Problems in the interpretation of small area analysis of epidemiological data: the case of cancer incidence in the West of Scotland

David J Hole, Douglas W Lamont

Abstract

Study objective—The aim was to examine the extent to which random variation alone will produce differences in observed incidence rates between small areas which will affect measures of spatial clustering and estimates of relative risk.

Design—This was a study of changes in the pattern of spatial concentration of cancer incidence over a five year time period. A comparison was made of observed incidence rates for 34 tumour sites with randomly generated values and, where possible, with expected values derived from known relative risks.

Setting—Twenty six local government districts in the west of Scotland.

Main results—A statistically significant relationship was observed between sample size and the stability of a summary measure of spatial concentration. Almost all observed highest mean rate ratios were within the 95% confidence interval of the simulated distribution of these values. In three cases examined, both observed and simulated highest/lowest rate ratios were larger than those expected on the basis of known exposures to risk.

Conclusions—In the absence of a prior hypothesis, small area analysis of epidemiological data for periods of less than 10 years will almost always give misleading results for all but the most common diseases.

Small area analysis of epidemiological data in the absence of a prior hypothesis has become common in recent years. The advent of sophisticated computer systems and graphics packages and the computerisation of census data down to enumeration district level has made possible the production of cancer incidence and mortality rates for ever smaller geographical units.\textsuperscript{1-4} It is our concern that the ease with which these data can be generated will encourage fruitless attempts to formulate hypotheses based on differences for which random variation may well be the major component.

Where small area analysis has been of use in suggesting possible risk factors, two conditions have been met. Firstly, widely differing incidence rates have occurred in areas which are close together geographically, and secondly the areas used have been of sufficient population size for observed differences in rates to be more than simply a function of random variation.

Methods

Cancer incidence data for the 26 local government districts of the west of Scotland (median population 79 000) were obtained from the Scottish cancer atlas\textsuperscript{3} for the period 1975–80 and from the West of Scotland Cancer Registry for 1980–85. In order to ensure the same degree of completeness of registration in both data sets (ie, allowing the same length of time from the latest date of diagnosis to the time at which analysis was undertaken) a small overlap in the two time periods was necessary.

Three approaches were used to examine the extent to which results from small area analysis can be influenced by random error alone. (1) Comparison was made of the stability of a summary measure of spatial concentration between the two time periods, over which little change would be expected. (2) Comparison was made of observed highest/mean ratios with simulated values derived from a distribution randomly generated on the assumption of no difference in rates between districts. Mean values were taken as being the rate for the area as a whole. The use of minimum values throughout would have made the ratios unstable for areas with very low rates and impossible to calculate for those with zero rates. (3) Comparison was made of observed and simulated highest/lowest ratios, with expected values calculated on the basis of different levels of exposure to risk between areas.

The summary measure of spatial concentration used was the statistic “D”, developed for use in the Scottish cancer atlas.\textsuperscript{5} This is defined as the average absolute difference in rank order of incidence between all possible pairs of spatially adjacent districts.\textsuperscript{3} Its sampling distribution is approximately normal about a mean of $(N+1)/3$ where $N$ is the total number of districts. Low values of $D$ are indicative of some degree of spatial clustering.
The sampling distribution of D for the west of Scotland was derived from 100 000 random assignments of ranks to each of the 26 local government districts in the region. This yielded a distribution with a mean of 9.01, in which 95% of the observations were within the range 7.55 to 10.32 (fig 1). Values less than 6.0 were encountered only twice in the course of 100 000 generations of the D statistic, and no value occurred which was greater than 12.0.

The simulation of rates for different areas was based on the assumption that the "true" incidence rate was the same for each district, equal to that of the west of Scotland as a whole. The expected number of cases for each district was then calculated. This provided the mean value for a Poisson distribution conditioning on area population size, similar to the approach used by Alexander et al whereby a site specific distribution of the D statistic was derived by allocating cases to districts according to population size. A value was chosen at random from each of these distributions from which a simulated rate could then be calculated, allowing highest/mean and highest/lowest ratios to be derived. This process was repeated 1000 times to derive a sampling distribution for these ratios.

Real differences between areas in the proportions of the population exposed to a known risk factor should be reflected in the incidence rates observed and thus in the ratio of the highest to the lowest of these rates. The age standardised incidence rate is a function of the proportion of the population exposed to a particular risk factor and the relative risk of contracting a given disease if thus exposed. This can be calculated from the formula

\[ r = \frac{r_n \times P_a}{RR \times r_a \times P_e} + \frac{(RR \times r_a \times P_e)}{P_m + RR \times P_e} \]

where \( RR \) = relative risk for disease; \( r \) = incidence rate in area; \( r_a \) = incidence rate among those not exposed to the risk factor; \( RR \times r_a \) = incidence rate among those exposed to the risk factor \( (r_e) \); \( P_a \) = proportion of population of area exposed to risk factor; and \( P_e \) = proportion of population of area not exposed to risk factor.

