*Journal of Epidemiology and Community Health* 1992; **46**: 573–576

# Leukaemia clusters in Great Britain. 2. Geographical concentrations

E G Knox, Estelle Gilman

**Abstract**

***Study objective*—The aim was to test a large set of childhood leukaemia and lymphoma registrations for the presence of short radius spacial clusters.**

***Design*—The study was a geographical cluster analysis.**

***Setting*—England, Wales and Scotland.**

***Patients*—All registrations for leukaemia and lymphoma between 1966 and 1983 in children aged 0 to 14 years were examined. The records included date and age of registration, sex, diagnosis, and the map reference of the postcode of residence. Of the 9411 registrations, 8888 were suitable for inclusion.**

***Main results*—There was a significant excess of case pair addresses separated by $<0{\cdot}5$ km. There was also a significant excess of pairs sharing the same postcode. Both findings were based upon comparison with random pairs of postcodes drawn from the Central Postcode Directory. Examination for clustering at this very short range was based upon a clear prior hypothesis derived from the results of a study of space-time interaction, reported in a companion paper.**

***Conclusions*—It is postulated that the space-time interaction and the geographical concentrations shown here result from a common epidemic process. The epidemiology of this disease is characterised by short range geographical concentrations, with temporal non-homogeneity superimposed. The findings exclude certain artefacts which remained unresolved in the space-time interaction study. The distributions almost certainly reflect biological processes, and the most probable explanation is in terms of an infective process.**

*J Epidemiol Community Health* 1992; **46**: 573–576

In two previous papers[1 2] we reported the occurrence of space-time clusters among registered leukaemias and lymphomas in children aged 0–14 years in England, Wales, and Scotland between 1966 and 1983. The space-time clusters occurred at very short ranges both in time and in space; and they were clearly detectable within intervals of 30 days, combined with distances of less than 0·5 km. The clusters occurred in numbers sufficient to reject the hypothesis that the spatial and temporal distributions of the events were independent of each other.

The material was carefully examined for the presence of several different potential artefacts. They included variation in the completeness of case ascertainment, population movements, and modifications of the date of registration of one case, stimulated by the detection of another close by. No direct evidence of such artefacts was found.

The temporal and distance scales of these space-time clusters were close to the resolutions of recording of the dates and of the map references; but there seemed to be no way in which this particular pattern could have been generated through inaccuracies of recording. The critical distances for detecting the interaction were less in high density regions than in moderate density regions; and in those zones with low population densities (eg, the south west of England) the interaction was not demonstrable. However, it was not clear whether this density dependence reflected a biological necessity or a statistical necessity—whether high density was a prerequisite of infective transmission, perhaps; or whether it simply increased the numbers of close geographical pairs available for analysis.

It is known already that the geographical distribution of childhood leukaemias is uneven, with substantial urban/rural differences,[3–6] and the vexed question of more localised clusters has been examined by many different investigators. Investigations into geographical clustering have generally been hindered by the absence of a prior hypothesis specifying the scale at which the sought for clusters might occur. In its absence, the question of clustering resolves into a series of separate scale dependent hypotheses, so that evidence of apparent clustering might appear in one of them purely through random elements in the sampling process. Here, however, the detection of a very short range space-time interaction permits formulation of a much more specific geographical question. It declares the exact distance against which the question of geographical heterogeneity can now be tested. The scale dependence specificity of the question asked, and its clear prior basis, therefore allows the use of closely focused search techniques—techniques which offer an enhanced likelihood of success should the hypothesis turn out to be true. This is the issue to which the present paper is directed.

## Methods

A full description of the data source is supplied in previous papers.[1 7 8] It consisted of a file of 8888 events, comprising all adequately specified and accurately located registrations of childhood leukaemias and non-Hodgkin lymphomas in England and Wales between 1966 and 1983. The data items used in the present analysis are the map references associated with the postcodes of addresses recorded at the time of registration. They were taken originally from the Central

**Department of Public Health and Epidemiology, University of Birmingham, Edgbaston, Birmingham B15 2TJ, United Kingdom**
E G Knox
E Gilman

Correspondence to:
Professor Knox

Accepted for publication
March 1992

Postcode Directory, where they were recorded to an accuracy of 100 m in England and Wales, and to an accuracy of 10 m in Scotland. These coordinates refer to the first address within each postcode. For a geographical area representing one third of the population of England and Wales an alternative and more precise map reference was made available by Pinpoint Analysis Ltd. It refers to the centroids of the postcodes and is presented to a resolution of one metre.

