Investigating Disease Clustering

Estelle A. Gilman, Birmingham

The existence of spatial patterns of disease occurrence, particularly if they can be demonstrated to have spatial correlations with social and environmental factors, may prove valuable in investigating the aetiology of a disease. The rationale is that if a disease is aggregated in space, its causes are likely to be so as well, and it may be possible to show what they are.

There are 3 major types of studies of geographical variation:

1) Studies whose aim is simply to describe geographical distribution of disease with respect to place of occurrence. The results of these studies are often presented as maps.

2) Ecological studies, or geographical correlation studies, where the aim is to describe geographical variation in disease in relation to corresponding variation in environmental factors. These studies can produce estimates of relative risk for different levels of exposure.

3) Small area studies. Unlike 1) and 2), which are usually carried out on a relatively large scale e.g. across countries, or counties within a country, small area studies examine disease risks which are much more spatially localised e.g. disease patterns in relation to proximity to an industrial installation, or tendency for disease to show small scale spatial clustering or aggregation. These analyses are carried out at the level of census enumeration district or less and are the most relevant for investigating disease clustering.

There are several types of small area studies, including:

i) studies of clustering as a general phenomenon

ii) studies of point sources of pollution

iii) studies of space-time clustering.

We've already had a session from John Bithell on statistical methods for analysing point source exposures, so I'll only be touching very briefly on those.

i) studies of clustering as a general phenomenon

Starting with some definitions of clusters and clustering. A cluster can be defined as an excess number of cases of disease in one small area or around a particular point source [10] also gave the following definition: 'a cluster is a geographically bounded group of occurrences of sufficient size and concentration to be unlikely to have occurred by chance.' Clustering is often defined as a general tendency for a more non-random or 'clumped' distribution of disease than would be expected to result from variations in population density and chance fluctuations [1].

Most cluster investigations start by choosing a particular area and comparing the observed number of cases with the number expected if the area had a similar incidence as some larger reference area. This approach makes the simple assumption that observed numbers should differ from expected numbers only because of Poisson sampling variation. But there are problems with this approach, including how to choose the area, the *post hoc* nature of the analysis, how to cope with extra-Poisson variation, trends of risk with distance from point source etc. The problem with *post hoc* identification of clusters is that, while it may demonstrate that cases can cluster, it doesn't allow determination of whether this is more than a chance occurrence.

A modification of the above approach is to assume that each case defines the centre of a cluster, and compare the number of other cases observed in circles of different radii around each point to the expected number. This is the basis for a useful test for clustering in general. Also, points which make the largest contribution to the overall test may be those involved in aetiologically significant clusters.

In looking for clustering as a general phenomenon, the methods I'm going to describe concentrate on localised clustering - not regional trends or large scale heterogeneity such as that between counties, or large scale variations in risk due to geographical changes in levels of known risk factors. While these variations are important, they are not relevant to the issue of detecting localised clustering, but it may be advisable to adjust for such large scale variations in risk when carrying out small area studies.

Cell count method

In a situation where there are many small areas, and a limited number of cases of a rare disease, some of the areas will have no cases at all, and a simple χ^2 test of homogeneity is not appropriate. An alternative is the Potthoff-Whittinghill test, which is a powerful test of the hypothesis of heterogeneous risks, distributed around a common value [12, 13]. This test can detect deviations from randomness, where the heterogeneity is either a result of contagion, i.e. preferential occurrence of excess cases in areas where there are already people with the disease; or is due to small excess risks in several areas.

Describing the situation in algebraic terms: suppose the study region is divided into a large number of small areas $\{A_i\}$, reference incidence rates r_j are available for the study region (these could be age-sex-specific or risk factor specific) as well as population counts n_{ij} for the population at risk in each small area, for each age, sex or other risk factor group. If Θ is the relative risk of the region containing A_i, then some accommodation of large scale trends and heterogeneity can be made by taking as the null hypothesis that the number of cases O_i observed in area A_i is Poisson distributed with mean equal to the number of expected cases E_i where $E_i = \Theta \Sigma_j[r_j n_{ij}]$.

The Potthoff-Whittinghill test statistic is

$$S = \Sigma_i[O_i (O_i - 1)/E_i]$$

where O_i = observed number of cases in small area i

E_i = expected number

and S is asymptotically approximately normal with mean

$$\mu = O_+(O_+ - 1)/E_+$$

and variance = $2(N-1)\ \mu/E_+$,

where N is the number of small areas and + denotes summation over all areas in the region of study.