Assuming that the "true" incidence rate of the disease in the non-exposed population is the same for all areas, the top/bottom ratio would be given by

\[ \frac{P_m + RR \times P_e}{P_m + RR \times P_e} \]

where the suffixes t and b refer to the areas with the top and bottom rates. This expected top/bottom ratio was compared with the observed ratio, and with that which would occur on the basis of chance alone (as produced by the simulation described above).

**Results**

**STABILITY OF D STATISTIC OVER TIME**

Figures 2 and 3 illustrate the contrasting spatial patterns produced by mapping incidence rates for two of the most and least common tumours in Scotland. Figure 2 shows the spatial distribution of areas of high and low incidence of lung cancer in men for 1980-85 compared with the situation five years earlier. The pattern is highly concentrated, with most cases per 100 000 population occurring in the Central Clydeside conurbation. This is reflected in a D statistic which is signifi-

---

**Figure 1** Sampling distribution of D statistic for west of Scotland based on 100 000 random permutations of district ranks

**Figure 2** World age standardised incidence of lung cancer in males—west of Scotland 1975-85 (quintiles)
cantly lower than would have been expected by chance and which also shows little change over time. In contrast, two of the five highest ranking
districts for cancer of the larynx in women between 1980 and 1985 (fig 3) were among the five
lowest ranking districts five years earlier. The patterns produced by mapping the incidence of
this tumour over time show no evidence of concentration, and are quite different from each
other and from the distribution of lung cancer, despite a common risk factor (cigarette smoking).

Table I shows the values of D obtained from analysis of the spatial distribution of incidence
rates at each of the two time periods for 16 tumour
sites in men and 18 in women. Six D statistics
achieve statistical significance at the 5% level for
1975–80 but only in one case (large bowel in
women) is there any support for a similar degree of clustering five years later. A statistically
significant relationship exists between stability of the D statistic over time and the total number of
cases available for analysis (fig 4). The largest
difference in D statistic (2-05 for lip cancer in
females) was produced by comparison of patterns
based on fewer than 50 cases recorded for the
whole of the west of Scotland. All of the nine sites
for which 1500 cases or more were available
recorded differences in the value of D of 0·73 or
less.

COMPARISON OF OBSERVED AND SIMULATED RATE RATIOS

Table II compares observed highest to mean rate
ratios for 1980–85 with simulated values based on
1000 randomly generated distributions of cases. For 28 of the 34 sites, observed highest/mean
ratios were within the 95% confidence limits of
those achieved by simulation, indicating that
random variation would more than account for
the range observed. Two sites (larynx and
stomach in females) showed less variation than
would be expected by chance, observed highest/
mean ratios being below the lower 95% confidence limit. In the four remaining sites the
observed value was significantly higher than the
simulated figure. These comprised two of the
most common (male lung and female breast) and
least common (male lip and female melanoma)
and the ratios were not correlated with sample
size, indicating a mixture of random variation and
genuine differences in incidence.

COMPARISON OF OBSERVED RATE RATIO WITH THAT
CALCULATED ON THE BASIS OF DIFFERENTIAL
EXPOSURE TO RISK

Table III shows expected rate ratios for lung,
bladder and breast cancer, calculated on the basis
of differences between areas in the prevalence of a history of cigarette smoking,\(^5\) employment in a pigments factory,\(^6\) and obesity.\(^7\)\(^9\)\(^10\) The final two columns present the ratios for the highest and lowest rates when only random variation is present (simulated ratio) and when random variation and genuine differences are combined (observed ratio). In all cases the expected ratios based on known risks and exposures are considerably smaller than those observed, and are also lower than simulated values.

**Discussion**

The collection and mapping of cancer incidence data for geographical areas has proved useful in generating hypotheses of aetiological factors when the areas have contained sufficiently large populations and the lifestyles or other risk factors involved have been quite distinct.\(^11\)\(^12\) The transfer of this approach to smaller areas still provides wide ranges in incidence rates but the role of random variation is considerably greater. Computationally sophisticated measures of spatial clustering have been devised in recognition of the fact that risk factors may cross administrative boundaries and to make full use of the data available.\(^13\) Our concern is that a proliferation of inductive studies at this small area level will detract from deductive investigations of localised environmental hazards, such as those which have been associated with nuclear installations.\(^14\)\(^15\) To this end, using data from the west of Scotland, we have attempted to show the shortcomings of small area analysis with no prior hypothesis.