The records of map references in our files were resolved to one metre in order to accommodate either format, and those original directory values which remained were simply padded out with zeros to represent single metres and tens of metres. This data set, together with a copy of the Central Postcode Directory (CPD), forms the main numerical basis of the present study.

We used two separate analytical techniques in order to compare the mutual proximity characteristics of addresses within the leukaemia/lymphoma register and within the postcode directory as a whole. The first technique was quantitative and based upon a frequency distribution of distances between all possible pairs of cases. This was compared with a similar distribution of a sample of all possible pairs of postcodes ("control" pairs) derived from the CPD. The second technique was a qualitative one, comparing the proportions of case pairs and "control" pairs which shared identical postcode coordinates. Our purposes in using more than one method were to try to overcome two basic uncertainties in the material. The first related to fundamental shortcomings of the CPD itself, which we describe below. The second arose from a restriction placed upon the registry data; we were allowed to inspect the map references associated with the registration postcodes, but not the postcodes themselves. The reasons for this restriction were to maintain confidentiality, and in order to honour undertakings made to clinicians responsible for registering cases under their care.

## Results

### DISTANCES APART

The frequency distribution of the distances between pairs of cases is given in table I (column 2). The coordinates were first rounded to a resolution of 100 m and the distribution was limited to pairs separated by 1000 days or less. This limit was chosen because this was the restriction within which the space-time interactions had been detected. (In fact, a secondary analysis involving all possible pairs, irrespective of time apart, gave closely similar results.)

The distribution of distances between case pairs was compared with one based upon random pairs of residential postcodes drawn from a subfile of the CPD. This subfile was limited to England, Wales, and Scotland and excluded "large users" such as factories, offices, and any coordinates marked with uncertainty flags. Coordinates were rounded to a resolution of 100 m, to match the resolution of the registrations. There was a relative excess, among case pairs, of separations less than or equal to 40 km. The greatest relative excess was at separations up to 0·5 km.

There were substantial logistical (and mechanical) difficulties in extracting a sufficient number of random pairs from a very large file to permit accurate case-control comparison at this short range. A secondary operation was therefore conducted in which random controls were selected from a version of the CPD which had been sorted first into ascending order of eastings and northings. For each random selection, all subsequent coordinates within 0·5 km were identified and counted, for example, 150 000 random selections generated 2 436 971 close pairs. Including the possibility that two controls could occur within the same postcode, this gave an estimated frequency of close addresses of 0·0244 close pairs per 1000 total pairs, compared with an expectation of 0·042 obtained from the random pair method, and the 0·069 observed. The 17 close control pairs in column 3 of table I therefore seem to represent an overestimate of the true expectation, and this confirms the reality of the relative excess among the short distance case pairs. However, exact comparison is flawed because some of the case coordinates had been modified using the Pinpoint index, so that the registers of coordinates for the cases and for the controls were not strictly comparable. Its seems unlikely that the registration list, with its refined resolution, could have generated false geographical granularities; indeed, the reverse seems more likely. However, a reservation remains, and it was for this reason that an alternative qualitative approach was developed.

*Table I Frequency distribution of inter-pair distances (km). Case and control coordinate pairs.*

| (1) Distance (d) apart (km) | (2) Case pairs (<1000 d apart)[a] n | per 1000 | (3) Random postcode (control) pairs n | per 1000 | (4) Case/control ratios (2)/(3) |
|---|---|---|---|---|---|
| 0·0<d <0·5 | 780 | 0·069 | 17 | 0·042[b] | 1·643[b] |
| 0·5<d <1·0 | 1456 | 0·128 | 43 | 0·108 | 1·185 |
| 1·0<d <2·0 | 4294 | 0·379 | 130 | 0·325 | 1·166 |
| 2·0<d <5·0 | 22 856 | 2·015 | 704 | 1·760 | 1·145 |
| 5·0<d <10·0 | 61 885 | 5·456 | 1656 | 4·140 | 1·318 |
| 10·0<d <15·0 | 80 992 | 7·141 | 2330 | 5·750 | 1·242 |
| 15·0<d <20·0 | 92 135 | 8·123 | 2616 | 6·540 | 1·242 |
| 20·0<d <40·0 | 398 289 | 25·117 | 12 352 | 30·880 | 1·137 |
| 40·0<d<100·0 | 1 522 628 | 134·248 | 52·720 | 131·800 | 1·018 |
| 100·0<d<200·0 | 3 164 764 | 279·034 | 115 588 | 288·970 | 0·966 |
| >200·0 | 5 991 792 | 528·255 | 211 874 | 529·685 | 0·997 |
| Total | 11 341 871 | 1000·000 | 400 000 | 1000·000 | 1·000 |