The above test is what is known as a cell count method, it uses geographical units, such as census enumeration districts, for which population data are readily available.

Distance or nearest neighbour methods

Other types of tests for clustering are known as distance methods or nearest neighbour methods. They involve the consideration of circles, the size of which may vary depending on local population density. Also, since population data are usually available only for pre-defined areas (ED, LA) circle populations are either estimates based on summing populations from whole EDs which fall within the circle, plus some proportion of the population of EDs which straddle the circle, or they're approximations based on aggregating whole EDs only into the rough shape of a circle.

The distance method described below does not require the population to be uniformly distributed in the absence of clustering. A group of controls is selected from the population at risk and statistics are based on

whether the nearest neighbour(s) to each case is another case or a control. Controls might be chosen from electoral lists, or if children are being studied, from birth registers. It is important that controls be a representative sample of the population with the same age and sex distribution as the cases. The null hypothesis is that the cases and controls are sampled at random from the same age- and sex-adjusted population.

In a method developed by Cuzick and Edwards [5] there are a set of case locations (x_1, x_2 .... x_n) and a set of control locations (x_{n+1}, ... x_{n+m}). The null hypothesis is that any location should be no more likely to be labelled as a case location than a control location, i.e. the set of case locations (x_1, x_2 .... x_n) is a random sample from the whole set of locations (x_1, ...x_{n+m}). The test of spatial aggregation is made by identifying, for each case location, its k nearest neighbours (usually k is a small integer such as 2 or 3) and counting how many of these are cases. In effect it's a count of the number of cases in area A_k, where the area is that needed to go up to kth nearest neighbour from the reference case location. The method tests for an unusual tendency for cases to have other cases as near neighbours. The test statistic is

$$T_k = \Sigma_{i \neq j} X_{ij} Y_{ij}$$

where $X_{ij} = 1$ if j is the label of the kth order neighbour of i, and 0 otherwise

$Y_{ij} = 1$ if i and j are cases, and 0 otherwise

The mean is $n(n-1)k/(m+n-1)$ and the variance can be computed or simulations can be used [5].

The Potthoff-Whittinghill and the Cuzick and Edwards tests are for globally testing for clustering, where no specific parametric form is assumed for the alternative process. If a test for generalised clustering is negative, it may be that expenditure of further resources investigating a single cluster can't really be justified.

ii) studies of point sources of pollution

The methods described above are not very powerful for detecting isolated clusters for a disease that doesn't have a general tendency for clustering, unless there is a cluster of overwhelming magnitude. In this situation, a better approach is to identify putative sources of risk. However, to do that will often involve post hoc identification of an association, and ideally a set of other, similar sources is needed on which to test the proposed association and so give the analyses statistical credibility.

To expand on this, it is difficult to deal with *post hoc* reports of disease excesses in the vicinity of a particular source of pollution. Often the suspicion of an excess in the local community has prompted an investigation, which then relates the excess to a previously unsuspected point source. This approach weakens the value of the resulting tests because one doesn't know how many clusters exist which haven't been reported, and the use of statistical testing in such a situation is, strictly speaking, invalid [2]. Ideally, the decision to evaluate the effects of a point source of pollution should be made without prior knowledge of the disease incidence in the locality. However, this is not usually possible unless the initial suspicion was raised at another similar site.

One way is to catalogue all major sources of potentially hazardous emissions, generate a hypothesis using data around one source, and then test it by examining the other sources of the same type. This requires investment of time and resources to create a database of emission sources and of population and disease data at an appropriate level of resolution to enable calculation of observed and expected numbers. In Britain such data resources are held by SAHSU (the Small Area Health Statistics Unit), an independent national facility for the investigation of routine health statistics near point sources of pollution [6]. But the problem of more than one type of emission

from each source or of different types of sources located close to each other, and of different levels of exposure from site to site, means that replication of sites can be difficult.

iii) studies of space-time clustering.

The term space-time clustering describes the situation where cases of a particular disease occur close together in space and close together in time, more often than would be expected due to chance. Usually the time and space criteria relate to date and place of disease onset, but for events such as congenital malformations, date and place of birth may be used.