Substantial change in the spatial pattern of incidence of different types of cancer due to changes in lifestyle or environment was not expected over such a short period as five years. Most changes which did occur in the chosen

### Table II: Observed and expected ratios between highest district rates and those for the west of Scotland as a whole, 1980-1985

<table>
<thead>
<tr>
<th>Tumour site</th>
<th>Observed incidence rates</th>
<th>Observed H/M ratio</th>
<th>Simulated H/M ratio (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Highest</td>
<td>Mean</td>
<td>Lowest</td>
</tr>
<tr>
<td>Thyroid</td>
<td>1.8</td>
<td>0.6</td>
<td>0.0</td>
</tr>
<tr>
<td>Lip</td>
<td>7.2</td>
<td>1.6</td>
<td>0.0</td>
</tr>
<tr>
<td>Hodgkin's disease</td>
<td>6.3</td>
<td>2.7</td>
<td>0.0</td>
</tr>
<tr>
<td>Melanoma</td>
<td>6.1</td>
<td>3.5</td>
<td>1.2</td>
</tr>
<tr>
<td>Testis</td>
<td>10.8</td>
<td>4.4</td>
<td>0.8</td>
</tr>
<tr>
<td>Oral/pharynx</td>
<td>8.1</td>
<td>5.5</td>
<td>1.3</td>
</tr>
<tr>
<td>Larynx</td>
<td>10.6</td>
<td>5.6</td>
<td>1.6</td>
</tr>
<tr>
<td>NH lymphoma</td>
<td>12.0</td>
<td>7.5</td>
<td>4.7</td>
</tr>
<tr>
<td>Leukaemia</td>
<td>11.5</td>
<td>7.9</td>
<td>3.3</td>
</tr>
<tr>
<td>Pancreas</td>
<td>12.0</td>
<td>8.4</td>
<td>4.7</td>
</tr>
<tr>
<td>Oesophagus</td>
<td>14.0</td>
<td>9.3</td>
<td>2.9</td>
</tr>
<tr>
<td>Bladder</td>
<td>30.2</td>
<td>20.8</td>
<td>12.4</td>
</tr>
<tr>
<td>Stomach</td>
<td>27.0</td>
<td>21.1</td>
<td>13.5</td>
</tr>
<tr>
<td>Prostate</td>
<td>36.4</td>
<td>24.2</td>
<td>18.3</td>
</tr>
<tr>
<td>Large bowel</td>
<td>45.2</td>
<td>34.0</td>
<td>23.8</td>
</tr>
<tr>
<td>Lung</td>
<td>133-4</td>
<td>100-1</td>
<td>57-2</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Thyroid</td>
<td>1.5</td>
<td>0.2</td>
<td>0.0</td>
</tr>
<tr>
<td>Lip</td>
<td>2.0</td>
<td>1.2</td>
<td>0.0</td>
</tr>
<tr>
<td>Hodgkin's disease</td>
<td>3.5</td>
<td>1.5</td>
<td>0.4</td>
</tr>
<tr>
<td>Oral/pharynx</td>
<td>4.5</td>
<td>1.9</td>
<td>0.0</td>
</tr>
<tr>
<td>Oesophagus</td>
<td>5.0</td>
<td>2.4</td>
<td>0.9</td>
</tr>
<tr>
<td>Leukaemia</td>
<td>8.5</td>
<td>4.6</td>
<td>2.6</td>
</tr>
<tr>
<td>Melanoma</td>
<td>8.8</td>
<td>4.9</td>
<td>2.6</td>
</tr>
<tr>
<td>NH lymphoma</td>
<td>18.4</td>
<td>5.6</td>
<td>3.0</td>
</tr>
<tr>
<td>Pancreas</td>
<td>8.1</td>
<td>5.6</td>
<td>2.4</td>
</tr>
<tr>
<td>Endometrium</td>
<td>8.5</td>
<td>6.1</td>
<td>3.7</td>
</tr>
<tr>
<td>Bladder</td>
<td>9.9</td>
<td>6.3</td>
<td>3.0</td>
</tr>
<tr>
<td>Stomach</td>
<td>10.2</td>
<td>7.1</td>
<td>4.1</td>
</tr>
<tr>
<td>Ovary</td>
<td>12.1</td>
<td>10.5</td>
<td>5.2</td>
</tr>
<tr>
<td>Cervix</td>
<td>16.8</td>
<td>11.5</td>
<td>8.4</td>
</tr>
<tr>
<td>Large bowel</td>
<td>19.6</td>
<td>12.4</td>
<td>5.6</td>
</tr>
<tr>
<td>Lung</td>
<td>36.7</td>
<td>25.9</td>
<td>16.9</td>
</tr>
<tr>
<td>Breast</td>
<td>43.9</td>
<td>31.0</td>
<td>18.8</td>
</tr>
</tbody>
</table>