[a] The case pair analysis was repeated for all possible pairs irrespective of the time interval. The results were almost exactly the same, with rates per 1000 pairs of 0·063 and 0·123 in the first two rows.
[b] Expected frequencies for distances up to 0·5 km. were re-estimated using a different method which provided larger numbers. A random set of 150 000 postcodes in an east/north sorted file of 1 372 359 postcodes was searched for subsequent entries within 0·5 km. There were 2 436 971. The proportion of close pairs was calculated as (150 000/2 + 2 436 971)/(150 000 × 1 372 359/2). This gives an expected value of 0·0244 per 1000 and a case-control ratio, within 0·5 km (all time intervals), of 2·582.

### SHARED POSTCODES

The qualitative approach for detecting clustering was based upon counting the numbers of pairs of cases sharing a single postcode map reference. A geographically sorted version of the registration file was scanned to identify case pairs and larger groups which shared identical map references. There were 128 postcode identical pairs, and an additional six sets of geographically identical triplets.

A published analysis of the CPD[9] identified 1 284 852 residential postcodes in Great Britain, so the 8888 disease coordinates are distributed at a mean density of 0·00692 per postcode. A simple Poisson model predicts that 30·3 postcodes should contain two events, and 0·07 should contain three

or more. The facts show a clear excess. However, there was a substantial hazard to this direct interpretation in that adjacent postcodes, within the directory, may sometimes have been allocated the same coordinates. It was therefore necessary to analyse the characteristics of the directory in more detail.

The CPD subfile, sorted in order of eastings and northings, was used to generate a frequency distribution of those coordinates which were shared by 1, 2, 3 . . . different postcodes. This was a more recent version of the CPD than that referred to by Wilson and Elliot[9] and the residential subfile contained 1 372 359 different postcodes. The resulting distribution is given in the first column of table II.

It is clear that there was a considerable degree of coordinate sharing and that the simple Poisson model is not reliable. There were only 886 108 separate map references, after rounding to 0·1 km, giving a mean of 1·55 postcodes per coordinate. Only 45·2% of the postcodes and 70·1% of the map references were uniquely linked to each other. At the extreme, 1737 map references were each attached to more than 10 separate postcodes, and three of these were each shared by over 200 different postcodes. Poisson expectations were therefore recalculated for each individual level of postcode grouping, the Poisson parameter having being adjusted for the number of shared postcodes. The third column of table II gives the expected numbers of pairs (or more) at each level, while the last two columns list the observations, and the ratios between observations and expectations. With this stratified model we would have expected about 74·6 postcode groups to contain two or more events, compared with the 134 observed. It should be noted that the excess is concentrated among the smaller groupings shown in the first three rows of the table; that is, the most accurately specified end of the distribution, and the elements least liable to uncontrolled error and artefact. The excesses are statistically significant in each of the first three rows of table II.

The CPD was searched for the map references corresponding with the coordinates attached to the cases. The groups were shown in each case to consist of adjacent postcodes. The ambiguities of location consisted entirely of errors of local resolution and there were no examples of errors arising from the false apposition of widely separated postcodes.

The excess pairs in the first row of table II result from cases sharing an individual postcode; but in the next two rows of the table only about one in four and one in eight, respectively, can probably be accounted for in this way. The remainder might span a single postcode boundary, or sometimes two. The scale of the clustering, corresponding with two to four postcodes, probably represents 80–100 people on average, including about 20 children. It is commensurate with the 500 m indicated by the earlier analysis, as shown in table I, and can be treated as an alternative demonstration of the same phenomenon.

All these calculations suppose that residential postcodes are uniform in terms of the numbers of households and children. We could find no published analysis which showed the actual degree of variation. However, we carried out a supplementary exercise in which the risk levels attached to each of the layers in table II were redistributed according to a Poisson distribution centred on the mean number of households per postcode. In effect the table was stratified into many more layers. The calculations were repeated, but the result was not much different. The overall expectation of map references sharing two or more registrations was now 76·1 instead of 74·6. A natural variability greater than that expressed by a Poisson distribution would presumably increase this expectation further, but probably not by any great amount. The conclusions are not changed.

