diseases. A cautious approach is therefore required before large research programmes are launched, as they may result in the expenditure of human and financial resources on unproductive activities: such activities may even unintentionally induce further anxiety in the population. It is difficult to give general guidelines and decision criteria for the pursuit of cluster investigations, but discussion of many of the issues involved can be found elsewhere.¹²

Because of the small contribution of cluster investigations to knowledge about the cause of disease, their scientific merits have been debated.⁹⁻¹³ Unless they deal with highly specific exposure-disease associations and high relative risks, cluster investigations per se are unlikely to prove fruitful. By contrast, for some questions—for example, the effects of incineration on human health—the geographical approach may be the only practicable method to study disease among the

large populations potentially exposed and to quantify possible risks associated with such exposure.¹⁴⁻¹⁵ Just as imaginative and original observations made on single clinical cases can inspire innovative thinking—despite such case studies rarely leading to new discoveries on their own—thoughtful, careful, and imaginative descriptive analyses of spatial occurrences of diseases do carry the potential to generate new knowledge and to inform us about disease causation and prevention. Geographical studies will also continue to have an important role in addressing the public's concerns about the possible effects of environmental pollution on human health. It is therefore essential that they are carried out to a high standard, with proper understanding of the limitations as well as the strengths of the method.

We thank the census, population, and health group of the Office for National Statistics (formerly Office for Population, Censuses and Surveys) for use of cancer registration data on which the figure is based. Thanks also to Mr Chris Grundy and Dr Peter Walls for producing the figure and Dr Michael Hills and Dr Helen Dolk for valuable discussions (all at the London School of Hygiene and Tropical Medicine) and to Dr Arne Poulsen (Vejle Public Health Authority, Denmark).

Funding: European Union Human Capital and Mobility Programme (grant number ERB CHBG CT 920161), and the Danish Medical Research Council (grant number 12-1799, 9305412).

Conflict of interest: None.

- 1 Elliott P. Investigation of disease risks in small areas. *Occupational and Environmental Medicine* 1995;52:785-9.
- 2 Elliott P, Cuzick J, English D, Stern R, eds. *Geographical and environmental epidemiology: methods for small-area studies*. Oxford: Oxford University Press, 1992.
- 3 Esteve J, Benhamou E, Raymond L. *Statistical methods in cancer research. Volume IV. Descriptive epidemiology*. Lyon: IACR Scientific Publications, 1994.
- 4 Waller LA, Jacquez GM. Disease models implicit in statistical tests of disease clustering. *Epidemiology* 1995;6:584-90.
- 5 Elliott P, Martuzzi M, Shaddick G. Spatial statistical methods in environmental epidemiology. *Statistical Methods in Medical Research* 1995;4:139-61.
- 6 Knox EG. Detection of clusters. In: Elliott P, ed. *Methodology of enquiries into disease clustering*. London: Small Area Health Statistics Unit, 1989:17-20.
- 7 Bulpitt CJ. Confidence intervals. *Lancet* 1987;i:494-7.
- 8 Gardner MJ. Review of reported increases of childhood cancer rates in the vicinity of nuclear installations in the UK. *Journal of the Royal Statistical Society, Series A* 1989;152:307-25.
- 9 Rothman KJ. A sobering start for the cluster busters' conference. *American Journal of Epidemiology* 1990;132(suppl):S6-S13.
- 10 Clayton D, Kaldor J. Empirical Bayes estimates of age-standardized relative risks for use in disease mapping. *Biometrics* 1987;43:671-82.
- 11 Smans M, Estève J. Practical approaches to disease mapping. In: Elliott P, Cuzick J, English D, Stern R, eds. *Geographical and environmental epidemiology: methods for small-area studies*. Oxford: Oxford University Press, 1992:141-50.
- 12 Centers for Disease Control. Guidelines for investigating clusters of health events. *Morbidity and Mortality Weekly Report* 1990;39(No RR-11):1-23.
- 13 Neutra RR. Counterpoint from a cluster buster. *American Journal of Epidemiology* 1990;132:1-8.
- 14 Elliott P, Hills M, Beresford J, Kleinschmidt I, Jolley D, Pattenden S, et al. Incidence of cancer of the larynx and lung near incinerators of waste solvents and oils in Great Britain. *Lancet* 1992;339:854-8.
- 15 Elliott P, Shaddick G, Kleinschmidt I, Jolley D, Walls P, Beresford J, et al. Cancer incidence near municipal solid waste incinerators in Great Britain. *Br J Cancer* 1996;73:702-10.

(Accepted 26 June 1996)

Correction

Caring for Older People: Homes and housing for elderly people

Due to an author's error, an address was wrong in this article by P Wanklyn (27 July, pp 218-21). The Abbeyfield Society is now at 53 Victoria Street, St Albans, Hertfordshire AL1 3UW (tel 01727 857536; fax 01727 846168; email abbeyf@geo2.poptel.org.uk).