\(H/M = \text{highest/mean}\)

*Significant at p < 0.05
measure of spatial concentration and in the incidence rates themselves could therefore be attributed to random variation. Even within one time period (1980–85) random error accounted for a large proportion of the variation observed (table II).

Ten of the 34 cancer sites presented in the Scottish cancer atlas yielded a D statistic significant at p<0·001. This level of significance was reached only twice when analysis was confined to the west of Scotland (table I). While some of the explanation for this disparity may be due to a greater variation in lifestyle or environment at a national rather than regional level, the influence of a greater proportion of very small local government districts in the national data set should not be underestimated. One additional case of cancer in an area of less than 25 000 population would result on average in an increase of at least 8·0 in the sex specific incidence rate per 100 000. Eleven of Scotland’s 56 local government districts (20%) contained fewer than 25 000 people at the 1981 census (table IV). These same areas recorded cancer incidence rates which were the highest in the country for 18 (53%) out of the 34 tumour sites presented.

Comparison of the spatial patterns produced by mapping on a relative (quintile) scale incidence rates for two of the largest and smallest tumour sites (cancer of the lung in males and of the larynx in females) illustrates clearly the problems involved in achieving a balance between a sufficiently detailed spatial framework (such as that provided by local government districts) and a sufficiently large number of cases in each (figs 2 and 3). The statistical reliability of any spatial analysis is dependent on the number and size of the spatial units used. Clayton and Kaldor have used a Bayesian approach to produce smoothed estimates of relative risk where the extent of smoothing is determined by the magnitude of the observed rate, its precision and (optionally) the size of the units used. Clayton and Kaldor have used a Bayesian approach to produce smoothed estimates of relative risk where the extent of smoothing is determined by the magnitude of the observed rate, its precision, and (optionally) the estimated underlying relative risk distribution. Our analysis has shown that unless study is confined to the larger cancer sites, or to areas with sizeable populations, random variation alone can explain most differences in incidence. The relationship between sample size and random variation for selected confidence intervals is shown in fig 5. Even at the 95% level, at least 60 cases would be required in any one district to prevent random error exceeding 50% of the observed rate. Account must also be taken of the undue influence on rates of very small populations—in order to achieve a maximum rate per case of 1·0 in every local government district in the west of Scotland, nine years’ data would be required.

On the basis of 1980–85 cancer incidence data, the following number of years’ cases would be required for each site in order to meet a requirement of an average (though not a minimum) of 60 cases per area for an analysis at the level of local government district in the west of Scotland:

- 1 year Lung (males)
- 2–5 years Breast and lung (females), large bowel, stomach, bladder (males), prostate
- 6–10 years Ovary, cervix, bladder (females), oesophagus (males), pancreas
- 11–20 years Oesophagus (females), endometrium, leukaemia, non-Hodgkin’s lymphoma, melanoma (females), larynx (males), oral cavity and pharynx (males)
- 20 years or more Melanoma (males), testis, oral cavity and pharynx (females), thyroid (females), lip (males), Hodgkin’s disease, larynx (females), thyroid (males), lip (females)

Table IV  Population distribution of local government districts in the west of Scotland and Scotland as a whole (1981 census)

<table>
<thead>
<tr>
<th>Population</th>
<th>Scotland</th>
<th>West of Scotland</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt; 10 000</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>&lt; 25 000</td>
<td>10 1 (18°)</td>
<td>1 (4°)</td>
</tr>
<tr>
<td>&lt; 50 000</td>
<td>21 (38°)</td>
<td>6 (23°)</td>
</tr>
<tr>
<td>&lt; 100 000</td>
<td>40 (71°)</td>
<td>17 (65°)</td>
</tr>
<tr>
<td>Total</td>
<td>56 (100°)</td>
<td>26 (100°)</td>
</tr>
</tbody>
</table>

*Sutherland, Skye and Lochalsh, Lochaber, Nairn, Badenoch and Strathspey, Tweeddale, Berwickshire, Stewartry, Orkney, and Shetland