*Table II Postcodes and postcode coordinates.*

| *No of postcodes within a "common coordinate" group*[a] | *Number of listed coordinates* | *Expected case clusters*[b] | | *Observed case clusters*[b] | | *Observed/ expected* |
|---|---|---|---|---|---|---|
| 1 | 620 989 | 12·970 | | 27 | | 2·08 |
| 2 | 156 300 | 13·002 | | 38 | | 2·92 |
| 3 | 60 912 | 11·352 | | 31 | | 2·73 |
| 4 | 24 973 | 8·238 | | 14 | | 1·70 |
| 5 | 10 513 | 5·411 | | 5 | | |
| 6 | 5143 | 3·785 | | 6 | | |
| 7 | 2529 | 2·522 | | 0 | | |
| 8 | 1477 | 1·916 | 10·963 | 3 | 11 | 1·00 |
| 9 | 920 | 1·504 | | 2 | | |
| 10 | 615 | 1·236 | | 0 | | |
| 11 | 395 | 0·956 | | 0 | | |
| 12 | 278 | 0·797 | | 1 | | |
| 13 | 240 | 0·805 | 3·430 | 0 | 2 | 0·58 |
| 14 | 128 | 0·496 | | 0 | | |
| 15 | 85 | 0·376 | | 1 | | |
| 16 | 107 | 0·536 | | 1 | | |
| 17 | 66 | 0·372 | | 0 | | |
| 18 | 57 | 0·359 | 1·918 | 0 | 2 | 1·04 |
| 19 | 47 | 0·328 | | 0 | | |
| 20 | 42 | 0·323 | | 1 | | |
| 21 | 24 | 0·203 | | 0 | | |
| 22 | 34 | 0·314 | | 0 | | |
| 23 | 25 | 0·251 | | 0 | | |
| 24 | 22 | 0·240 | | 1 | | |
| 25 | 7 | 0·082 | 1·880 | 0 | 1 | 0·53 |
| 26 | 10 | 0·190 | | 0 | | |
| 27 | 10 | 0·136 | | 0 | | |
| 28 | 12 | 0·175 | | 0 | | |
| 29 | 10 | 0·156 | | 0 | | |
| 30 | 8 | 0·133 | | 0 | | |
| Over 30 | 125 | 9·509 | | 3 | | – |
| Total | 886 108 | 74·560 | | 134 | | 1·717 |

[a] The group size is the number of postcodes shown as sharing a single map reference, at 0·1 km resolution. Thus 620 989 map references were attached to a unique postcode, while 156 300 map references were each attached to two different postcodes—312 600 postcodes altogether. A total of 886 108 map references, attached to 1 372 359 postcodes, gives an overall mean of 1·549 postcodes per given map reference.
[b] A case cluster is a pair, a triplet, or a greater aggregation of registrations.

## Discussion

The results of our analyses supply clear evidence that registrations of childhood leukaemias and lymphomas tended to occur in close proximity to each other, with an excess of pairs and triplets within a range of about 500 m, and with a strong tendency to share common or adjacent postcodes. Clustering on so fine a scale as this is unlikely to be detected by general search methods designed without reference to a previously justified level of scale dependency. The use of such scale specific search techniques was justified here, as seldom elsewhere, by the prior demonstration of space-time clusters within these same distance limits.

The main technical problems of this analysis sprang from (a) the variable resolution of the

geographical coordinates recorded in the Central Postcode Directory, (b) the partial and not clearly specifiable "corrections" introduced to the geographical coordinates of the leukaemia registrations, and (c) the "censoring" of the registration postcodes on political/ethical grounds. These access restrictions imposed the need to devise two separate pragmatic approaches to cluster detection, one quantitative and one qualitative, and enforced a major reanalysis of the grouping characteristics of the Central Postcode Directory. The main source of inaccuracy was where several adjacent postcodes shared a common map reference, thus reducing the precision with which interpair distances could be defined. It also disturbed the simple probability calculus on which we might estimate random expectations. The results of the two modes of analysis were concordant and highly significant, but some uncertainties will necessarily remain until the original postcodes of the leukaemia and lymphoma registrations are eventually released.

The absolute excess of case pairs sharing individual postcodes or small groups of postcodes amounted to about 60 pairs and triplets. The quantitative analysis—with its uncertainties—suggested an excess of 200–300 pairs separated by distances up to 0·5 km. In our previous paper there were 35 pairs within 0·5 km *and* 30 days of each other, showing an excess of 10 over the random expectation. The space-time excess is therefore contained within the geographical excess, suggesting that both demonstrations are reflections of a common epidemic process which might be represented as a pattern of transient and intermittent hot spots.