Space-time clustering is different to spatial clustering - where cases are distributed in a non-random way across a geographical area, after allowing for the underlying population distribution.. It is also distinct from temporal clustering - where cases are distributed in a non-random way with respect to time period, an example would be a disease where cases show a marked seasonal distribution.

Why look for space-time clustering? It's of particular interest because its presence may indicate that the disease involves exposure to infectious agents; or that there is transmission from case to case; or that clusters of cases within short distances and short time intervals of each other may be related to intermittent toxic releases of some sort to the environment. To explain space-time clustering one is looking for factors or agents which show the same pattern as the disease cases, and since there are not likely to be many which do, the analysis is potentially very revealing.

The Knox method

The simplest method of analysis is the Knox method [9]. To use the method the only data needed are the location of each case in space (e.g. grid reference of place of onset) and its location in time (e.g. date of onset). Population denominators are not needed. All possible pairs of cases are assembled, their distances apart in space and time calculated, and pairs classified as to whether or not they fall within specified space and time intervals. Essentially the analysis involves a table of the form:

Space interval	Time interval ≤t	>t	All times
≤s	a	b	a+b
>s	c	d	c+d
All distances	a+c	b+d	n(n-1)/2

where

n = number of cases

n(n-1)/2 = total number of possible pairs

a = observed number of pairs whose members are distance s or less apart and time t or less apart.

The observed number of pairs within short space and short time intervals is compared with the number expected if the space and time intervals between pairs are independent of each other. The null hypothesis is that cases which occur close together in space should be no more likely to occur close together in time than other cases, and vice versa. Assuming cases are rare, independent events, distributed as a Poisson variable, the expected number of close pairs, Exp[a], is calculated as:

$$Exp[a] = (a+b)(a+c)2/(n(n-1)).$$

If the observed number of close pairs significantly exceeds the expected number, this suggests that cases occur close together more often than would be expected due to chance and that some other mechanism may be involved. The significance of the departure is tested using d, where:

$$d = (a-Exp[a])/sqrt(variance[a])$$

and is distributed as the standard normal deviate. The variance of the number of

close pairs is calculated by permutation [11].

How do we decide what is a 'short' space and 'short' time interval? Often there is no clear hypothesis to test, and so a data set is examined over a range of space and time intervals, and many significance tests are carried out. For example, in an analysis of data on childhood cancers [7], 7 space intervals (<1 km, <2 km, <3 km, <4 km, <5 km, <10 km, <20 km) and 12 time intervals were used (same month, <1, <2, <3, <4, <5, <6, <9, <12, <18, <24, <48 months). However, the problem of multiple significance testing is not as large as it may seem, since the 84 tests are not independent of one another, e.g. pairs of cases diagnosed within 2 km and 3 months of each other are also included in pairs of cases diagnosed within 3 km and 4 months, and so on. This correlation between the results of tests means that the true significance level is not greatly affected [4, 8]. In this situation, adjusting results for multiple significance testing using the Bonferroni correction (dividing the critical significance level by the number of tests performed, giving P=0.05/X as the critical P value for X independent tests at the 0.05 level of significance [3]) would be too conservative. Rodrigues et al. [14], in an application of the Knox method to space-time clustering of Sudden Infant Death Syndrome births, adjusted the critical P value for the most significant individual test by allowing a factor of 2 for multiple testing within the same subset and dividing by the number of separate subsets of data examined. An alternative approach would be to divide the data into 2 sets, one for hypothesis generation, and one for hypothesis testing, and this was the approach used for the analyses of the childhood cancer data [7].

There is still the problem of which dates and places to perform the analysis on. Here some knowledge of the disease process is needed. For example, for disease which are diagnosed shortly after birth, and have their origins *in utero*, e.g. congenital malformations, clearly it is more sensible to use date and place of birth rather than date and place of diagnosis of the condition. The ideal is to choose the date and location which are as near as possible to the actual disease-producing event. For conditions which are recognised soon after birth, which probably arose during foetal development, date and place of birth are often the nearest one can get. For other diseases occurring later in life, date and place of disease onset may be the only option. In such cases the sensitivity of the method will be affected if there has been migration between the time of disease initiation and the time of its recognition. Sensitivity will also be reduced if the latent period between disease initiation and recognition is variable, this applies even if cases haven't moved between initiation and diagnosis. In all these situations, cases which were clustered at their time of disease initiation may not appear to be clustered at their time of diagnosis. So the chances of detecting space-time clustering are best if the date and place chosen are as near as possible to the actual initiating event.