Table III  Comparison of observed, simulated and expected highest to lowest incidence rate ratios for local government districts with highest and lowest exposure to selected risks

<table>
<thead>
<tr>
<th>Tumour site</th>
<th>Risk factor</th>
<th>% Exposed (High)</th>
<th>Relative risk</th>
<th>Rate ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lung (m)</td>
<td>Ever smoked cigarettes</td>
<td>83</td>
<td>9·1</td>
<td>1·2</td>
</tr>
<tr>
<td>Breast (f)</td>
<td>Body mass index &gt; 29 kg/m²</td>
<td>20</td>
<td>2·7</td>
<td>1·1</td>
</tr>
<tr>
<td>Bladder (m)</td>
<td>Employed in pigments factory</td>
<td>3</td>
<td>5·3</td>
<td>1·1</td>
</tr>
</tbody>
</table>

Exp = expected; Sim = simulated; Obs = observed

Exposure data refer to a sample of west Scotland districts. Relative risk and threshold value for body mass index in relation to breast cancer were derived from unpublished data for Renfrew and Paisley. Estimates of the proportion of women with body mass index greater than or equal to 29 kg/m² were calculated by assuming the same frequency distribution about the mean in the highest and lowest sample district as for Scotland as a whole. Observed and simulated rate ratios are for 1980–85 and are for the west of Scotland.

Figure 5  Relationship between random variation and sample size for selected confidence intervals
Obviously for most purposes such amounts of data would be difficult to obtain. There is also a greater chance of confounding changes taking place within too long a time period, although this is likely to be much less serious a problem than the production of meaningless statistics calculated over too short a time scale. Twenty years' data were deemed necessary for the analysis of cancer mortality at county level in the United States.\textsuperscript{17}

Observed differences between highest and lowest recorded rates at a local government district level were far in excess of estimated ratios based on actual patterns of exposure to risk (table III). While such excesses were to be expected in the cases of breast and bladder cancer, where the proportions of the total risk attributable to the factors selected were low (18-25 and 0-11\% respectively), a substantial difference between observed and expected ratios also occurred in relation to lung cancer, where cigarette smoking accounted for 85-87\% of the total risk in all districts for which exposure data were available. The greater the number of cases, the greater was the difference between simulated and observed values, ie, the lower the proportion of random variation present in the observed pattern.

For a given relative risk, levels of exposure would have to be extremely high and the hazard highly localised before the effect of the population attributable risk on incidence rates in small areas was sufficient to exceed random error. Examples of these would include the identification 20 years ago of nasal cancer among woodworkers in High Wycombe\textsuperscript{18} and present excesses of mesothelioma in former shipbuilding workers in Clydebank.\textsuperscript{19} Employment in a west of Scotland pigments factory the work force of which comprised 3\% of the local population carried a relative risk of 5.3 for carcinoma of the bladder in men (table III). This level of risk and exposure would indicate an expected ratio between highest and lowest incidence rates of 1:13, considerably below the simulated ratio for the period 1980-1985 of 1.9, which in turn accounted for almost 80\% of the observed ratio of 2:4. In the case of lung cancer, the expected ratio is still slightly below the simulated figure, although random variation accounts for much less of the observed rate ratio.

The conclusions drawn in this paper do not in themselves invalidate the small area approach. It is however essential that the effects of random variation are taken into account. Case-control studies, by considering groups of individuals with a common characteristic rather than examining heterogeneous groups on a geographical basis, will almost always offer a more effective way of testing hypotheses and identifying risks.

We are grateful to Alison Stewart of the Department of Audivisual Services, Stobhill Hospital, Glasgow, for technical assistance in the production of the illustrations. A preliminary report on this study was presented to the 8th Annual Meeting of the Scottish Epidemiology Group, Dundee, June 1988.

The views expressed in this paper are those of the authors alone and are not necessarily those of the Greater Glasgow Health Board.

References


Problems in the interpretation of small area analysis of epidemiological data: the case of cancer incidence in the West of Scotland.
D J Hole and D W Lamont

*J Epidemiol Community Health* 1992 46: 305-310
doi: 10.1136/jech.46.3.305

Updated information and services can be found at:
http://jech.bmj.com/content/46/3/305

**Email alerting service**

*These include:*

Receive free email alerts when new articles cite this article. Sign up in the box at the top right corner of the online article.

Notes

To request permissions go to:
http://group.bmj.com/group/rights-licensing/permissions

To order reprints go to:
http://journals.bmj.com/cgi/reprintform

To subscribe to BMJ go to:
http://group.bmj.com/subscribe/