We speculated in our previous paper[2] that space-time clustering might have arisen because an epidemic communicable disease had provoked haematological examinations, and that this might have triggered the near simultaneous diagnosis of latent leukaemias already existing within the population. This could certainly account for space-time interaction—asynchronous temporal clustering in different localised areas—but it could not in itself account for geographical clustering among events aggregated over a prolonged period of time. If the geographical clustering and the space-time clustering are indeed aspects of the same phenomenon, as must seem likely, then this particular artefact is excluded.

It is not easy to suggest a unitary mechanism which might explain both findings, or to say whether it might relate to childhood leukaemias and lymphomas as a whole, or to a particular aetiological subset. The space-time interaction suggests a disease-promoting event close to the date of registration, a circumstance identifying it with clinical presentation rather than with a primary oncogenic stimulus. This result suggests an immune reaction by an already transformed lymphocyte clone, triggered by an infective exposure. The pure geographical concentrations then suggest that this might be a response to *re*infection, and that clustered primary infections may also have arisen from common sources. It is difficult to reconcile the apparent absence of repetitions within sibships or within twin pairs with any mechanism involving chains of person to person spread of an infective agent. If the combined spatial and space-time clustering is indeed based upon an infective process, it more probably represents a non-household low risk exposure through a mechanism which does not give rise to retransmission.

We are grateful to Dr G J Draper and the Childhood Cancer Research Group for providing the data on which these analyses are based. Thanks are also due to the cancer registries and the UK Children's Cancer Study Group who originally provided the cancer registrations.

1 Gilman EA, Knox EG. Temporal-spatial distribution of childhood leukaemias and non-Hodgkin lymphomas in Great Britain. In: Draper G, ed. *The geographical epidemiology of childhood leukaemia and non-Hodgkin lymphomas in Great Britain, 1966–83.* (Studies on Medical and Population Subjects, No 53.) London: OPCS, 1991.
2 Knox EG, Gilman EA. Leukaemia clusters in Great Britain: 1. Space-time interactions. *J Epidemiol Community Health* 1992; **46**: 566–72.
3 Knox EG, Stewart AM, Gilman EA, Kneale GW. Background radiation and childhood cancer. *J Radiol Protect* 1988; **8**: 9–18.
4 Stiller CA, Draper GJ, Vincent TJ, O'Connor CM. Incidence rates nationally and in administratively defined areas. In: Draper G, ed. *The geographical epidemiology of childhood leukaemia and non-Hodgkin lymphomas in Great Britain, 1966–83.* (Studies on Medical and Population Subjects, No 53.) London: OPCS, 1991.
5 Draper GJ, Vincent TJ, O'Connor CM, Stiller CA. Socioeconomic factors and variations in incidence rates between County Districts. In: Draper G, ed. *The geographical epidemiology of childhood leukaemia and non-Hodgkin lymphomas in Great Britain, 1966–83.* (Studies on Medical and Population Subjects, No 53.) London: OPCS, 1991.
6 Rodrigues L, Hills M, McGale P, Elliott P. Socioeconomic factors in relation to childhood leukaemia and non-Hodgkin lymphomas: an analysis based on small area statistics for census tracts. In: Draper G, ed. *The geographical epidemiology of childhood leukaemia and non-Hodgkin lymphomas in Great Britain, 1966–83.* (Studies on Medical and Population Subjects, No 53.) London: OPCS, 1991.
7 Stiller CA, O'Connor CM, Vincent TJ, Draper GJ. The National Registry of Childhood Tumours and the leukaemia/lymphoma data for 1966–83. In: Draper G, ed. *The geographical epidemiology of childhood leukaemia and non-Hodgkin lymphomas in Great Britain 1966–83.* (Studies on Medical and Population Subjects, No 53.) London: OPCS, 1991.
8 Elliott P, McGale P, Vincent TJ. Description of population data and definitions of areas In: Draper G, ed. *The geographical epidemiology of childhood leukaemia and non-Hodgkin lymphomas in Great Britain, 1966–83.* (Studies on Medical and Population Subjects, No 53.) London: OPCS, 1991.
9 Wilson PR, Elliot DJ. An evaluation of the Postcode Address File as a sampling frame and its use within OPCS. *J R Stat Soc A* 1987; **150**: 230–40.