Possible artefactual causes of clustering must also be considered when interpreting any clustering patterns found in the data, e.g. incomplete ascertainment of cases, or clustered recognition of cases, with more complete ascertainment in some areas or some periods than others.

Summary of small area studies

To summarise some of the main points of small area studies. The typical cluster investigation usually starts with an assessment of whether the number of cases in the area under study is truly in excess of the number expected. This involves selecting a geographical sampling frame and a time period. These might be biased (data dependent), but even so the observed to expected ratio of cases may not be much above unity.

Quickly establishing this may be all that is needed to eliminate concern over the cluster and to provide reassurance that disease rates are not elevated. When an elevated rate is found, potential biasing and confounding factors need to be considered, e.g. socioeconomic confounding, biases due to choice of area, time period, age groups etc. The extent to which the disease excess depends on the choice of these should be examined. If there is no pre-determined source or focus of risk, tests for generalised clustering should be used over a wider area than that which provided the original concern. If there is a suspected source of risk, methods appropriate to analysing point sources should be used. Following this, if the cluster is still considered 'real' the next step could be examination of other, similar areas for excesses e.g. around similar point sources if the cluster arose through suspicion of some point source location. This allows the hypothesis to be strengthened or rejected before embarking on detailed studies which may be expensive and time consuming, need careful planning, and for which numbers may still be too small for meaningful results.

Conclusions

To conclude this brief introduction to small area methods of investigating disease clustering, one must be aware that the results of these types of investigation need careful interpretation. Their usefulness lies in generating hypotheses, and further epidemiological studies, of a different design, e.g. case-control, will usually be needed to investigate the issues raised by them.

References

1. Alexander F.E. and Cuzick, J., 1992. Methods for the assessment of disease clusters. In Elliott P., Cuzick J., English D., Stern R. (eds.) 1992. Geographical and Environmental Epidemiology. Methods for Small-Area Studies. WHO, Oxford University Press., pp 238-250.

2. Bithell J.F., 1992. Statistical methods for analysing point-source exposures. In Elliott P., Cuzick J., English D., Stern R. (eds.) 1992. Geographical and Environmental Epidemiology. Methods for Small-Area Studies. WHO, Oxford University Press., pp 221-230.

3. Bland J.M. and Altman D.G., 1995. Multiple significance tests: the Bonferroni method. Brit. Med. J., 310, 170.

4. Chen R., Mantel N. and Klingberg M.A., 1984. A study of 3 techniques for time-space clustering in Hodgkin's disease. Statistics in Medicine, 3, 173-184.

5. Cuzick J. and Edwards R., 1990. Spatial clustering for inhomogeneous populations. J.R. Statist. Soc. (B) No.1, 73-104.

6. Elliott P., Kleinschmidt I. and Westlake A.J., 1992. Use of routine data in studies of point sources of environmental pollution. In Elliott P., Cuzick J., English D., Stern R. (eds.) 1992. Geographical and Environmental Epidemiology. Methods for Small-Area Studies. WHO, Oxford University Press., pp 205-220.

7. Gilman E.A. and Knox E.G., 1995. Childhood cancers: space-time distribution in Britain. J. Epidemiol. and Comm. Health, 49, 158-163.

8. Glass A.G. and Mantel N., 1969. Lack of space-time clustering of childhood leukemia in Los Angeles County 1960-64. Cancer Res., 29, 1995-2001.

9. Knox E.G., 1964. Epidemiology of childhood leukaemia in Northumberland and Durham. Brit. J. Prev. Soc. Med., 18, 17-24.

10. Knox E.G., 1989. Detection of clusters. In Elliott, P. (ed.) Methodology of enquiries into disease clustering. Small Area Health Statistics Unit, London.

11. Mantel N., 1967. The detection of disease clustering and a generalised regression approach. Cancer Research , 27, 209-220.

12. Potthoff R.F. and Whittinghill M., 1966a. Testing for homogeneity: I. The binomial and multinomial distribution. Biometrika, 53, 167-182.

13. Potthoff R.F. and Whittinghill M., 1966b. Testing for homogeneity: II. The Poisson distribution. Biometrika, 53, 183-190.

14. Rodrigues L.C., Marshall T., Murphy M. and Osmond C., 1992. Space time clustering of births in SIDS: do perinatal infections play a role? Int. J. Epidemiology, 21, 714-719.